Simulation of Rapidly Varying and Dry Bed Flow using a Finite Element Volume Method for the Serre Equations.

Jordan Pitt

October 2018

A thesis submitted for the degree of Doctor of Philosophy
of the Australian National University

To my mother and father who have provided me with everything.

Declaration

The work in this thesis is my own except where otherwise stated.

Jordan Pitt

Acknowledgements

I would like to thank my lead supervisor Professor Stephen Roberts for his insight, suggestions and time spent improving my research. Dr Chris Zoppou who put a tremendous amount of time and effort into reading and editing my work. The remainder of my supervisor panel Professor Markus Hegland and Professor John Urbas.

I would also like to thank our fellow researchers who provided us with experimental data:

- Dr David George, Cascades Volcano Observatory, U.S. Geological Survey for providing the digitised data for the rectangular depression experiment.
- Professor Sedar Beji, Department of Naval Architecture and Ocean Engineering, Istanbul Technical University, for providing the data for the periodic waves over a submerged bar experiment.
- Dr Volker Roeber, Department of Physical Oceanography, University of Hawai‘i at Mānoa for providing the data for the solitary wave over a fringing reef experiment.

Abstract

Recent research in numerical wave modelling has focused on developing computational methods for solving non-linear, dispersive wave equations as an extension to the well understood computational methods for solving the nonlinear shallow water wave equations. These equations contain extra terms that allow for dispersion; better modelling the behaviour of actual water waves. An interesting example of these nonlinear dispersive equations from the viewpoint of modelling water waves are the Serre equations; striking a balance between its accuracy as a model for water waves and the computational expense required to solve it as compared to the shallow water wave equations.

In this work an efficient and robust numerical method for the one-dimensional Serre equations was developed. This method uses a consistent polynomial representation of the solution over the cells to approximate the Serre equations which makes it suitable for unstructured meshes and parrallelisable. This method was extended to allow for the recovery of the lake at rest steady state and the simulation of flow over dry beds.

The convergence and dispersion properties of the method was determined using a linear analysis. The method was validated against analytic and forced solutions of the Serre equations, demonstrating its convergence properties. Finally, it was validated against experimental data for a wide array of physical scenarios, establishing its utility as a physical model. All these analyses and validations were conducted whilst comparing the method to previous methods to demonstrate its strengths and weaknesses. Overall the developed method was found to be the most robust whilst being adequately accurate and is therefore the preferred numerical method for the Serre equations.

Contents

Chapter 1

Introduction

A significant portion of the world's people and critical infrastructure is located near the coast, with shipping being the dominant method of trade in our globalised world. While the ocean provides many opportunities it also poses significant hazards, for instance; tsunamis and storm surges. Furthermore, the dynamics of ocean waves significantly impacts our understanding of other physical phenomena; such as the breakup of sea-ice and the erosion of beaches. Therefore, accurate modelling of ocean waves is important to our society.

The physics of water is well understood being an application of Newtons second law; with their governing partial differential equations initially presented by Euler [1]. These equations were then extended to include viscosity, producing the full Navier-Stokes equations. Numerical methods [2, 3, 4] have been developed to solve these full water equations; however due to the complexity of these partial differential equations such methods can only accurately resolve fluid behaviour over small scales and not the scale required to model a tsunami over an ocean basin.

For this reason the central focus of water wave modelling has been simplified water wave theories that approximate the behaviour of the free surface of water governed by the Euler or Navier-Stokes equations. The most popular class of these approximate water wave theories are the shallow water wave theories where the characteristic water depth h_0 is far smaller than the characteristic wave length λ_0, so that $\sigma = h_0/\lambda_0 \ll 1$. If we neglect all terms of order $\mathcal{O}(\sigma^2)$ then the full Euler equations reduce to the Shallow Water Wave Equations (SWWE) [5] which describe fully nonlinear non-dispersive waves. Retaining higher powers of σ leads to a class of equations known as 'Boussinesq-type' equations. Boussinesq-type equations are then classified by the powers of σ they retain and their retained

nonlinearity; which is based on the size of $\epsilon = a_0/h_0$ which measures the characteristic amplitude of the waves a_0 against h_0. These wave models form a spectrum with the SWWE being the simplest and most restrictive model and the Boussinesq-type models retaining the highest powers of σ and largest ϵ being the most complex and least restrictive. The Serre equations are one particular Boussniesq-type equation that retains all terms of order $\mathcal{O}(\sigma^4)$ and makes no assumption on the size of ϵ [5] and is thus the most appropriate model for water waves for the $\mathcal{O}(\sigma^4)$ class of Boussinesq-type equations.

There has previously been a significant amount of research into developing large scale, efficient and robust computational methods for the SWWE [6, 7, 8]. The SWWE neglect all terms of order $\mathcal{O}(\sigma^2)$ in the Euler equations and so do not capture all water wave behaviour; in particular dispersion. Recent research [9, 10] has highlighted the need for dispersive wave models for the evolution of tsunamis. For the purposes of ocean wave modelling the Serre equations are the best placed [5]; retaining high-order σ terms and allowing for any size of ϵ. Hence the overarching goal of our research is the development of large-scale, efficient and robust computational models for the Serre equations for the purposes of ocean wave modelling.

1.1 Objectives of the Thesis

In view of the overarching goal, the primary motivation of this thesis was the development of a numerical method for the one-dimensional Serre equations that is robust to steep gradients in the free surface and dry beds and can be readily extended to the two dimensional Serre equations using unstructured meshes.

This goal was achieved through the development of the method described by Zoppou [11]. This method was extended to adequately handle dry beds and its finite difference method was replaced with a finite element method, making the method more suitable for unstructured meshes.

This method was then assessed with a linear analysis, a validation against analytic and forced solutions and finally with a validation against experimental results. At all stages of this assessment the method is compared to at least one other method; demonstrating its strengths and weaknesses. Overall, the method is found to be superior to the others and satisfies all the desired properties constituting the main goal of the thesis.

1.2 Original Contribution of the Thesis

My research made the following original contributions to the field:

- Observation and justification of a new structure in the solution of the Serre equations to the dam-break problem.
- Development and description of the well balanced second-order finite element volume method that can handle dry beds.
- Extension of the second-order finite difference volume method to allow for dry beds.
- Implementation of the third-order finite difference volume method.
- A linear analysis of convergence for all developed hybrid finite volume methods and some finite difference methods.
- An analysis of the dispersion properties of all named methods extending previous work by allowing for non-zero mean fluid velocity and analysing the total dispersion error.
- Validation of the method using forced solutions where all terms of the Serre equations are present for wet and dry beds.
- Comparison of numerical solutions of the Serre equations and experimental results in the presence of dry beds and with wave breaking.

1.2.1 Publications

My research contributed to the following publications in chronological order.

A SOLUTION OF THE CONSERVATION LAW FORM OF THE SERRE EQUATIONS

Australia and New Zealand Industrial and Applied Mathematics Journal (2016)
C. Zoppou, S.G. Roberts and J. Pitt

Abstract:

The nonlinear and weakly dispersive Serre equations contain higher-order dispersive terms. These include mixed spatial and temporal derivative flux terms

which are difficult to handle numerically. These terms can be replaced by an alternative combination of equivalent temporal and spatial terms, so that the Serre equations can be written in conservation law form. The water depth and new conserved quantities are evolved using a second-order finite-volume scheme. The remaining primitive variable, the depth-averaged horizontal velocity, is obtained by solving a second-order elliptic equation using simple finite differences. Using an analytical solution and simulating the dam-break problem, the proposed scheme is shown to be accurate, simple to implement and stable for a range of problems, including flows with steep gradients. It is only slightly more computationally expensive than solving the shallow water wave equations.

My Contribution:

The greater computational cost was calculated from my numerical methods.

Numerical solution of the fully non-linear weakly dispersive serre equations for steep gradient flows

Applied Mathematical Modelling (2017)

C. Zoppou, J. Pitt and S.G. Roberts

Abstract:

We demonstrate a numerical approach for solving the one-dimensional non-linear weakly dispersive Serre equations. By introducing a new conserved quantity the Serre equations can be written in conservation law form, where the velocity is recovered from the conserved quantities at each time step by solving an auxiliary elliptic equation. Numerical techniques for solving equations in conservative law form can then be applied to solve the Serre equations. We demonstrate how this is achieved. The system of conservation equations are solved using the finite volume method and the associated elliptic equation for the velocity is solved using a finite difference method. This robust approach allows us to accurately solve problems with steep gradients in the flow, such as those generated by discontinuities in the initial conditions.

The method is shown to be accurate, simple to implement and stable for a range of problems including flows with steep gradients and variable bathymetry.

My Contribution:

The methods, linear dispersion analysis and numerical solutions were all produced by me; these results were then written up into this paper by my coauthors.

Importance of Dispersion for Shoaling Waves

22nd International Congress on Modelling and Simulation (2017)
J. Pitt, C. Zoppou and S.G. Roberts

Abstract:

A tsunami has four main stages of its evolution; in the first stage the tsunami is generated, most commonly by seismic activity near subduction zones. The second stage is the tsunamis propagation through the ocean far from the coast, where variation in bathymetry is slight and gradual. The third stage is the shoaling and interaction of the tsunami with bathymetry as it approaches the coastline. Finally the tsunami reaches and inundates the shore. For our purposes the hydrodynamic models we are interested in deal with the final three stages of the evolution of a tsunami.

The propagation of a tsunami with wavelength λ through water that is H deep is well understood when $\lambda/H \leq 1/20$, which we call shallow water as noted by Sorensen (2006). The wavelengths for tsunamis range from a few to hundreds of kilometres, while the maximum water depth is $11km$ at the Marianas trench, so that most tsunamis occur in shallow water. This stage of tsunami behaviour is adequately modelled using the shallow water wave equations. Current research into tsunamis focuses around more complex approximations to the Euler equations for the third and fourth stages. In this paper we focused on the Serre equations as they are considered a very good model for fluid behaviour up to the shoreline, and they reduce to the shallow water wave equations for large wavelengths.

Although more complicated, the Serre equations provide a better description of the fluid behaviour than the shallow water wave equations and are therefore more computationally expensive to solve numerically. In particular for the methods of this work, we find that the Serre equations have a run-time 50% longer than our equivalent finite volume method for the shallow water wave equations in the one dimensional case. To simulate tsunamis as efficiently as possible it is important to know when using the more complicated Serre equations leads to more accurate predictions of the evolution of a tsunami than the shallow water wave equations. To investigate this we have numerically simulated a laboratory

experiment of periodic waves propagating over a submerged bar, and the propagation of a small amplitude wave up a gradual linear slope using both the Serre and the shallow water wave equations.

The results of these simulations demonstrated that the Serre and shallow water wave equations produce similar results for shoaling waves when the wavelength is large compared to the water depth. This is not surprising as this is the regime under which the shallow water wave equations are derived. However, outside this regime the shallow water wave equations are a poor model for wave shoaling and propagation, poorly approximating the shape and maximum height of waves. Furthermore we demonstrate that for steep waves generated by shoaling, the shallow water wave equations can underestimate the arrival time and amplitude of an incoming wave. These results suggest that for a tsunami it is sufficient to use the shallow water wave equations in stages two and some of stage three, even for large changes in bathymetry. Although dispersive equations such as the Serre equations are required to accurately capture fluid behaviour in stages three and four nearer to the coastline, particularly when wavelengths are short or waves are steep. Since the Serre equations represent only a moderate increase in run-times this suggests that our inundation models should be based on them.

My Contribution:

This paper was produced by me with the support of my coauthors based on my own work.

Behaviour of the Serre equations in the presence of steep gradients revisited

Wave Motion (2018)

J.P.A. Pitt, C. Zoppou and S.G. Roberts

Abstract:

We use numerical methods to study the short term behaviour of the Serre equations in the presence of steep gradients because there are no known analytical solutions for these problems. In keeping with the literature we study a class of initial condition problems that are a smooth approximation to the initial conditions of the dam-break problem. This class of initial condition problems allow us to observe the behaviour of the Serre equations with varying steepness of the initial conditions. The numerical solutions of the Serre equations are justified by

demonstrating that as the resolution increases they converge to a solution with little error in conservation of mass, momentum and energy independent of the numerical method. We observe and justify four different structures of the converged numerical solutions depending on the steepness of the initial conditions. Two of these structures were observed in the literature, with the other two not being commonly found in the literature. The numerical solutions are then used to assess how well the analytical solution of the shallow water wave equations captures the mean behaviour of the solution of the Serre equations for the dam-break problem in the short term. Lastly the numerical solutions are compared to asymptotic results in the literature to approximate the depth and location of the front of an undular bore.

My Contribution:

This paper was produced by me with the support of my coauthors based on my own work.

1.3 Organisation of the Thesis

Chapter 2 proceeds by presenting the one-dimensional Serre equations in conservation law form with a source term, describing their dispersion and conservation properties and then summarising the main results of my investigation into the behaviour of the Serre equations in the presence of steep gradients in the free surface [12].

This is followed by Chapter 3 which describes in detail the developed method. In this thesis the results of other numerical methods are also provided; descriptions of these methods can be found in the literature [13, 12]. However, since the developed method is the preferred method only its description is provided in this thesis.

Chapter 4 provides a linear analysis of the convergence and dispersion properties of the developed finite element volume method in detail. The analysis begins with the linearised Serre equations over a horizontal bed and then derives the evolution matrix; from which the convergence and dispersion properties of the methods can be studied. The results of the linear analysis are also provided for all the methods used by Pitt et al. [12].

The convergence and conservation properties of the numerical methods of Pitt et al. [12] are then assessed in Chapter 5 using analytic and forced solutions of the Serre equations. While Chapter 6 validates the numerical methods against

experimental results.

Finally, Chapter 7 summarises the major contributions and findings of the thesis and presents a summary of the future work for this method.

Chapter 2

The Serre Equations

In this chapter the Serre equations are introduced and their relevant properties are presented.

The Serre equations are a system of partial differential equations that describe the free-surface waves of fluids. They are an approximation to the Euler equations [1]; describing waves in shallow water when the characteristic depth of the water h_0 is far smaller than the characteristic wavelength of the waves λ_0 so that the shallowness parameter $\sigma = h_0/\lambda_0 \ll 1$. Typically, water is considered to be shallow when $\sigma \leq 1/20$ [14].

The Serre equations for one-dimensional flows over horizontal beds were first derived by Serre [15] using asymptotic expansion then later derived using depth integration by Su and Gardner [16]; they are equivalent to the Green-Naghdi equations [17] derived using the theory of directed fluid sheets. The Serre equations were then extended to spatially varying bathymetry by Seabra-Santos et al. [18].

The Serre equations are fully nonlinear and thus applicable across the entire range of characteristic wave amplitudes a_0 which are usually summarised using the nonlinearity parameter $\epsilon = a_0/h_0$. The fluid described by the Serre equations possesses a non hydrostatic pressure distribution and is dispersive in nature, as are real fluids. Furthermore, the dispersion relationship of linear waves of the Serre equations well approximates the linear wave theory for the Euler equations [19]. For these reasons the Serre equations are considered one of the best approximate water wave models [5, 20].

In this chapter we present the one-dimensional Serre equations and a reformulation of these equations into conservation law form. We then present the relevant properties of the Serre equations that will be used to assess the validity

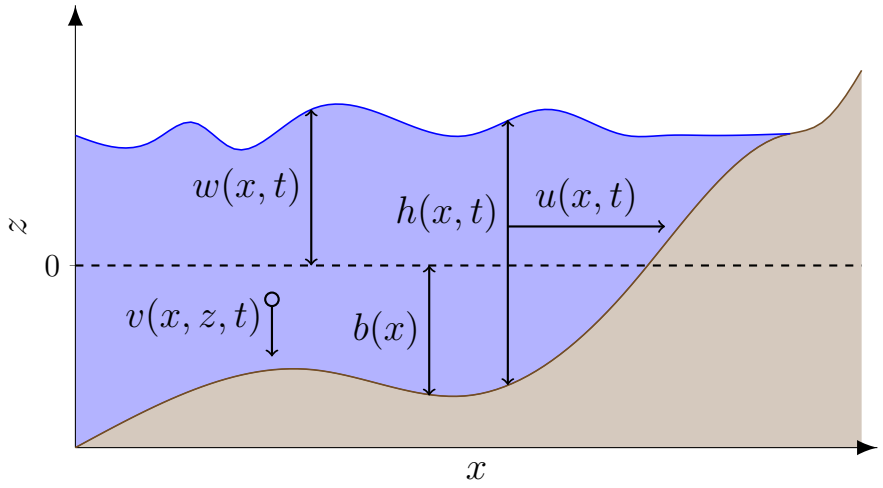


Figure 2.1: Diagram demonstrating the quantities used to describe the fluid (▫) and the bed (▫) for the Serre equations.

of the developed numerical methods, ending with the contribution of my research to understanding the behaviour of steep gradients in the free-surface.

2.1 The Equations

In this thesis we take the Serre equations as derived from the depth-integration approach [16, 18]. Given the extent of the literature already available for the derivation of these equations we will only introduce the relevant quantities and present the equations. Under the depth-integration approach the one-dimensional Serre equations describe the behaviour of unsteady free surface fluid flow for an inviscid fluid with constant density ρ, neglecting wave-breaking and bottom friction. The primitive variables are the height $h(x,t)$ of the free surface above a stationary bed profile given by $b(x)$ and the depth average horizontal velocity $u(x,t)$ which are all demonstrated in Figure 2.1.

Additionally we define the stage $w(x,t) = h(x,t) + b(x)$ which gives the absolute location of the free surface. The derivation is similar to that of the Shallow Water Wave Equations (SWWE) [21], except for the Serre equations the vertical velocity $v(x,z,t)$ varies linearly with depth and is given by [11]

$$v(x,z,t) = u\frac{\partial b}{\partial x} - (z-b)\frac{\partial u}{\partial x}. \tag{2.1}$$

Unlike the SWWE the vertical velocity of the Serre equations is not zero throughout the depth of water consequently, the Serre equations possess a non-hydrostatic

pressure distribution

$$p(x,z,t) = \underbrace{\rho g\left(h+b-z\right)}_{\text{hydrostatic pressure}} + \rho\left(h+b-z\right)\Psi + \frac{1}{2}\rho\left(h^2-\left[z-b\right]^2\right)\Phi \tag{2.2}$$

where

$$\Psi = \frac{\partial b}{\partial x}\left(\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x}\right) + u^2\frac{\partial^2 b}{\partial x^2}, \tag{2.3a}$$

$$\Phi = \frac{\partial u}{\partial x}\frac{\partial u}{\partial x} - u\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial x \partial t}. \tag{2.3b}$$

By depth integrating the Euler equations [16, 11] with a no-slip condition at the bed, a free surface condition at the free surface, the vertical velocity relation (2.1) and the pressure distribution (2.2) we obtain the depth integrated approximation to the conservation of mass and momentum equations

$$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} = 0, \tag{2.4a}$$

$$\frac{\partial (uh)}{\partial t} + \frac{\partial}{\partial x}\left(u^2 h + \frac{gh^2}{2} + \frac{h^2}{2}\Psi + \frac{h^3}{3}\Phi\right) + \frac{\partial b}{\partial x}\left(gh + h\Psi + \frac{h^2}{2}\Phi\right) = 0 \tag{2.4b}$$

When $\Phi = \Psi = 0$ the Serre equations reduce to the SWWE where the vertical velocity is zero, so that only the hydrostatic part of the pressure is present and there is no dispersion.

Due to the presence of the Φ and Ψ terms the Serre equations are more difficult to solve analytically and numerically than the SWWE. The primary reason for this is that whilst the SWWE are hyperbolic the Serre equations are neither hyperbolic nor parabolic. Furthermore, the Serre equations are not in conservation law form due to the presence of temporal derivatives in Φ and Ψ, although they are derived from conservation equations.

For a horizontal bed $\partial b/\partial x = 0$, $\Psi = 0$ and so the Serre equations reduce to

$$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} = 0, \tag{2.5a}$$

$$\frac{\partial (uh)}{\partial t} + \frac{\partial}{\partial x}\left(u^2 h + \frac{gh^2}{2} + \frac{h^3}{3}\Phi\right) = 0. \tag{2.5b}$$

For horizontal beds the Serre equations are more challenging to solve analytically and numerically than the SWWE.

2.1.1 Alternative form of the Serre Equations

A major hurdle for developing numerical methods for the Serre equations is the presence of the temporal derivative in Φ and Ψ (2.3). By rewriting the Serre equations and introducing a new conserved quantity G [22, 11, 23] the Serre equations can be written in conservation law form with a source term

$$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} = 0, \tag{2.6a}$$

$$\begin{aligned} &\frac{\partial G}{\partial t} + \frac{\partial}{\partial x}\left(uG + \frac{gh^2}{2} - \frac{2}{3}h^3 \left[\frac{\partial u}{\partial x}\right]^2 + h^2 u \frac{\partial u}{\partial x}\frac{\partial b}{\partial x}\right) \\ &\qquad \underbrace{+ \frac{1}{2}h^2 u \frac{\partial u}{\partial x}\frac{\partial^2 b}{\partial x^2} - hu^2 \frac{\partial b}{\partial x}\frac{\partial^2 b}{\partial x^2} + gh\frac{\partial b}{\partial x}}_{\text{source term}} = 0. \end{aligned} \tag{2.6b}$$

with

$$G = hu\left(1 + \frac{\partial h}{\partial x}\frac{\partial b}{\partial x} + \frac{1}{2}h\frac{\partial^2 b}{\partial x^2} + \left[\frac{\partial b}{\partial x}\right]^2\right) - \frac{\partial}{\partial x}\left(\frac{1}{3}h^3 \frac{\partial u}{\partial x}\right). \tag{2.7}$$

which resembles h multiplied by the irrotationality [24, 25].

This conservation law form makes the Serre equations well suited to be numerically solved using the finite volume method for the conservation of h and G equations, provided one can solve for u given h and G.

For a horizontal bed $\partial b/\partial x = 0$ the conservation law form of the Serre equations using the new quantity G is

$$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} = 0, \tag{2.8a}$$

$$\frac{\partial G}{\partial t} + \frac{\partial}{\partial x}\left(uG + \frac{gh^2}{2} - \frac{2}{3}h^3 \left[\frac{\partial u}{\partial x}\right]^2\right) = 0 \tag{2.8b}$$

with

$$G = hu - \frac{\partial}{\partial x}\left(\frac{1}{3}h^3 \frac{\partial u}{\partial x}\right). \tag{2.8c}$$

2.2 Properties of the Serre Equations

The Serre equations possess a number of desirable properties for the modelling of water waves; in particular their conservation of fundamental quantities and dispersion relation. If a numerical method accurately approximates the Serre equations then the numerical method should reproduce the conservation and dispersion properties of the Serre equations. In this thesis these properties and the analytic solutions of the Serre equations are used to assess the veracity of the numerical method.

To complement the available analytic solutions, the Serre equations are modified to force certain solutions using a source term, which is called a forced solution. These forced solutions will be used to assess the validity of the numerical methods for a wider array of flow scenarios then possible given the limited number of analytic solutions currently available for the Serre equations.

Finally the results of Pitt et al. [26] for the behaviour of the Serre equations in the presence of steep gradients are summarised. These results demonstrate that the developed numerical method is robust in the presence of steep gradients; one of the main objectives of the thesis.

2.2.1 Conservation Properties

A quantity is conserved if the total amount of a quantity q in a closed system remains constant in time.

Definition 2.1. The total amount of a quantity q in a system occurring on the interval $[a, b]$ at time t is

$$\mathcal{C}_q(t) = \int_a^b q(x,t)\,dx.$$

Using this notation conservation of a quantity q means that $\mathcal{C}_q(0) = \mathcal{C}_q(t)$ for all t.

Integrating the Serre equations in both non-conservation law form (2.4) and conservation law form (2.6) for a closed system we get that h, uh and G are conserved by the Serre equations. Additionally, the Green-Naghdi equations [17] which are equivalent to the Serre equations for one dimensional flows were derived by conserving the energy

$$\mathcal{H}(x,t) = \frac{1}{2}\left(gh\left(h+2b\right) + hu^2 + \frac{h^3}{3}\left[\frac{\partial u}{\partial x}\right]^2 + u^2 h\left[\frac{\partial b}{\partial x}\right]^2 - uh^2\frac{\partial u}{\partial x}\frac{\partial b}{\partial x}\right). \quad (2.9)$$

Therefore, the one dimensional Serre equations should also conserve $\mathcal{H}$. The energy $\mathcal{H}$ is the sum of the gravitational potential energy, the horizontal kinetic energy and the vertical kinetic energy which integrated over the depth of water are

$$\begin{aligned}
&\frac{1}{2}\int_b^{h+b} gz\, dx = \frac{1}{2}gh\left(h+2b\right),\\
&\frac{1}{2}\int_b^{h+b} u^2\, dx = \frac{1}{2}hu^2,\\
&\frac{1}{2}\int_b^{h+b} v^2\, dx = \frac{1}{2}\left(\frac{h^3}{3}\left[\frac{\partial u}{\partial x}\right]^2 + u^2h\left[\frac{\partial b}{\partial x}\right]^2 - uh^2\frac{\partial u}{\partial x}\frac{\partial b}{\partial x}\right),
\end{aligned}$$

respectively. Where the vertical velocity v in the Serre equations is given by (2.1). For horizontal beds $\mathcal{H}$ is the Hamiltonian of the Serre equations [27].

For the system to be closed the flux terms of the equations for h and uh (2.4) at the boundaries must cancel and the integral of the source term over the domain must vanish.

2.2.2 Dispersion Properties

The dispersion properties of wave equations are primarily studied through linearising the equations, assuming periodic wave solutions and then deriving a relationship between the frequency ω and the wave number k of these solutions. For the Serre equations the dispersion relation [23] is

$$\omega^{\pm} = Uk \pm k\sqrt{gH}\sqrt{\frac{3}{(kH)^2+3}} \tag{2.10}$$

where U and H are the mean velocity and height of the fluid respectively and the subscript $\pm$ denotes the positive and negative branches of the dispersion relation. Barthélemy [19] compared this dispersion relation to that of the linear theory of water waves and demonstrated its utility when k is small. However, when k is large the difference between the dispersion relation of the Serre equations and that of the linear water wave theory increases.

From the dispersion relation (2.10) the phase velocity $v_p^{\pm} = \omega^{\pm}/k$ and the group velocity $v_g^{\pm} = \partial\omega^{\pm}/\partial k$ can be written in terms of the wave number as

$$v_p^{\pm} = U \pm \sqrt{gH}\sqrt{\frac{3}{(kH)^2+3}}, \tag{2.11a}$$

$$v_g^{\pm} = U \pm \sqrt{gH}\left(\sqrt{\frac{3}{(kH)^2+3}} \mp (kH)^2\sqrt{\frac{3}{\left([kH]^2+3\right)^3}}\right). \tag{2.11b}$$

Since both the phase and group velocities depend on the wave number, waves of different wavelengths travel at different speeds meaning the Serre equations describe dispersive waves.

Fortunately, the phase velocity and the group velocity of waves are bounded, since as $k \to 0$ then $v_p^{\pm}$ and $v_g^{\pm} \to U \pm \sqrt{gH}$ and as $k \to \infty$ then $v_p^{\pm}$ and $v_g^{\pm} \to U$. Therefore, we have that

$$U - \sqrt{gH} \leq v_p^- \leq U \leq v_p^+ \leq U + \sqrt{gH}, \tag{2.12a}$$
$$U - \sqrt{gH} \leq v_g^- \leq U \leq v_g^+ \leq U + \sqrt{gH}. \tag{2.12b}$$

2.2.3 Analytic Solutions

Few analytic solutions have been discovered for the Serre equations. There is a travelling wave solution for a horizontal bed [28] and the stationary lake at rest solution for arbitrary bathymetry.

Solitary Travelling Wave Solution

The Serre equations admit a travelling wave solution that propagates at a constant speed without deformation due to a balance between the nonlinear and dispersive effects. Unlike the Euler equations this travelling wave solution has a closed form

$$h(x,t) = a_0 + a_1 \text{sech}\left(\kappa\left[x - ct\right]\right), \tag{2.13a}$$

$$u(x,t) = c\left(1 - \frac{a_0}{h(x,t)}\right), \tag{2.13b}$$
$$b(x) = 0 \tag{2.13c}$$

with

$$\kappa = \frac{\sqrt{3a_1}}{2a_0\sqrt{(a_0+a_1)}},$$

$$c = \sqrt{g(a_0+a_1)}.$$

From these equations G and the total amounts of the conserved quantities can be derived, these are presented in Appendix A for reference.

This solitary wave solution has an amplitude of a_1, an infinite wavelength and propagates on water a_0 deep. It is one particular example of a family of smooth periodic travelling wave solutions [28]. However, these solitary wave solutions are not true solitons, due to their inelastic collisions with one another [29].

This analytic solution can only be reproduced with the appropriate order of accuracy if all terms of the Serre equations with a horizontal bed (2.8) are adequately approximated by the numerical method. Furthermore, since this solution is maintained by a balance between nonlinear and dispersive effects it tests the balance of these effects in the numerical method. Therefore, this analytic solution is a good test for assessing the accuracy of numerical methods for solving the Serre equations with a horizontal bed (2.8).

Lake at Rest

The lake at rest solution is a rudimentary stationary solution of the Serre equations that exists for all bathymetry $b(x)$, due to a balance between the hydrostatic pressure distribution and the forcing of the bed slope. The lake at rest solution is

$$h(x,t) = \max\left\{a_0 - b(x), 0\right\}, \tag{2.14a}$$

$$u(x,t) = 0, \tag{2.14b}$$

$$G(x,t) = 0. \tag{2.14c}$$

It represents a quiescent body of water with a horizontal water surface or stage $w(x,t) = h(x,t) + b(x)$ over any bathymetry. The maximum function is included for the water depth to allow for dry regions of the bed when $b(x) > a_0$. We write these quantities in terms of $b(x)$ as this solution holds for all bed profiles. The corresponding total amounts of h and $\mathcal{H}$ can be calculated from (A.3) and (A.4). While the total amounts of u and G are given by (A.2).

Since $w(x,t)$ is constant when $h > 0$ then $\partial w/\partial x = \partial h/\partial x + \partial b/\partial x = 0$ and $u = 0$ so that the Serre equations (2.6) reduce to

$$\frac{\partial h}{\partial t} = 0,$$

$$\frac{\partial G}{\partial t} = 0.$$

Therefore, G and h are constant in time and so is u and thus the solution is stationary.

For naive numerical methods of the Serre equations the hydrostatic pressure and bed slope terms do not completely cancel causing numerical solutions of an initially still lake to produce nonphysical velocities, degrading their convergence to the solution. To combat this, modifications are made to the flux and source term approximations in the finite volume method so that these terms do completely cancel, leading to a so called 'Well-Balanced' method. This analytic solution then provides a test for the effectiveness of these well balancing modifications for the numerical methods.

2.2.4 Forced Solutions

The known analytic solutions of the Serre equations provide a stringent test when the bed is horizontal, as all terms in the equations are non-zero and vary in space and time. For varying bathymetry there is only the lake at rest solution where all terms are constant in time and some vanish. Therefore, the accuracy of the approximations of all terms of the Serre equations in the numerical method is not adequately assessed using only the available analytic solutions.

The verification of the order of accuracy of the numerical methods for transient solutions with varying bathymetry requires the use of forced solutions. To do this we select some particular functions for all of the primitive quantities; h, u and b which we denote by h^*, u^* and b^* respectively. From these functions we calculate

$$S_h = -\frac{\partial h^*}{\partial t} - \frac{\partial (u^* h^*)}{\partial x},$$

$$S_G = -\frac{\partial G^*}{\partial t} - \frac{\partial}{\partial x}\left(u^* G^* + \frac{g\left[h^*\right]^2}{2} - \frac{2}{3}\left[h^*\right]^3 \left[\frac{\partial u^*}{\partial x}\right]^2 + \left[h^*\right]^2 u^* \frac{\partial u^*}{\partial x}\frac{\partial b^*}{\partial x}\right)$$

$$-\frac{1}{2}\left[h^*\right]^2 u^* \frac{\partial u^*}{\partial x}\frac{\partial^2 b^*}{\partial x^2} + h^* [u^*]^2 \frac{\partial b^*}{\partial x}\frac{\partial^2 b^*}{\partial x^2} - g h^* \frac{\partial b^*}{\partial x}.$$

Now h^*, u^* and b^* will be solutions of the forced Serre equations in conservation

law form with a source term

$$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} + S_h = 0, \tag{2.15a}$$

$$\frac{\partial G}{\partial t} + \frac{\partial}{\partial x}\left(uG + \frac{gh^2}{2} - \frac{2}{3}h^3 \left[\frac{\partial u}{\partial x}\right]^2 + h^2 u \frac{\partial u}{\partial x}\frac{\partial b}{\partial x}\right) \tag{2.15b}$$

$$+\frac{1}{2}h^2 u \frac{\partial u}{\partial x}\frac{\partial^2 b}{\partial x^2} - hu^2 \frac{\partial b}{\partial x}\frac{\partial^2 b}{\partial x^2} + gh\frac{\partial b}{\partial x} + S_G = 0.$$

These forced Serre equations are then numerically solved by solving the Serre equations (2.6) with the analytic values of S_h and S_G given h^*, u^* and b^*. So that, the only error for these numerical solutions of the forced Serre equations is the error produced by the numerical methods used to solve the Serre equations.

2.2.5 Behaviour of Steep Gradients

To ensure that the developed numerical methods are robust, their capability to handle initial condition problems with quantities possessing discontinuities must be tested. One group of these initial condition problems that has been of particular interest to the water wave community is the dam-break problem [28, 22, 30, 31, 32]; where a body of water is initially still with a discontinuous jump in its surface between two depth values. So that

$$h(x,0) = \begin{cases} h_l & x < x_0 \\ h_r & x \geq x_0 \end{cases}, \tag{2.16a}$$

$$u(x,0) = 0, \tag{2.16b}$$

$$G(x,0) = 0, \tag{2.16c}$$

$$b(x) = 0. \tag{2.16d}$$

where h_l and h_r are the height to the left and right of x_0, respectively.

Currently, these dam-break problems (2.16) have no known analytic solutions for the Serre equations. However, some insight into the behaviour of the evolution of these initial condition problems has been gained from asymptotic [28] and linear [33] analyses.

There have also been a number of numerical solutions to dam-break problems presented in the literature [28, 22, 30, 31, 32] which have used a variety of numerical methods. Some of these numerical methods cannot handle discontinuous

initial conditions [28, 30, 31, 32] and so smooth approximations to the initial conditions (2.16) were employed. The variety of numerical approaches has lead to different conclusions about the behaviour of the evolution of dam-break problems in the Serre equations in the literature. To resolve these differences a comprehensive review of a particular dam-break problem with a variety of numerical methods and smoothing of the initial conditions was performed [12].

The relevant results garnered from the asymptotic [28] and linear [33] analyses for the evolution of the dam-break problem are presented here, followed by a summary of the results [12], which constituted a significant portion of my research.

Asymptotic and Linear Results

The asymptotic analysis of El et al. [28] used Whitham modulation to study the evolution of dispersive shock waves of the Serre equations as $t \to \infty$. Because, a dispersive shock wave is generated in the evolution of the dam-break problem in the Serre equations; these results are very useful. In particular, they provide a relationship between the beginning heights of the dam-break problem h_l and h_r and the amplitude A^+ and speed S^+ of the initial wave in the produced dispersive wave train

$$\frac{\Delta}{(A^+ + 1)^{1/4}} - \left(\frac{3}{4 - \sqrt{A^+ + 1}}\right)^{21/10} \left(\frac{2}{1 + \sqrt{A^+ + 1}}\right)^{2/5} = 0 \tag{2.17a}$$

$$S^+ = \sqrt{g\,(A^+ + 1)} \tag{2.17b}$$

where

$$\Delta = \frac{\left(\sqrt{\frac{h_l}{h_r}} + 1\right)^2}{4h_r} \tag{2.18}$$

These estimates were found to agree well with numerical simulations provided that $\Delta < 1.43$ [28].

The linear analysis studies the behaviour of the linearised Serre equations to garner insights about the full nonlinear Serre equations (2.4). One of the key results of linear analysis is the dispersion relation (2.10), which was used by Dougalis et al. [33] to argue for a separation of dispersive wave trains in the Serre equations due to the separation of the negative and positive branches of the phase and group velocities. This implies that the structure of dispersive shock waves of the Serre equations should also be two separate dispersive wave trains.

Numerical Solutions for the Smoothed Dam-break Problem

To resolve the differences present in the literature a variety of numerical methods were used to solve the most common class of smoothed versions of the dam-break problem initial conditions (2.16) which are given by

$$h(x,0) = h_r + \frac{h_l - h_r}{2}\left(1 + \tanh\left(\frac{x_0 - x}{\alpha}\right)\right), \tag{2.19a}$$

$$u(x,0) = 0, \tag{2.19b}$$

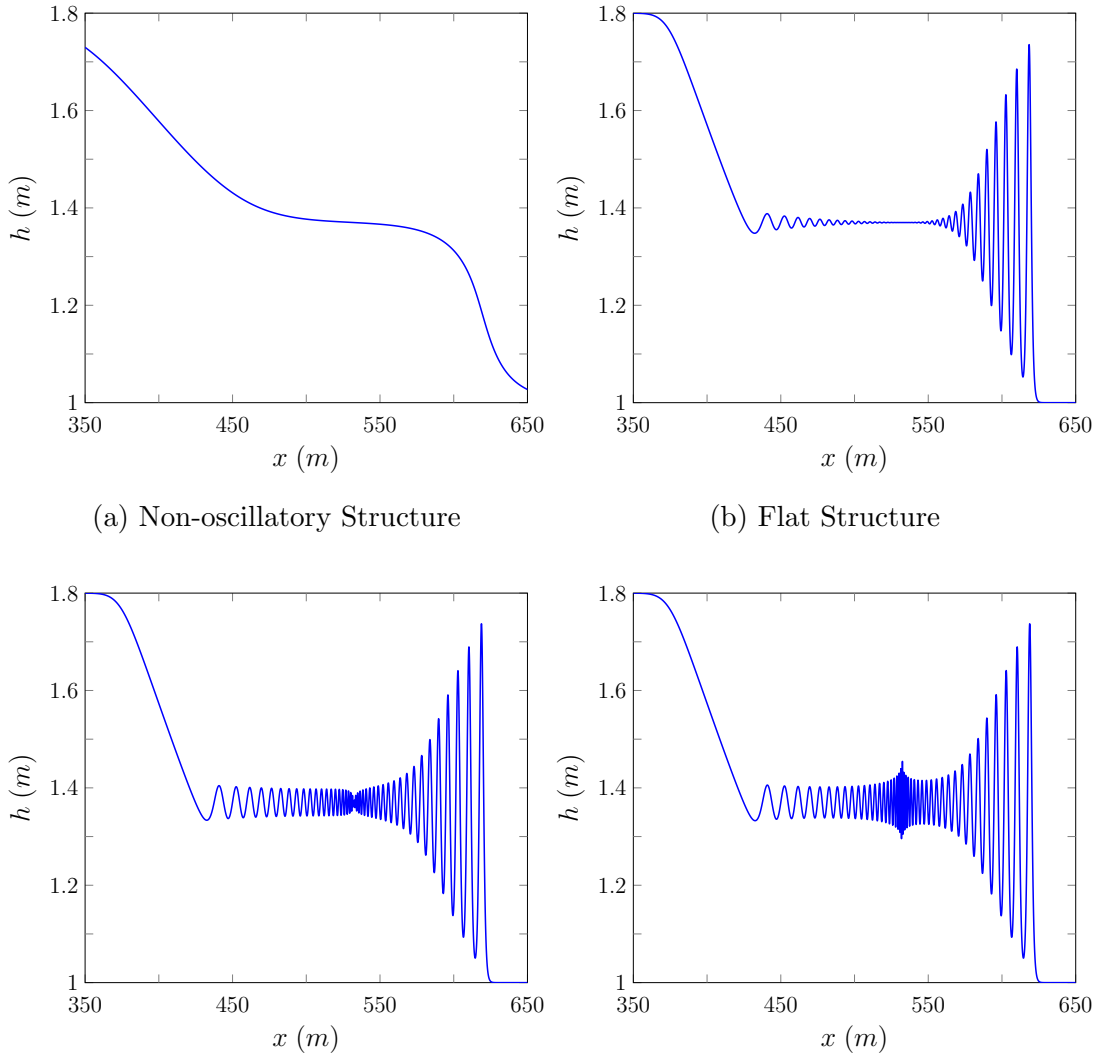
$$G(x,0) = 0, \tag{2.19c}$$

$$b(x) = 0 \tag{2.19d}$$

where α controls the width of the transition from h_l to h_r and thus the steepness of the initial gradients in the water surface. This was dubbed the smoothed dam-break problem and most of the numerical simulations were focused on the case where $h_l = 1.8m$, $h_r = 1m$ and $x_0 = 500m$ with a final time of $t = 30s$. The smoothing parameter α and the resolution of the methods were varied to investigate their influence on the observed behaviour of the numerical solution. Four structures were observed in the numerical solutions; the non-oscillatory structure, the flat structure, the node structure and the growth structure. Example numerical solutions at $t = 30s$ for the mentioned h_l and h_r values and $x_0 = 500m$ demonstrating the observed structures are shown in Figure 2.2.

The growth structure was consistently observed in numerical solutions of the smoothed dam-break problem as $\alpha \to \infty$ for high-order accurate methods on high resolution grids and thus well represents the structure of the solution of the Serre equations for this dam-break problem at $t = 30s$. For this particular dam-break problem, the observation of other behaviours at $t = 30s$ is caused by; small α values which overly smooth the initial conditions, low-order numerical schemes introducing large diffusive errors and low numerical resolutions which cannot resolve the higher frequency waves observed in the growth structure. These structures exist on a spectrum where the severity of these effects determine the observed behaviour. So that, the most severe damping effects produced the non-oscillatory structure and the least severe effects produced the justified growth structure. These effects explained the observations of different structures previously present in the literature [28, 22, 30, 31, 32].

The differences in the observed structures are primarily driven by the different internal structures of the dispersive shock wave, so that for the flat, node and growth structure in Figure 2.2 the front of the dispersive wave trains are



(a) Non-oscillatory Structure

(b) Flat Structure

(c) Node Structure

(d) Growth Structure

Figure 2.2: Comparison of the different structures observed in numerical solutions displayed by Pitt et al. [12].

essentially identical. Therefore, the results of numerical solutions that haven't resolved all the internal structure present in the growth structure still agree well with the Whitham modulation results (2.17) of El et al. [28].

The amplitude of waves at the centre of the growth structure decay over time, resulting in the observation of the flat structure when t is large. These results agree with the linear argument put forth by Dougalis et al. [33]. This indicates that for smaller times the nonlinear terms of the Serre equations play a significant role in the evolution of steep gradients, while for long times the linear terms are dominant and thus a separation of the dispersive wave trains is observed.

In this chapter the Serre equations and their relevant properties were given together with a summary of the main results about evolution of steep gradients in the free surface.

Chapter 3

Finite Element Volume Method

In this chapter the finite element volume method is described in detail.

A variety of numerical methods have been used to solve the Serre equations; from complete finite difference methods [34, 28] and finite element methods [30, 23, 31] to combinations of finite difference and finite volume methods [22, 13]. Splitting techniques have also been employed, most commonly to split the Serre equations into their nonlinear and dispersive parts; resulting in an elliptic operator for the dispersive part and the SWWE for the nonlinear part [35, 29, 36].

For the purposes of development of a numerical method for the two-dimensional Serre equations with variable bathymetry methods that make use of the conservation law form of the Serre equations (2.6) [22, 23, 13] are the most promising. The primary reason for this is that these methods are robust and extend well to unstructured meshes with complex geometries which are the meshes most desirable for modelling physical scenarios. Secondly, to properly handle the elliptic operator produced by the nonlinear and dispersive splitting requires overly restrictive assumptions about the smoothness of the physical quantities, particularly the water depth.

I have developed an extension of the Finite Difference Volume Methods (FDVM) [22, 13] that uses a finite element method in place of the finite difference method. This second-order Finite Element Volume Method (FEVM) which will be referred to as FEVM_2 was a main objective of the Thesis; it consists of two main parts a Finite Element Method (FEM) to solve (2.7) and a Finite Volume Method (FVM) to solve (2.6) hence its name. Making use of these two methods results in a numerical method with a number of desirable properties. It is robust in the presence of discontinuities in the free surface [12], robust during the wetting and drying of beds, has a consistent polynomial representation over the cells and finally all the

terms of the finite volume method can be calculated only knowing the quantities inside the cell. These last two points mean that this method is the best suited of the variant hybrid finite volume methods [13] to solve the two-dimensional Serre equations on unstructured meshes with parallelised code.

In addition to the FEVM_2 the first- and second-order FDVM of Le Métayer et al. [22] and Zoppou et al. [13] were reproduced, these methods will be referred to as FDVM_1 and FDVM_2 respectively. Furthermore the third-order extension FDVM_3 was implemented during my research. I have also reproduced the second-order naive finite difference method [37] and the finite difference method of El et al. [28]; which I refer to as $\mathcal{D}$ and $\mathcal{W}$ respectively. Descriptions of these methods have already been published [13, 12] and so are omitted from the thesis.

In this chapter we introduce the notation for the numerical grids and then describe the second-order Finite Element Volume Method FEVM_2 in detail.

3.1 Notation for Numerical Grids

To produce FEVM_2, time and space will be discretised in different ways; time is broken up into time levels separated by constant durations Δt and space is broken up into cells of constant width Δx. The FEVM can be extended to allow for varying Δt and Δx values, with this description restricted to the constant case for simplicity. The notation for time is quite simple; from an initial time t^0 we define the n^{th} time level where $n \in \mathbb{N}$ to be

$$t^n = t^0 + n\Delta t.$$

The goal of FEVM_2 is to update the quantities at the current time level t^n to the next time level t^{n+1}, solving the equations.

The notation for space is a bit more complicated; as we require definitions of multiple locations inside the cells. The cells are defined by their midpoints; which are given from a starting location x_0, so that the midpoint of the j^{th} cell where $j \in \mathbb{N}$ is

$$x_j = x_0 + j\Delta x.$$

Other points inside the j^{th} cell can be defined in relation to the midpoint so that

$$x_{j+s} = x_j + s\Delta x$$

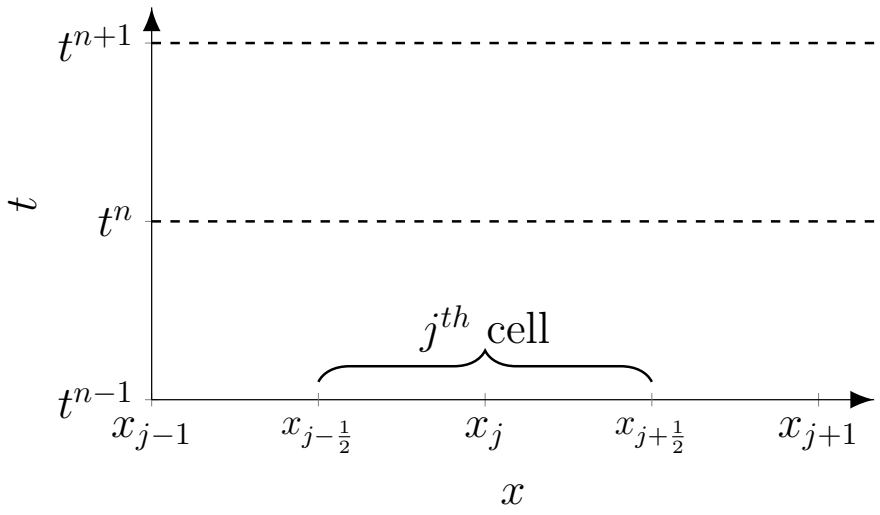


Figure 3.1: The numerical grid in space and time.

where $s \in \left[-\frac{1}{2}, \frac{1}{2}\right] \subset \mathbb{R}$, although for our purposes we restrict ourselves to rational values of s. Using this notation the j^{th} cell spans $\left[x_{j-1/2}, x_{j+1/2}\right]$. These discretisations in space and time result in the grids displayed in Figure 3.1

The temporal and spatial grid notation naturally extends to our quantities of interest, for example, for a general quantity q

$$q_j^n = q(x_j, t^n).$$

These are the nodal values of q. Since the FEVM uses a FVM the cell averages of the quantities are also required. For each cell we define the average of a quantity at time level t^n

$$\overline{q}_j^n = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} q(x, t^n)\, dx$$

over the j^{th} cell.

In the FEVM we reconstruct quantities at various points inside the cell from the cell average values. At the cell edges $x_{j\pm1/2}$, two reconstructions are possible from each of the neighbouring cells, we distinguish between the two possible reconstructions using superscripts. For example, for the cell edge $x_{j+1/2}$ and a general quantity q, there is the reconstructed $q^-_{j+1/2}$ from the leftward j^{th} cell and the reconstructed value $q^+_{j+1/2}$ from the rightward $(j+1)^{th}$ cell.

3.2 Structure Overview

To describe the FEVM we first present an overview of the evolution step and then provide the details for each component. We begin our evolution step with

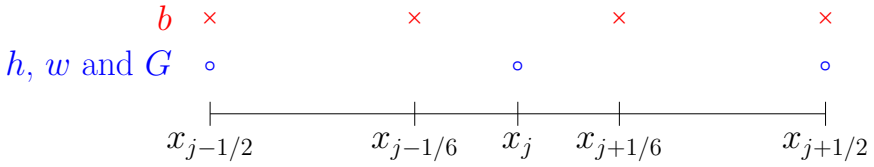


Figure 3.2: Location of the reconstructions for h, w, G and b inside the j^{th} cell.

all the cell averages for h, w and G at time t^n and all the nodal values of b being known. We write these as vectors from the starting 0^{th} to the final m^{th} cell in the following way

$$\overline{\boldsymbol{h}}^n = \begin{bmatrix} \overline{h}_0^n \\ \overline{h}_1^n \\ \vdots \\ \overline{h}_m^n \end{bmatrix}, \qquad \overline{\boldsymbol{w}}^n = \begin{bmatrix} \overline{w}_0^n \\ \overline{w}_1^n \\ \vdots \\ \overline{w}_m^n \end{bmatrix}, \qquad \overline{\boldsymbol{G}}^n = \begin{bmatrix} \overline{G}_0^n \\ \overline{G}_1^n \\ \vdots \\ \overline{G}_m^n \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{bmatrix}.$$

The evolution step proceeds by reconstructing the quantities over the cell, calculating the fluid velocity, approximating the flux, approximating the source term, updating the cell averages and then applying second-order time stepping.

(i) Reconstruction: The locations for the reconstruction of all the quantities in the j^{th} cell are displayed in Figure 3.2. The quantities h, w and G are reconstructed at $x_{j-1/2}$, x_j and $x_{j+1/2}$ from their cell average values using the second-order reconstruction operators $\mathcal{R}^+_{j-1/2}$, $\mathcal{R}_j$ and $\mathcal{R}^-_{j+1/2}$ respectively. While the bed profile b in the j^{th} cell is reconstructed at $x_{j-1/2}$, $x_{j-1/6}$, $x_{j+1/6}$ and $x_{j+1/2}$ from its nodal values using the fourth-order reconstruction operators $\mathcal{B}_{j-1/2}$, $\mathcal{B}_{j-1/6}$, $\mathcal{B}_{j+1/6}$ and $\mathcal{B}_{j+1/2}$ respectively.

So that for a generic quantity q representing h, w and G and the bed b we have

$$q^{\pm}_{j\pm1/2} = \mathcal{R}^{\pm}_{j\pm1/2}\left(\overline{\boldsymbol{q}}^n\right), \qquad b_{j\pm1/2} = \mathcal{B}_{j\pm1/2}\left(\boldsymbol{b}\right),$$
$$q_j = \mathcal{R}_j\left(\overline{\boldsymbol{q}}^n\right), \qquad b_{j\pm1/6} = \mathcal{B}_{j\pm1/6}\left(\boldsymbol{b}\right).$$

To keep the notation simple the time superscript is omitted from the reconstructed quantities. This generates the vectors of these quantities recon-

structed for every cell; $\hat{\boldsymbol{h}}$, $\hat{\boldsymbol{w}}$, $\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{b}}$ at time t^n which are

$$\hat{\boldsymbol{h}} = \begin{bmatrix} h^+_{-1/2} \\ h_0 \\ h^-_{1/2} \\ \vdots \\ h^-_{m+1/2} \end{bmatrix}, \quad \hat{\boldsymbol{w}} = \begin{bmatrix} w^+_{-1/2} \\ w_0 \\ w^-_{1/2} \\ \vdots \\ w^-_{m+1/2} \end{bmatrix}, \quad \hat{\boldsymbol{G}} = \begin{bmatrix} G^+_{-1/2} \\ G_0 \\ G^-_{1/2} \\ \vdots \\ G^-_{m+1/2} \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{b}} = \begin{bmatrix} b_{-1/2} \\ b_{-1/6} \\ b_{1/6} \\ b_{1/2} \\ \vdots \\ b_{m+1/2} \end{bmatrix}.$$

(ii) Fluid Velocity: The remaining unknown quantity, the depth averaged fluid velocity, u is calculated at $x_{j-1/2}$, x_j and $x_{j+1/2}$ in each cell by solving the elliptic equation (2.7) with a second-order FEM. We denote the solution map of the FEM by $\mathcal{G}$, which takes $\hat{\boldsymbol{h}}$, $\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{b}}$ as inputs. So that

$$\hat{\boldsymbol{u}} = \begin{bmatrix} u_{-1/2} \\ u_0 \\ u_{1/2} \\ \vdots \\ u_{m+1/2} \end{bmatrix} = \mathcal{G}\left(\hat{\boldsymbol{h}}, \hat{\boldsymbol{G}}, \hat{\boldsymbol{b}}\right).$$

(iii) Flux Across Cell Interfaces: We calculate the temporally averaged fluxes $F^n_{j-1/2}$ and $F^n_{j+1/2}$ across the cell boundaries $x_{j-1/2}$ and $x_{j+1/2}$ using $\mathcal{F}_{j-1/2}$ and $\mathcal{F}_{j+1/2}$, so that

$$F^n_{j\pm1/2} = \mathcal{F}_{j\pm1/2}\left(\hat{\boldsymbol{h}}, \hat{\boldsymbol{G}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{u}}\right).$$

(iv) Source Terms: We calculate the source term contribution to the cell average of a quantity over a time step; S^n_j with the operator $\mathcal{S}$

$$S^n_j = \mathcal{S}_j\left(\hat{\boldsymbol{h}}, \hat{\boldsymbol{w}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{u}}\right).$$

(v) Update Cell Averages: We update the cell average values from time t^n to t^{n+1} with a forward Euler approximation, resulting in a method that is second-order in space and first-order in time.

(vi) Second-Order SSP Runge-Kutta Method: We repeat steps (i)-(v) and use SSP Runge-Kutta time stepping to obtain $\overline{\boldsymbol{h}}$ and $\overline{\boldsymbol{G}}$ at t^{n+1} with second-order accuracy in space and time.

3.2.1 Reconstruction

We now provide details for the reconstruction of h, w, G and b in the j^{th} cell at the locations shown in Figure 3.2. For h, w and G the reconstructions are performed from the cell averages. While b is reconstructed from the nodal values.

Reconstruction of the h, w and G

We reconstruct h, w and G with piecewise linear functions over a cell from neighbouring cell averages. Since h, w and G use the same reconstruction operators we demonstrate them for a general quantity q. For the j^{th} cell we reconstruct the values of q at $x_{j-1/2}$, x_j and $x_{j+1/2}$ in the following way

$$q^+_{j-1/2} = \mathcal{R}^+_{j-1/2}\left(\overline{\boldsymbol{q}}\right) = \bar{q} - \frac{\Delta x}{2} d_j, \tag{3.1a}$$

$$q_j = \mathcal{R}_j\left(\overline{\boldsymbol{q}}\right) = \bar{q}, \tag{3.1b}$$

$$q^-_{j+1/2} = \mathcal{R}^-_{j+1/2}\left(\overline{\boldsymbol{q}}\right) = \bar{q} + \frac{\Delta x}{2} d_j \tag{3.1c}$$

where

$$d_j = \text{minmod}\left(\theta \frac{\overline{q}_j - \overline{q}_{j-1}}{\Delta x}, \frac{\overline{q}_{j+1} - \overline{q}_{j-1}}{2\Delta x}, \theta \frac{\overline{q}_{j+1} - \overline{q}_j}{\Delta x}\right) \tag{3.2}$$

with $\theta \in [1, 2]$. The choice of the θ parameter changes the diffusion introduced by the reconstruction. When $\theta = 1$ the reconstruction introduces the most diffusion and is equivalent to the minmod reconstruction [38]. When $\theta = 2$ the reconstruction introduces the least diffusion and is equivalent to the monotized central reconstruction [39].

Definition 3.1. The minmod function

$$\text{minmod}\left(a_0, a_1, \dots\right) := \begin{cases} \min\left\{a_i\right\} & a_i > 0 \text{ for all } i \\ \max\left\{a_i\right\} & a_i < 0 \text{ for all } i \\ 0 & \text{otherwise} \end{cases}$$

takes a list of $a_i \in \mathbb{R}$. If all elements have the same sign then minmod returns the element with smallest absolute value, otherwise it returns zero.

The nonlinear limiting used to calculate d_j ensures that the reconstruction of h, w and G inside the cell is Total Variation Diminishing (TVD) [40], hence it does not introduce non-physical oscillations. The TVD property of this reconstruction is achieved by constraining the slope d_j to zero near local extrema, resulting in a piecewise constant reconstruction which is TVD. Away from local extrema d_j

will be the gradient with the smallest absolute value, making our reconstruction second-order accurate.

The reconstruction operator $\mathcal{R}_j$ is second-order accurate regardless of the presence of local extrema. This can be seen through the error analysis of the midpoint quadrature rule [41] for which we have that

$$\overline{q}_j = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} q \, dx = q_j + \mathcal{O}\left(\Delta x^2\right). \tag{3.3}$$

Reconstruction of the Bed Profile

For the bed profile we require a reconstruction that is at least second-order for b, $\partial b/\partial x$ and $\partial^2 b/\partial x^2$. To accomplish this b is reconstructed with a cubic polynomial $C_j(x)$ centred around x_j

$$C_j(x) = c_0 \left(x - x_j\right)^3 + c_1 \left(x - x_j\right)^2 + c_2 \left(x - x_j\right) + c_3. \tag{3.4}$$

By forcing $C_j(x)$ to pass through the nodal values b_{j-2}, b_{j-1}, b_{j+1} and b_{j+2} we get

$$\begin{bmatrix} -8\Delta x^3 & 4\Delta x^2 & -2\Delta x & 1 \\ -\Delta x^3 & \Delta x^2 & -\Delta x & 1 \\ \Delta x^3 & \Delta x^2 & \Delta x & 1 \\ 8\Delta x^3 & 4\Delta x^2 & 2\Delta x & 1 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} b_{j-2} \\ b_{j-1} \\ b_{j+1} \\ b_{j+2} \end{bmatrix}.$$

Solving this we get the polynomial coefficients for $C_j(x)$

$$c_0 = \frac{-b_{j-2} + 2b_{j-1} - 2b_{j+1} + b_{j+2}}{12\Delta x^3},$$

$$c_1 = \frac{b_{j-2} - b_{j-1} - b_{j+1} + b_{j+2}}{6\Delta x^2},$$

$$c_2 = \frac{b_{j-2} - 8b_{j-1} + 8b_{j+1} - b_{j+2}}{12\Delta x},$$

$$c_3 = \frac{-b_{j-2} + 4b_{j-1} + 4b_{j+1} - b_{j+2}}{6}.$$

We require a continuous bed profile across the cell edges and so we average the two reconstructions from the adjacent cells. Therefore, our reconstruction of the

bed profile in the j^{th} cell is the cubic which takes these values

$$b_{j-1/2} = \mathcal{B}_{j-1/2}(\boldsymbol{b}) = \frac{1}{2}\left(C_j(x_{j-1/2}) + C_{j-1}(x_{j-1/2})\right), \tag{3.5a}$$

$$b_{j-1/6} = \mathcal{B}_{j-1/6}(\boldsymbol{b}) = C_j(x_{j-1/6}), \tag{3.5b}$$

$$b_{j+1/6} = \mathcal{B}_{j+1/6}(\boldsymbol{b}) = C_j(x_{j+1/6}), \tag{3.5c}$$

$$b_{j+1/2} = \mathcal{B}_{j+1/2}(\boldsymbol{b}) = \frac{1}{2}\left(C_j(x_{j+1/2}) + C_{j+1}(x_{j+1/2})\right). \tag{3.5d}$$

3.2.2 Fluid Velocity

The elliptic equation that relates the conserved variables h and G and the bed profile b to the primitive variable u was given in (2.7). For the FEM we begin with the weak form of (2.7) with a test function v over the spatial domain Ω which is

$$\int_\Omega Gv\,dx = \int_\Omega uh\left(1 + \frac{\partial h}{\partial x}\frac{\partial b}{\partial x} + \frac{1}{2}h\frac{\partial^2 b}{\partial x^2} + \left[\frac{\partial b}{\partial x}\right]^2\right)v - \frac{\partial}{\partial x}\left(\frac{1}{3}h^3\frac{\partial u}{\partial x}\right)v\,dx.$$

Integrating by parts with zero Dirichlet boundary conditions we get

$$\int_\Omega Gv\,dx = \int_\Omega uh\left(1 + \left[\frac{\partial b}{\partial x}\right]^2\right)v\,dx + \int_\Omega \frac{1}{3}h^3\frac{\partial u}{\partial x}\frac{\partial v}{\partial x}\,dx - \int_\Omega \frac{1}{2}uh^2\frac{\partial b}{\partial x}\frac{\partial v}{\partial x}\,dx. - \int_\Omega \frac{1}{2}h^2\frac{\partial b}{\partial x}\frac{\partial u}{\partial x}v\,dx. \tag{3.6}$$

By assuming that time is fixed, making all the functions only functions in space, this formulation implies that by ensuring that G, h, b and $\partial b/\partial x$ have finite integrals over Ω, then u and $\partial u/\partial x$ must have finite integrals too. Since we require $\partial u/\partial x$ to be well defined to approximate the fluxes and the source term (2.6) and thus have finite integrals we will assume that for each time t that $h, G \in \mathbb{L}^2(\Omega)$ and $b \in \mathbb{W}^{1,2}(\Omega)$ so that $u \in \mathbb{W}^{1,2}(\Omega)$. See Appendix B for a precise definition of $\mathbb{L}^2(\Omega)$ and $\mathbb{W}^{1,2}(\Omega)$.

We simplify (3.6) by performing the integration over the cells and then summing the integrals together to get the equation for the entire domain

$$\sum_j \left(\int_{x_{j-1/2}}^{x_{j+1/2}} \left[\left(uh\left(1 + \left[\frac{\partial b}{\partial x}\right]^2\right) - \frac{1}{2}h^2\frac{\partial b}{\partial x}\frac{\partial u}{\partial x} - G\right)v + \left(\frac{1}{3}h^3\frac{\partial u}{\partial x} - \frac{1}{2}uh^2\frac{\partial b}{\partial x}\right)\frac{\partial v}{\partial x}\right]dx\right) = 0 \tag{3.7}$$

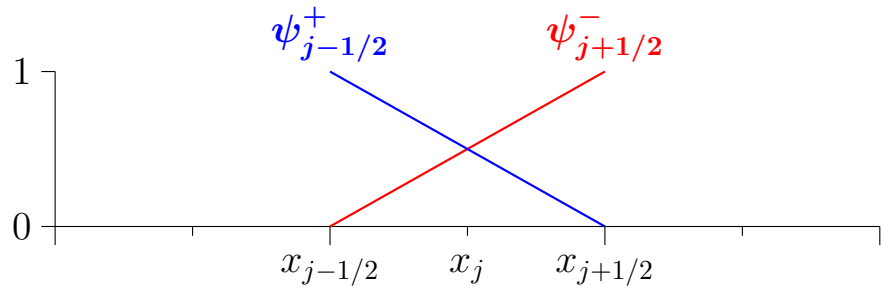


Figure 3.3: Discontinuous linear basis functions over a cell.

which holds for all test functions v. The next step is to replace the functions for the quantities h, G, b and u with their corresponding basis function approximations.

Basis Function Approximations

For h and G we use the basis functions ψ (B.1) which are linear inside a cell and zero elsewhere and so are not continuous as shown in Figure 3.3. This is consistent with our reconstruction which is second-order accurate inside the cell and possesses discontinuities at the cell edges. Since these basis functions are in $\mathbb{L}^2(\Omega)$ our basis function approximations to h and G are in the appropriate function space.

From the basis functions ψ we have the following representation for h and G in our FEM written for the generic quantity q

$$q = \sum_j \left(q^+_{j-1/2}\psi^+_{j-1/2} + q^-_{j+1/2}\psi^-_{j+1/2}, \right). \tag{3.8}$$

To calculate the flux and source terms in (2.6b) we require a locally calculated second-order approximation to the first derivative of u. To do this we require a quadratic representation of u in each cell and since we desire $u \in \mathbb{W}^{1,2}(\Omega)$, this representation will be continuous across the cell edges $x_{j\pm1/2}$. Therefore, we use the continuous quadratic basis functions $\phi_{j\pm1/2}$ and ϕ_j (B.2) depicted in Figure 3.4.

From the basis functions ϕ our basis function approximation to u is

$$u = u_{-1/2}\phi_{-1/2} + \sum_j \left(u_j\phi_j + u_{j+1/2}\phi_{j+1/2} \right) \tag{3.9}$$

For the source term of the evolution of G equation (2.6b) we require a local approximation to the second derivative of the bed that is also second-order accurate. To allow for an appropriate second derivative of the bed profile, b must

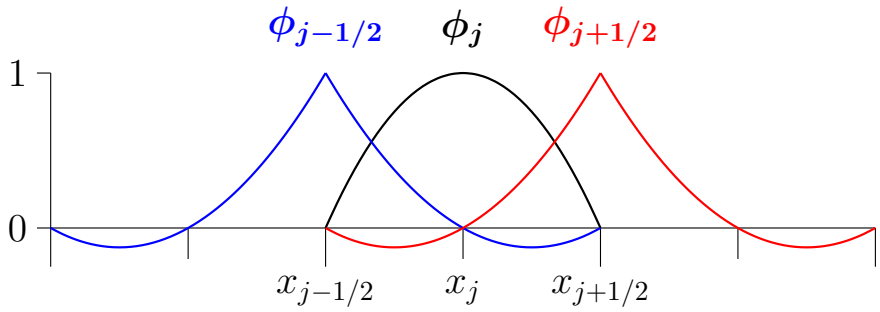


Figure 3.4: Continuous piecewise quadratic basis functions over a cell.

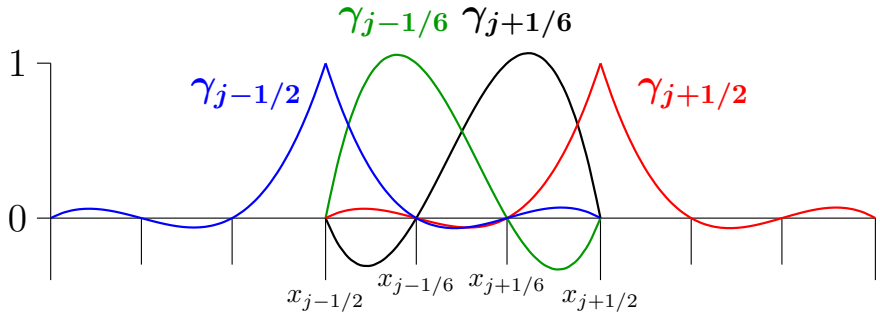


Figure 3.5: Continuous piecewise cubic basis functions over a cell.

be a member of $\mathbb{W}^{2,2}(\Omega)$ which is smoother than required by the elliptic equation (3.6). We choose the cubic basis functions γ (B.3) which are continuous across the cell edges, as the bed profile will be continuous. These basis functions are shown in Figure 3.5 and from them we get our basis function approximation to b

$$b = b_{-1/2}\gamma_{-1/2} + \sum_j \left(b_{j-1/6}\gamma_{j-1/6} + b_{j+1/6}\gamma_{j+1/6} + b_{j+1/2}\gamma_{j+1/2} \right). \tag{3.10}$$

Calculation of Element-wise Matrices

The integral equation (3.7) holds for all v. However, since our solution space has the basis functions ϕ it is sufficient to satisfy (3.7) for all ϕ to generate the solution. Since only the basis functions $\phi_{j-1/2}$, ϕ_j and $\phi_{j+1/2}$ are non-zero over

the j^{th} cell we can calculate the j^{th} term in the sum (3.7) like so

$$\int_{x_{j-1/2}}^{x_{j+1/2}} \left(\left[uh \left(1 + \left[\frac{\partial b}{\partial x} \right]^2 \right) - \frac{1}{2} h^2 \frac{\partial b}{\partial x} \frac{\partial u}{\partial x} - G \right] \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} + \left[\frac{1}{3} h^3 \frac{\partial u}{\partial x} - \frac{1}{2} h^2 \frac{\partial b}{\partial x} u \right] \frac{\partial}{\partial x} \left(\begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} \right) \right) dx \quad (3.11)$$

where we use our finite element approximations for h (3.8), G (3.8), u (3.9) and b (3.10). This integral can be generalised by moving to the natural reference ξ-space, as the basis functions which are non-zero in one element are just translations of the non-zero basis functions in another element. The mapping from the x-space to the ξ-space where $\xi \in [-1, 1]$ is

$$x = x_j + \xi \frac{\Delta x}{2}.$$

Making the change of variables from x to ξ in (3.11) we get

$$\frac{\Delta x}{2} \int_{-1}^{1} \left(\left[uh \left(1 + \frac{4}{\Delta x^2} \left[\frac{\partial b}{\partial \xi} \right]^2 \right) - \frac{2}{\Delta x^2} h^2 \frac{\partial b}{\partial \xi} \frac{\partial u}{\partial \xi} - G \right] \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} + \frac{4}{\Delta x^2} \left[\frac{1}{3} h^3 \frac{\partial u}{\partial \xi} - \frac{1}{2} h^2 \frac{\partial b}{\partial \xi} u \right] \frac{\partial}{\partial \xi} \left(\begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} \right) \right) d\xi.$$

We will demonstrate the rest of the process for the uh term as an example with the remaining integrals provided online (https://sites.google.com/view/jordanpitt/phd-thesis-resources/finite-element-integrals). The uh term is

$$\frac{\Delta x}{2} \int_{-1}^{1} uh \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} d\xi.$$

Since the integral is computed over $[-1, 1]$, there are only a few non-zero contri-

butions from the finite element approximations to h and u, so we have

$$\frac{\Delta x}{2}\int_{-1}^{1}\Bigg(\left(u_{j-1/2}\phi_{j-1/2}+u_j\phi_j+u_{j+1/2}\phi_{j+1/2}\right)\times\left(h^+_{j-1/2}\psi^+_{j-1/2}+h^-_{j+1/2}\psi^-_{j+1/2}\right)\begin{bmatrix}\phi_{j-1/2}\\ \phi_j\\ \phi_{j+1/2}\end{bmatrix}\Bigg)d\xi$$

$$=\frac{\Delta x}{2}\Bigg(h^+_{j-1/2}\int_{-1}^{1}\psi^+_{j-1/2}\begin{bmatrix}\phi_{j-1/2}\phi_{j-1/2} & \phi_j\phi_{j-1/2} & \phi_{j+1/2}\phi_{j-1/2}\\ \phi_{j-1/2}\phi_j & \phi_j\phi_j & \phi_{j+1/2}\phi_j\\ \phi_{j+1/2}\phi_{j-1/2} & \phi_{j+1/2}\phi_j & \phi_{j+1/2}\phi_{j+1/2}\end{bmatrix}d\xi$$

$$+h^-_{j+1/2}\int_{-1}^{1}\psi^-_{j+1/2}\begin{bmatrix}\phi_{j-1/2}\phi_{j-1/2} & \phi_j\phi_{j-1/2} & \phi_{j+1/2}\phi_{j-1/2}\\ \phi_{j-1/2}\phi_j & \phi_j\phi_j & \phi_{j+1/2}\phi_j\\ \phi_{j+1/2}\phi_{j-1/2} & \phi_{j+1/2}\phi_j & \phi_{j+1/2}\phi_{j+1/2}\end{bmatrix}d\xi\Bigg)\begin{bmatrix}u_{j-1/2}\\ u_j\\ u_{j+1/2}\end{bmatrix}.$$

Calculating the integrals of all the basis function combinations we get

$$\frac{\Delta x}{2}\int_{-1}^{1}uh\begin{bmatrix}\phi_{j-1/2}\\ \phi_j\\ \phi_{j+1/2}\end{bmatrix}d\xi=$$

$$\frac{\Delta x}{60}\begin{bmatrix}7h^+_{j-1/2}+h^-_{j+1/2} & 4h^+_{j-1/2} & -h^+_{j-1/2}-h^-_{j+1/2}\\ 4h^+_{j-1/2} & 16h^+_{j-1/2}+16h^-_{j+1/2} & 4h^-_{j+1/2}\\ -h^+_{j-1/2}-h^-_{j+1/2} & 4h^-_{j+1/2} & h^+_{j-1/2}+7h^-_{j+1/2}\end{bmatrix}\begin{bmatrix}u_{j-1/2}\\ u_j\\ u_{j+1/2}\end{bmatrix}. \tag{3.12}$$

Assembly of the Global Matrix

By combining all the matrices generated by the integral of each of the u terms we get the contribution of the j^{th} cell to the stiffness matrix $\mathbf{A}_j$. Likewise all the integrals of the remaining term Gv in (3.7) generate the element wise vector $\boldsymbol{g}_j$. These element wise matrices and vectors are then assembled into the global stiffness matrix $\mathbf{A}$ and the global right hand-side term $\boldsymbol{g}$ thus (3.7) is rewritten as

$$\mathbf{A}\hat{\boldsymbol{u}}=\boldsymbol{g}. \tag{3.13}$$

This is a penta-diagonal matrix equation which can be solved by direct banded matrix solution techniques such as those of Press et al. [42] to obtain

$$\hat{\boldsymbol{u}} = \mathcal{G}\left(\hat{\boldsymbol{h}}, \hat{\boldsymbol{G}}, \hat{\boldsymbol{b}}\right) = \mathbf{A}^{-1}\boldsymbol{g} \tag{3.14}$$

as desired.

3.2.3 Flux Across the Cell Interfaces

We use the method of Kurganov et al. [43] to calculate the flux across a cell interface. This method was employed because it can handle discontinuities across the cell boundary and only requires an estimate of the maximum and minimum wave speeds. This is precisely the situation for the Serre equations which do not have a known expression for the characteristics but do possess estimates on the maximum and minimum wave speeds (2.12a).

Only the calculation of the flux term $F_{j+1/2}$ is demonstrated as the process to calculate the flux term $F_{j-1/2}$ is identical but with different cells. For a general quantity q the method of Kurganov et al. [43] is

$$F_{j+\frac{1}{2}} = \frac{a^+_{j+\frac{1}{2}} f\left(q^-_{j+\frac{1}{2}}\right) - a^-_{j+\frac{1}{2}} f\left(q^+_{j+\frac{1}{2}}\right)}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}} + \frac{a^+_{j+\frac{1}{2}}\, a^-_{j+\frac{1}{2}}}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}}\left(q^+_{j+\frac{1}{2}} - q^-_{j+\frac{1}{2}}\right) \tag{3.15}$$

where $a^+_{j+\frac{1}{2}}$ and $a^-_{j+\frac{1}{2}}$ are given by bounds on the wave speed. Applying the wave speed bounds (2.12a) we obtain

$$a^-_{j+\frac{1}{2}} = \min\left\{0\ ,\ u^-_{j+1/2} - \sqrt{gh^-_{j+1/2}}\ ,\ u^+_{j+1/2} - \sqrt{gh^+_{j+1/2}}\right\}, \tag{3.16}$$

$$a^+_{j+\frac{1}{2}} = \max\left\{0\ ,\ u^-_{j+1/2} + \sqrt{gh^-_{j+1/2}}\ ,\ u^+_{j+1/2} + \sqrt{gh^+_{j+1/2}}\right\}. \tag{3.17}$$

The flux functions $f(q^-_{j+\frac{1}{2}})$ and $f(q^+_{j+\frac{1}{2}})$ are evaluated using the reconstructed values $q^-_{j+\frac{1}{2}}$ and $q^+_{j+\frac{1}{2}}$ from the j^{th} and $(j+1)^{th}$ cell respectively. From the continuity equation (2.6a) we have

$$f\left(h^\pm_{j+\frac{1}{2}}\right) = u^\pm_{j+1/2} h^\pm_{j+1/2}.$$

For the evolution of G equation (2.6b) we have

$$\begin{aligned} f\left(G^\pm_{j+\frac{1}{2}}\right) = u^\pm_{j+1/2} G^\pm_{j+1/2} + \frac{g}{2}\left(h^\pm_{j+1/2}\right)^2 - \frac{2}{3}\left(h^\pm_{j+1/2}\right)^3 \left[\left(\frac{\partial u}{\partial x}\right)^\pm_{j+1/2}\right]^2 \\ + \left(h^\pm_{j+1/2}\right)^2 u^\pm_{j+1/2}\left(\frac{\partial u}{\partial x}\right)^\pm_{j+1/2}\left(\frac{\partial b}{\partial x}\right)^\pm_{j+1/2}. \end{aligned} \tag{3.18}$$

The quantities $h^+_{j-1/2}$, $h^-_{j+1/2}$, $G^+_{j-1/2}$, $G^-_{j+1/2}$ were calculated during the reconstruction and the FEM provided $u^\pm_{j+1/2} = u_{j+1/2}$ as u is continuous across the cell boundaries.

Approximations to $\left(\frac{\partial b}{\partial x}\right)^\pm_{j+1/2}$ and $\left(\frac{\partial u}{\partial x}\right)^\pm_{j+1/2}$ are now required to calculate the flux (3.18).

Calculation of Derivatives

To calculate the derivatives in u and b we use the basis function approximation to these quantities in the FEM to define the reconstruction polynomial of these quantities over a cell. For u we have the quadratic $P^u_j(x)$ that passes through $u_{j-1/2}$, u_j and $u_{j+1/2}$ while for b we have the cubic $P^b_j(x)$ that passes through $b_{j-1/2}$, $b_{j-1/6}$, $b_{j+1/6}$ and $b_{j+1/2}$. So we have

$$P^u_j(x) = p^u_0 (x - x_j)^2 + p^u_1 (x - x_j) + p^u_2, \tag{3.19a}$$
$$P^b_j(x) = p^b_0 (x - x_j)^3 + p^b_1 (x - x_j)^2 + p^b_2 (x - x_j) + p^b_3, \tag{3.19b}$$

For $P^u_j(x)$ we obtain

$$\begin{aligned} p^u_0 &= \frac{u_{j-1/2} - 2u_j + u_{j+1/2}}{2\Delta x^2}, \\ p^u_1 &= \frac{-u_{j-1/2} + u_{j+1/2}}{\Delta x}, \\ p^u_2 &= u_j. \end{aligned}$$

While for $P^b_j(x)$ we obtain

$$p^b_0 = \frac{-9b_{j-1/2} + 27b_{j-1/6} - 27b_{j+1/6} + 9b_{j+1/2}}{2\Delta x^3},$$

$$p^b_0 = \frac{9b_{j-1/2} - 9b_{j-1/6} - 9b_{j+1/6} + 9b_{j+1/2}}{4\Delta x^2},$$

$$p^b_0 = \frac{b_{j-1/2} - 27b_{j-1/6} + 27b_{j+1/6} - b_{j+1/2}}{8\Delta x},$$

$$p^b_0 = \frac{-b_{j-1/2} + 9b_{j-1/6} + 9b_{j+1/6} - b_{j+1/2}}{16}.$$

Taking the derivative of the polynomials (3.19) we get

$$\frac{\partial}{\partial x}P_j^u(x) = 2p_0^u\,(x - x_j) + p_1^u,$$
$$\frac{\partial}{\partial x}P_j^b(x) = 3p_0^b\,(x - x_j)^2 + 2p_1^b\,(x - x_j) + p_2^b.$$

This gives a second-order approximation to the derivative of u and b at $x_{j+1/2}$ for the j^{th} cell. The process for the $(j+1)^{th}$ cell is the same and we get

$$\left(\frac{\partial u}{\partial x}\right)^-_{j+1/2} = \frac{\partial}{\partial x}P_j^u(x_{j+1/2}), \tag{3.20a}$$
$$\left(\frac{\partial u}{\partial x}\right)^+_{j+1/2} = \frac{\partial}{\partial x}P_{j+1}^u(x_{j+1/2}), \tag{3.20b}$$
$$\left(\frac{\partial b}{\partial x}\right)^-_{j+1/2} = \frac{\partial}{\partial x}P_j^b(x_{j+1/2}), \tag{3.20c}$$
$$\left(\frac{\partial b}{\partial x}\right)^+_{j+1/2} = \frac{\partial}{\partial x}P_{j+1}^b(x_{j+1/2}). \tag{3.20d}$$

Therefore, we possess all the terms needed to calculate the approximation to the flux (3.15) for both h and G, as desired. However, to ensure that the FEVM is well balanced and recovers the lake at rest steady state solution, these fluxes must be modified.

Well Balancing

To recover the lake at rest steady state solution we follow the work of Audusse et al. [44], who accomplished this for the SWWE. Earlier we have demonstrated that this process can be extended to the Serre equations [37]. To enforce well balancing the reconstruction of h is modified at the cell edges in the following way; first calculate

$$\dot{b}^-_{j+1/2} = w^-_{j+1/2} - h^-_{j+1/2} \qquad \text{and} \qquad \dot{b}^+_{j+1/2} = w^+_{j+1/2} - h^+_{j+1/2}. \tag{3.21}$$

Find the maximum

$$\ddot{b}_{j+1/2} = \max\left\{\dot{b}^-_{j+1/2}, \dot{b}^+_{j+1/2}\right\}$$

then define

$$\ddot{h}^-_{j+1/2} = \max\left\{0, w^-_{j+1/2} - \ddot{b}_{j+1/2}\right\}, \tag{3.22a}$$
$$\ddot{h}^+_{j+1/2} = \max\left\{0, w^+_{j+1/2} - \ddot{b}_{j+1/2}\right\}. \tag{3.22b}$$

This generates the vector $\ddot{\boldsymbol{h}}$

$$\ddot{\boldsymbol{h}} = \begin{bmatrix} \ddot{h}^{+}_{-1/2} \\ h_0 \\ \ddot{h}^{-}_{1/2} \\ \vdots \\ \ddot{h}^{-}_{m+1/2} \end{bmatrix}$$

which we use to calculate the flux term $F_{j+1/2}$ in (3.15) for h and G. Applying the same process but with different cells we obtain $F_{j-1/2}$ and we have

$$F^n_{j\pm1/2} = \mathcal{F}_{j\pm1/2}\left(\ddot{\boldsymbol{h}}, \hat{\boldsymbol{G}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{u}}\right). \tag{3.23}$$

for the evolution of h and G equations as desired.

3.2.4 Source Terms

To evolve the Serre equations (2.6) to the next time level, we require an approximation to the source term at the cell centre x_j which we denote as S_j. Equation (2.6a) has no source term, therefore we just present the calculation of the source term for equation (2.6b).

Following the work of Audusse et al. [44], we split our approximation to S^n_j into the centred source term S_{ci} and the corrective interface source terms $S^-_{j+\frac{1}{2}}$ and $S^+_{j+\frac{1}{2}}$

$$S^n_j = S^-_{j+\frac{1}{2}} + \Delta x S_{ci} + S^+_{j-\frac{1}{2}}.$$

Where S_{ci} is the naive source term approximation and $S^-_{j+\frac{1}{2}}$ and $S^+_{j+\frac{1}{2}}$ are correction terms that ensure that the flux and source term cancel for the lake at rest solution.

We calculate the centred source term using

$$S_{ci} = -\frac{1}{2}\left(h_j\right)^2 u_j \left(\frac{\partial u}{\partial x}\right)_j \left(\frac{\partial^2 b}{\partial x^2}\right)_j + h_j \left(u_j\right)^2 \left(\frac{\partial b}{\partial x}\right)_j \left(\frac{\partial^2 b}{\partial x^2}\right)_j - g h_j \left(\frac{\partial b}{\partial x}\right)_j.$$

Where we use h_j from the reconstruction process (3.1) and u_j from the solution of the elliptic equation (3.14). To calculate the derivatives we employ our polynomial representations of u and b inside a cell (3.19). However, to ensure that the terms cancel properly for a lake at rest we modify our approximation to $\dfrac{\partial b}{\partial x}$ to use

$\dot{b}^-_{j+1/2}$ and $\dot{b}^+_{j+1/2}$ from (3.21). Therefore, the following approximations are used to calculate S_{ci}

$$\left(\frac{\partial u}{\partial x}\right)_j = \frac{\partial}{\partial x} P^u_j(x_j), \tag{3.24a}$$

$$\left(\frac{\partial b}{\partial x}\right)_j = \frac{\dot{b}^-_{j+1/2} - \dot{b}^+_{j-1/2}}{\Delta x}, \tag{3.24b}$$

$$\left(\frac{\partial^2 b}{\partial x^2}\right)_j = \frac{\partial^2}{\partial x^2} P^b_j(x_j). \tag{3.24c}$$

To ensure well balancing the corrective interface source terms

$$S^-_{j+\frac{1}{2}} = \frac{g}{2}\left(\ddot{h}^-_{j+\frac{1}{2}}\right)^2 - \frac{g}{2}\left(h^-_{j+\frac{1}{2}}\right)^2,$$

$$S^+_{j-\frac{1}{2}} = \frac{g}{2}\left(h^+_{j-\frac{1}{2}}\right)^2 - \frac{g}{2}\left(\ddot{h}^+_{j-\frac{1}{2}}\right)^2$$

are also added. These corrective terms make use of $h^-_{j+\frac{1}{2}}$ and $h^+_{j+\frac{1}{2}}$ obtained from the reconstruction (3.1) and the modified values $\ddot{h}^-_{j+\frac{1}{2}}$ and $\ddot{h}^+_{j+\frac{1}{2}}$ from (3.22). Combining the centred and interface source terms our approximation to the source term for G is

$$S^n_j = \boldsymbol{\mathcal{S}}_j\left(\hat{\boldsymbol{h}}, \ddot{\boldsymbol{h}}, \hat{\boldsymbol{w}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{u}}\right) = S^-_{j+\frac{1}{2}} + \Delta x S_{ci} + S^+_{j-\frac{1}{2}}. \tag{3.25}$$

3.2.5 Update Cell Averages

Applying a forward Euler approximation with our approximation to the flux and source terms we get that

$$\overline{q}^{n+1}_j = \overline{q}^n_j + \frac{\Delta t}{\Delta x}\left(F^n_{j+\frac{1}{2}} - F^n_{j-\frac{1}{2}} + S^n_j\right) \tag{3.26}$$

where $F^n_{j+\frac{1}{2}}$, $F^n_{j-\frac{1}{2}}$ and S^n_j are all calculated using the quantities at time t^n. This update formula is first-order in time.

3.2.6 Second-Order SSP Runge-Kutta Method

To increase the order of accuracy in time we can employ the strong stability preserving Runge-Kutta method [45] which is a convex combination of the first-

order time steps (3.26) in the following way

$$\overline{q}_j^{(1)} = \overline{q}_j^n + \frac{\Delta t}{\Delta x}\left(F^n_{j+\frac{1}{2}} - F^n_{j-\frac{1}{2}} + S^n_j\right), \tag{3.27a}$$

$$\overline{q}_j^{(2)} = \overline{q}_j^{(1)} + \frac{\Delta t}{\Delta x}\left(F^{(1)}_{j+\frac{1}{2}} - F^{(1)}_{j-\frac{1}{2}} + S^{(1)}_j\right), \tag{3.27b}$$

$$\overline{q}_j^{n+1} = \frac{1}{2}\left(\overline{q}_j^n + \overline{q}_j^{(2)}\right). \tag{3.27c}$$

This results in a time stepping method that preserves the stability of the first-order method (3.26) and is second-order accurate in time. Since all the spatial approximations are second-order accurate, the steps (i-vi) should result in a second-order accurate FEVM for the Serre equations, as desired.

3.3 CFL condition

To ensure the stability of our FEVM we use the Courant-Friedrichs-Lewy (CFL) condition [46] which is necessary for stability. The CFL condition ensures that time steps are small enough so that information is only transferred between neighbouring cells. For the Serre equations the CFL condition is

$$\Delta t \leq \frac{Cr}{\max_j\left\{a^{\pm}_{j+1/2}\right\}}\Delta x \tag{3.28}$$

where $a^{\pm}_{j+1/2}$ are the wave-speed bounds used in the flux approximation (3.17) and $0 \leq Cr \leq 1$ is the Courant number. Typically, we use the conservative $Cr = 0.5$ for our numerical experiments.

3.4 Boundary Conditions

To numerically model the Serre equations over finite spatial domains we must enforce boundary conditions at the left and right edge of the domain; $x_{-1/2}$ and $x_{m+1/2}$ respectively. We have only developed Dirichlet boundary conditions for the FEVM, which we enforce using ghost cells located outside the domain boundaries. These ghost cells contain the complete representation of their respective quantities over the cell. For h, w, G and u only one ghost cell at each boundary is required, while for b we require two ghost cells at each boundary. The ghost

cells for h, w and G written for a generic quantity q are

$$\hat{\boldsymbol{q}}_{-1} = \begin{bmatrix} q^+_{-3/2} \\ q_{-1} \\ q^-_{-1/2} \end{bmatrix}, \qquad \hat{\boldsymbol{q}}_{m+1} = \begin{bmatrix} q^+_{m+1/2} \\ q_{m+1} \\ q^-_{m+3/2} \end{bmatrix}.$$

For u and b the ghost cells are

$$\hat{\boldsymbol{u}}_{-1} = \begin{bmatrix} u_{-3/2} \\ u_{-1} \\ u_{-1/2} \end{bmatrix}, \qquad \hat{\boldsymbol{u}}_{m+1} = \begin{bmatrix} u_{m+1/2} \\ u_{m+1} \\ u_{m+3/2} \end{bmatrix},$$

$$\hat{\boldsymbol{b}}_{-2} = \begin{bmatrix} b_{-5/2} \\ b_{-13/6} \\ b_{-11/6} \\ b_{-3/2} \end{bmatrix}, \quad \hat{\boldsymbol{b}}_{-1} = \begin{bmatrix} b_{-3/2} \\ b_{-7/6} \\ b_{-5/6} \\ b_{-1/2} \end{bmatrix}, \quad \hat{\boldsymbol{b}}_{m+1} = \begin{bmatrix} b_{m+1/2} \\ b_{m+5/6} \\ b_{m+7/6} \\ b_{m+3/2} \end{bmatrix}, \quad \hat{\boldsymbol{b}}_{m+2} = \begin{bmatrix} b_{m+3/2} \\ b_{m+11/6} \\ b_{m+13/6} \\ b_{m+5/2} \end{bmatrix}.$$

To ensure that the solution of u by (3.14) agrees with the boundary conditions $\hat{\boldsymbol{u}}_{-1}$ and $\hat{\boldsymbol{u}}_m$ the element-wise stiffness matrices $\mathbf{A}_0$ and $\mathbf{A}_m$ and vectors $\boldsymbol{g}_0$ and $\boldsymbol{g}_m$ must be modified in the following way

$$\mathbf{A}_0 = \begin{bmatrix} 1 & 0 & 0 \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \qquad \boldsymbol{g}_0 = \begin{bmatrix} u_{-1/2} \\ g_1 \\ g_2 \end{bmatrix}, \tag{3.29}$$

$$\mathbf{A}_m = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix}, \qquad \boldsymbol{g}_m = \begin{bmatrix} g_0 \\ g_1 \\ u_{m+1/2} \end{bmatrix}. \tag{3.30}$$

These are then assembled with the other element contributions in the global stiffness matrix $\mathbf{A}$ and right hand side vector $\boldsymbol{g}$ in (3.13).

3.5 Dry Beds

Dry beds are handled adequately by all steps of the FEVM in their current form, except the FEM for u. For the elliptic solver the dry bed presents two issues; when h and G are small then small errors in h and G can produce large errors in

u leading to instabilities and when $h = 0$ the stiffness matrix $\mathbf{A}$ (3.14) becomes singular.

The issue of large errors in u when h is small also arises when solving the SWWE; due to $u = (uh)/h$ being undefined as uh and h go to zero. For the Serre equations with horizontal beds when $h \ll 1$ from (2.8c) we have

$$G = uh + \mathcal{O}\left(h^2\right). \tag{3.31}$$

Since $h \ll 1$ we neglect the $\mathcal{O}\left(h^2\right)$ terms, and thus when h is small G is equal to the momentum uh, and the challenges posed by $h \to 0$ for the SWWE and the Serre equations are equivalent. Therefore, we can apply the dry bed handling techniques from the SWWE to the Serre equations; in particular a desingularisation transformation [47].

These desingularisation transforms act by modifying the calculation of u given h and uh to avoid the singularity as the numerator and denominator go to zero, hence their name. The simplest such transformation is

$$u = \frac{(uh)h}{h\left(h + h_{base}\right)} \tag{3.32}$$

where h_{base} is some small chosen parameter. The error introduced by this transformation is smallest when h_{base} is smallest. However, as noted by Kurganov and Petrova [47] small values of h_{base} lead to large numerical errors in the calculation of u. To avoid such errors h_{base} can be made larger or following Kurganov and Petrova [47] different desingularisation transforms can be employed. For the main purpose of this thesis; the validation tests reported in Chapter 5 we found the simpler transformation with small values of h_{base} more useful, keeping in mind that large numerical errors in u were possible for small values of h.

To adapt the calculation of u in (3.32) to the elliptic equation (2.7) we view it as a transformation of the quantity h which is equivalent to

$$h \to h\left(\frac{h + h_{base}}{h}\right). \tag{3.33}$$

This transformation is ill-defined when $h = 0$ so we also add in a small term h_{tol} to the denominator; this h_{tol} also serves as our cut-off value with any cells with $h < h_{tol}$ being considered dry. Therefore, our transformation for the reconstructed

values of h in the finite element method is

$$h^+_{j-1/2} = h^+_{j-1/2}\left(\frac{h^+_{j-1/2} + h_{base}}{h^+_{j-1/2} + h_{tol}}\right), \tag{3.34a}$$

$$h^-_{j+1/2} = h^-_{j+1/2}\left(\frac{h^-_{j+1/2} + h_{base}}{h^-_{j+1/2} + h_{tol}}\right) \tag{3.34b}$$

where on the right hand side are the reconstructed values of h from (3.1) and the left hand side are the values of h used to defined the basis functions of the finite element method (3.8). This transformation is applied to all terms in the FEM avoiding the singularity as $h \to 0$; and in the case where $G = uh$ the transformation is equivalent to (3.32) for the SWWE.

Even with the transform (3.34), the matrix $\mathbf{A}$ can become singular. The methods of Zoppou et al. [13] made use of direct banded matrix solvers such as the Thomas algorithm [48] to solve (3.14) which rely on non-singular matrices and so are unsuitable when $h = 0$. This was resolved by employing an LU decomposition algorithm described by Press et al. [42]. This algorithm solves banded matrix problems using an LU decomposition with partial pivoting, which inserts small non-zero pivots when the pivots value is below some tolerance value p_{tol}. It does this while also keeping the banded matrix structure, and so is not as memory intensive as a standard LU decomposition. Typically we set $p_{tol} = 10^{-20}$ allowing the matrix solver to accurately invert $\mathbf{A}$ and thus solve (3.14) when $h = 0$.

Finally, after solving (3.14) using the LU decomposition algorithm of Press et al. [42] where the transformation (3.34) has been applied to the reconstructed values of h we possess an approximation to u in the presence of dry beds. Additionally to avoid numerical errors becoming dominant when h is very small we place a cut-off on h past which $h = G = u = 0$ and the cells are properly dry; this is given by h_{tol}. This drying of the cells is performed for the whole cell based

on the cell average value of h so that if $\overline{h}_j \leq h_{tol}$ then

$$\begin{aligned}
&h^+_{j-1/2} = 0 && G^+_{j-1/2} = 0 && w^+_{j-1/2} = b_{j-1/2} \\
&h_j = 0 && G_j = 0 && w_j = b_j, \\
&h^-_{j+1/2} = 0 && G^-_{j+1/2} = 0 && w^-_{j+1/2} = b_{j+1/2}, \\
\\
&u_{j-1/2} = 0 && \text{if} && h_{j-1} \leq h_{tol} \\
&u_j = 0 \\
&u_{j+1/2} = 0 && \text{if} && h_{j+1} \leq h_{tol}
\end{aligned}$$

this drying procedure occurs after the solution of (3.14). In the numerical experiments the typical values used were $h_{tol} = 10^{-12}$ and $h_{base} = 10^{-8}$.

In this chapter FEVM$_2$ was described, including the details for the well balancing and dry bed handling procedures.

Chapter 4

Linear Analysis of the Numerical Methods

In this chapter a linear analysis is used to study the convergence and dispersion properties of the numerical methods.

An important property of a numerical method is convergence. Convergence guarantees that as the spatial and temporal resolution of a numerical method is increased, then the numerical solution approaches the solution of the partial differential equations they approximate.

For linear partial differential equations the Lax-equivalence theorem states that a numerical method is convergent if and only if it is stable and consistent [49]. A numerical scheme is consistent if the error introduced by the numerical method over a time step approaches zero as the spatial and temporal resolution increases. A numerical method is stable if the errors from previous time steps are not amplified by subsequent time steps.

Another important attribute of a numerical method modelling dispersive wave equations, such as the Serre equations is its dispersion properties. The dispersion relation of a system determines the phase and group velocity of travelling waves in that system. The Serre equations possess a dispersion relation that well approximates the dispersion relation given by linear theory for water waves [19]. Therefore, how well the dispersion relation of a numerical method approximates the dispersion relation of the Serre equations is of particular interest.

We analysed the convergence and the dispersion properties of the complete numerical methods for the linearised Serre equations with a horizontal bed; simultaneously analysing the spatial and temporal approximations. The effect of variations in the bed and nonlinear terms are important when studying the con-

vergence properties of our methods for solving the full Serre equations. However, these effects greatly increase the complexity of the convergence analysis. We therefore, restrict ourselves to the study of the linearised Serre with a horizontal bed to offer some insight into the convergence properties of the complete numerical methods without having to deal with this additional complexity. In general, we would expect that a numerical method that has poor convergence properties for the linearised Serre equations with a horizontal bed will also have poor convergence properties when the bed and nonlinear terms are included. The dispersion properties of the Serre equations are derived from the linearised Serre equations with a horizontal bed [13], therefore the presented analysis of the dispersion properties of the numerical methods is a complete analysis.

The linear analyses of convergence and dispersion properties rely on establishing a relationship of the form

$$\begin{bmatrix} h \\ G \end{bmatrix}_j^{n+1} = \mathbf{E} \begin{bmatrix} h \\ G \end{bmatrix}_j^{n} \tag{4.1}$$

where $\mathbf{E}$ is the evolution matrix relating the conserved quantities h and G at time level t^n with the conserved quantities at time level t^{n+1}, which is independent of n and j. The evolution matrix $\mathbf{E}$ is obtained in the analyses by propagating Fourier modes through the numerical scheme. By analysing the properties of $\mathbf{E}$ we can determine the convergence and dispersion properties of its associated numerical method.

We derive $\mathbf{E}$ in (4.1) for FEVM_2 and perform the convergence and dispersion analysis. We then present the results of these analyses for all the other numerical methods named in this thesis.

4.1 Linearised Serre Equations with a Horizontal Bed

The Serre equations with a horizontal bed (2.5) are linearised by considering waves as small perturbations $\delta \times \eta(x,t)$ and $\delta \times \mu(x,t)$ on a flow with a mean height H and a mean velocity U respectively. So we have

$$h(x,t) = H + \delta\eta(x,t) + \mathcal{O}\left(\delta^2\right), \tag{4.2a}$$

$$u(x,t) = U + \delta\mu(x,t) + \mathcal{O}\left(\delta^2\right), \tag{4.2b}$$

where $\delta \ll 1$. These waves are relatively small so terms of order δ^2 are negligible. We substitute (4.2) into the Serre equations and neglect terms of order δ^2 to obtain

$$\frac{\partial(\delta\eta)}{\partial t} + H\frac{\partial(\delta\mu)}{\partial x} + U\frac{\partial(\delta\eta)}{\partial x} = 0, \tag{4.3a}$$

$$H\frac{\partial(\delta\mu)}{\partial t} + gH\frac{\partial(\delta\eta)}{\partial x} + UH\frac{\partial(\delta\mu)}{\partial x} - \frac{H^3}{3}\left(U\frac{\partial^3(\delta\mu)}{\partial x^3} + \frac{\partial^3(\delta\mu)}{\partial x^2\partial t}\right) = 0 \tag{4.3b}$$

and for G

$$G = UH + U\delta\eta + H\delta\mu - \frac{H^3}{3}\frac{\partial^2(\delta\mu)}{\partial x^2}. \tag{4.3c}$$

Absorbing the δ factor into corresponding η and μ terms and rewritting these equations in conservation law form for η and G we obtain

$$\frac{\partial\eta}{\partial t} + \frac{\partial}{\partial x}(H\mu + U\eta) = 0, \tag{4.4a}$$

$$\frac{\partial G}{\partial t} + \frac{\partial}{\partial x}(UG + UH\mu + gH\eta) = 0 \tag{4.4b}$$

where

$$G = UH + U\eta + H\mu - \frac{H^3}{3}\frac{\partial^2\mu}{\partial x^2}. \tag{4.4c}$$

4.2 Evolution Matrix

To derive the evolution matrix, $\mathbf{E}$ we study the behaviour of (4.4) when η and μ are Fourier modes. A Fourier mode $q(x,t)$ is

$$q(x,t) = q(0,0)e^{i\left(\omega^{\pm}t + kx\right)} \tag{4.5}$$

where k is the wavenumber, $\omega^{\pm}$ is the frequency (2.10) and i is the imaginary number. The Fourier modes are the eigenfunctions of these linearised Serre equations (4.4). Since the eigenfunctions form a basis of the solution space, their dispersion and convergence properties are inherited by all solutions of (4.4). Therefore, it is sufficient to only study the convergence and dispersion properties for Fourier mode solutions captured by the evolution matrix $\mathbf{E}$.

A consequence of a quantity q being a Fourier mode represented on a uniform temporal and spatial grid is that for any real numbers m and l we have

$$q_{j+l}^{n+m} = q_j^n e^{i\left(m\omega^{\pm}\Delta t + lk\Delta x\right)}. \tag{4.6}$$

Because η and μ are Fourier modes then so is G. Furthermore, the cell averages of these quantities $\overline{\eta}$, $\overline{\mu}$ and $\overline{G}$ are Fourier modes as well.

In addition, the operators $\mathcal{R}^+_{j-1/2}$, $\mathcal{R}_j$, $\mathcal{R}^-_{j+1/2}$, $\mathcal{G}$, $\mathcal{F}_{j-1/2}$ and $\mathcal{F}_{j+1/2}$ from Chapter 3 will only vary with H, U, k, $\omega^{\pm}$, Δx and Δt and hence be independent of j and n. By combining these operators the evolution matrix $\mathbf{E}$ can be derived for FEVM$_2$ for the linearised Serre equations with a horizontal bed. Since all the constituent operators of $\mathbf{E}$ are independent of j and n then $\mathbf{E}$ will also be independent of j and n, as desired. We will now derive expressions for all these operators, following the structure of the method laid out in Section 3.2. Since the linearised Serre equations with a horizontal bed have no source terms step (iv), which approximates the source terms will be skipped.

4.2.1 Reconstruction

Given $\overline{\boldsymbol{\eta}}$ and $\overline{\boldsymbol{G}}$ at t^n the first step of our numerical method is to reconstruct η and G inside the j^{th} cell at $x_{j-1/2}$, x_j and $x_{j+1/2}$ using $\mathcal{R}^+_{j-1/2}$, $\mathcal{R}_j$ and $\mathcal{R}^-_{j+1/2}$ from (3.1). Since η and G are Fourier modes and therefore smooth we do not require non-linear limiters to ensure our scheme is TVD and so we use the slope $d_j = \left(-\bar{q}_{j-1} + \bar{q}_{j+1}\right) / \left(2\Delta x\right)$ in the reconstruction. Applying (4.6) to the reconstructions (3.1) with the centred slope approximation we obtain

$$q^+_{j-\frac{1}{2}} = \overline{q}_j - \frac{-\overline{q}_j e^{-ik\Delta x} + \overline{q}_j e^{ik\Delta x}}{4} = \left(1 - \frac{i\sin\left(k\Delta x\right)}{2}\right)\overline{q}_j = \mathcal{R}^+_{j-1/2}\overline{q}_j \tag{4.7a}$$

$$q_j = \bar{q}_j = \mathcal{R}_j \bar{q}_j, \tag{4.7b}$$

$$q^-_{j+\frac{1}{2}} = \overline{q}_j + \frac{-\overline{q}_j e^{-ik\Delta x} + \overline{q}_j e^{ik\Delta x}}{4} = \left(1 + \frac{i\sin\left(k\Delta x\right)}{2}\right)\overline{q}_j = \mathcal{R}^-_{j+1/2}\overline{q}_j. \tag{4.7c}$$

Note that these reconstructions operators $\mathcal{R}^+_{j-1/2}$, $\mathcal{R}_j$ and $\mathcal{R}^-_{j+1/2}$ are independent of j.

4.2.2 Fluid Velocity

To calculate $\mu_{j+1/2}$ we use a second-order FEM. We begin our FEM for (4.4c) with its weak formulation, obtained by multiplying (4.4c) by a test function v and integrating over the spatial domain Ω

$$\int_\Omega Gv\,dx = UH\int_\Omega v\,dx + U\int_\Omega \eta v\,dx + H\int_\Omega \mu v\,dx + \frac{H^3}{3}\int_\Omega \frac{\partial \mu}{\partial x}\frac{\partial v}{\partial x}\,dx.$$

The FEM then proceeds for (4.4) as in Chapter 3 for the nonlinear Serre equations with a horizontal bed (2.8). So that G has the basis functions $\psi^+_{j-1/2}$ and $\psi^-_{j+1/2}$ (B.1), which means our approximation to G is linear inside a cell with discontinuous jumps at the cell edges. For v and μ the basis functions $\phi_{j-1/2}$, ϕ_j and $\phi_{j+1/2}$ (B.2) are used so that v and our approximation to μ are quadratic polynomials inside a cell and are continuous across the cell edges.

Given the detailed description of the method in Chapter 3, we will just present the element wise matrix $\mathbf{A}_j$ and vector $\boldsymbol{g}_j$ for (4.4)

$$\mathbf{A}_j = H\frac{\Delta x}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix} + \frac{H^3}{9\Delta x}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix},$$

$$\boldsymbol{g}_j = \frac{\Delta x}{6}\left(\begin{bmatrix} G^+_{j-1/2} \\ 2G^+_{j-1/2} + 2G^-_{j+1/2} \\ G^-_{j+1/2} \end{bmatrix} - UH\begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix} - U\begin{bmatrix} \eta^+_{j-1/2} \\ 2\eta^+_{j-1/2} + 2\eta^-_{j+1/2} \\ \eta^-_{j+1/2} \end{bmatrix}\right).$$

To calculate the intercell flux we require μ at $x_{j+1/2}$, and therefore we only need to solve the FEM approximation that relates all the quantities at $x_{j+1/2}$. From the element wise matrices and vectors for the j and $(j+1)^{th}$ cells the equation that relates all the quantities at $x_{j+1/2}$ is

$$\begin{aligned}\frac{\Delta x}{6}\left(G^-_{j+1/2} + G^+_{j+1/2}\right) = & \\ & \frac{\Delta x}{6}2UH + \frac{\Delta x}{6}U\left(\eta^-_{j+1/2} + \eta^+_{j+1/2}\right) \\ & + \left(H\frac{\Delta x}{30}\left[-\mu_{j-1/2} + 2\mu_j + 8\mu_{j+1/2} + 2\mu_{j+1} - \mu_{j+3/2}\right]\right. \\ & \left. + \frac{H^3}{9\Delta x}\left[\mu_{j-1/2} - 8\mu_j + 14\mu_{j+1/2} - 8\mu_{j+1} + \mu_{j+3/2}\right]\right).\end{aligned}$$

Using (4.7) and (4.6), we obtain

$$\begin{aligned}\frac{\Delta x}{6}\left(\mathcal{R}^-_{j+1/2}+\mathcal{R}^+_{j+1/2}\right)\overline{G}_j = \\ \frac{\Delta x}{6}2UH+\frac{\Delta x}{6}U\left(\mathcal{R}^-_{j+1/2}+\mathcal{R}^+_{j+1/2}\right)\overline{\eta}_j \\ +\left(H\frac{\Delta x}{30}\left[-e^{-ik\Delta x}+2e^{-ik\frac{\Delta x}{2}}+8+2e^{ik\frac{\Delta x}{2}}-e^{ik\Delta x}\right]\right. \\ \left.+\frac{H^3}{9\Delta x}\left[e^{-ik\Delta x}-8e^{-ik\frac{\Delta x}{2}}+14-8e^{ik\frac{\Delta x}{2}}+e^{ik\Delta x}\right]\right)\mu_{j+1/2}.\end{aligned}$$

Rearranging the equation we have that

$$\mu_{j+1/2}=\mathcal{G}^\eta\overline{\eta}_j+\mathcal{G}^G\overline{G}_j+\mathcal{G}^c \tag{4.8}$$

where

$$\mathcal{G}^\eta=\frac{-U\frac{\Delta x}{6}\left(\mathcal{R}^-_{j+1/2}+\mathcal{R}^+_{j+1/2}\right)}{H\frac{\Delta x}{30}\left(4\cos\left(\frac{k\Delta x}{2}\right)-2\cos\left(k\Delta x\right)+8\right)+\frac{H^3}{9\Delta x}\left(-16\cos\left(\frac{k\Delta x}{2}\right)+2\cos\left(k\Delta x\right)+14\right)},$$

$$\mathcal{G}^G=\frac{\frac{\Delta x}{6}\left(\mathcal{R}^-_{j+1/2}+\mathcal{R}^+_{j+1/2}\right)}{H\frac{\Delta x}{30}\left(4\cos\left(\frac{k\Delta x}{2}\right)-2\cos\left(k\Delta x\right)+8\right)+\frac{H^3}{9\Delta x}\left(-16\cos\left(\frac{k\Delta x}{2}\right)+2\cos\left(k\Delta x\right)+14\right)},$$

$$\mathcal{G}^c=\frac{-2UH\frac{\Delta x}{6}}{H\frac{\Delta x}{30}\left(4\cos\left(\frac{k\Delta x}{2}\right)-2\cos\left(k\Delta x\right)+8\right)+\frac{H^3}{9\Delta x}\left(-16\cos\left(\frac{k\Delta x}{2}\right)+2\cos\left(k\Delta x\right)+14\right)}$$

which do not depend on n or j as desired.

4.2.3 Flux Across the Cell Interfaces

The average intercell flux $F_{j+1/2}$ is approximated using (3.15). For the linearised Serre equations we have the wave speed bounds (2.12a), so that

$$a^-_{j+1/2}=\min\left\{0,U-\sqrt{gH}\right\}\quad\text{and}\quad a^+_{j+1/2}=\max\left\{0,U+\sqrt{gH}\right\}. \tag{4.9}$$

This method has three different approximations to $F_{j+1/2}$ depending on the Froude number $Fr=\dfrac{U}{\sqrt{gH}}$; (i) supercritical flow to the left where $Fr<-1$,

(ii) critical and subcritical flow in both directions where $-1 \leq Fr \leq 1$ and (iii) supercritical flow to the right where $Fr > 1$. We will derive the flux operators for each of these cases separately.

Left Supercritical Flow $Fr < -1$:

For left supercritical flow; $Fr < -1$ and therefore $U + \sqrt{gH} < 0$ so we have from (4.9) that $a^-_{j+1/2} = U - \sqrt{gH}$ and $a^+_{j+1/2} = 0$. For these values the flux approximation reduces to the upwind approximation

$$F_{j+\frac{1}{2}} = f\left(q^+_{j+\frac{1}{2}}\right) \tag{4.10}$$

for a general quantity q.

Substituting the flux function from the continuity equation (4.4a) into the flux approximation we obtain

$$F^\eta_{j+\frac{1}{2}} = H\mu_{j+1/2} + U\eta^+_{j+1/2}$$

since μ is continuous $\mu_{j+1/2} = \mu^+_{j+1/2} = \mu^-_{j+1/2}$.

Using the FEM for $\mu_{j+1/2}$ (4.8) and the reconstruction (4.7) we have

$$\begin{aligned} F^\eta_{j+\frac{1}{2}} &= H\left(\mathcal{G}^G\overline{G}_j + \mathcal{G}^\eta\overline{\eta}_j + \mathcal{G}^c\right) + U\eta^+_{j+1/2} \\ &= \left(H\mathcal{G}^\eta + U\mathcal{R}^+_{j+1/2}\right)\overline{\eta}_j + H\mathcal{G}^G\overline{G}_j + H\mathcal{G}^c \end{aligned}$$

This can be written as coefficients for $\overline{\eta}_j$ and $\overline{G}_j$ like so

$$F^\eta_{j+\frac{1}{2}} = \mathcal{F}^{\eta,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{\eta,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{\eta,c}_{j+\frac{1}{2}} \tag{4.11a}$$

where

$$\mathcal{F}^{\eta,G}_{j+\frac{1}{2}} = H\mathcal{G}^G \tag{4.11b}$$

$$\mathcal{F}^{\eta,\eta}_{j+\frac{1}{2}} = H\mathcal{G}^\eta + U\mathcal{R}^+_{j+1/2} \tag{4.11c}$$

$$\mathcal{F}^{\eta,c}_{j+\frac{1}{2}} = H\mathcal{G}^c \tag{4.11d}$$

Substituting the flux function for the G equation (4.4b) into the flux approximation (4.10) we obtain

$$F^G_{j+\frac{1}{2}} = UG^+_{j+1/2} + UH\mu_{j+1/2} + gH\eta^+_{j+1/2}.$$

Using the FEM (4.8) to calculate $\mu_{j+1/2}$ and our interface reconstruction (4.7) we have

$$F^G_{j+\frac{1}{2}} = UG^+_{j+1/2} + UH\left(\mathcal{G}^G\overline{G}_j + \mathcal{G}^\eta\overline{\eta}_j + \mathcal{G}^c\right) + gH\eta^+_{j+1/2}$$

which can be rewritten as

$$F^G_{j+\frac{1}{2}} = \mathcal{F}^{G,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{G,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{G,c}_{j+\frac{1}{2}} \tag{4.12a}$$

where

$$\mathcal{F}^{G,G}_{j+\frac{1}{2}} = U\mathcal{R}^+_{j+1/2} + UH\mathcal{G}^G, \tag{4.12b}$$

$$\mathcal{F}^{G,\eta}_{j+\frac{1}{2}} = UH\mathcal{G}^\eta + gH\mathcal{R}^+_{j+1/2}, \tag{4.12c}$$

$$\mathcal{F}^{G,c}_{j+\frac{1}{2}} = UH\mathcal{G}^c. \tag{4.12d}$$

Subcritical Flow $-1 \leq Fr \leq 1$:

When the flow is subcritical we have $-1 \leq Fr \leq 1$, which means that $a^-_{j+1/2} = U - \sqrt{gH}$ and $a^+_{j+1/2} = U + \sqrt{gH}$. Therefore, the flux approximation (3.15) becomes

$$\begin{aligned} F_{j+\frac{1}{2}} = &\frac{U}{2\sqrt{gH}}\left[f\left(q^-_{j+\frac{1}{2}}\right) - f\left(q^+_{j+\frac{1}{2}}\right)\right] + \frac{1}{2}\left[f\left(q^-_{j+\frac{1}{2}}\right) + f\left(q^+_{j+\frac{1}{2}}\right)\right] \\ &+ \frac{U^2 - gH}{2\sqrt{gH}}\left[q^+_{j+\frac{1}{2}} - q^-_{j+\frac{1}{2}}\right]. \end{aligned} \tag{4.13}$$

Substituting in the flux function for η given by (4.4a) we get

$$\begin{aligned} F^\eta_{j+\frac{1}{2}} = &\frac{U}{2\sqrt{gH}}\left(H\mu_{j+1/2} + U\eta^-_{j+\frac{1}{2}} - H\mu_{j+1/2} - U\eta^+_{j+\frac{1}{2}}\right) \\ &+ \frac{1}{2}\left(H\mu_{j+1/2} + U\eta^-_{j+\frac{1}{2}} + H\mu_{j+1/2} + U\eta^+_{j+\frac{1}{2}}\right) \\ &+ \frac{U^2 - gH}{2\sqrt{gH}}\left(\eta^+_{j+\frac{1}{2}} - \eta^-_{j+\frac{1}{2}}\right). \end{aligned}$$

Using the reconstruction factors (4.7) and the FEM solver factors (4.8) and rearranging we get

$$F^\eta_{j+\frac{1}{2}} = \mathcal{F}^{\eta,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{\eta,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{\eta,c}_{j+\frac{1}{2}} \tag{4.14a}$$

where

$$\mathcal{F}^{\eta,G}_{j+\frac{1}{2}} = H\mathcal{G}^G \tag{4.14b}$$

$$\mathcal{F}^{\eta,\eta}_{j+\frac{1}{2}} = H\mathcal{G}^\eta + \frac{U}{2}\left[\mathcal{R}^-_{j+1/2} + \mathcal{R}^+_{j+1/2}\right] - \frac{\sqrt{gH}}{2}\left[\mathcal{R}^+_{j+1/2} - \mathcal{R}^-_{j+1/2}\right] \tag{4.14c}$$

$$\mathcal{F}^{\eta,c}_{j+\frac{1}{2}} = H\mathcal{G}^c \tag{4.14d}$$

For the flux function of G (4.4b) the flux approximation (4.13) becomes

$$\begin{aligned} F^G_{j+\frac{1}{2}} = &\frac{U}{2\sqrt{gH}}\left(UG^-_{j+\frac{1}{2}} + UH\mu_{j+1/2} + gH\eta^-_{j+\frac{1}{2}} - UG^+_{j+\frac{1}{2}} - UH\mu_{j+1/2} - gH\eta^+_{j+\frac{1}{2}}\right) \\ &+ \frac{1}{2}\left(UG^-_{j+\frac{1}{2}} + UH\mu_{j+1/2} + gH\eta^-_{j+\frac{1}{2}} + UG^+_{j+\frac{1}{2}} + UH\mu_{j+1/2} + gH\eta^+_{j+\frac{1}{2}}\right) \\ &+ \frac{U^2 - gH}{2\sqrt{gH}}\left(G^+_{j+\frac{1}{2}} - G^-_{j+\frac{1}{2}}\right). \end{aligned}$$

By using the reconstruction factors (4.7) and the elliptic solver (4.8) we get

$$F^G_{j+\frac{1}{2}} = \mathcal{F}^{G,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{G,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{G,c}_{j+\frac{1}{2}} \tag{4.15a}$$

where

$$\mathcal{F}^{G,G}_{j+\frac{1}{2}} = UH\mathcal{G}^G + +\frac{U}{2}\left[\mathcal{R}^-_{j+1/2} + \mathcal{R}^+_{j+1/2}\right] - \frac{\sqrt{gH}}{2}\left[\mathcal{R}^+_{j+1/2} - \mathcal{R}^-_{j+1/2}\right], \tag{4.15b}$$

$$\mathcal{F}^{G,\eta}_{j+\frac{1}{2}} = \frac{U\sqrt{gH}}{2}\left[\mathcal{R}^-_{j+1/2} - \mathcal{R}^+_{j+1/2}\right] + UH\mathcal{G}^\eta + \frac{gH}{2}\left[\mathcal{R}^-_{j+1/2} + \mathcal{R}^+_{j+1/2}\right], \tag{4.15c}$$

$$\mathcal{F}^{G,c}_{j+\frac{1}{2}} = UH\mathcal{G}^c. \tag{4.15d}$$

Right Supercritical Flow $Fr > 1$**:**

When the flow is flowing to the right and supercritical we have $Fr > 1$, which means that $a^-_{j+1/2} = 0$ and $a^+_{j+1/2} = U + \sqrt{gH}$. This is very similar to the left supercritical case, except instead of $\mathcal{R}^+_{j+1/2}$ we have $\mathcal{R}^-_{j+1/2}$ in our flux approximation for a general quantity (3.15) which reduces to

$$F_{j+\frac{1}{2}} = f\left(q^-_{j+\frac{1}{2}}\right). \tag{4.16}$$

Substituting in the flux function into (4.4a) and (4.4b) we obtain

$$F^\eta_{j+\frac{1}{2}} = \mathcal{F}^{\eta,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{\eta,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{\eta,c}_{j+\frac{1}{2}} \tag{4.17a}$$

where

$$\mathcal{F}^{\eta,G}_{j+\frac{1}{2}} = H\mathcal{G}^G, \tag{4.17b}$$

$$\mathcal{F}^{\eta,\eta}_{j+\frac{1}{2}} = H\mathcal{G}^\eta + U\mathcal{R}^-_{j+1/2}, \tag{4.17c}$$

$$\mathcal{F}^{\eta,c}_{j+\frac{1}{2}} = H\mathcal{G}^c \tag{4.17d}$$

and

$$F^G_{j+\frac{1}{2}} = \mathcal{F}^{G,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{G,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{G,c}_{j+\frac{1}{2}} \tag{4.18a}$$

where

$$\mathcal{F}^{G,G}_{j+\frac{1}{2}} = U\mathcal{R}^+_{j+1/2} + UH\mathcal{G}^G, \tag{4.18b}$$
$$\mathcal{F}^{G,\eta}_{j+\frac{1}{2}} = UH\mathcal{G}^\eta + gH\mathcal{R}^-_{j+1/2}, \tag{4.18c}$$
$$\mathcal{F}^{G,c}_{j+\frac{1}{2}} = UH\mathcal{G}^c. \tag{4.18d}$$

4.2.4 Update Cell Averages

We have obtained the operators for the flux functions for all cases, supercrticial, critical and subcritical flow. Substituting the appropriate flux approximation into the forward Euler step, (3.26) we get

$$\overline{\eta}_j^{n+1} = \overline{\eta}_j^n - \frac{\Delta t}{\Delta x}\left(\left[\mathcal{F}^{\eta,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{\eta,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{\eta,c}_{j+\frac{1}{2}}\right] - \left[\mathcal{F}^{\eta,\eta}_{j-\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{\eta,G}_{j-\frac{1}{2}}\overline{G}_j + \mathcal{F}^{\eta,c}_{j-\frac{1}{2}}\right]\right),$$
$$\overline{G}_j^{n+1} = \overline{G}_j^n - \frac{\Delta t}{\Delta x}\left(\left[\mathcal{F}^{G,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{G,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{G,c}_{j+\frac{1}{2}}\right] - \left[\mathcal{F}^{G,\eta}_{j-\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{G,G}_{j-\frac{1}{2}}\overline{G}_j + \mathcal{F}^{G,c}_{j-\frac{1}{2}}\right]\right).$$

Since $\mathcal{F}^{\eta,\eta}_{j-\frac{1}{2}} = e^{-ik\Delta x}\mathcal{F}^{\eta,\eta}_{j+\frac{1}{2}}$, $\mathcal{F}^{\eta,G}_{j-\frac{1}{2}} = e^{-ik\Delta x}\mathcal{F}^{\eta,G}_{j+\frac{1}{2}}$, $\mathcal{F}^{G,\eta}_{j-\frac{1}{2}} = e^{-ik\Delta x}\mathcal{F}^{G,\eta}_{j+\frac{1}{2}}$ and $\mathcal{F}^{G,G}_{j-\frac{1}{2}} = e^{-ik\Delta x}\mathcal{F}^{G,G}_{j+\frac{1}{2}}$ we have

$$\overline{\eta}_j^{n+1} = \overline{\eta}_j^n - \frac{\Delta t}{\Delta x}\left(\left[1 - e^{-ik\Delta x}\right]\left[\mathcal{F}^{\eta,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{\eta,G}_{j+\frac{1}{2}}\overline{G}_j\right]\right),$$
$$\overline{G}_j^{n+1} = \overline{G}_j^n - \frac{\Delta t}{\Delta x}\left(\left[1 - e^{-ik\Delta x}\right]\left[\mathcal{F}^{G,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{G,G}_{j+\frac{1}{2}}\overline{G}_j\right]\right).$$

This can be written in matrix form as

$$\begin{aligned}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n+1} &= \begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n} - \left(1 - e^{-ik\Delta x}\right)\frac{\Delta t}{\Delta x}\begin{bmatrix}\mathcal{F}^{\eta,\eta} & \mathcal{F}^{\eta,G}\\ \mathcal{F}^{G,\eta} & \mathcal{F}^{G,G}\end{bmatrix}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n}\\ &= \left(\mathbf{I} - \Delta t\mathbf{F}\right)\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n}\end{aligned} \tag{4.19}$$

for a single Euler step which is first-order in time. To increase the order of accuracy in time we use SSP Runge-Kutta time stepping which makes use of a convex combination of multiple Euler steps.

4.2.5 Second-Order SSP Runge-Kutta Method

To achieve second-order accurate time stepping, the second-order SSP Runge-Kutta scheme (3.27) is used. This scheme uses the following convex combination of the Euler steps (4.19)

$$\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{(1)} = (\mathbf{I} - \Delta t\mathbf{F}) \begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n}, \tag{4.20a}$$

$$\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{(2)} = (\mathbf{I} - \Delta t\mathbf{F}) \begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{(1)}, \tag{4.20b}$$

$$\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n+1} = \frac{1}{2}\left(\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n} + \begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{(2)} \right). \tag{4.20c}$$

Substituting (4.20a) and (4.20b) into (4.20c) we can write this in terms of the flux matrix $\mathbf{F}$ and our cell averages at t^n as

$$\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n+1} = \frac{1}{2}\left(\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n} + (\mathbf{I} - \Delta t\mathbf{F})^2 \begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n} \right).$$

Expanding $(\mathbf{I} - \Delta t\mathbf{F})^2$ we get

$$\begin{aligned}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n+1} &= \left(\mathbf{I} - \Delta t\mathbf{F} + \frac{1}{2}\Delta t^2\mathbf{F}^2\right) \begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n} \\ &= \mathbf{E} \begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n}\end{aligned} \tag{4.21}$$

which is in the desired form of (4.1).

This is the evolution matrix $\mathbf{E}$ for FEVM$_2$. The matrix $\mathbf{E}$ is dependent on the flux matrix $\mathbf{F}$ and therefore will depend on the Froude number. The Froude number is constant over time in this analysis and so we can investigate supercritical, subcritical and critical flow individually.

Both the convergence and dispersion analysis then proceed by studying the properties of the evolution matrix $\mathbf{E}$ for FEVM$_2$, FDVM$_1$, FDVM$_2$, FDVM$_3$, $\mathcal{D}$ and $\mathcal{W}$. The evolution matrices for FDVM$_1$, FDVM$_2$, FDVM$_3$ can be derived following the derivation of the evolution matrix of FEVM$_2$ using the expressions for its constituent operators in Appendix C. For $\mathcal{D}$ and $\mathcal{W}$ the evolution matrices are (C.1) and (C.2) respectively. We begin with the convergence analysis.

4.3 Convergence Analysis

We apply the Lax-equivalence theorem to demonstrate the convergence of our numerical methods by establishing their consistency and stability. We use a Von Neumann stability analysis to demonstrate stability. Consistency is demonstrated for the Fourier modes (4.5) solutions which form a basis of the solution space of the linearised Serre equations. Together these stability and consistency conditions imply convergence of the numerical method under the L_2 norm.

4.3.1 Stability

For a numerical method to be stable we must ensure that errors from previous time steps are not amplified by the current time step. To accomplish this we must ensure

$$\rho\left(\mathbf{E}\right) \leq 1 \tag{4.22}$$

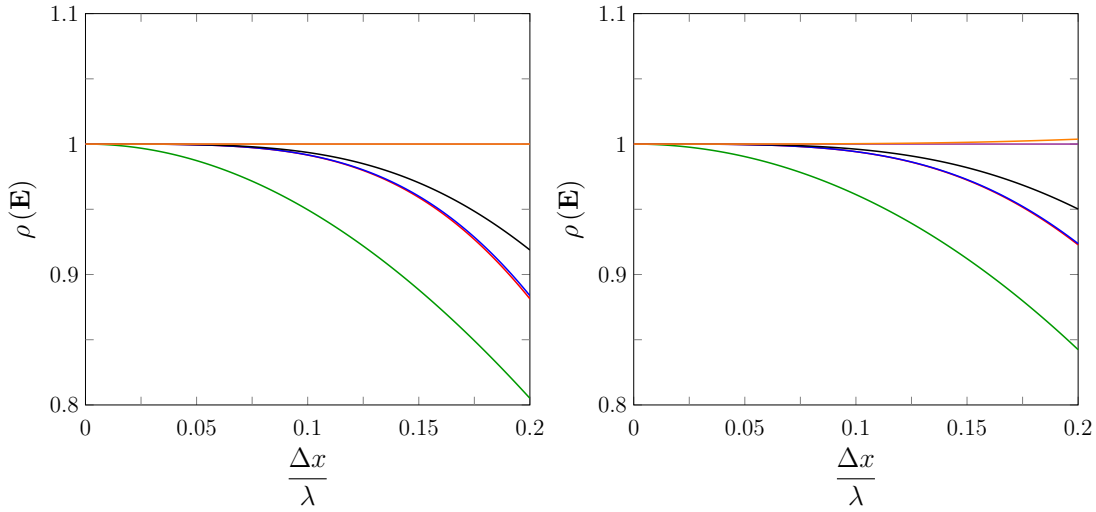
where $\rho\left(\mathbf{E}\right)$ is the spectral radius of $\mathbf{E}$. Since $\mathbf{E}$ was derived for our methods by using Fourier modes, this condition implies Von Neumann stability.

We calculated $\rho\left(\mathbf{E}\right)$ numerically for various values of Δx, Δt, k, H and U to check if (4.22) holds. We summarised our results in Figure 4.1 which is a plot of $\rho\left(\mathbf{E}\right)$ against $\Delta x/\lambda$ for representative values of k, H and U; where $\lambda = 2\pi/k$ is the wavelength. We used g $= 9.81 m/s^2$ and chose $\Delta t = 0.5/\left(U + \sqrt{gH}\right)\Delta x$ to satisfy the CFL condition (3.28). This is the common choice of Δt in our numerical experiments.

The behaviour of $\rho\left(\mathbf{E}\right)$ for $H = 1m$, $k = \frac{\pi}{10} m^{-1}$ and $U = 0m/s$ and $1m/s$ is shown in Figure 4.1 and is representative of the behaviour for all other values of H, k and U. For these k and H values the shallowness parameter $\sigma = \frac{1}{20}$ and so the Serre equations are applicable [19].

In Figure 4.1 it can be seen that all methods have $\rho\left(\mathbf{E}\right) \leq 1$ for $U = 0m/s$ and are therefore stable. The two finite difference methods overlap and have $\rho\left(\mathbf{E}\right) = 1$ for all Δx values, while the FDVM_2 and the FEVM_2 also overlap with $\rho\left(\mathbf{E}\right) < 1$. However, when $U \neq 0m/s$ the method $\mathcal{W}$ has $\rho\left(\mathbf{E}\right) > 1$ for all Δx values and is therefore unstable. All other methods have $\rho\left(\mathbf{E}\right) \leq 1$, retaining their stability when $U \neq 0m/s$.

We observed the same results for a wide range of k, H and U values in particular, all methods except $\mathcal{W}$ were stable for any combination of these variables. While $\mathcal{W}$ was only stable when $U = 0m/s$.



(a) $U = 0$ (b) $U = 1$

Figure 4.1: Spectral radius of $\mathbf{E}$ for FDVM$_1$ (–), FDVM$_2$(–), FEVM$_2$ (–), FDVM$_3$(–), $\mathcal{D}$ (–) and $\mathcal{W}$ (–). With $H = 1m$ and $k = \frac{\pi}{10}$.

4.3.2 Consistency

For a numerical method to be consistent the error introduced by the method for a single time step must approach zero as the spatial and temporal resolution is increased. To demonstrate convergence, it is enough to demonstrate consistency for the eigenfunctions of the linearised Serre equations, which are the Fourier modes. Therefore, we can demonstrate consistency by investigating the evolution matrix $\mathbf{E}$. The error introduced for a single time step from t^n to t^{n+1}, $\mathcal{T}^n$ is

$$\mathcal{T}^n = \mathbf{E}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^n - \begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n+1}. \tag{4.23}$$

To ensure consistency we must have that $\lim_{\Delta x,\Delta t\to 0}\|\mathcal{T}^n\| = 0$ for all n. Taking the norm of both sides of (4.23) and using (4.6) we obtain

$$\|\mathcal{T}^n\| = \left\|\mathbf{E}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^n - e^{i\omega^{\pm}\Delta t}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^n\right\|.$$

Using the matrix norm induced by the vector norm we have that

$$\|\mathcal{T}^n\| \leq \left\|\mathbf{E} - e^{i\omega^{\pm}\Delta t}\mathbf{I}\right\|\left\|\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^n\right\|. \tag{4.24}$$

Since $\overline{\eta}_j^n$ and $\overline{G}_j^n$ are finite and independent of Δx and Δt, if $\lim_{\Delta x,\Delta t\to 0}\left\|\mathbf{E}-e^{i\omega^{\pm}\Delta t}\mathbf{I}\right\|=0$ then $\lim_{\Delta x,\Delta t\to 0}\|\mathcal{T}^n\|=0$ as desired.

We calculated the Taylor series of $\mathbf{E}-e^{i\omega^{\pm}\Delta t}\mathbf{I}$ for all the numerical methods for all flow scenarios; subcritical, critical and supercritical flows. Since the results are the same for ω^+ and ω^- we only report the results for ω^+. We have reported the lowest order Δx and Δt terms of the Taylor series in Tables 4.1 and 4.2 for FDVM_1, Table 4.3 for FDVM_2, Table 4.4 for the FEVM_2, Tables 4.5 and 4.6 for FDVM_3, Table 4.7 for $\mathcal{D}$ and Table 4.8 for $\mathcal{W}$.

We observe for all the methods that the Taylor series of all the elements of $\mathbf{E}-e^{i\omega^{+}\Delta t}\mathbf{I}$ have a factor of Δt. So that for all methods

$$\begin{aligned}\left\|\mathbf{E}-e^{i\omega^{+}\Delta t}\mathbf{I}\right\| &= \|\Delta t\left(\mathbf{M}_0+\mathcal{O}\left(\Delta t\right)\right)\|\\ &= |\Delta t|\,\|\mathbf{M}_0+\mathcal{O}\left(\Delta t\right)\|\\ &\leq |\Delta t|\left(\|\mathbf{M}_0\|+\|\mathcal{O}\left(\Delta t\right)\|\right)\end{aligned}$$

where $\mathbf{M}_0$ is some matrix.

Choosing a particular vector norm and its induced matrix norm it is clear from Tables 4.1-4.8 that $\mathbf{M}_0$ is independent of Δt and finite so that as $\Delta t\to 0$ then $|\Delta t|\left(\|\mathbf{M}_0\|+\|\mathcal{O}\left(\Delta t\right)\|\right)\to 0$ and therefore $\left\|\mathbf{E}-e^{i\omega^{+}\Delta t}\mathbf{I}\right\|\to 0$. Therefore, for all the numerical methods $\lim_{\Delta x,\Delta t\to 0}\|\mathcal{T}^n\|=0$ and so all the numerical methods are consistent for Fourier mode solutions implying consistency for all solutions as desired.

4.4 Dispersion Analysis

To study the dispersion properties of the numerical method, we must calculate the dispersion relation of the numerical method that relates the frequency $\widetilde{\omega}^{\pm}$ to the wavenumber k. Making use of (4.6) in (4.21) we get that for an exact method

$$\mathbf{E}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^n = e^{i\omega^{\pm}\Delta t}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^n. \tag{4.25}$$

Therefore, the evolution matrix $\mathbf{E}$ of an exact method has the eigenvalues $e^{i\omega^{+}\Delta t}$ and $e^{i\omega^{-}\Delta t}$ where $\omega^{\pm}$ are the positive and negative branches of the dispersion relation of the linearised Serre equautions (2.10). For approximate numerical methods the dispersion relation denoted $\widetilde{\omega}^{\pm}$ can be calculated by taking the

Element	Lowest Order Δx Term of Error for FDVM_1		
	$Fr < -1$	$-1 < Fr < 1$	$Fr > 1$
$E_{0,0} - e^{i\omega^+\Delta t}$	$\frac{1}{2}k^2U\Delta t\Delta x$	$-\frac{1}{2}\sqrt{gH}k^2\Delta t\Delta x$	$-\frac{1}{2}k^2U\Delta t\Delta x$
$E_{0,1}$	$\frac{1}{2}gHk^2\Delta t\Delta x$	$\frac{3+\beta}{4\beta^2}ik^3\Delta t\Delta x^2$	$\frac{1}{2}gHk^2\Delta t\Delta x$
$E_{1,0}$	$-\frac{1}{2}\sqrt{gH}k^2\Delta t\Delta x$	$-\frac{1}{2}\sqrt{gH}k^2\Delta t\Delta x$	$-\frac{1}{2}\sqrt{gH}k^2\Delta t\Delta x$
$E_{1,1} - e^{i\omega^+\Delta t}$	$\frac{1}{2}k^2U\Delta t\Delta x$	$-\frac{1}{2}\sqrt{gH}k^2\Delta t\Delta x$	$-\frac{1}{2}k^2U\Delta t\Delta x$

Table 4.1: Lowest order Δx term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega^+\Delta t}\mathbf{I}$. Here $\beta = 3 + k^2H^2$.

Element	Lowest Order Δt Term of Error for FDVM_1
$E_{0,0} - e^{i\omega^+\Delta t}$	$\frac{\sqrt{3gH\beta} + 3U}{\beta}ik\Delta t$
$E_{0,1}$	$-\frac{3}{\beta}ik\Delta t$
$E_{1,0}$	$\left(-gH + \frac{3U^2}{\beta}\right)ik\Delta t$
$E_{1,1} - e^{i\omega^+\Delta t}$	$\frac{\sqrt{3gH\beta} - 3U}{\beta}ik\Delta t$

Table 4.2: Lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega^+\Delta t}\mathbf{I}$ for all values of Fr. Here $\beta = 3 + k^2H^2$.

Element	Lowest Order Term of Error for FDVM_2	
	Δx	Δt
$E_{0,0} - e^{i\omega^+\Delta t}$	$-\dfrac{i\left(27+9H^2k^2+H^4k^4\right)}{12\beta^2}Uk^3\Delta x^2$	$\dfrac{\sqrt{3gH\beta}+3U}{\beta}ik\Delta t$
$E_{0,1}$	$\dfrac{3+\beta}{4\beta^2}ik^3\Delta t\Delta x^2$	$-\dfrac{3}{\beta}ik\Delta t$
$E_{1,0}$	$-\left(gH+\dfrac{3U^2}{\beta}+\dfrac{9U^2}{\beta^2}\right)\dfrac{k^3}{12}\Delta t\Delta x^2$	$\left(-gH+\dfrac{3U^2}{\beta}\right)ik\Delta t$
$E_{1,1} - e^{i\omega^+\Delta t}$	$\dfrac{-9+H^2k^2\beta}{\beta^2}\dfrac{k^3}{12}iU\Delta t\Delta x^2$	$\dfrac{\sqrt{3gH\beta}-3U}{\beta}ik\Delta t$

Table 4.3: Lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega^+\Delta t}\mathbf{I}$ for all values of Fr. Here $\beta = 3 + k^2H^2$.

Element	Lowest Order Term of Error for FEVM_2	
	Δx	Δt
$E_{0,0} - e^{i\omega^+\Delta t}$	$-\dfrac{i\left(54+45H^2k^2+10H^4k^4\right)}{120\beta^2}Uk^3\Delta t\Delta x^2$	$\dfrac{\sqrt{3gH\beta}+3U}{\beta}ik\Delta t$
$E_{0,1}$	$\dfrac{\beta-3}{\beta^2}\dfrac{ik^3}{40}\Delta t\Delta x^2$	$-\dfrac{3}{\beta}ik\Delta t$
$E_{1,0}$	$-\left(gH-\dfrac{15U^2}{\beta}+\dfrac{9U^2}{\beta}\right)\dfrac{k^3}{120}\Delta t\Delta x^2$	$\left(-gH+\dfrac{3U^2}{\beta}\right)ik\Delta t$
$E_{1,1} - e^{i\omega^+\Delta t}$	$\dfrac{126+75H^2k^2+10H^4k^4}{\beta^2}\dfrac{k^3}{120}iU\Delta t\Delta x^2$	$\dfrac{\sqrt{3gH\beta}-3U}{\beta}ik\Delta t$

Table 4.4: Lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega^+\Delta t}\mathbf{I}$ for all values of Fr. Here $\beta = 3 + k^2H^2$.

Element	Lowest Order Δx Term of Error for FDVM$_3$		
	$Fr < -1$	$-1 < Fr < 1$	$Fr > 1$
$E_{0,0} - e^{i\omega^+\Delta t}$	$\frac{1}{12}k^4U\Delta t\Delta x^3$	$-\frac{1}{12}\sqrt{gH}k^4\Delta t\Delta x^3$	$-\frac{1}{12}k^4U\Delta t\Delta x^3$
$E_{0,1}$	$\frac{1}{4\beta}iUk^5\Delta t^2\Delta x^3$	$\frac{\sqrt{gH}}{4\beta}ik^5\Delta t^2\Delta x^3$	$-\frac{1}{4\beta}iUk^5\Delta t^2\Delta x^3$
$E_{1,0}$	$\frac{1}{12}gHk^4\Delta t^2\Delta x^3$	$-\frac{1}{12}\sqrt{gH}k^4\Delta t\Delta x^3$	$-\frac{1}{12}gHk^4\Delta t^2\Delta x^3$
$E_{1,1} - e^{i\omega^+\Delta t}$	$\frac{1}{12}k^4U\Delta t\Delta x^3$	$-\frac{1}{12}\sqrt{gH}k^4\Delta t\Delta x^3$	$-\frac{1}{12}k^4U\Delta t\Delta x^3$

Table 4.5: Lowest order Δx term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega^+\Delta t}\mathbf{I}$. Here $\beta = 3 + k^2H^2$.

Element	Lowest Order Δt Term of Error for FDVM$_3$
$E_{0,0} - e^{i\omega^+\Delta t}$	$\frac{\sqrt{3gH\beta} + 3U}{\beta}ik\Delta t$
$E_{0,1}$	$-\frac{3}{\beta}ik\Delta t$
$E_{1,0}$	$\left(-gH + \frac{3U^2}{\beta}\right)ik\Delta t$
$E_{1,1} - e^{i\omega^+\Delta t}$	$\frac{\sqrt{3gH\beta} - 3U}{\beta}ik\Delta t$

Table 4.6: Lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega^+\Delta t}\mathbf{I}$ for FDVM$_3$ for all values of Fr. Here $\beta = 3 + k^2H^2$.

Element	Lowest Order Term of Error for $\mathcal{D}$	
	Δx	Δt
$E_{0,0} - e^{i\omega^+ \Delta t}$	$\frac{ik^3}{3} U \Delta t \Delta x^2$	$\sqrt{\frac{3gH}{\beta}} 2ik\Delta t$
$E_{0,1}$	$\frac{iHk^3}{3} \Delta t \Delta x^2$	$-2Hik\Delta t$
$E_{1,0}$	$\frac{ig\,(3+\beta)}{2\beta^2} k^3 \Delta t \Delta x^2$	$-\frac{6igk}{\beta} \Delta t$
$E_{1,1} - e^{i\omega^+ \Delta t}$	$\frac{ik^3}{3} U \Delta t \Delta x^2$	$\sqrt{\frac{3gH}{\beta}} 2ik\Delta t$

Table 4.7: Lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega^{\pm} \Delta t}\mathbf{I}$ for all values of Fr. Here $\beta = 3 + k^2 H^2$.

Element	Lowest Order Term of Error for $\mathcal{W}$	
	Δx	Δt
$E_{0,0} - e^{i\omega^+ \Delta t}$	$\frac{ik^3}{6} U \Delta t \Delta x^2$	$\sqrt{\frac{3gH}{\beta}} ik \Delta t$
$E_{0,1}$	$\frac{iHk^3}{6} \Delta t \Delta x^2$	$-Hik\Delta t$
$E_{1,0}$	$\frac{ig\,(3+\beta)}{2\beta^2} k^3 \Delta t \Delta x^2$	$-\frac{6igk}{\beta} \Delta t$
$E_{1,1} - e^{i\omega^+ \Delta t}$	$\frac{ik^3}{3} U \Delta t \Delta x^2$	$\sqrt{\frac{3gH}{\beta}} 2ik \Delta t$

Table 4.8: Lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega^+ \Delta t}\mathbf{I}$ for all values of Fr. Here $\beta = 3 + k^2 H^2$.

eigenvalues of its evolution matrix $\lambda^{\pm}$ like so

$$\widetilde{\omega}^{\pm} = \frac{1}{i\Delta t} \log\left[\lambda^{\pm}\right].$$

By comparing $\widetilde{\omega}^{\pm}$ with the analytic $\omega^{\pm}$ given by the linearised Serre equations (2.10) we can determine the error in the dispersion relation for the numerical method. The real part of $\widetilde{\omega}^{\pm}$ determines the speed of a wave, while the imaginary part determines the change in amplitude. For the frequency of the linearised Serre equations $\omega^{\pm}$ the imaginary part is zero and so the amplitude of waves are constant in time. We only present the results for the positive branch of the dispersion relation comparing $\widetilde{\omega}^{+}$ and ω^{+} as the behaviour of the negative branch is very similar.

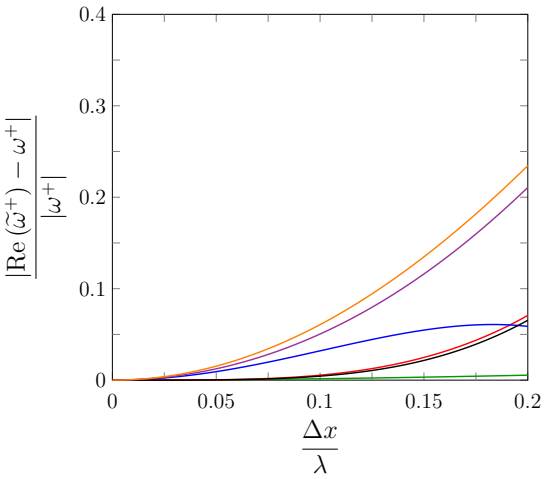
The relative error in the dispersion relation was plotted against $\Delta x/\lambda$ for representative values of H, U and k. Where $\lambda = 2\pi/k$ is the wavelength of the wave with wave number k. We used $g = 9.81m/s^2$ and chose $\Delta t = 0.5/\left(U + \sqrt{gH}\right)\Delta x$ to satisfy the CFL condition (3.28).

In Figures 4.2 and 4.3 we present the plots for $kH = \pi/10$ for shallow water as $\sigma = kH/2\pi = 1/20$ and so the Serre equations are appropriate. We present the real and imaginary errors separately to isolate the errors in the speed and amplitude of the wave for the numerical method. The total error is also reported as a measure of the overall error in the dispersion relation of the numerical method.

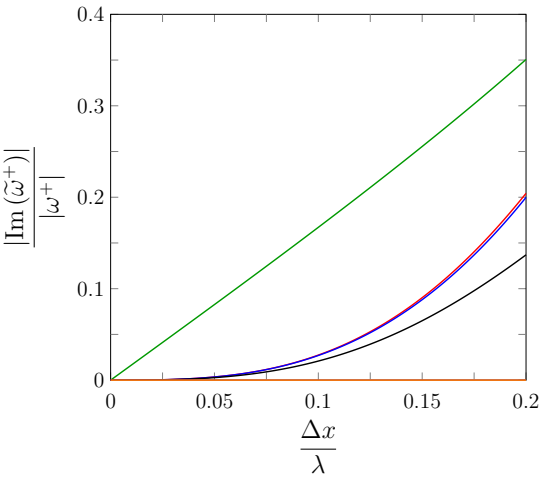
From Figures 4.2 and 4.3 we can see that all methods approximate the dispersion relation of the Serre equations well with the approximation improving as $\Delta x \to 0$, as expected.

For the real part of the dispersion error the FEVM and the FDVM outperform the two finite difference methods and therefore will better approximate the speed of waves of the linearised Serre equations. However, for the amplitude of waves the roles are reversed with the two finite difference methods either dilating the waves very little or not at all. When taking both effects into account with the total error we see that the FDVM_1 has the largest dispersion error followed by $\mathcal{W}$, $\mathcal{D}$, FEVM_2, FDVM_2 and finally FDVM_3 has the lowest dispersion error. So that the size of the total dispersion error is mainly determined by the order of accuracy of the numerical scheme. Whilst hybrid methods perform better than finite difference methods of the same order.

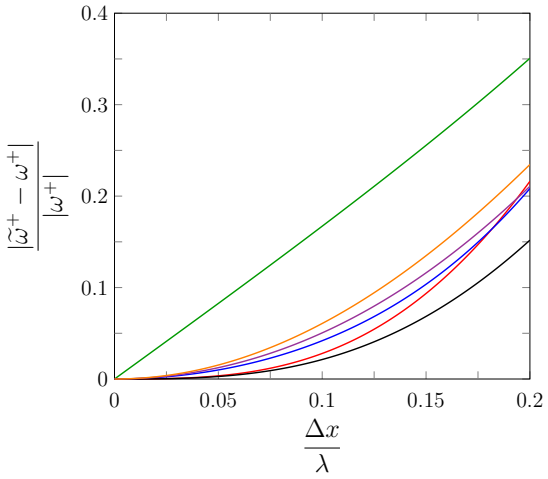
Figures 4.2 and 4.3 furthermore demonstrate that FDVM_2 is superior to FEVM_2 not just for the complete dispersion error, but its real and imaginary parts individually as well. Therefore, FDVM_2 should more accurately model the speed and amplitude of waves.



(a) Real Part

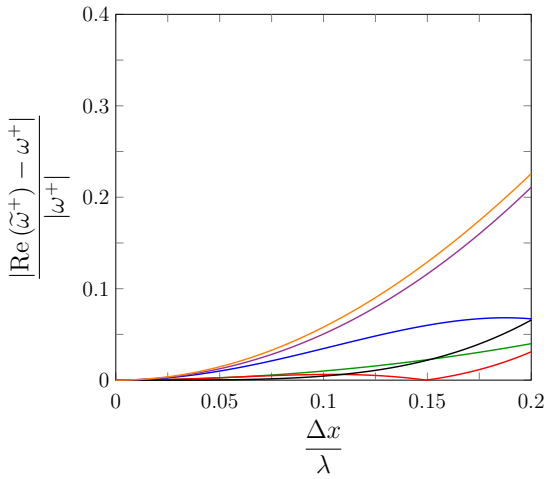


(b) Imaginary Part

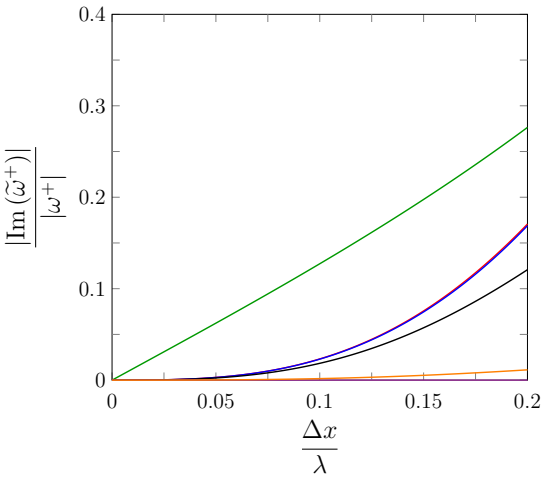


(c) Real and Imaginary Parts

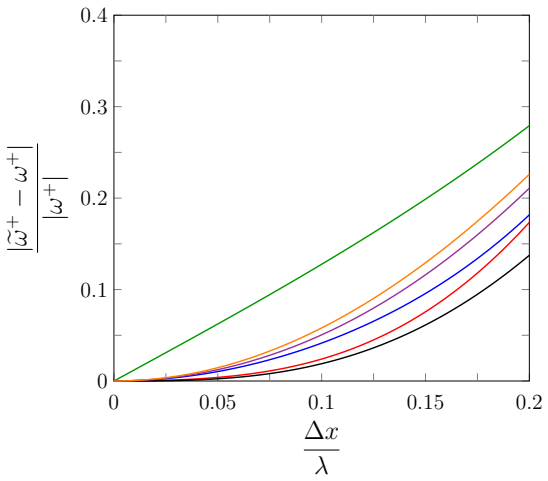
Figure 4.2: Relative dispersion error when $H = 1m$, $k = \frac{\pi}{10}$ and $U = 0m/s$ for FDVM_1 (–), FDVM_2(–), FEVM_2 (–), FDVM_3 (–), $\mathcal{D}$ (–) and $\mathcal{W}$ (–).



(a) Real part



(b) Imaginary part



(c) Real and Imaginary Parts

Figure 4.3: Relative dispersion error when $H = 1m$, $k = \frac{\pi}{10}$ and $U = 1m/s$ for FDVM_1 (–), FDVM_2(–), FEVM_2 (–), FDVM_3 (–), $\mathcal{D}$ (–) and $\mathcal{W}$ (–).

We observed similar results across a wide array of k, H and U values. However, as kH is increased the distinction between FDVM_2 and FEVM_2 becomes less pronounced. This can be seen in Figure 4.4 where $kH = 2.5$ and $\sigma = 5/4\pi > 1/20$ where the water is no longer shallow.

These kH values are the same as those by Filippini et al. [36], and our results are similar for the real part of the dispersion error. Our FDVM and the FEVM compare favourably with the methods described and analysed by Filippini et al. [36]. Furthermore, we extended their work by allowing for non-zero values of U, combining the spatial and temporal approximations and examining the imaginary and complete error in the dispersion relation.

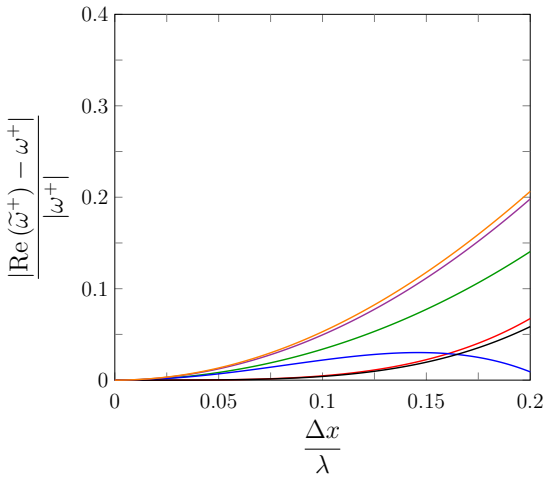
Figure 4.5 demonstrates that the results of the real part of the dispersion error is slightly different if we allow for non-zero values of U. For example the non-zero value of U significantly changes the real part of the dispersion error for FDVM_1 when $kH = 2.5$. Therefore, for some methods allowing for non-zero values of U can have a significant impact on the conclusions drawn from the dispersion analysis. Furthermore, taking the imaginary part of the dispersion error into account is important as $\omega^{\pm}$ determines not only the speed of waves but also their amplitude. For instance the FDVM_1 performs very well for the real part of the dispersion error and poorly for the imaginary part, and so false conclusions about the accuracy of the method could be drawn from only considering the real part of the dispersion error.

The Taylor series expansion of $\widetilde{\omega}^{\pm}$ was also derived for all the numerical methods. We have compiled the lowest order terms of the Taylor series for $\widetilde{\omega}^{+} - \omega^{+}$ in Table 4.9 when $-1 \leq Fr \leq 1$ for the FDVM and FEVM. In Table 4.9 it is clear that these schemes estimated ω^{+} with the expected order of accuracy in both space and time. This was also the case for ω^{-}.

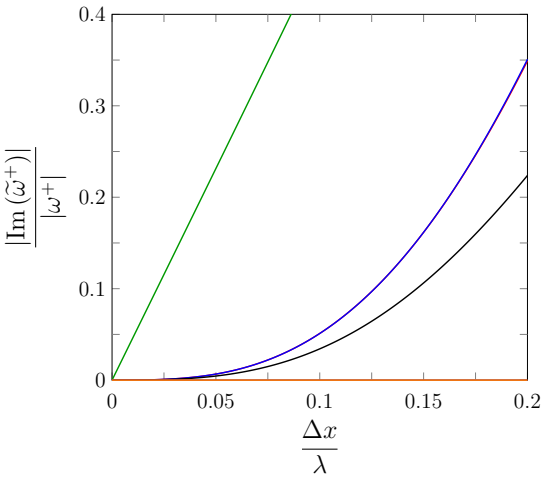
We also present the lowest order terms of the Taylor series for $\widetilde{\omega}^{+} - \omega^{+}$ for both $Fr < -1$ and $Fr > 1$ in Table 4.10. We only present the errors that are different from those reported in Table 4.9, this was only the case for the spatial error of the first- and third–order numerical methods. From these tables it is clear that the FDVM and the FEVM retain their order of accuracy when approximating ω^{+}, this was also the case for ω^{-}.

Finally we present the lowest order terms of the Taylor series for $\widetilde{\omega}^{+} - \omega^{+}$ for the finite difference methods in Table 4.11. These methods do not change depending on the value of the physical quantities. The two finite difference methods retain their order of accuracy in space and time when approximating ω^{+}.

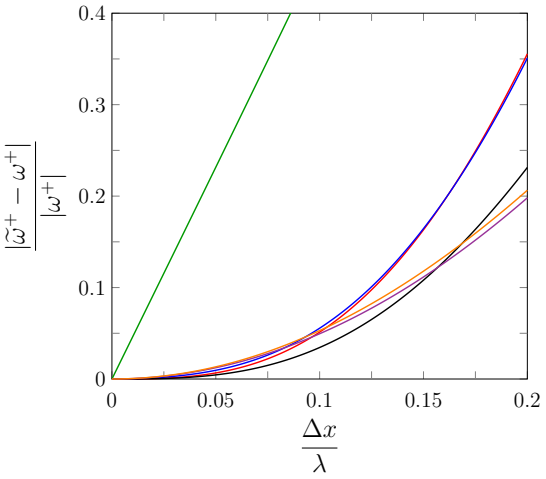
Because all methods were demonstrated to have the expected order of accuracy



(a) Real part

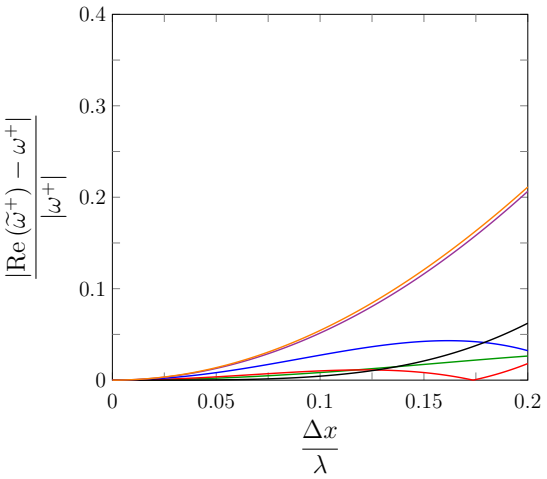


(b) Imaginary part

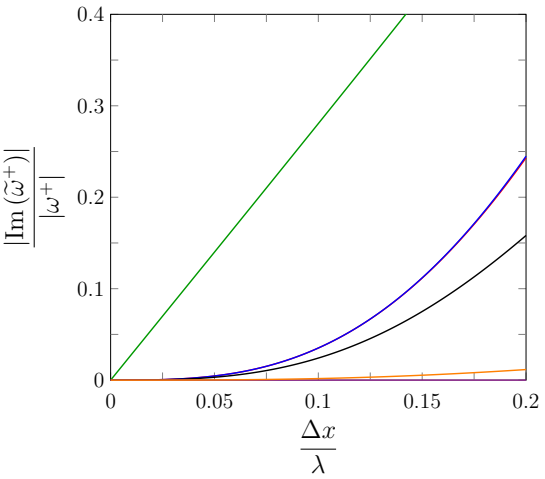


(c) Real and Imaginary Parts

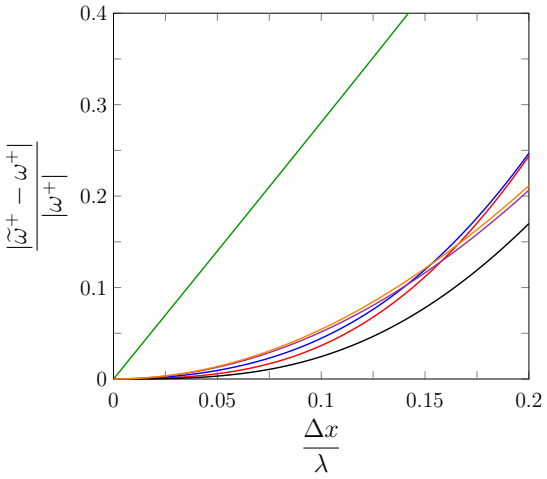
Figure 4.4: Relative dispersion error when $H = 1m$, $k = 2.5$ and $U = 0m/s$ for FDVM_1 (–), FDVM_2(–), FEVM_2 (–), FDVM_3 (–), $\mathcal{D}$ (–) and $\mathcal{W}$ (–).



(a) Real part



(b) Imaginary part



(c) Real and Imaginary Parts

Figure 4.5: Relative dispersion error when $H = 1m$, $k = 2.5$ and $U = 1m/s$ for FDVM_1 (–), FDVM_2(–), FEVM_2 (–), FEVM_3 (–), $\mathcal{D}$ (–) and $\mathcal{W}$ (–).

Scheme	Lowest Order Term of $\widetilde{\omega}^+ - \omega^+$	
	Δx	Δt
FDVM_1	$-\left(2\sqrt{gH} - \sqrt{\frac{3U}{\beta}}\right)\frac{ik^2}{4}\Delta x$	$\frac{i\left(\omega^+\right)^2}{2}\Delta t$
FDVM_2	$\frac{2\beta U - 3\sqrt{3gH\beta}}{\beta^2}\frac{k^3}{24}\Delta x^2$	$-\frac{\left(\omega^+\right)^3}{6}\Delta t^2$
FEVM_2	$\left(U + \frac{\left(42 + 15k^2H^2\right)\sqrt{3gH\beta}}{20\beta^2}\right)\frac{k^3}{12}\Delta x^2$	$-\frac{\left(\omega^+\right)^3}{6}\Delta t^2$
FDVM_3	$-\left(2\sqrt{gH} - \sqrt{3\beta}U\right)\frac{ik^4}{24}\Delta x^3$	$-\frac{i\left(\omega^+\right)^4}{24}\Delta t^3$

Table 4.9: Lowest order term for $\widetilde{\omega}^+ - \omega^+$ for all FDVM and the FEVM. With $-1 \leq Fr \leq 1$ and $\beta = 3 + H^2k^2$.

in approximating $\omega^{\pm}$ for the linearised Serre equations this implies that for small Δx values the order of accuracy will be the primary driver of the dispersion error.

In this chapter the convergence and dispersion properties of the numerical methods were studied using a linear analysis. The results of this analysis demonstrated the superiority of the higher-order accurate hybrid finite volume methods over the finite difference methods.

Scheme	Lowest Order Δx Term of $\widetilde{\omega}^+ - \omega^+$	
	$Fr < -1$	$Fr > 1$
FDVM$_1$	$-\left(2U + \sqrt{\frac{3gH}{\beta}}\right)\frac{ik^2}{4}\Delta x$	$\left(2U + \sqrt{\frac{3gH}{\beta}}\right)\frac{ik^2}{4}\Delta x$
FDVM$_3$	$-\left(2U + \sqrt{\frac{3gH}{\beta}}\right)\frac{ik^4}{24}\Delta x^3$	$\left(2U + \sqrt{\frac{3gH}{\beta}}\right)\frac{ik^4}{24}\Delta x^3$

Table 4.10: Lowest order spatial term for $\widetilde{\omega}^+ - \omega^+$ for all FDVM and the FEVM for supercritical Froude numbers where different from Table 4.9. With $\beta = 3 + H^2k^2$.

Scheme	Lowest Order Term of $\widetilde{\omega}^+ - \omega^+$	
	Δx	Δt
$\mathcal{D}$	$-\left(U + \frac{(4 + H^2k^2)\sqrt{3gH\beta}}{4\beta^2}\right)\frac{k^3}{3}\Delta x^2$	$-\frac{(\omega^+)^3}{3}\Delta t^2$
$\mathcal{W}$	$\left(U + \frac{(4 + H^2k^2)\sqrt{3gH\beta}}{4\beta^2}\right)\frac{k^3}{3}\Delta x^2$	$\left(\beta U^2\left[9\sqrt{3gH\beta} + 4\beta U\right] + 3gH^2\left[\sqrt{3gH\beta} + 6\beta U\right]\right)\frac{k^3}{18\beta^2}\Delta t^2$

Table 4.11: Lowest order term of $\widetilde{\omega}^+ - \omega^+$ for $\mathcal{D}$ and $\mathcal{W}$.

Chapter 5

Numerical Validation

In this chapter analytic and forced solutions are used to validate the numerical methods.

To verify that the numerical methods have the expected convergence and conservation properties we make use of the analytic and forced solutions described in Chapter 2. To assess these properties we first introduce the measures of convergence and conservation for a numerical solution. These measures are then used to assess all numerical methods for the solitary travelling wave solution and the lake at rest solution for FDVM_2 and FEVM_2; which are the only methods in this thesis that currently incorporate varying bathymetry.

Finally we validate FDVM_2 and FEVM_2 using forced solutions which test the accuracy of their approximations to all terms in the Serre equations. Since forced solutions introduce source terms to the Serre equations (2.15) they no longer conserve the conserved quantities of the Serre equations in the general case. Therefore, the forced solutions are only used to assess the convergence properties of these numerical methods.

5.1 Measuring Convergence and Conservation

The convergence of the numerical methods is studied by comparing their numerical solutions to the analytic solutions and forced solutions of the Serre equations. While conservation is investigated by comparing the total amount of a conserved quantity in a numerical solution at some time with the total amount of that quantity present in the initial conditions. We introduce notation for these measures and describe their calculation here, beginning with convergence.

5.1.1 Measure of Convergence

By measuring the relative difference between the numerical and analytic solutions as Δx varies, the convergence of the numerical methods can be investigated. To measure the relative difference we use the L_1 vector norm; to compare the numerical and analytic solutions at the numerical grid locations x_j at the end of the simulations. For a quantity q, the vector of its values $\boldsymbol{q}$ at the grid locations x_j and the corresponding numerical solution at those locations $\boldsymbol{q^*}$; the L_1 norm is

$$L_1(\boldsymbol{q}, \boldsymbol{q^*}) = \begin{cases} \dfrac{||\boldsymbol{q^*} - \boldsymbol{q}||_1}{||\boldsymbol{q}||_1} & ||\boldsymbol{q}||_1 > 0 \\ \\ ||\boldsymbol{q^*}||_1 & ||\boldsymbol{q}||_1 = 0. \end{cases} \tag{5.1}$$

5.1.2 Measures of Conservation

The conservation properties of the methods are established by calculating the total amount of a conserved quantity in the numerical solution $\mathcal{C}^* (\boldsymbol{q^*})$ at the end of the simulation and comparing it to the total amount of that quantity for the initial conditions $\mathcal{C} (q(x, 0))$, derived analytically. Again a relative measure is used;

$$C_1(q, \boldsymbol{q^*}) = \begin{cases} \dfrac{|\mathcal{C}^* (\boldsymbol{q^*}) - \mathcal{C} (q(x, 0)) |}{|\mathcal{C} (q(x, 0)) |} & |\mathcal{C} (q(x, 0)) | > 0 \\ \\ |\mathcal{C}^* (\boldsymbol{q^*}) | & |\mathcal{C} (q(x, 0)) | = 0 \end{cases} \tag{5.2}$$

where $\mathcal{C}^* (\boldsymbol{q^*})$ was calculated using 3 point Gaussian quadrature over the j^{th} cell and summing these cell integrals for all j. The three points needed to perform the Gaussian quadrature were calculated by interpolating the j^{th} cell using a quartic polynomial that fits the nodal values q_{j-2}, q_{j-1}, q_j, q_{j+1} and q_{j+2}. The Gaussian quadrature using three points is 5^{th} order accurate and interpolation by quartics is 5^{th} order accurate for the quantity q and 4^{th} order accurate for its spatial derivative $\partial q/\partial x$. Since all methods are third-order accurate or less, the error introduced by the calculation of $\mathcal{C}^* (\boldsymbol{q^*})$ for the mass, momentum, G and $\mathcal{H}$ will be dominated by the error introduced by the numerical solvers.

In some cases $\mathcal{C} (q(x, 0))$ may be difficult to derive analytically. In this case we compare $\mathcal{C}^* (\boldsymbol{q^*})$ with $\mathcal{C}^* (\boldsymbol{q}^0)$; where $\boldsymbol{q}^0$ is the vector of the quantity at the grid locations used as the initial conditions of our numerical method. Comparing

these we get

$$C_1^*(\boldsymbol{q}^0, \boldsymbol{q}^*) = \begin{cases} \dfrac{|\mathcal{C}^*(\boldsymbol{q}^*) - \mathcal{C}^*(\boldsymbol{q}^0)|}{|\mathcal{C}^*(\boldsymbol{q}^0)|} & |\mathcal{C}^*(\boldsymbol{q}^0)| > 0 \\ \\ |\mathcal{C}^*(\boldsymbol{q}^*)| & |\mathcal{C}^*(\boldsymbol{q}^0)| = 0. \end{cases} \tag{5.3}$$

5.2 Analytic Solution for Horizontal Bed

To assess the ability of our numerical methods to solve the Serre equations with a horizontal bed we use the solitary travelling wave solution (2.13) described in Chapter 2. This is a particular member of the family of periodic travelling wave solutions [28], but all these solutions except the trivial stationary one provide a similar test for the numerical methods and so it is sufficient to only study the solitary travelling wave solution.

For the solitary wave analytic solution all the terms in (2.8) must be adequately approximated by the numerical method to properly reproduce the analytic solution. Therefore, this analytic solution serves as a benchmark for the ability of a numerical method to accurately solve the Serre equations with a horizontal bed for smooth solutions.

For our numerical tests we used the solitary travelling wave solution (2.13) with $a_0 = 1m$, $a_1 = 0.7m$ and $g = 9.81m/s^2$ at $t = 0s$ as the initial conditions. The spatial domain was $[-250m, 250m]$ and the problem was solved until $t = 50s$. This was done for a range of Δx values that had the following form; $\Delta x = 100/2^k m$ with $k \in [6, \ldots, 19]$. The CFL condition was satisfied with CFL number $Cr = 0.5$ by setting $\Delta t = Cr\Delta x/\sqrt{g(a_0 + a_1)}$. For FDVM_2 and FEVM_2 $\theta = 1.2$ was used as the limiting parameter in the generalised minmod limiter (3.2). While FDVM_3 used a Koren limiter, with no parameter.

For the parameters $a_0 = 1m$ and $a_1 = 0.7m$ the nonlinearity parameter is $\epsilon = a_1/a_0 = 0.7$; this is large but beneath most of the well known breaking thresholds for water waves $\epsilon \leq 0.8$ [50]. Because ϵ is large the nonlinear effects are large and therefore so are the balancing dispersive effects making this particular analytic solution a rigorous test of the numerical methods. For this spatial domain and final time $t = 50s$ there is no interaction of the wave with the boundary, therefore the Dirichlet boundary conditions were appropriate.

5.2.1 Results for Solitary Travelling Wave Solution

An example numerical solution with $\Delta x = 100/2^{11} m \approx 0.049m$ from all methods was plotted in Figure 5.1 against the analytic solution at $t = 50s$. We have only plotted an illustrative amount of the points in the numerical solution. From these plots it is clear that FDVM_1 performs significantly worse than the higher-order methods at reproducing the analytic solution, even for this relatively fine grid where the wave is captured by more than 200 cells. This is primarily due to the numerical diffusion introduced by the method, which has caused the wave in the numerical solution to decrease in amplitude and widen significantly. The higher-order numerical methods all accurately replicate the analytic solution, with insignificant visual differences in these plots due to the high resolution of the grid.

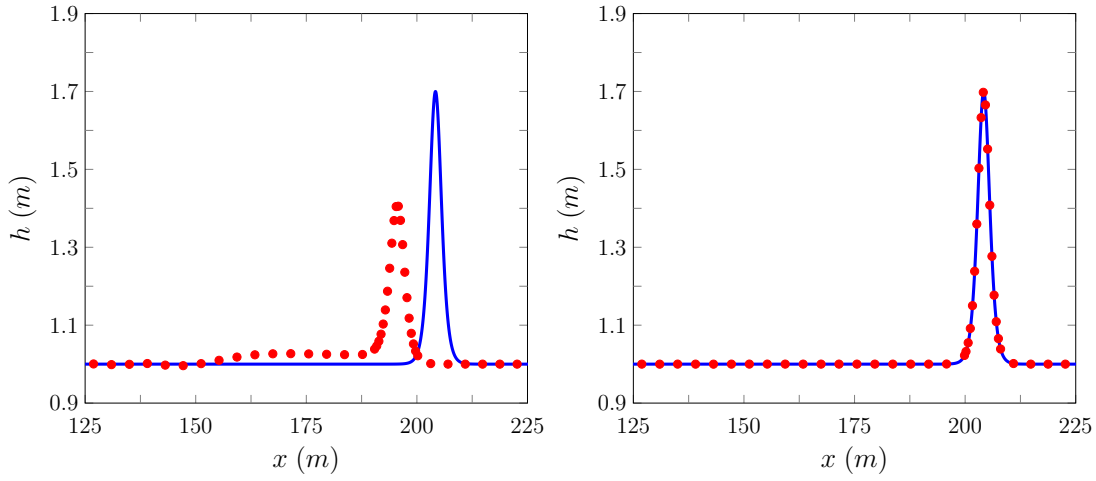
The L_1 norm was calculated for h, u and G for all numerical solutions and was plotted against Δx for all numerical methods in Figure 5.2. From these plots it is clear that all numerical methods are convergent. The rate at which the numerical solutions converge to the analytic solution over Δx is determined by the order of accuracy of the numerical scheme. All methods demonstrate the expected order of accuracy given the order of accuracy of the approximations used in the method; which agrees with the results of the linear analysis in Chapter 4.

All methods more accurately reproduced the analytic solution for h than either G or u across all Δx values. This is due to the simplicity of h's evolution equation (2.6a) compared to the evolution equation of G (2.6b); with the error in u being dominated by the error in G.

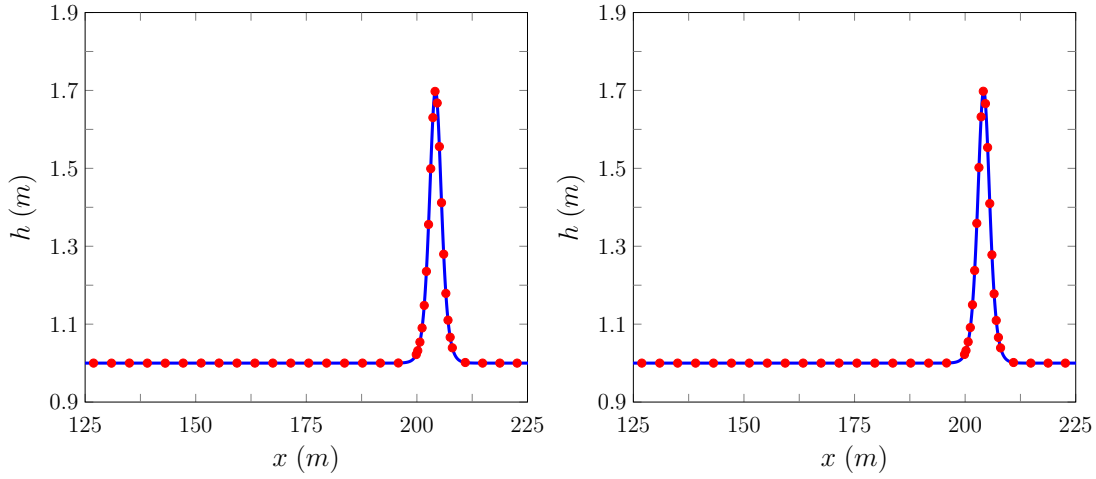
Increasing the order of accuracy of our numerical methods leads to smaller errors when comparing two methods for the same Δx value, as Figure 5.2 clearly demonstrates. This is consistent with the example numerical solution in Figure 5.1, where the lowest order accuracy scheme, FDVM_1 had the poorest reproduction of the analytic solution. However, the benefit of increasing the order of accuracy is significantly diminished for the the third accurate FDVM_3 over the second-order FEVM_2 and FDVM_2.

For the second-order methods we find that FDVM_2 consistently produces the smallest L_1 error followed by FEVM_2, $\mathcal{W}$ and $\mathcal{D}$. The difference between the FDVM_2 and FEVM_2 is significant with errors of FEVM_2 being 2 to 4 times larger than FDVM_2. Therefore, FDVM_2 is reproducing the solitary wave solution more accurately than FEVM_2.

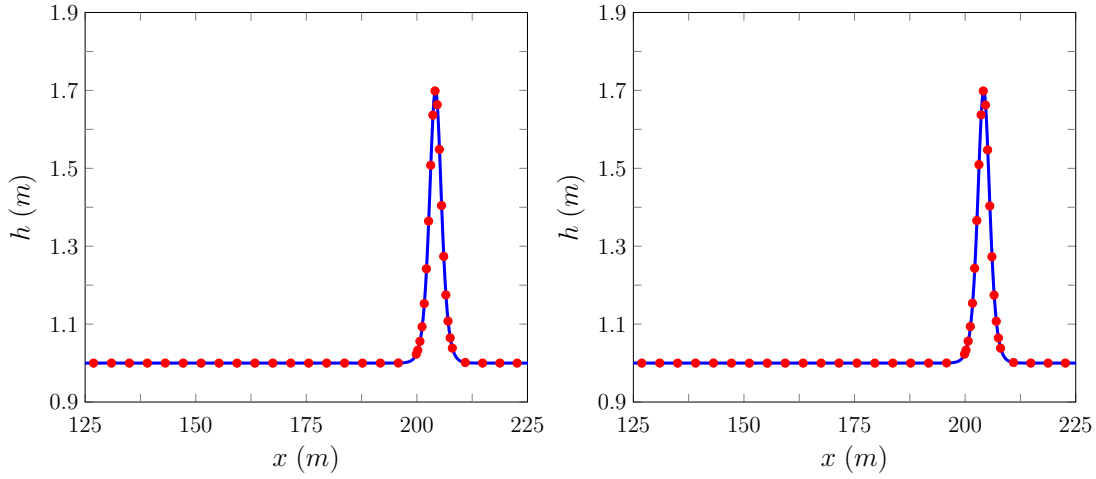
The finite difference methods produce very similar errors which are twice as large as the errors from FEVM_2. Additionally, the round-off effects which increase



(a) FDVM_1 (b) FDVM_2

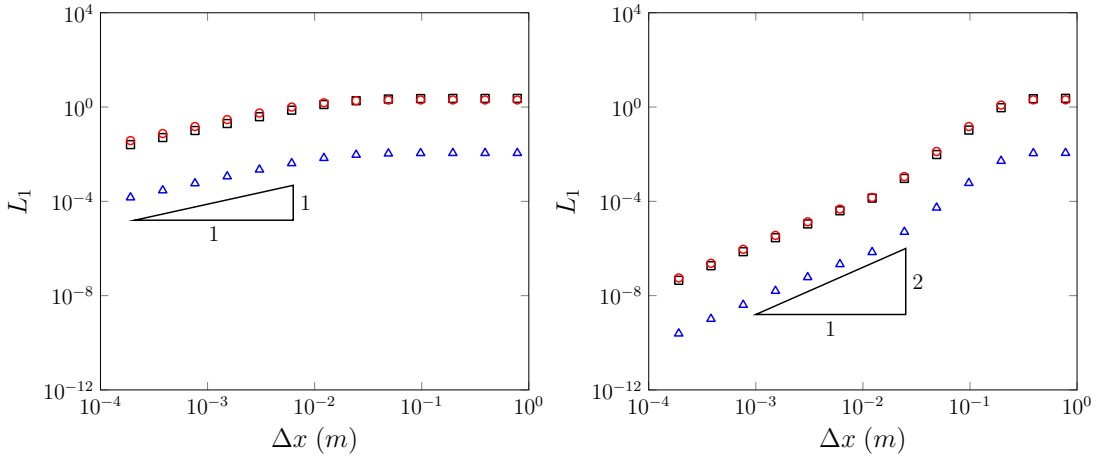


(c) FEVM_2 (d) FDVM_3



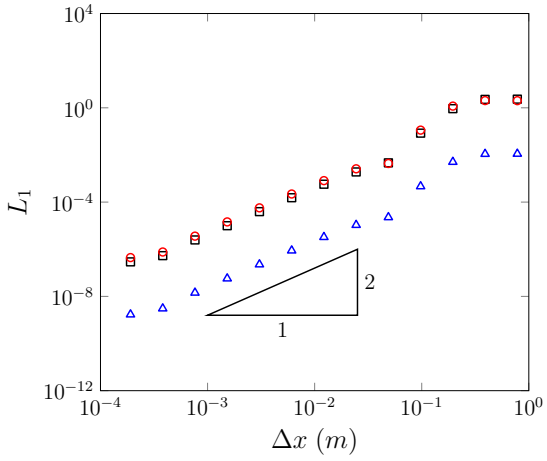
(e) $\mathcal{D}$ (f) $\mathcal{W}$

Figure 5.1: Comparison of the analytic solution (–) and numerical solution with $\Delta x = 100/2^{11} m$ (•) for the soliton problem at $t = 50s$ for all methods.

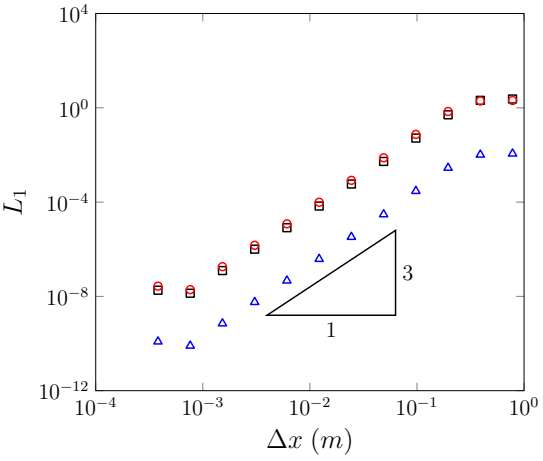


(a) FDVM_1

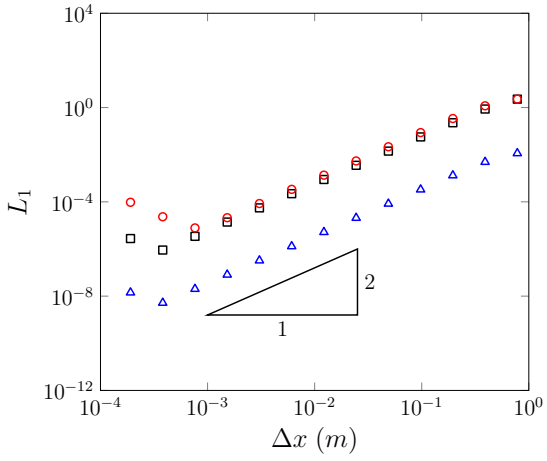
(b) FDVM_2



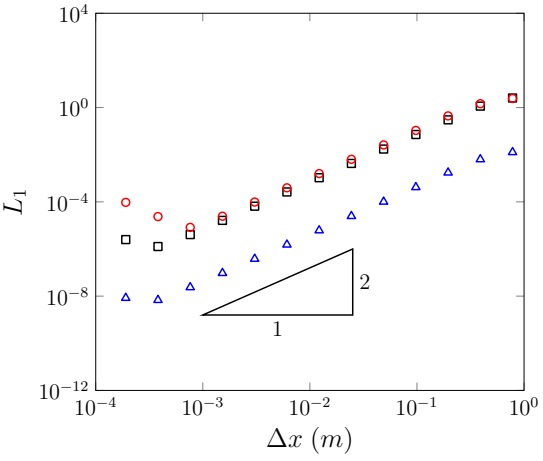
(c) FEVM_2



(d) FDVM_3



(e) $\mathcal{D}$



(f) $\mathcal{W}$

Figure 5.2: Convergence plots as measured by the L_1 norm for h (△), u (□) and G (○) for the soliton problem for all methods.

the L_1 error of the finite difference methods occur for larger Δx values than the hybrid finite volume methods.

The error in conservation C_1 was calculated for all methods using the analytic expressions for the total amounts of the conserved quantities in the initial conditions (A.1). The error in conservation was plotted against the spatial resolution in Figure 5.3. These results demonstrate that due to the use of the finite volume methods for h and G, both are conserved at round-off error for all the hybrid finite volume methods as expected. While the finite difference methods only conserved h at round-off error because the employed finite difference method for the continuity equation (2.6a) is a conservative method.

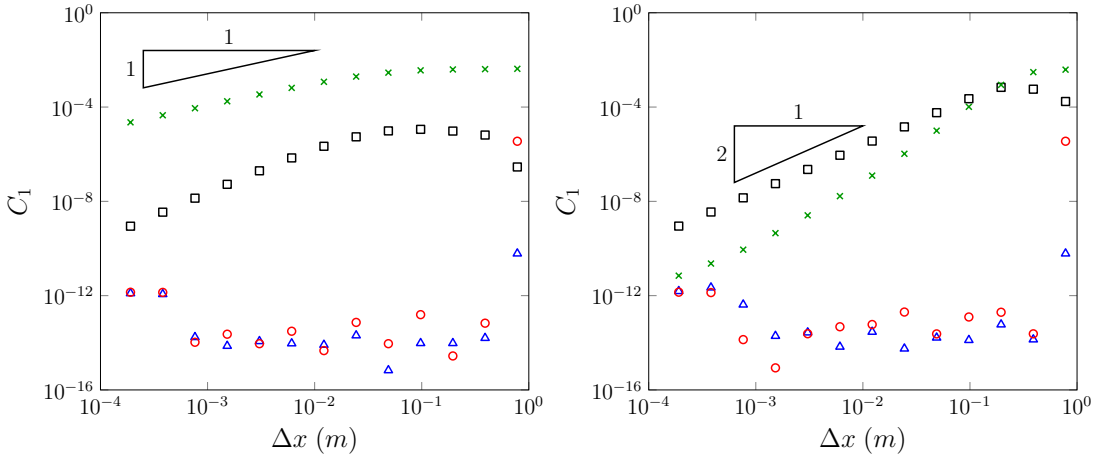
No methods conserve the energy $\mathcal{H}$ or the momentum uh at machine precision. Since none of the methods were designed to have these as conserved variables this is not surprising, although the error in conservation of all methods for these quantities does inherit the order of accuracy of the convergence of the numerical method or better, as expected.

For small Δx values the round-off errors become dominant, particularly for the finite difference methods. Interestingly, FDVM_3 has an accumulation of round-off error in the conservation error for h and G as Δx decreases. This was found to be caused by the Runge-Kutta coefficients [13] which in the last step are $1/3$ and $2/3$ which cannot be exactly represented in floating point, so that every time step accumulates a small conservation errors of machine precision size leading to the observed increase as Δx becomes small and the number of time steps increases.

These results demonstrate the need for higher-order accurate schemes to accurately approximate the Serre equations. Furthermore, they suggest that second-order accuracy is sufficient, with third-order accurate schemes showing only a slight improvement. Finally they also demonstrate the utility of these hybrid FVM for conserving h and G, as desired. Given these results, only FEVM_2 and FDVM_2 have been extended to allow for variable bathymetry and dry beds. Consequently, the rest of the results in this chapter and Chapter 6 will only be for these numerical methods.

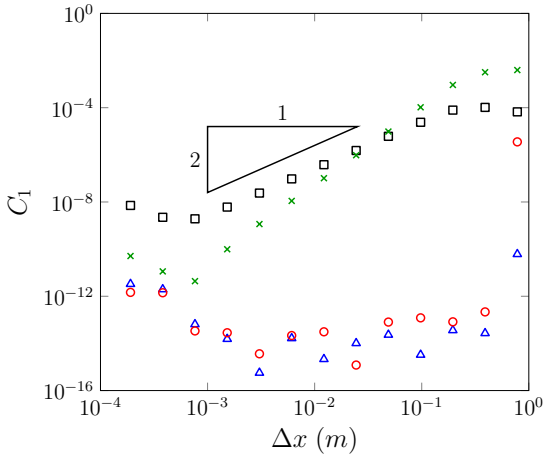
5.3 Analytic Solution for Variable Bathymetry

To verify the validity of our numerical methods for the Serre equations with variable bathymetry and assess the well balancing method we compare various numerical solutions to the lake at rest analytic stationary solution given by (2.14).

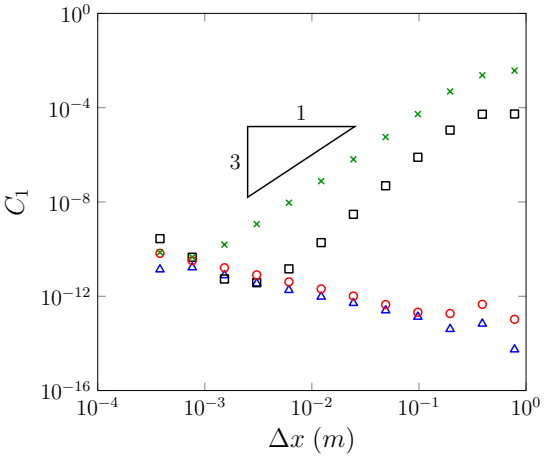


(a) FDVM_1

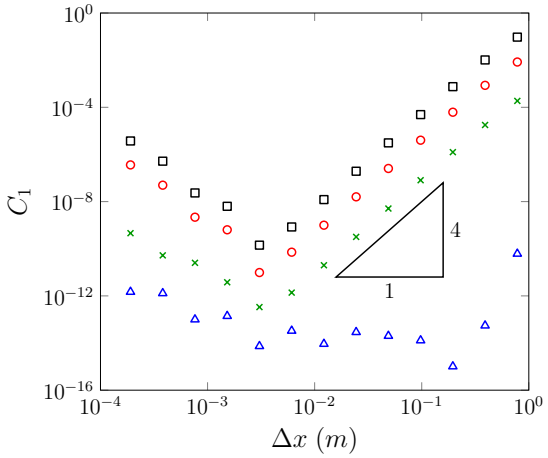
(b) FDVM_2



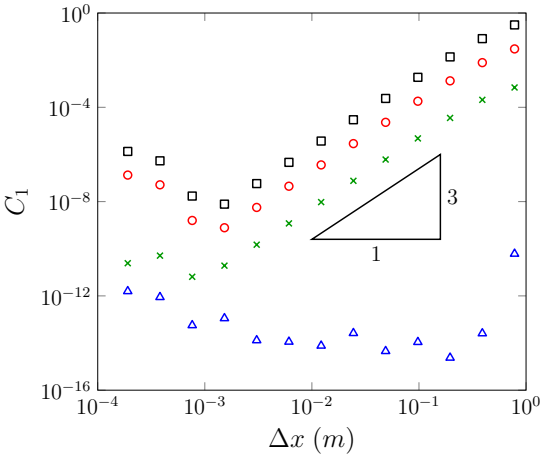
(c) FEVM_2



(d) FDVM_3



(e) $\mathcal{D}$



(f) $\mathcal{W}$

Figure 5.3: Conservation plots as measured by C_1 for h (△), uh (□), G (○) and $\mathcal{H}$ (×) for the soliton problem for all methods.

The particular lake at rest solution (2.14) associated with the bed profile

$$b(x) = a_1 \sin(a_2 x) \tag{5.4}$$

was chosen for this validation to ensure that all terms with derivatives of the bed were tested. To demonstrate the capability of the methods in the presence of dry and wet beds the parameter values $a_0 = 0m$, $a_1 = 1m$ and $a_2 = 2\pi/50m^{-1}$ were chosen. These parameter values result in a periodic bed where water with a constant stage submerges the troughs of the bed while the peaks of the bed are dry.

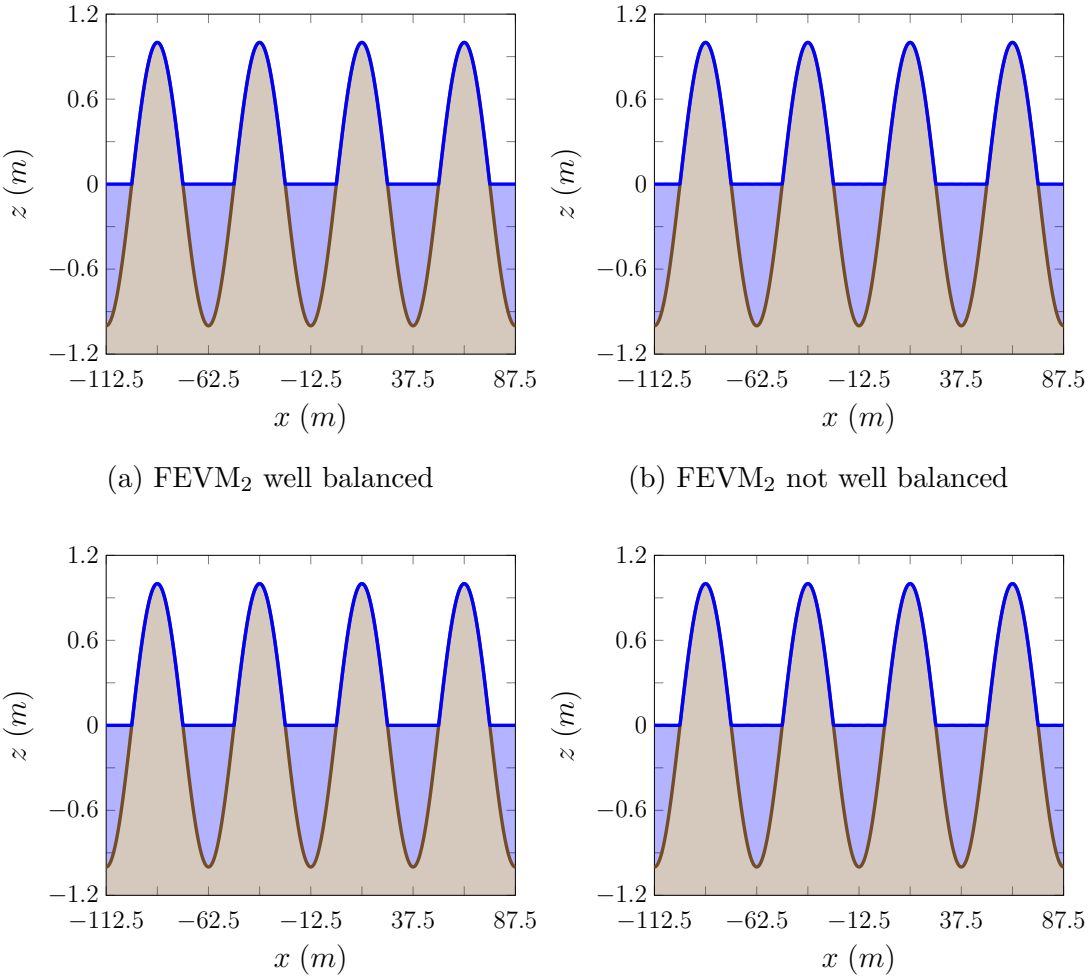
For the numerical solutions the spatial domain was $x \in [-112.5m, 87.5m]$ and the final time was $t = 10s$, with the standard gravitational acceleration $g = 9.81m/s^2$. The spatial resolution of the method was varied so that $\Delta x = 100/2^k m$ with $k \in [8, \ldots, 17]$ and the CFL condition (3.28) was satisfied by having $\Delta t = Cr\Delta x/\sqrt{g}$ with condition number $Cr = 0.5$. The standard limiting parameter $\theta = 1.2$ was used in the generalised minmod limiter, (3.2) for both FEVM_2 and FDVM_2. Dirichlet boundary conditions were used at both ends as the analytic solution is stationary.

The numerical methods are assessed by using the specified lake at rest solution as initial conditions and comparing the numerical solutions of FEVM_2 and FDVM_2 at $t = 10s$ to the analytic solution, which are the initial conditions. To demonstrate the utility of the well balancing method the results from two versions of FEVM_2 and FDVM_2 are presented, where the well balancing method described in Chapter 3 is and isn't employed.

5.3.1 Results for Lake at Rest

Example numerical solutions with $\Delta x = 100/2^{10} m \approx 0.0977m$ at $t = 10s$ for all versions of FEVM_2 and FDVM_2 are given in Figure 5.4. The numerical solutions in these figures are indistinguishable from the analytic solutions at this scale and so the analytic solutions have been omitted from the plots.

Examination of the L_1 errors depicted in Figure 5.5 reveals that only the well balanced methods have accurately recovered the analytic solution. With both well balanced versions of the methods reproducing h, G and u precisely, accounting for round-off errors. This is most clear for h as it is consistently around the machine epsilon value. While for G and u their error is increasing due to an accumulation of the round-off errors for each cell and time step; hence their second-order increase as $\Delta x \to 0$.



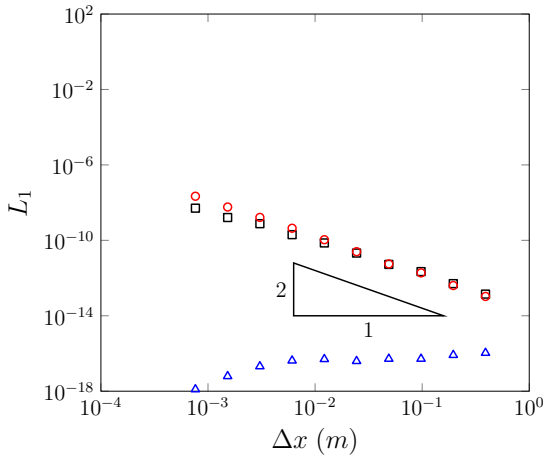
(a) FEVM$_2$ well balanced

(b) FEVM$_2$ not well balanced

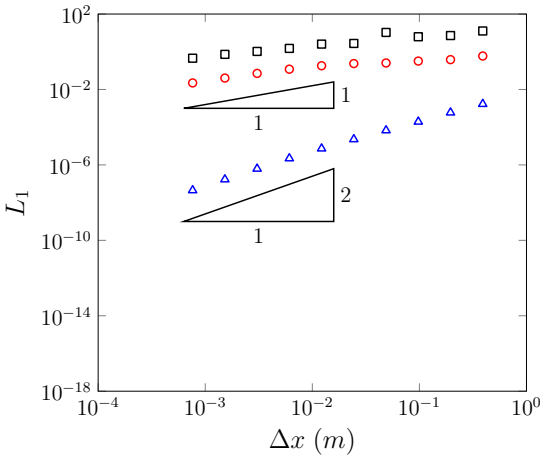
(c) FDVM$_2$ well balanced

(d) FDVM$_2$ not well balanced

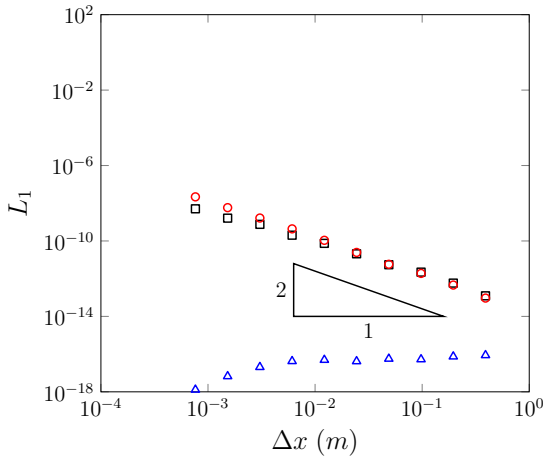
Figure 5.4: Plot of numerical solutions for w (■) and b (■) with $\Delta x = 100/2^{10}m$ for the lake at rest problem at $t = 10s$ for all methods.



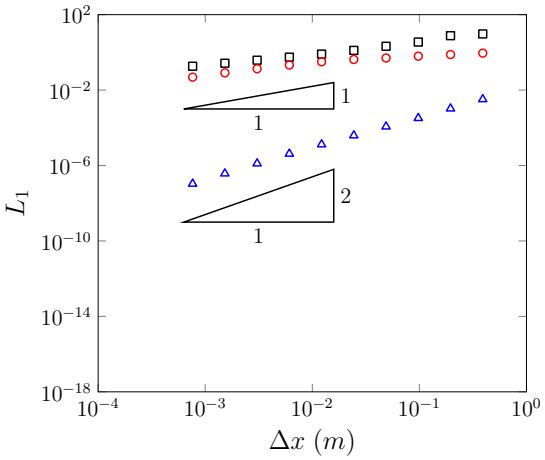
(a) FEVM$_2$ well balanced



(b) FEVM$_2$ not well balanced



(c) FDVM$_2$ well balanced



(d) FDVM$_2$ not well balanced

Figure 5.5: Convergence plots as measured by the L_1 norm for h (△), u (□) and G (○) for the lake at rest problem at $t = 10s$ for all methods.

For methods without well balancing; the errors are significantly larger, yet they are converging to the analytic solution. However, the convergence of these methods has lost an order of accuracy in u and G; with only first-order convergence in these quantities and not the expected second-order accuracy observed for h.

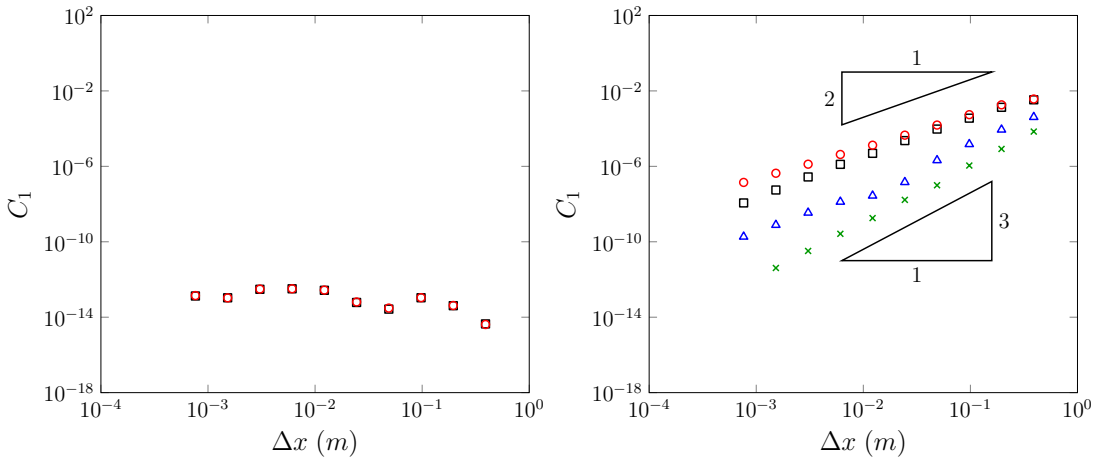
Using the expressions in Appendix A for the total amounts of the conserved variables C_1 was calculated for all numerical solutions with the results plotted in Figure 5.6. The error in conservation of these methods affirms the superiority of the well-balanced version of the methods. For the well balanced FEVM_2 the conservation of h and $\mathcal{H}$ has zero error and so could not be represented on the log-log plot, so that all errors in conservation are at machine precision or lower. For the well-balanced FDVM_2 the error in h vanishes when Δx is small, although the error in $\mathcal{H}$ does not. Therefore, the well balanced version of FEVM_2 outperforms the FDVM_2 in this instance. While for the methods with no well balancing none of the quantities are conserved at machine precision.

These results demonstrate the need for the well-balancing for both numerical methods, as it is only with their inclusion that the lake at rest steady state can be accurately reproduced.

5.4 Forced Solutions

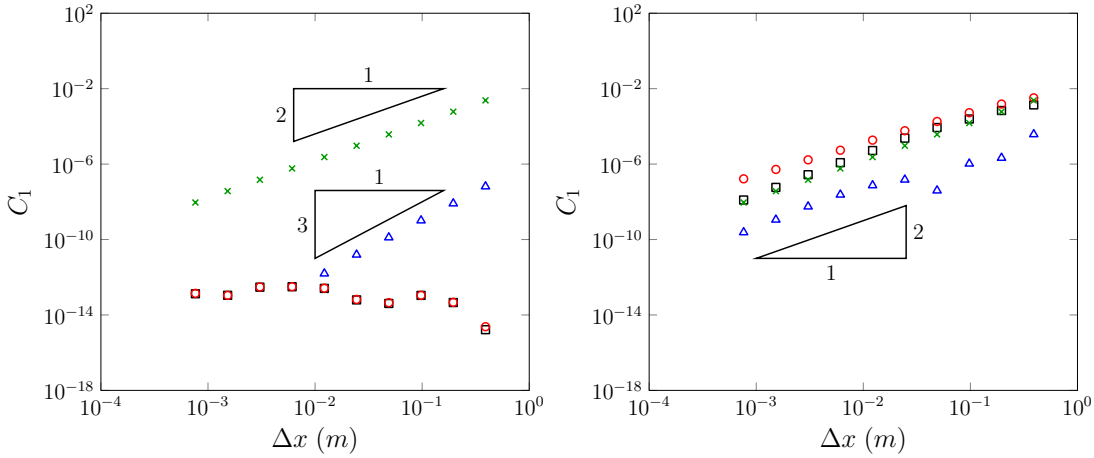
There are currently no known analytic solutions for the Serre equations that possess varying bathymetry and non-zero velocities. Therefore, the previous analytic solution validations do not provide a stringent test for all terms present in the Serre equations. To remedy this the forced solutions introduced in Chapter 2 were used. Since the source terms in the modified Serre equations, (2.15) can be determined and accounted for analytically, the only source of error in the numerical solutions reproduction of the forced solutions are the numerical methods themselves and thus the theoretical second-order accuracy of FEVM_2 and FDVM_2 should be recovered.

We performed validation tests for two forced solutions; one with a finite water depth everywhere and the other with a dry bed to validate and compare the numerical solutions in both situations. To ensure that all terms of the Serre



(a) FEVM$_2$ well balanced

(b) FEVM$_2$ not well balanced



(c) FDVM$_2$ well balanced

(d) FDVM$_2$ not well balanced

Figure 5.6: C_1 error against Δx for h (△), uh (□), G (○) and $\mathcal{H}$ (×) for the lake at rest problem at $t = 10s$ for all methods.

equations were accurately approximated in the numerical method the functions

$$h^*(x,t) = a_0 + a_1 \exp\left(-\frac{\left[(x - a_2 t) - a_3\right]^2}{2a_4}\right), \tag{5.5a}$$

$$u^*(x,t) = a_5 \exp\left(-\frac{\left[(x - a_2 t) - a_3\right]^2}{2a_4}\right), \tag{5.5b}$$

$$b^*(x) = a_6 \sin\left(a_7 x\right) \tag{5.5c}$$

for the primitive variables were chosen. These functions produce an a_1 high Gaussian bump for h and u that travels at a fixed speed a_2 over a periodic bed. For nontrivial choices of the parameters a_i all terms in the Serre equations vary in space and time and so all terms must be accurately approximated by the numerical method to adequately reproduce the forced solution.

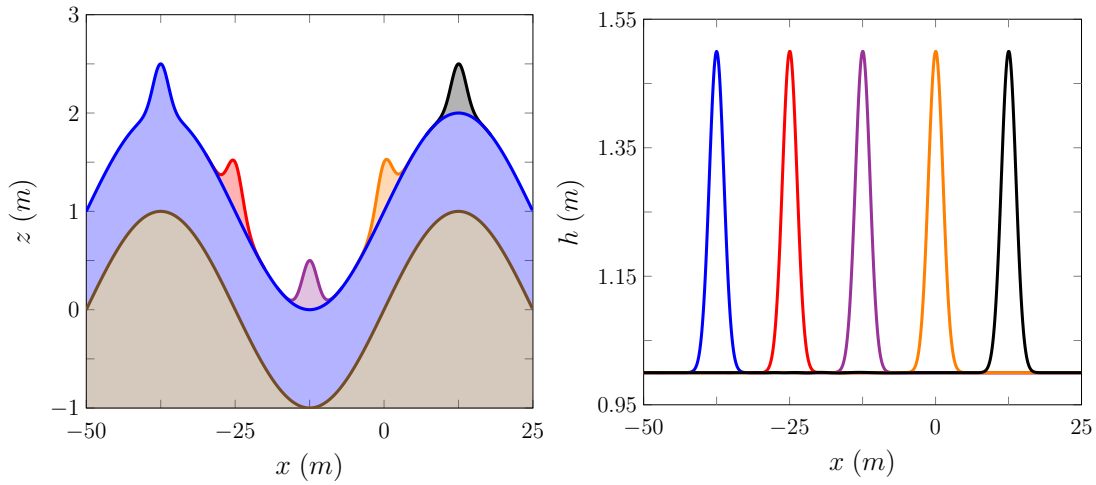
Both validation studies used $a_1 = 0.5m$, $a_2 = 2\pi/\left(10a_7\right) m/s$, $a_3 = -3\pi/\left(2a_7\right) m$, $a_4 = \pi/(16a_7)m$, $a_5 = 0.5m/s$, $a_6 = 1.0m$ and $a_7 = \pi/25m^{-1}$ with $a_0 = 1m$ for the finite water depth forced solution and $a_0 = 0m$ for the dry bed forced solution. These parameter values results in a Gaussian bump that has a width much smaller than the wavelength of the bed profile and travels precisely one wavelength in $10s$.

The domain of the numerical solutions was $x \in [-112.5m, 87.5m]$ with $t \in [0s, 10s]$. The standard gravitational acceleration $g = 9.81m/s^2$ was used. The spatial resolution of numerical methods was varied like so $\Delta x = 100/2^k m$ with $k \in [8, \ldots, 16]$. To satisfy the CFL condition, (3.28) the temporal resolution $\Delta t = Cr\Delta x/\left(a_2 + a_5 + \sqrt{g\left(a_0 + a_1\right)}\right)$ was chosen with condition number $Cr = 0.5$. The value $\theta = 1.2$ was used in the generalised minmod limiter (3.2) for both FEVM_2 and FDVM_2 and Dirichlet boundary conditions were applied at the boundaries of the domain.

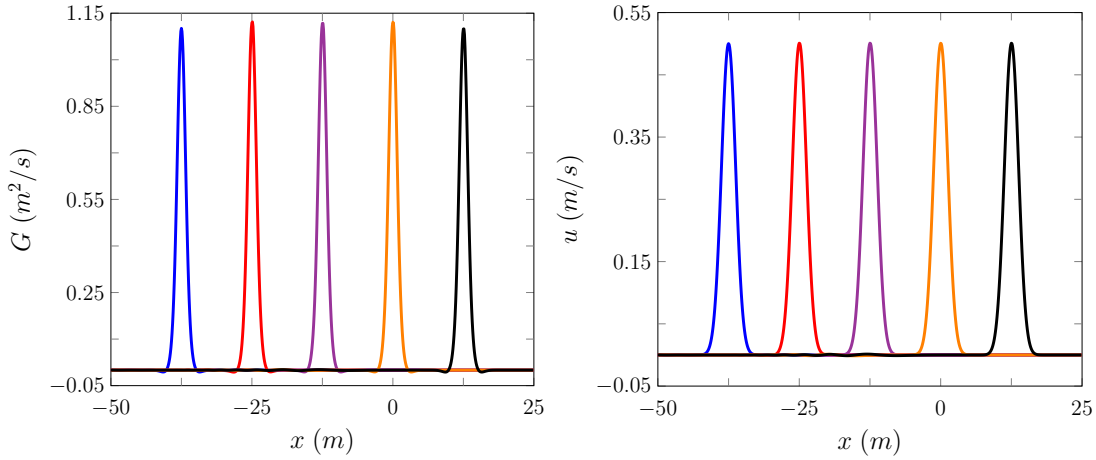
5.4.1 Results for Finite Water Depth

For the finite water depth case where $a_0 = 1m$ an example of the numerical solutions of FEVM_2 and FDVM_2 are given in Figures 5.7 and 5.8 respectively for $\Delta x = 100/2^{10} m \approx 0.0977m$ at various times. The numerical solutions and the forced solutions are identical at all times for these scales, accurately reproducing the forced solution as it travels over the bed.

The L_1 error of h, u and G for the FEVM_2 and FDVM_2 are given in Figure 5.9. Both methods recover the expected second-order accuracy. Since the source term of the modified Serre equations is added analytically and all terms must

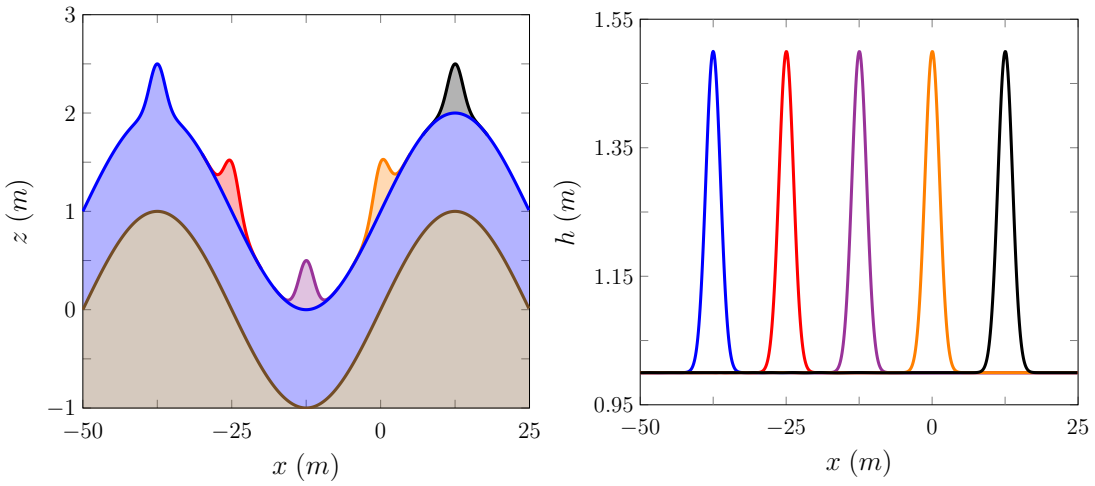


(a) w and b (□) (b) h



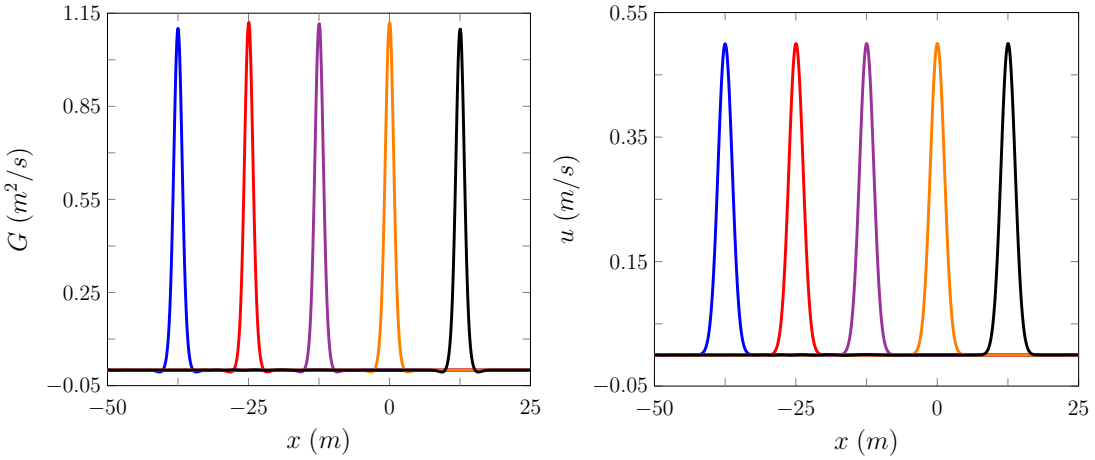
(c) G (d) u

Figure 5.7: Plots of w, b, h, G and u produced by FEVM$_2$ with $\Delta x = 100/2^{10}m$ at $t = 0s$ (– / □), $2.5s$ (–/ □), $5.0s$ (– / □), $7.5s$ (–/ □), $10.0s$ (– / □) of the finite water depth forced solution problem, where $a_0 = 1m$.



(a) w and b (□)

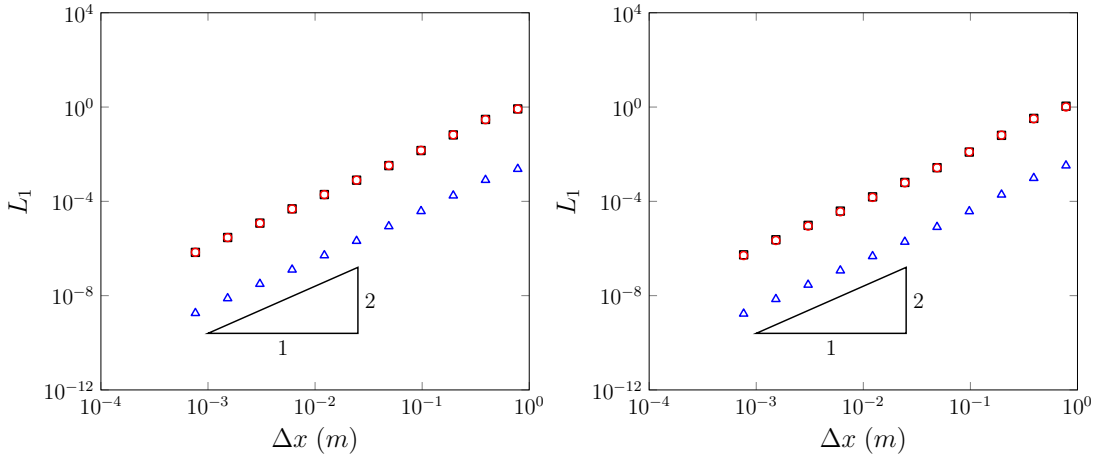
(b) h



(c) G

(d) u

Figure 5.8: Plots of w, b, h, G and u produced by FDVM$_2$ with $\Delta x = 100/2^{10}m$ at $t = 0s$ (– / □), $2.5s$ (–/ □), $5.0s$ (– / □), $7.5s$ (–/ □), $10.0s$ (– / □) of the finite water depth forced solution problem, where $a_0 = 1m$.



(a) FEVM$_2$ (b) FDVM$_2$

Figure 5.9: Convergence plots as measured by the L_1 norm for h (△), u (□) and G (○) for the finite water forced solution problem for FEVM and FDVM at $t = 10s$.

be accurately approximated by the method for this forced solution, these results demonstrate that our scheme is second-order accurate for all terms when the bed is wet everywhere, as desired.

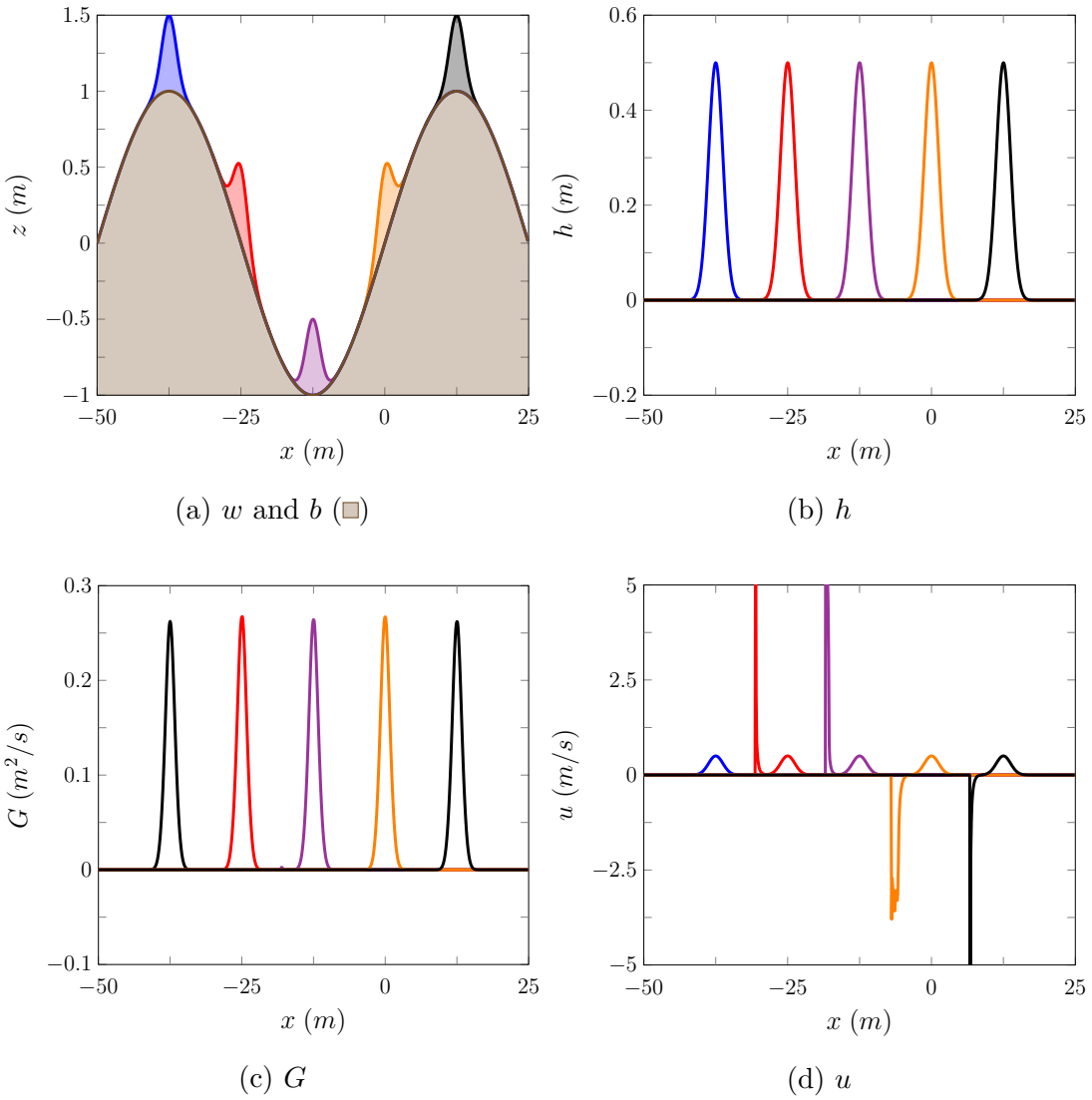
5.4.2 Results with Dry Beds

To demonstrate the capability of the methods to handle wetting and drying of beds, a series of numerical simulations of the forced solutions (5.5a) where $a_0 = 0m$ were conducted using both FEVM$_2$ and FDVM$_2$.

Example numerical solutions demonstrating the evolution of the wave are given in Figure 5.10 for FEVM$_2$ and Figure 5.11 FDVM$_2$ with $\Delta x = 100/2^{10}m \approx 0.0977m$ at various times. The methods accurately reproduce the analytic solution for the stage w, h and G. However, both fail to accurately reproduce u when h is small, particularly behind the wave.

These large errors in u when h is small are caused by the particular choices $h_{base} = 10^{-8}$ and $h_{tol} = 10^{-12}$ used in the desingularisation transformation applied to the elliptic solver, (3.14). By choosing larger values of these quantities the errors in u can be significantly damped. However, if h_{base} and h_{tol} are larger they dominate the L_1 errors for larger Δx values making the convergence less obvious. This trade-off is present in all desingularisation transforms.

For our purposes the chosen desingularisation transform (3.34) with small h_{base} and h_{tol} values are sufficient, resulting in large observed errors in u when h



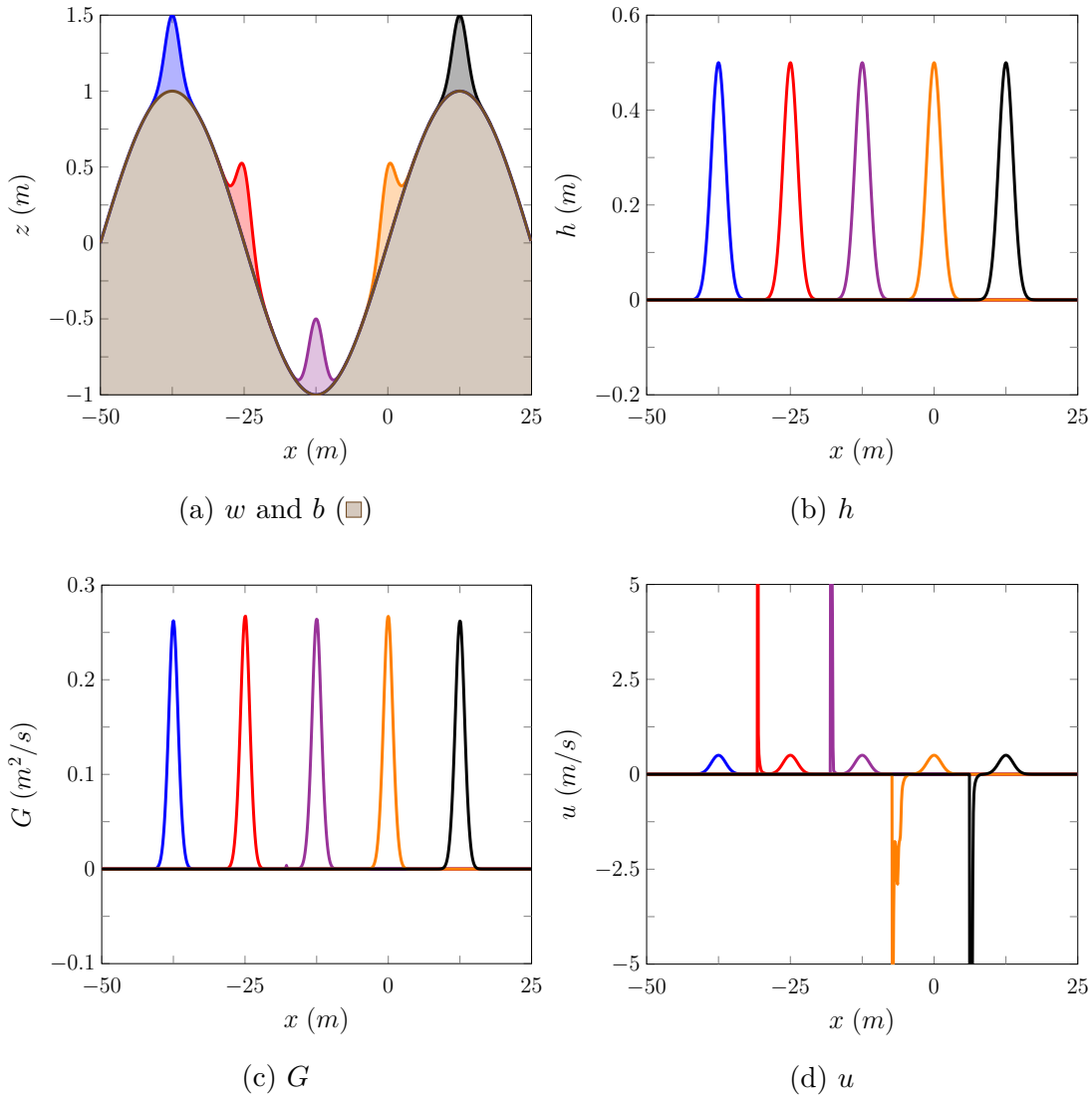
(a) w and b (□)

(b) h

(c) G

(d) u

Figure 5.10: Plots of w, b, h, G and u produced by FEVM$_2$ with $\Delta x = 100/2^{10}m$ at $t = 0s$ (– / □), $2.5s$ (–/ □), $5.0s$ (– / □), $7.5s$ (–/ □), $10.0s$ (– / □) of the dry bed forced solution problem, where $a_0 = 0m$.



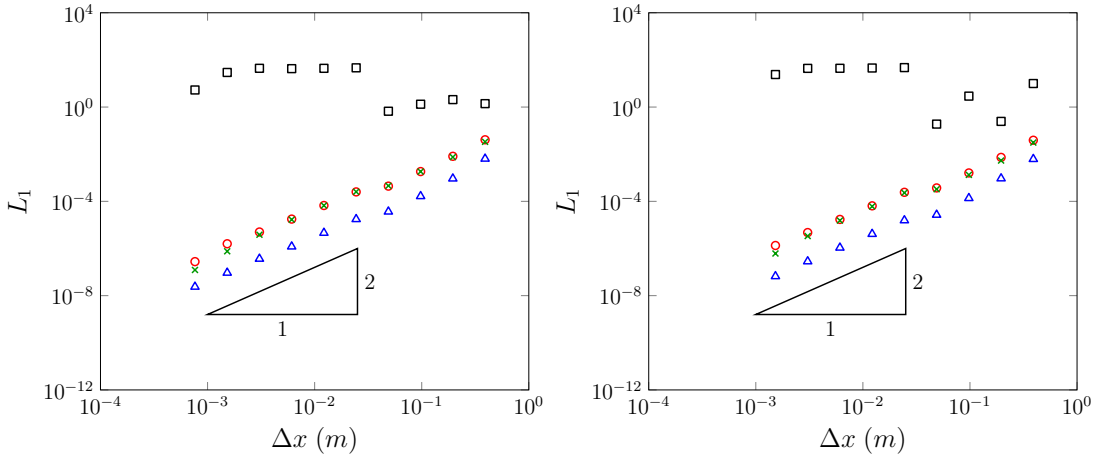
(a) w and b (□)

(b) h

(c) G

(d) u

Figure 5.11: Plots of w, b, h, G and u produced by FDVM$_2$ with $\Delta x = 100/2^{10}m$ at $t = 0s$ (– / □), $2.5s$ (–/ □), $5.0s$ (– / □), $7.5s$ (–/ □), $10.0s$ (– / □) of the dry bed forced solution problem, where $a_0 = 0m$.



(a) FEVM$_2$ (b) FDVM$_2$

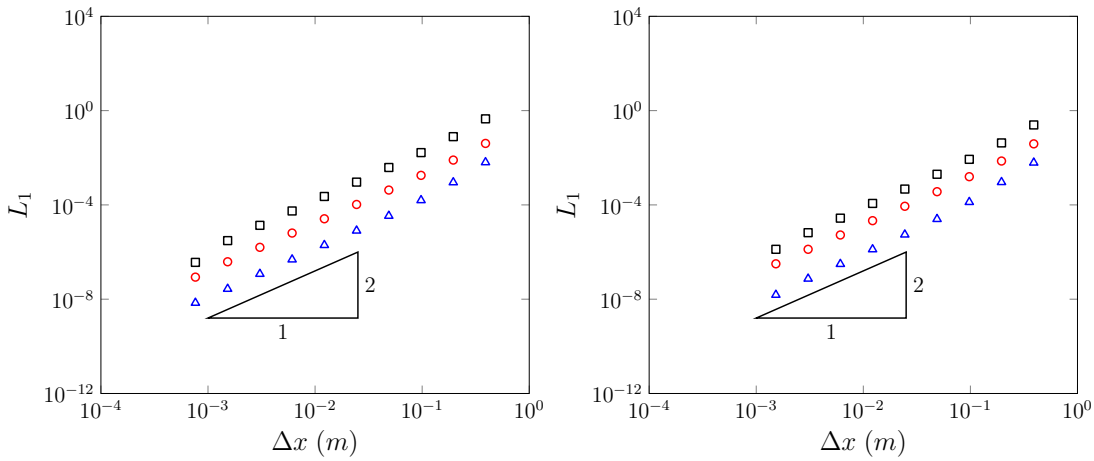
Figure 5.12: Convergence plots as measured by the L_1 norm for h (△), u (□), uh (×) and G (○) for the dry bed forced solution problem for FEVM and FDVM at $t = 10s$.

is small.

The L_1 errors for h, u, uh and G for both methods are given in Figure 5.12. Both methods exhibit second-order convergence in all the quantities except u. This is because all the flux and source terms of the Serre equations (2.6) only depend on u multiplied by some power of h; so that the large errors in u when h is small do not translate to significant errors in G, h or uh. Indeed by restricting the L_1 errors to compare only the regions where $h > 10^{-3}m$ as in Figure 5.13, we recover the expected second-order accuracy in all quantities.

Therefore, these methods can accurately handle the dry bed problem, even with small h_{base} and h_{tol} values, although in such cases the velocity may have large errors in regions where h is small.

In this chapter the analytic and forced solutions were used to assess the numerical methods. It was found that the hybrid finite volume methods performed better than the finite difference methods and that second-order methods were sufficient to accurately reproduce the analytic and forced solutions of the Serre equations.



(a) FEVM$_2$ (b) FDVM$_2$

Figure 5.13: Convergence plots for regions where $h > 10^{-3}m$ as measured by the L_1 norm for h (△), u (□) and G (○) for the dry bed forced solution problem for FEVM and FDVM at $t = 10s$.

Chapter 6

Experimental Validation

In this chapter the second-order hybrid finite volume methods are assessed using experimental data.

The numerical methods $FDVM_2$ and $FEVM_2$ are experimentally validated by comparing their numerical solutions to experimental data. The chosen experiments allow the methods capability to model a variety of physical situations to be tested. These situations include the presence of steep gradients in the flow, the interaction of strong dispersive waves with varying bathymetry, shoaling and wave breaking and finally the wetting and drying of a beach. Thus, the ability of these methods to reproduce all the experimental results well strongly demonstrates their capability to model all physical situations very well.

6.1 Evolution of Rectangular Depression

A series of experiments studying the evolution of rectangular depressions and thus steep gradients in the free-surface was conducted by Hammack and Segur [51]. These experiments were performed in a wave tank that was $0.394m$ wide, $31.6m$ long and $0.61m$ high. The rectangular depressions were generated using a piston $0.61m$ long with its left edge against the wave tank wall. The $0.1m$ deep water is initially stationary with a horizontal free surface and the piston in the up position. The experiment begins when the piston suddenly moves down. This creates a sudden depression in the water surface, generating waves that are recorded at wave gauges located at $0m$, $5m$, $10m$, $15m$ and $20m$ from the right edge of the piston. A diagram of the longitudinal section of the wave-tank with the wave gauge locations is given in Figure 6.1.

These experiments provide a good benchmark for the capability of the nu-

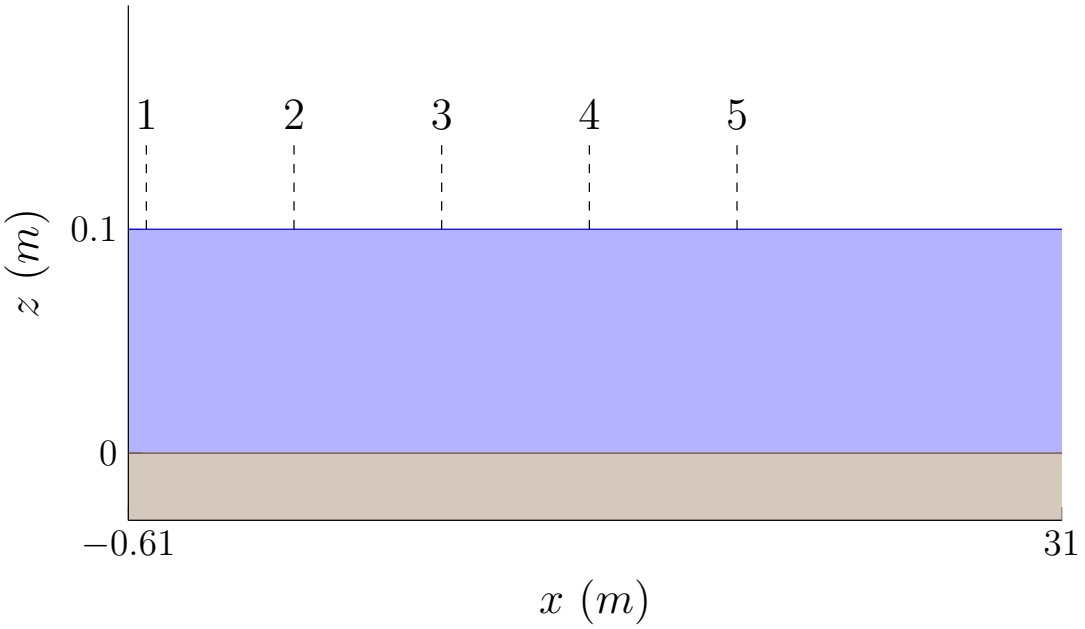


Figure 6.1: Diagram demonstrating the water (■) and the bed (■) for the Segur experiments, with the wave gauge locations marked.

merical method to accurately model problems with steep gradients in the free surface. These experiments are affected by bed friction and viscosity and the inability of the piston and water to move vertically instantaneously. Since the Serre equations do not contain viscosity, bed friction and we use discontinuous initial conditions we expect numerical solutions of the Serre equations to produce many more oscillations in the dispersive wave trains than are observed experimentally [12].

Hammack and Segur [51] report the results for two different initial depression depths $0.01m$ and $0.03m$, resulting in the nonlinearity parameters $\epsilon = 0.1$ and $\epsilon = 0.3$ respectively. Since these nonlinearity parameters are relatively small there was no breaking of waves throughout the experiment.

This experiment was modelled numerically using the reflected problem, with the wall as the axis of symmetry. In the numerical experiments the domain is $[-60m, 60m]$ and the experiment is run for $50s$ with $g = 9.81m/s^2$. For the spatial resolution we set $\Delta x = 0.01m$ to satisfy the CFL condition, (3.28) $\Delta t = 0.5\Delta x/\sqrt{g\ 0.1}$. The limiting parameter $\theta = 1.2$ was used in the reconstruction in FEVM_2 and FDVM_2.

6.1.1 Results for $0.01m$ Rectangular Depression

Plots comparing the numerical and experimental wave gauge data for the $0.01m$ rectangular depression are displayed in Figures 6.2 and 6.4 for FEVM_2 and FDVM_2 respectively. We present this data using the same dimensionless scales as reported in the original paper [51]. Tables 6.1 and 6.2 are also provided, which record the conservation of all the quantities.

The numerical solutions agree well with the experimental results; particularly for the front of the dispersive wave train. While all the conserved quantities have indeed been conserved very well by the methods.

The numerical solutions produce larger and consequently faster waves and observe oscillations in the depression which are not observed in the experimental data of wave gauge 1. Moreover, as expected the methods produce many more oscillations than were observed experimentally. These discrepancies can be attributed to the lack of viscosity and bed friction in the Serre equations (2.6). Furthermore, it is highly likely the experiment produced some smooth approximation to a discontinuous jump in the water depth with the down-stroke of the piston. Such a smoothing of the initial conditions will significantly attenuate the higher frequency waves in the generated dispersive wave train [12]. Given these challenges the numerical methods do a very good job of replicating the experimental behaviour.

Both FEVM_2 and FDVM_2 have produced visually identical results at this scale and have demonstrated very good conservation of all the quantities see, Tables 6.1 and 6.2. Given the extensive review of these methods [12] for steep gradient problems, this indicates that these solutions are indicative of true solutions of the Serre equations which are capable of reproducing experimental results.

6.1.2 Results for $0.03m$ Rectangular Depression

The wave gauge data for the numerical and experimental results for the evolution of the $0.03m$ rectangular depression are displayed in Figures 6.5 and 6.6 for FEVM_2 and FDVM_2 respectively. These results are reported using the same dimensionless scales as the original paper [51]. The conservation of all the conserved quantities are given in Tables 6.3 and 6.3 for FEVM_2 and FDVM_2 respectively.

Both methods again reproduce the overall behaviour of this experiment very well. Because the rectangular wave is deeper, this experiment provides a more rigorous test for the numerical methods. However, increasing the depth also strengthens the causes of the discrepancy between the experimental results and

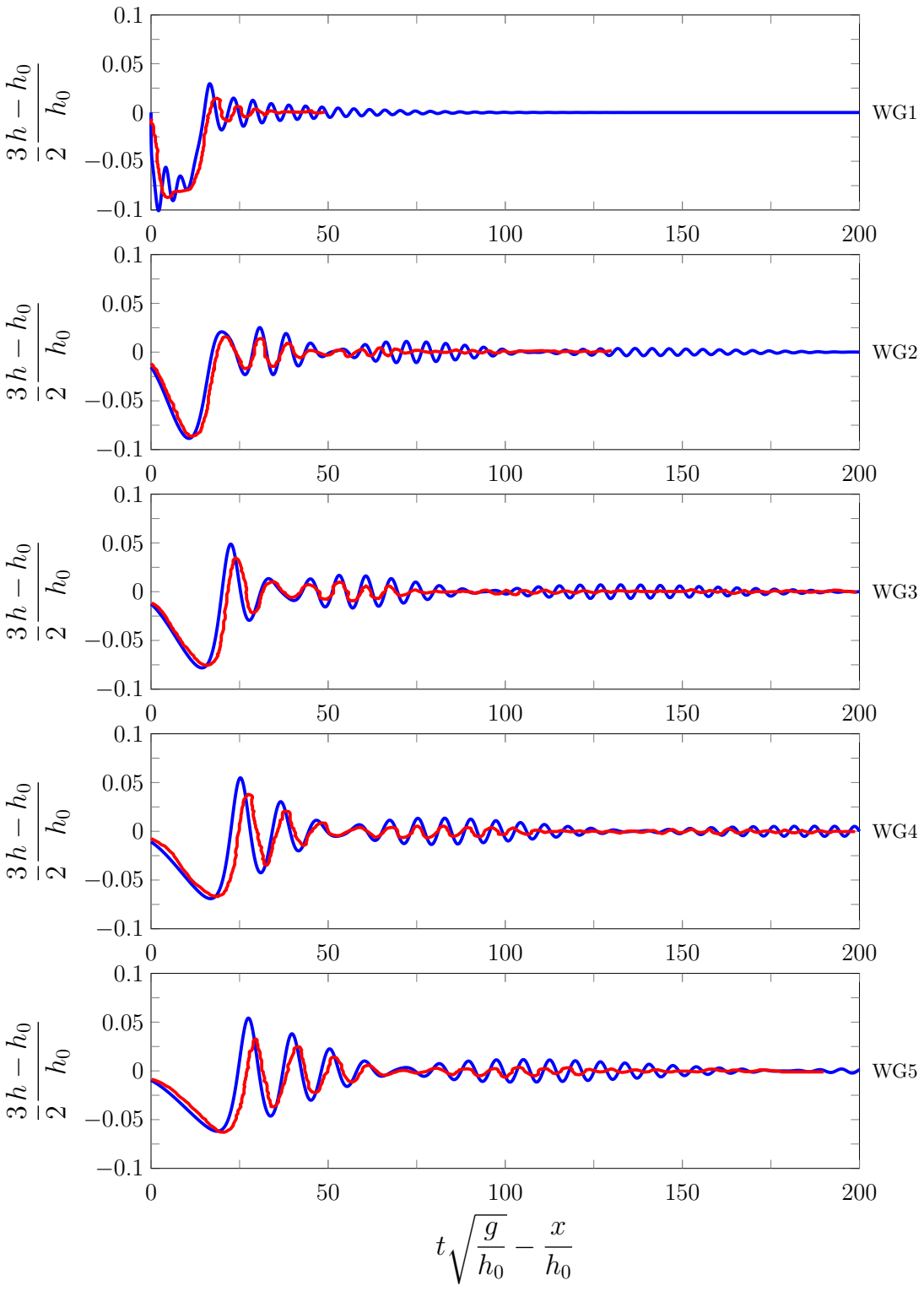


Figure 6.2: Comparison of experimental wave gauge data (–) and numerical results (–) of FEVM$_2$ for the $0.01m$ rectangular depression.

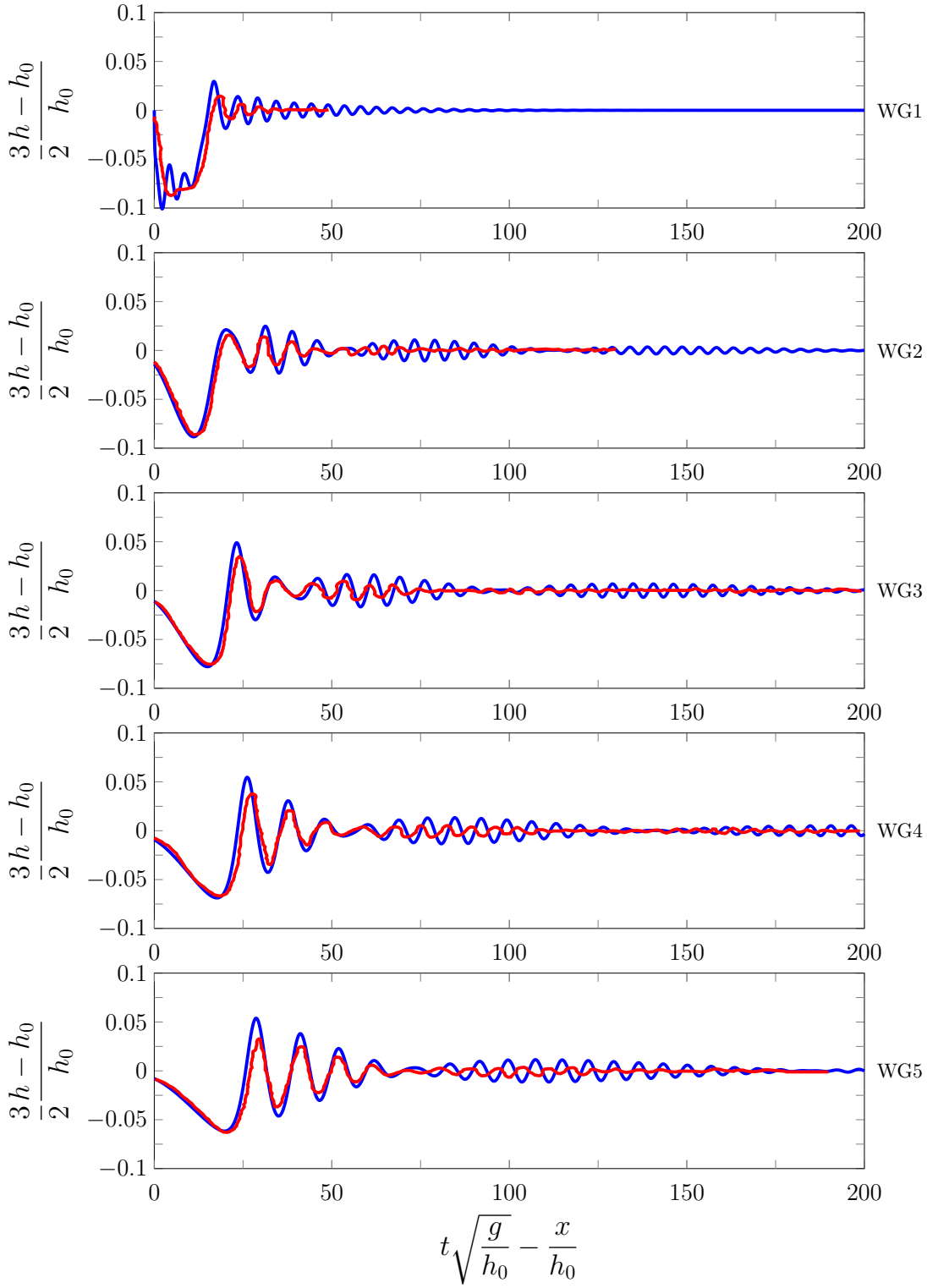


Figure 6.3: FDVM

Figure 6.4: Comparison of experimental wave gauge data (–) and numerical results (–) of FDVM$_2$ for the $0.01m$ rectangular depression.

Quantity	$\mathcal{C}^*(\boldsymbol{q}^0)$	$\mathcal{C}^*(\boldsymbol{q}^*)$	$\mathcal{C}_1^*(\boldsymbol{q}^0, \boldsymbol{q}^*)$
h	11.9888	11.9888	0
uh	0	7.44×10^{-18}	7.44×10^{-18}
G	0	1.56×10^{-18}	1.56×10^{-18}
$\mathcal{H}$	5.8751	5.8751	5.70×10^{-6}

Table 6.1: Initial and final total amounts and the conservation error for all conserved quantities for FEVM_2 numerical solution of the $0.01m$ rectangular depression.

Quantity	$\mathcal{C}^*(\boldsymbol{q}^0)$	$\mathcal{C}^*(\boldsymbol{q}^*)$	$\mathcal{C}_1^*(\boldsymbol{q}^0, \boldsymbol{q}^*)$
h	11.9888	11.9888	0
uh	0	-1.19×10^{-17}	-1.19×10^{-17}
G	0	-8.05×10^{-18}	-8.05×10^{-18}
$\mathcal{H}$	5.8751	5.8751	6.27×10^{-6}

Table 6.2: Initial and final total amounts and the conservation error for all conserved quantities for FDVM_2 numerical solution of the $0.01m$ rectangular depression.

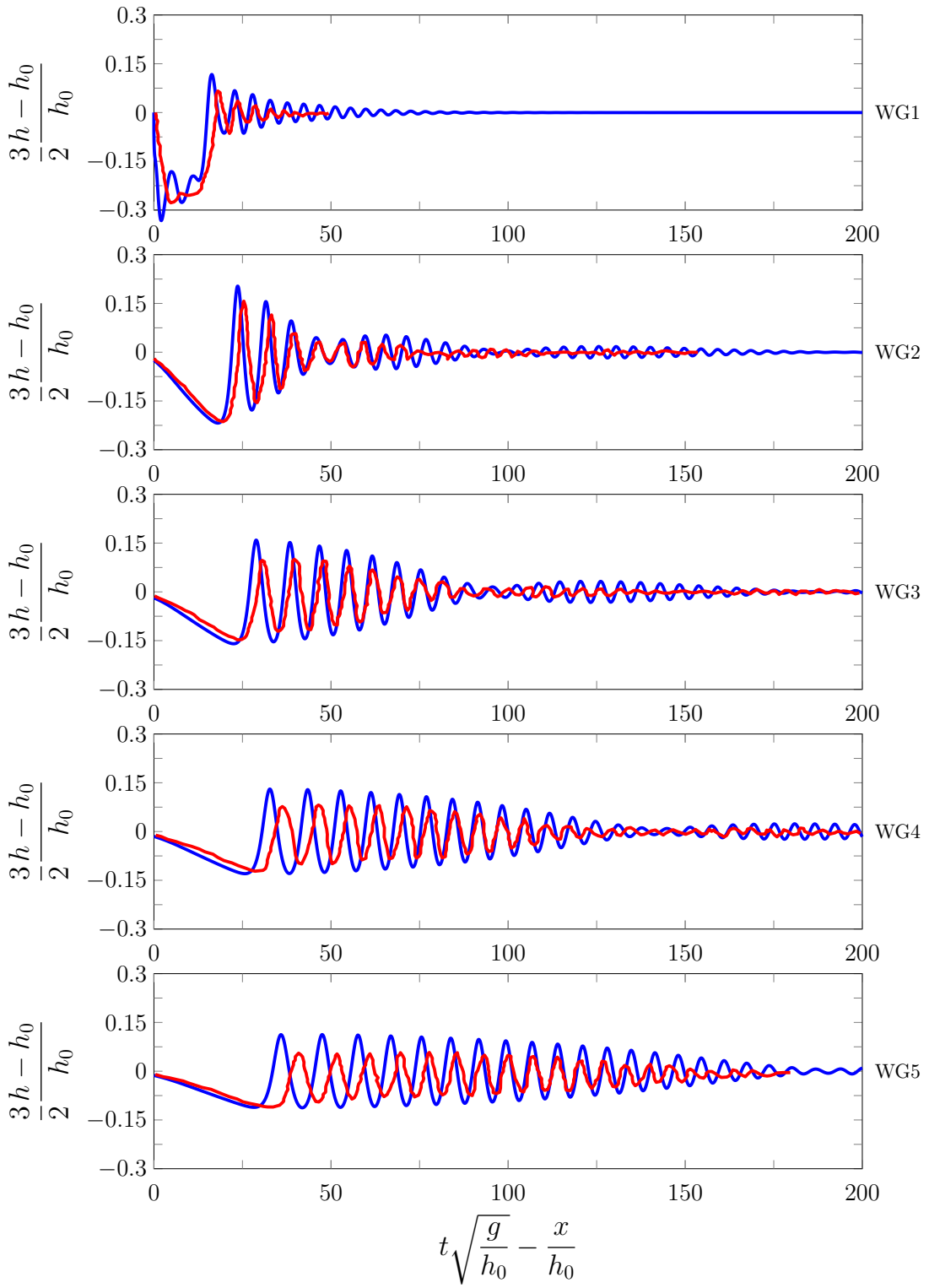


Figure 6.5: Comparison of experimental wave gauge data (–) and numerical results (–) of FEVM$_2$ for the $0.03m$ rectangular depression.

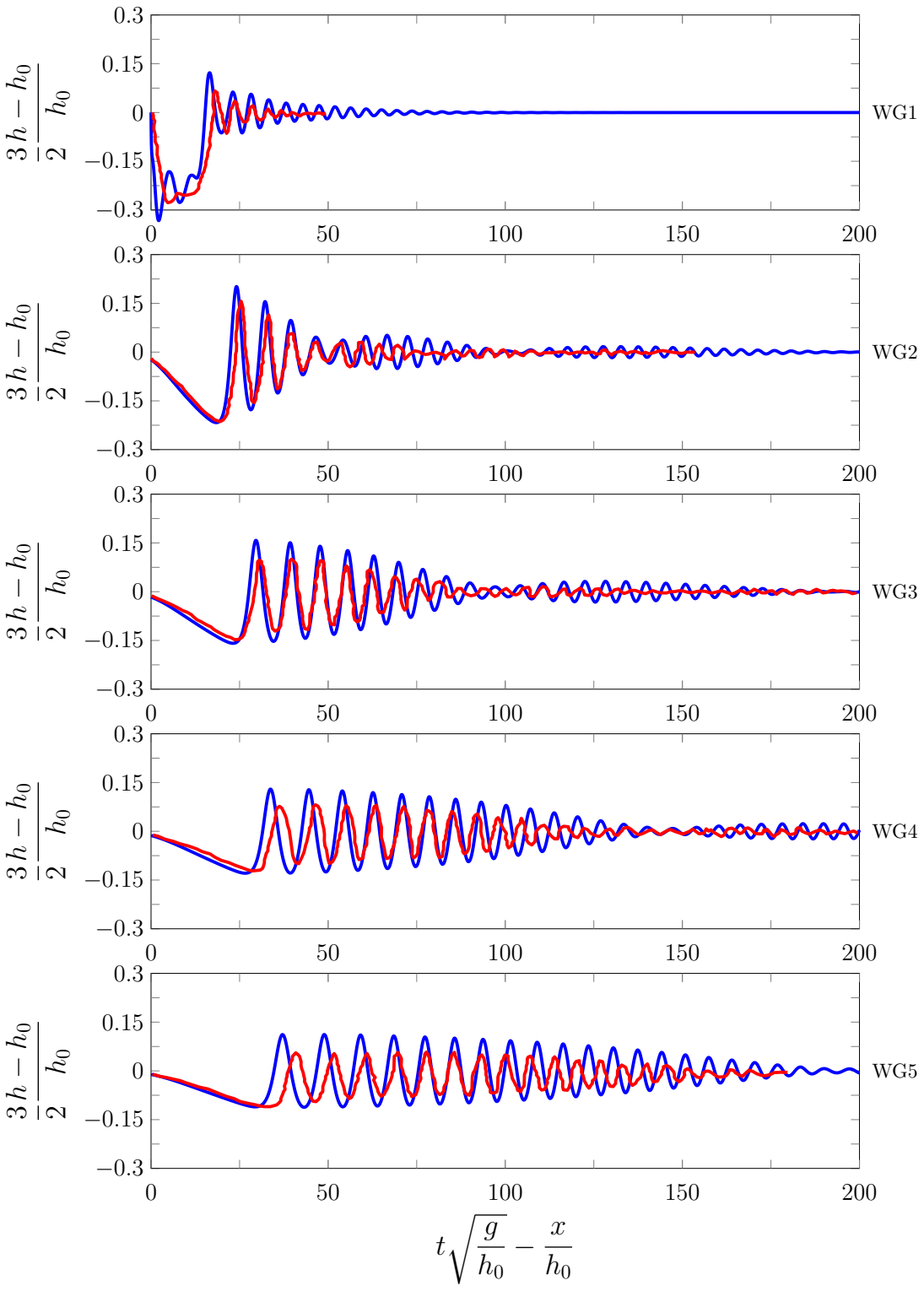


Figure 6.6: Comparison of experimental wave gauge data (–) and numerical results (–) of FDVM$_2$ for the $0.03m$ rectangular depression.

Quantity	$\mathcal{C}^*(\boldsymbol{q}^0)$	$\mathcal{C}^*(\boldsymbol{q}^*)$	$\mathcal{C}_1^*(\boldsymbol{q}^0, \boldsymbol{q}^*)$
h	11.9644	11.9644	0
uh	0	-7.75×10^{-17}	-7.75×10^{-17}
G	0	-3.33×10^{-16}	-3.33×10^{-16}
$\mathcal{H}$	5.8560	5.8552	1.24×10^{-4}

Table 6.3: Initial and final total amounts and the conservation error for all conserved quantities for FEVM_2 numerical solution of the $0.03m$ rectangular depression.

the numerical solutions of the Serre equations. This can be seen most acutely for the amplitude and speed of the generated waves.

Since the rectangular depression is larger the numerical methods have a larger error in conservation for all the quantities as compared to the $0.01m$ rectangular depression except mass; which is conserved exactly. For G and momentum these errors are around machine epsilon and can be disregarded, so that only the conservation of energy is significantly effected. Even with this larger error, all quantities are still well conserved by the numerical methods, suggesting that the numerical solutions well approximate the true solutions of the Serre equations.

These experiments have been well replicated by the numerical methods, and given the resolution and error in conservation and the extensive study summarised in Chapter 2; these results demonstrate the accuracy of the numerical methods in the presence of steep gradients in the free surface.

6.2 Periodic Waves Over A Submerged Bar

Beji and Battjes conducted a series of experiments investigating the effect of submerged bars on the propagation of periodic waves [52, 53]. The behaviour of these experiments were mainly driven by the dispersion properties of the waves and their interaction with variations in bathymetry. Therefore, these experiments serve as a benchmark for the ability of the numerical schemes to accurately model the interaction of variable bathymetry and dispersive waves. For our purposes we will focus on the monochromatic wave experiments of Beji and Battjes [53].

The experiments of Beji and Battjes [53] were conducted in a wave tank $37.7m$

Quantity	$\mathcal{C}^*(\boldsymbol{q}^0)$	$\mathcal{C}^*(\boldsymbol{q}^*)$	$\mathcal{C}_1^*(\boldsymbol{q}^0, \boldsymbol{q}^*)$
h	11.9644	11.9644	0
uh	0	-9.09×10^{-17}	-9.09×10^{-17}
G	0	-1.16×10^{-16}	-1.16×10^{-16}
$\mathcal{H}$	5.8560	5.8552	1.30×10^{-4}

Table 6.4: Initial and final total amounts and the conservation error for all conserved quantities for FDVM_2 numerical solution of the $0.03m$ rectangular depression.

long, $0.8m$ wide and $0.75m$ high. A diagram of the longitudinal section of the wave tank is given in Figure 6.7. There are seven wave gauges at the following locations; $5.7m$, $10.5m$, $12.5m$, $13.5m$, $14.5m$, $15.7m$ and $17.3m$. Waves are generated from a piston-type wave maker located at $0m$ and travel on the initially still water $0.4m$ deep to the right, over the submerged trapezoidal bar and are absorbed by a sloping beach.

Two sinusoidal monochromatic non-breaking wave experiments were conducted. A low frequency one with a wavelength $\lambda \approx 3.69m$ and a period of $T = 2s$, and a high frequency one with $\lambda \approx 2.05m$ and a period of $T = 1.25s$. Both experiments had a wave amplitude of $0.01m$ and so both had the same small non-linearity parameter $\epsilon = 0.01/0.4 = 0.025$.

We numerically simulated these experiments over the spatial domain $[5.7m, 150m]$ with $\Delta x = 0.1/2^4 m \approx 0.0063m$ and $\Delta t = Sp/2^5 s \approx 0.0012s$ where $Sp = 0.039s$ is the experimental sampling period. These Δx and Δt values satisfy the CFL condition, (3.28). In our numerical experiments only the submerged trapezoidal bar is present, and the sloping beach is replaced with a very long horizontal bed that ensures that we do not observe any effects from the Dirichlet boundary conditions at the downstream boundary.

To simulate the incoming waves at the upstream boundary we used the first wave gauge as our left boundary condition together with linear extrapolation to calculate the other required h values in the left ghost cell. The velocity boundary conditions were calculated from the height values in the same way as Beji and Battjes [53]

$$u(x,t) = \sqrt{gh_0}\,\frac{h(x,t) - h_0}{h(x,t)}.$$

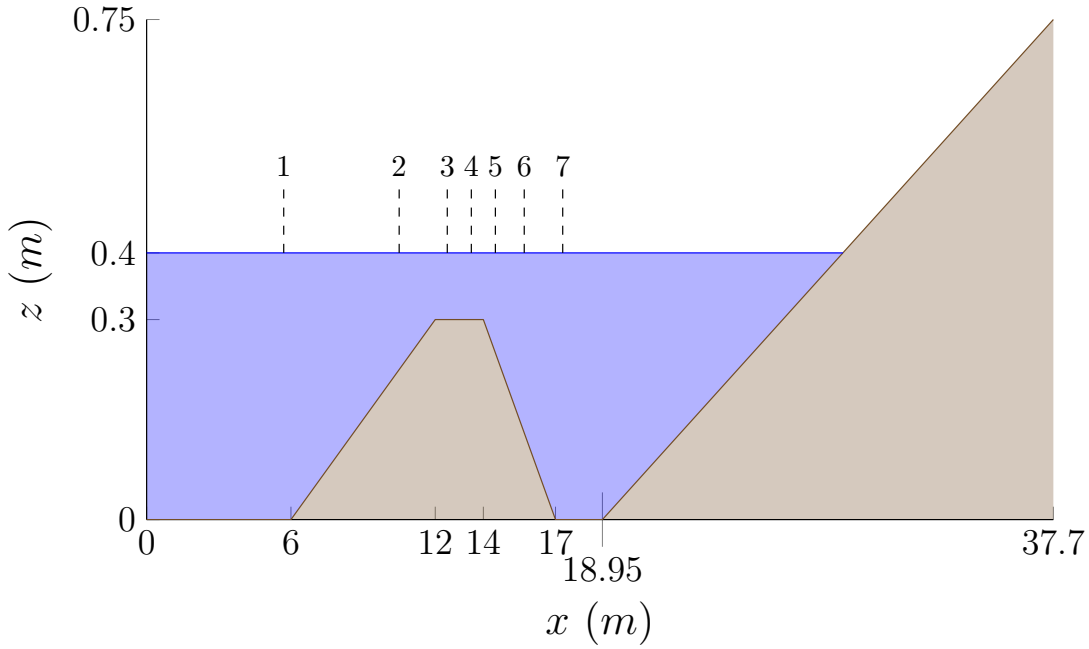


Figure 6.7: The flume configuration with water (■) and the bed (■) for the Beji experiments, with the wave gauge locations marked.

Finally the boundary conditions for G were calculated using the boundary conditions for h and u.

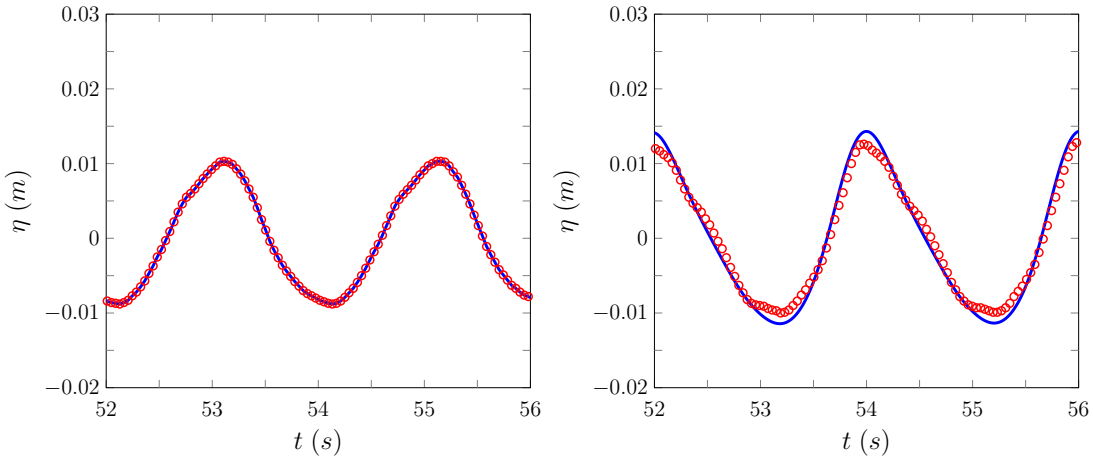
We shall now present our numerical results for the low and high frequency experiments.

6.2.1 Low Frequency Results

A comparison of the wave heights η of the experimental and numerical results are located in Figures 6.8 and 6.9 for FEVM_2 and Figures 6.10 and 6.11 for FDVM_2. These numerical schemes both produce identical results for all wave gauges and so this benchmark does not help us discriminate between these two methods.

These results demonstrate the ability of these numerical methods to recreate the experimental results, particularly for wave gauge 1 to 5 where the agreement between experimental and numerical results is best. Results at these gauges validate the numerical schemes for simulating shoaling of dispersive waves as these wave gauges are all located on the windward side of the submerged bar where shoaling occurs in the experiment.

The numerical results for wave gauges 6 and 7 on the leeward side capture some of the wave behaviour but their agreement with the experiments results is



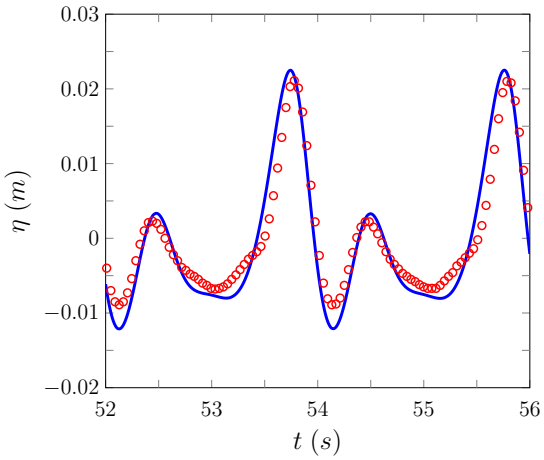
(a) Wave Gauge 1

(b) Wave Gauge 2

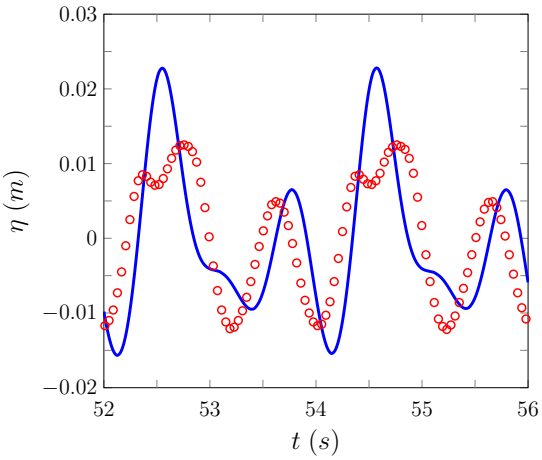
(c) Wave Gauge 3

(d) Wave Gauge 4

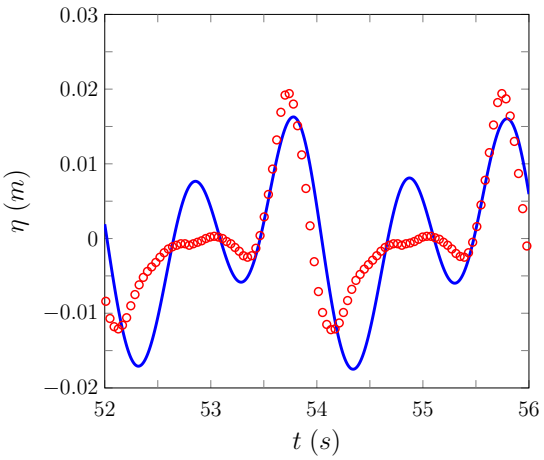
Figure 6.8: Comparison of the wave heights η of the numerical results for the FEVM$_2$ (–) and the experimental results (○) for wave gauges 1 - 4 for the low frequency experiment.



(a) Wave Gauge 5

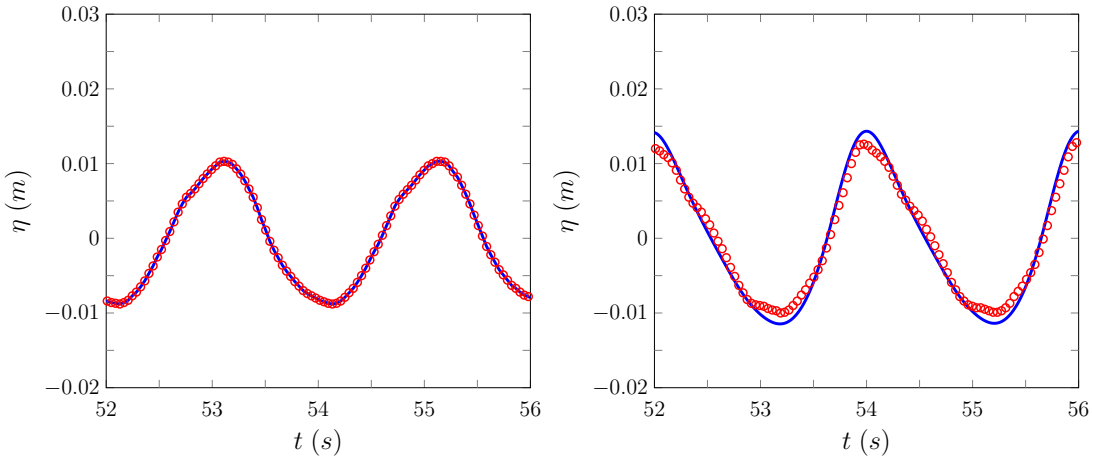


(b) Wave Gauge 6



(c) Wave Gauge 7

Figure 6.9: Comparison of the wave heights η of the numerical results for the FEVM$_2$ (–) and the experimental results (○) for wave gauges 5 - 7 for the low frequency experiment.



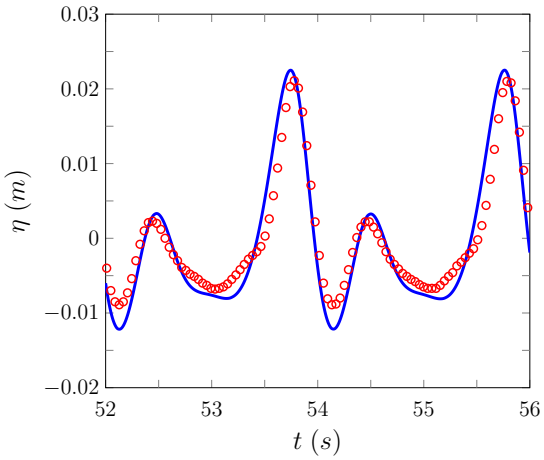
(a) Wave Gauge 1

(b) Wave Gauge 2

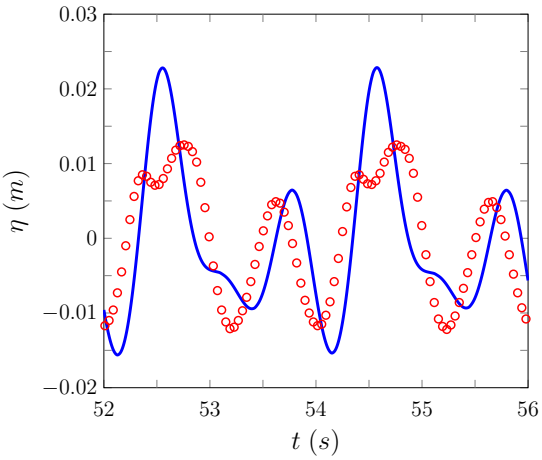
(c) Wave Gauge 3

(d) Wave Gauge 4

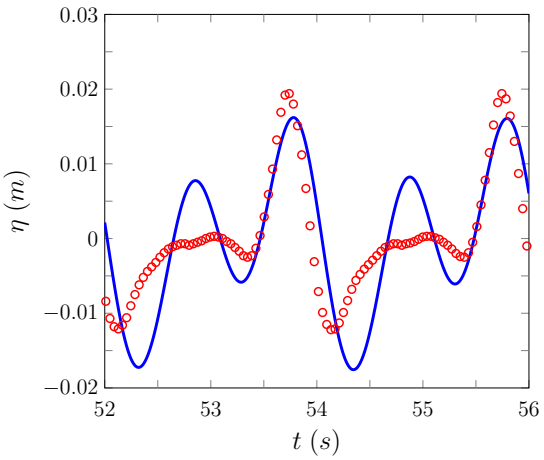
Figure 6.10: Comparison of the wave heights η of the numerical results for the FDVM$_2$ (–) and the experimental results (○) for wave gauges 1 - 4 for the low frequency experiment.



(a) Wave Gauge 5



(b) Wave Gauge 6



(c) Wave Gauge 7

Figure 6.11: Comparison of the wave heights η of the numerical results for the FDVM$_2$ (–) and the experimental results (○) for wave gauges 5 - 7 for the low frequency experiment.

much worse. The inadequacy of the numerical results here appears to be due to the discrepancy between the dispersion properties of the Serre equations and actual water waves [53, 54].

The dispersion terms in the Serre equations are vital to recreating the experimental results for wave gauges 2 to 5, as non-dispersive equations such as the SWWE are not capable of accurately simulating this experiment [26].

6.2.2 High Frequency Results

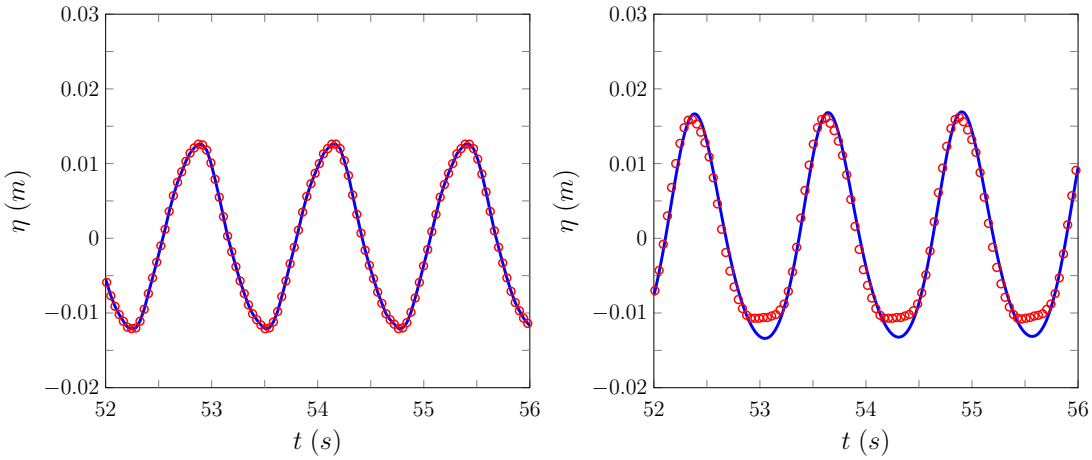
The wave heights of the experimental and numerical results are given in Figures 6.12 and 6.13 for FEVM_2. While the results for FDVM_2 are given in Figures 6.14 and 6.15. As for the low frequency experiment FEVM_2 and FDVM_2 produce identical results for all wave gauges at this scale and so this benchmark does not discriminate between these two methods.

As in the low frequency experiment we observe that the numerical results perform well on the windward side of the slope for wave gauges 1 to 4 but perform poorly for the leeward side of the slope for wave gauges 5 to 7. With the high frequency experiment we see the divergence between the numerical and experimental results earlier than the low frequency experiment, so that now wave gauge 5 which is on the leeward side exhibits a significant difference between the numerical and experimental results. As in the low frequency example this is caused by the difference in the dispersion relations of the Serre equations and the linear theory for water waves [53, 54]. Because the difference between the dispersion relation of the Serre equations and water waves is largest for higher frequency and therefore for shorter waves [19] the earlier divergence between experimental and numerical results is not surprising.

These numerical results for the FDVM_2 and FEVM_2 agree well with other numerical results for weakly dispersive equations without improved dispersion properties for the simulation of periodic waves over a submerged bar in the literature [53, 54, 23, 55]. Therefore, without changing the underlying partial differential equations, our numerical methods perform as well as other numerical schemes in the literature at recreating the experimental results of Beji and Battjes [53].

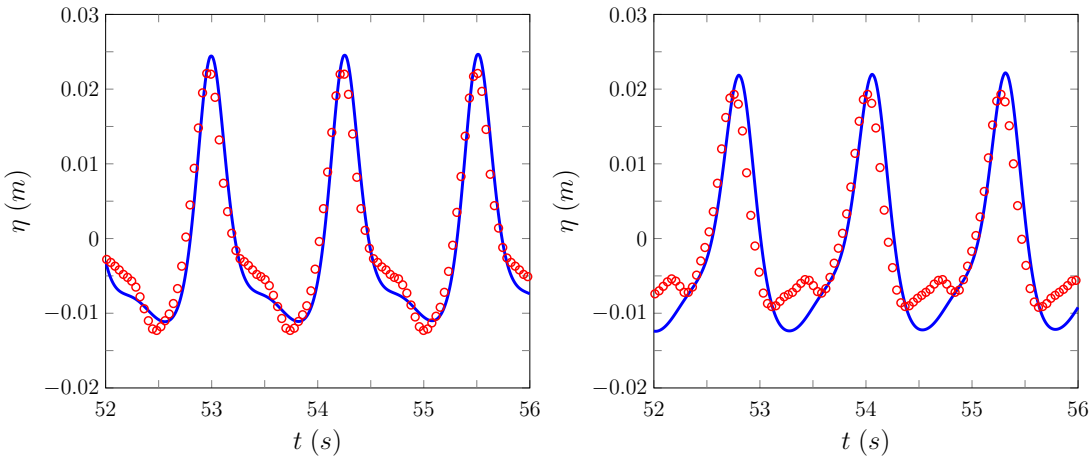
6.3 Solitary Wave Over a Fringing Reef

To study the evolution of waves on fringing reefs a series of experiments were conducted by Roeber [56]. These experiments were performed in a wave tank



(a) Wave Gauge 1

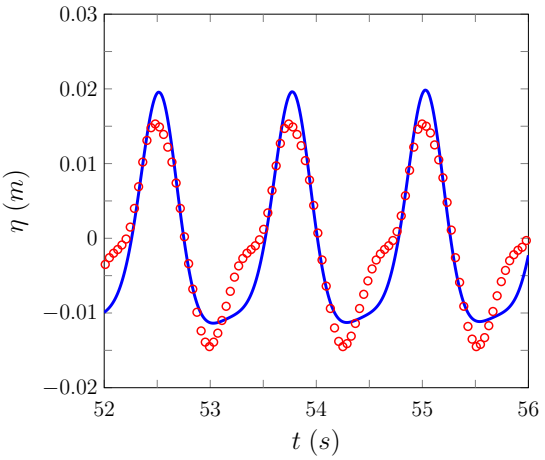
(b) Wave Gauge 2



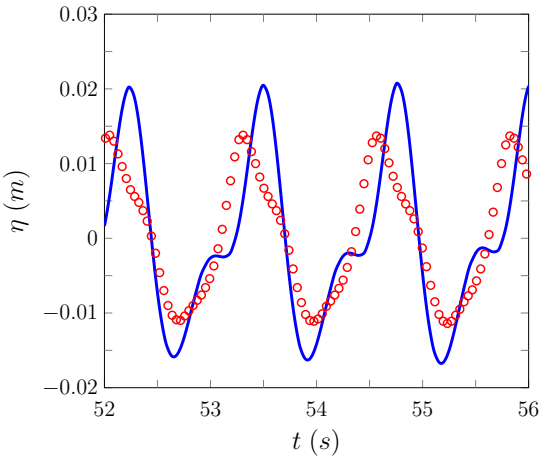
(c) Wave Gauge 3

(d) Wave Gauge 4

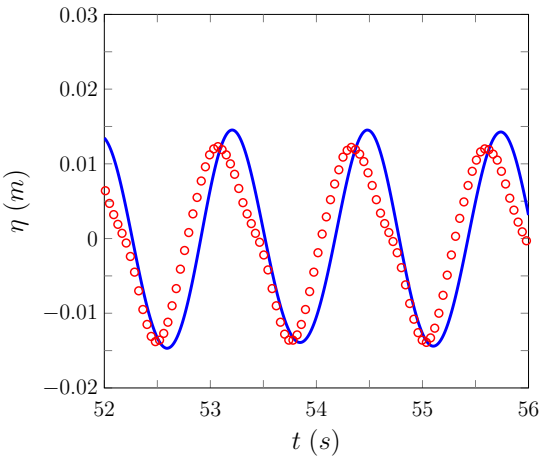
Figure 6.12: Comparison of the wave heights η of the numerical results for the FEVM$_2$ (–) and the experimental results (○) for wave gauges 1 - 4 for the low frequency experiment.



(a) Wave Gauge 5

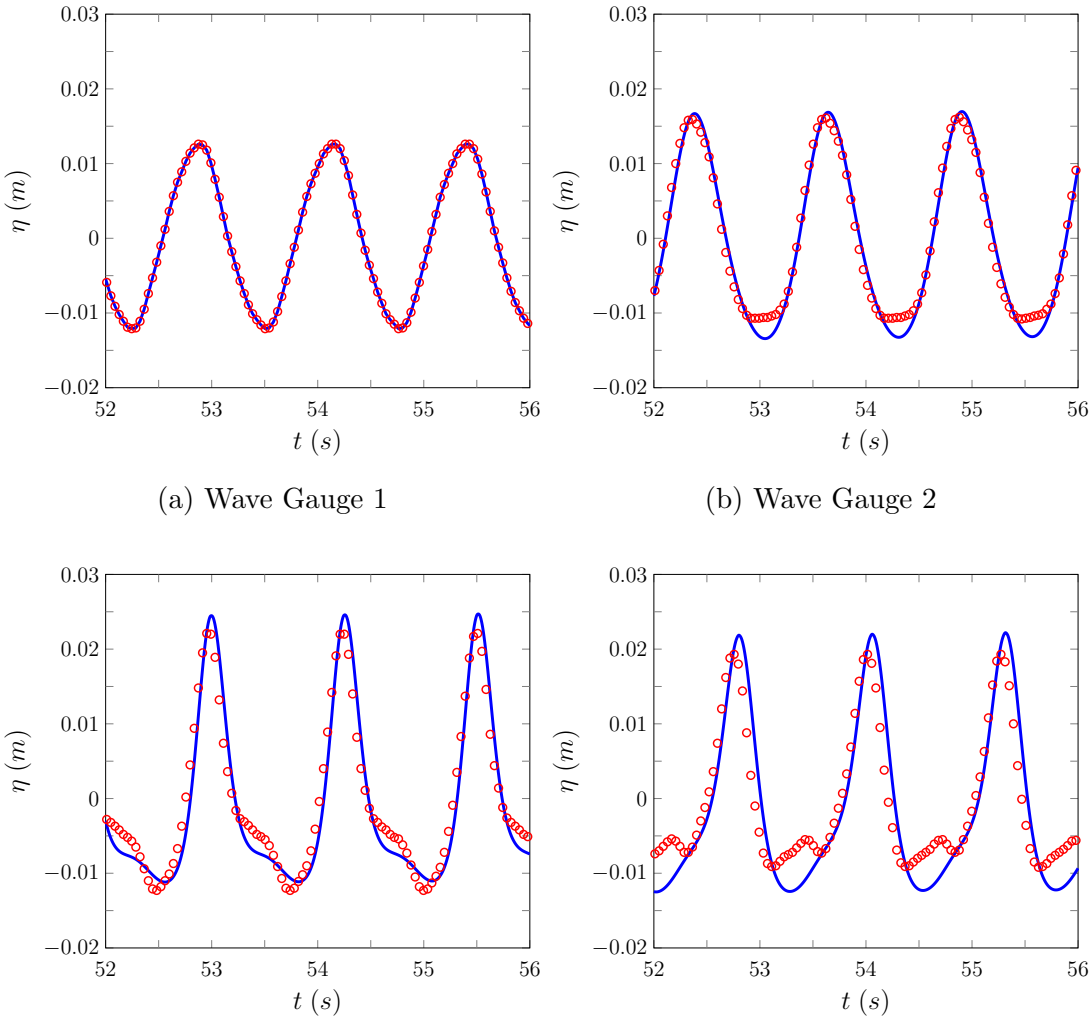


(b) Wave Gauge 6



(c) Wave Gauge 7

Figure 6.13: Comparison of the wave heights η of the numerical results for the FEVM$_2$ (–) and the experimental results (○) for wave gauges 5 - 7 for the high frequency experiment.



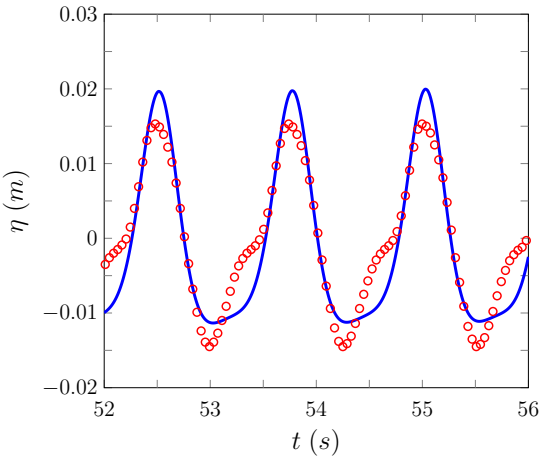
(a) Wave Gauge 1

(b) Wave Gauge 2

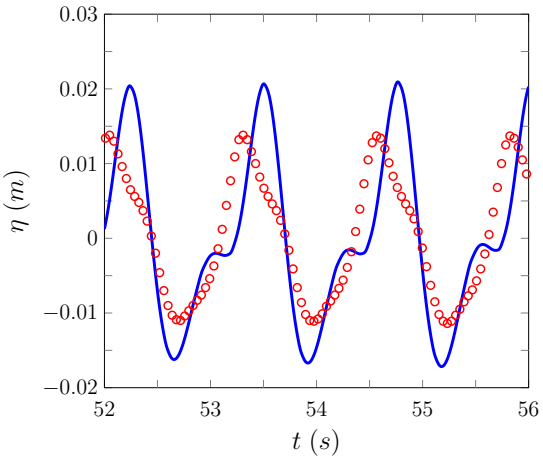
(c) Wave Gauge 3

(d) Wave Gauge 4

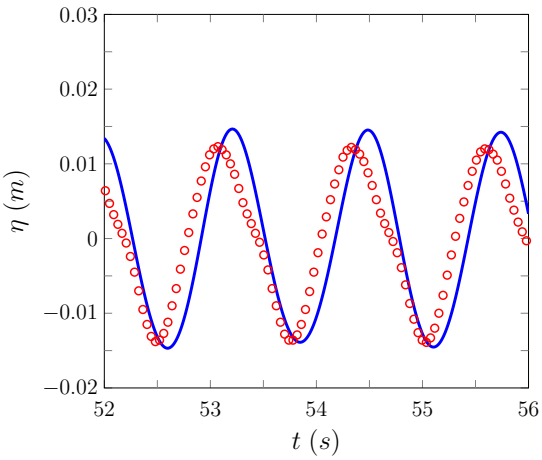
Figure 6.14: Comparison of the wave heights η of the numerical results for the FDVM$_2$ (–) and the experimental results (○) for wave gauges 1 - 4 for the high frequency experiment.



(a) Wave Gauge 5



(b) Wave Gauge 6



(c) Wave Gauge 7

Figure 6.15: Comparison of the wave heights η of the numerical results for the $FDVM_2$ (–) and the experimental results (○) for wave gauges 5 - 7 for the high frequency experiment.

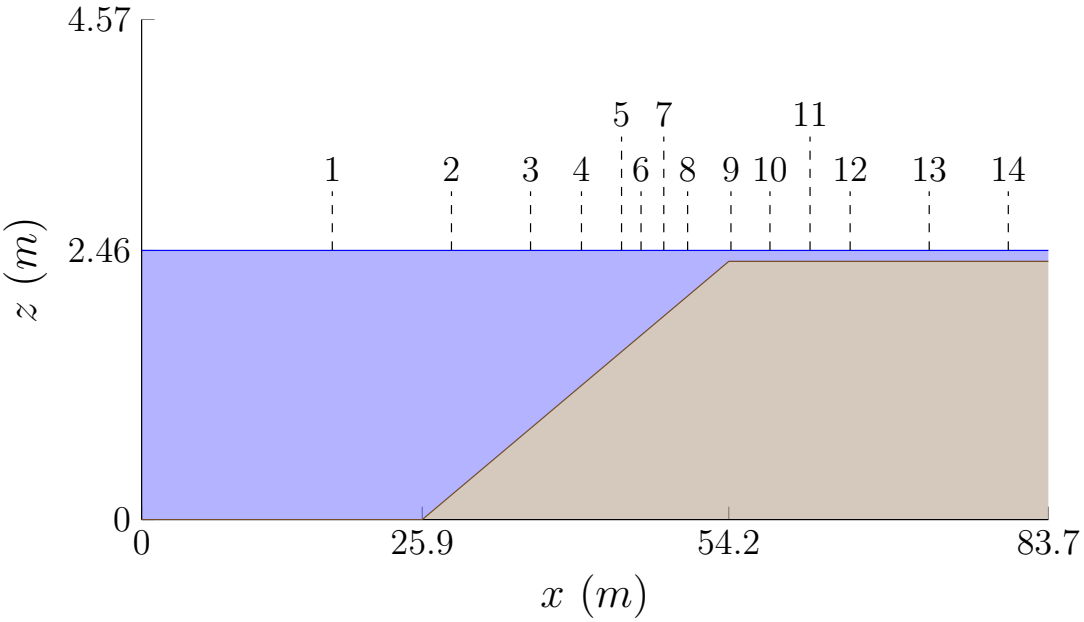


Figure 6.16: Diagram demonstrating the water (■) and the ground (■) for the solitary wave over a fringing reef experiment, with the wave gauge locations marked.

$3.66m$ wide, $83.7m$ long and $4.57m$ high with a removable bed that allowed for the wide range of experiments reported by Roeber [56]. We have computationally modelled the experiment with the bathymetry displayed in Figure 6.16, where a solitary wave is generated from the wave maker at $0m$ and is recorded at the wave gauges $17.6m$, $28.6m$, $35.9m$, $40.6m$, $44.3m$, $46.1m$, $48.2m$, $50.4m$, $54.4m$, $58.0m$, $61.7m$, $65.4m$, $72.7m$ and $80.0m$ downstream of the wave maker.

This experiment investigates the behaviour of waves with high nonlinearity $\epsilon \approx 1.23/2.46 = 0.5$ as it shoals over a linear bed into a very shallow body of water. Given the high nonlinearity of this wave, it is not surprising that as it shoals it becomes a plunging breaker by $t \approx 32s$ with an elliptical air cavity observed at $t \approx 33s$ [56]. As with other depth averaged equations, the Serre equations cannot naively model breaking waves so this experiment is not an entirely appropriate test of the numerical methods, particularly after $t = 32s$.

This experiment was numerically modelled on the domain $[17.6m, 400m]$ and was run until $t = 60s$ after which the reflections from the downstream end of the tank become significant in the experiment. The beginning of the domain was chosen so that wave gauge 1 could be used as the left boundary conditions, where the technique for the boundary condition in section 6.2 was employed.

The spatial resolution was $\Delta x = 0.025m$ and the temporal resolution was $\Delta t = Sp/8s = 0.0025$ where $Sp = 0.02s$ was the sampling period of the wave gauges, these spatial and temporal resolutions satisfy the CFL condition (3.28). The right edge of the domain used Dirichlet boundary conditions, since the domain was so large no effects from the downstream boundary were observed throughout the numerical simulation.

6.3.1 Results

The wave gauge results comparing the numerical and experimental data are displayed in Figures 6.17, 6.18 and 6.19 for FEVM_2 and 6.20 and 6.21 for FDVM_2.

Both methods accurately reproduce the shoaling of the solitary wave, particularly in wave gauges 1 through 8 which record the wave before breaking begins. The behaviour of the trailing waves is not as well replicated, with the numerical solutions overestimating their amplitude and speed as in the previous experiments. The reflected wave can also be observed in the wave gauges and since the numerical simulation did not have reflective boundaries these waves are not replicated in their solution.

When breaking begins the numerical solutions perform much worse as expected; most notably FDVM_2 becomes unstable and the solution blows up. Because of this the numerical solution of FDVM_2 was only plotted until $t = 34s$. The instability is caused by the appearance of a very steep gradient with a large jump in the water depth compared to the depth of water that surrounds it as the wave breaks. The FEVM_2 method does not suffer from these instability issues, but due to the limitations of the Serre equations does produce a dispersive wave train with amplitudes far exceeding the observed amplitudes of the experiment.

Given the limitations of the underlying Serre equations the results for FEVM_2 are robust and accurately model the shoaling of the solitary wave. However, these results indicate the need for more accurate handling of breaking waves to be able to accurately model physical situations.

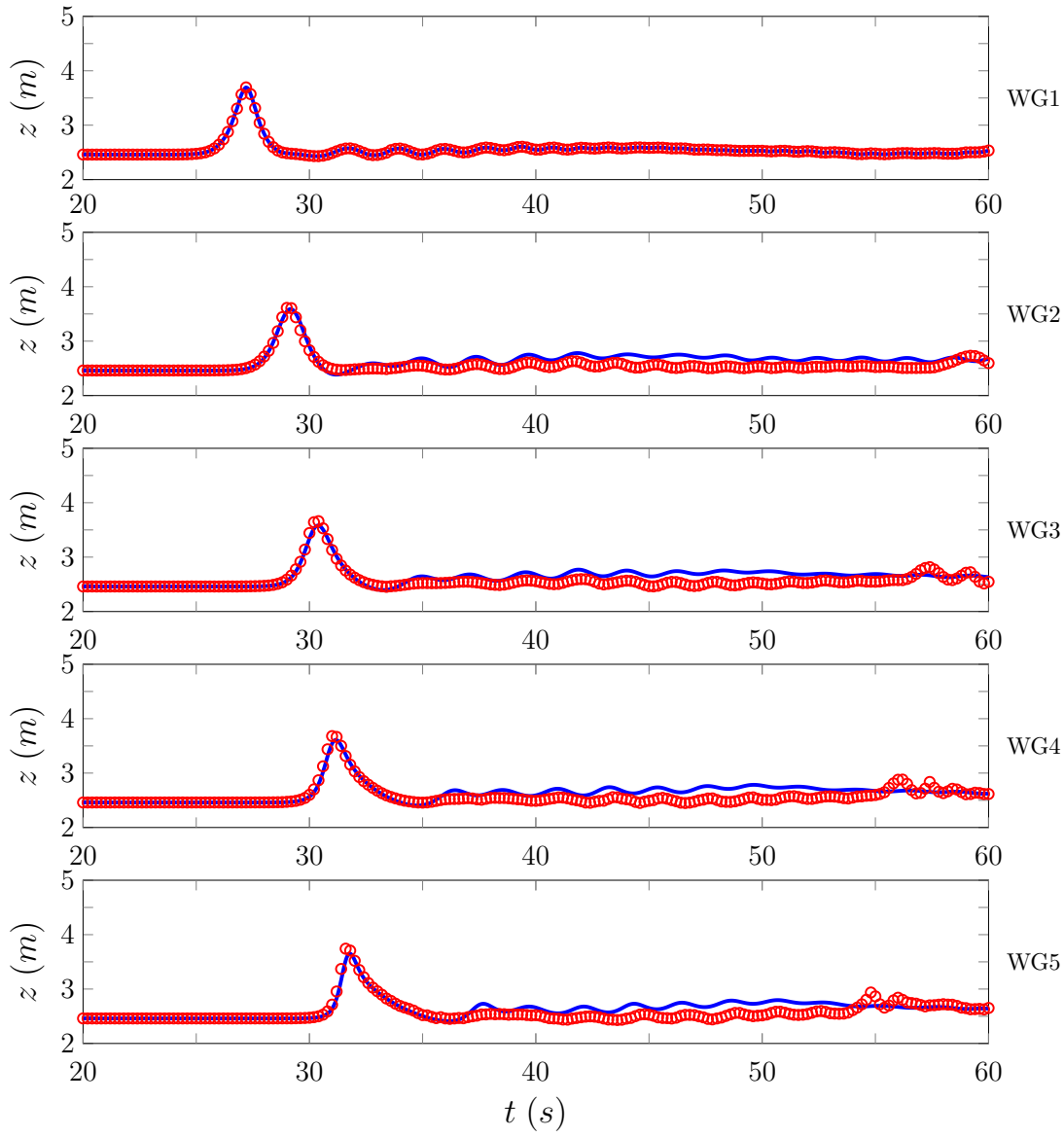


Figure 6.17: Comparison of the experimental (○) and numerical (–) wave gauge data produced by FEVM_2 for gauges 1 to 5.

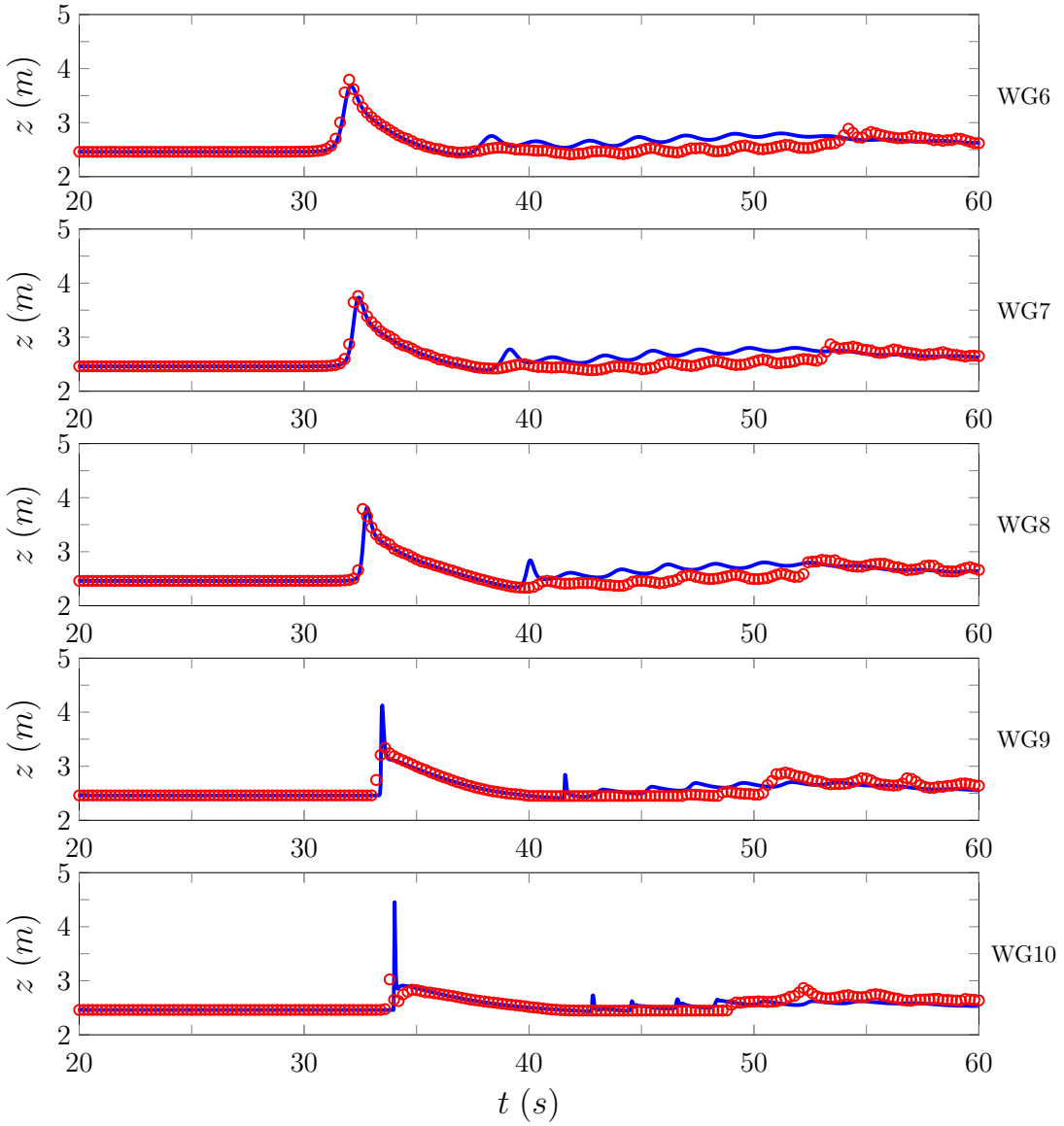


Figure 6.18: Comparison of the experimental (○) and numerical (–) wave gauge data produced by FEVM_2 for gauges 6 to 10.

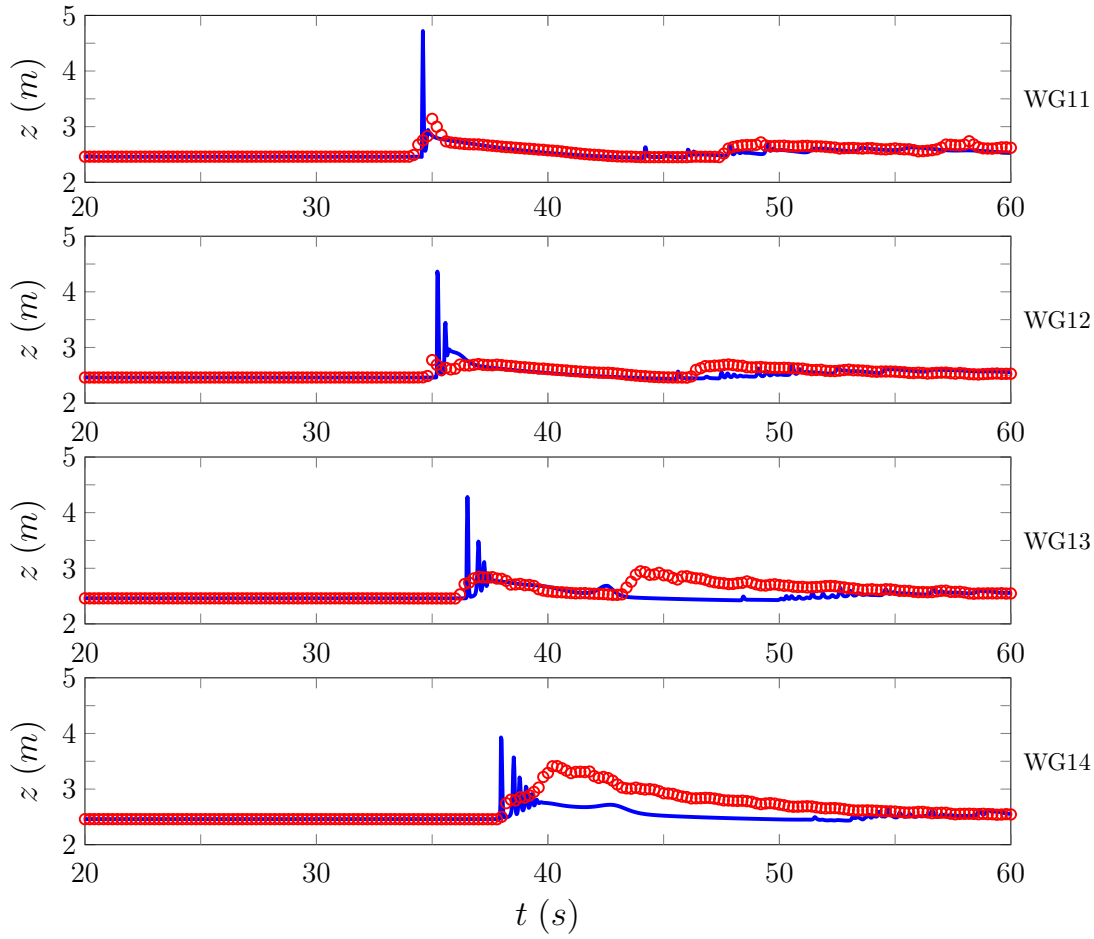


Figure 6.19: Comparison of the experimental (○) and numerical (—) wave gauge data produced by FEVM_2 for gauges 11 to 14.

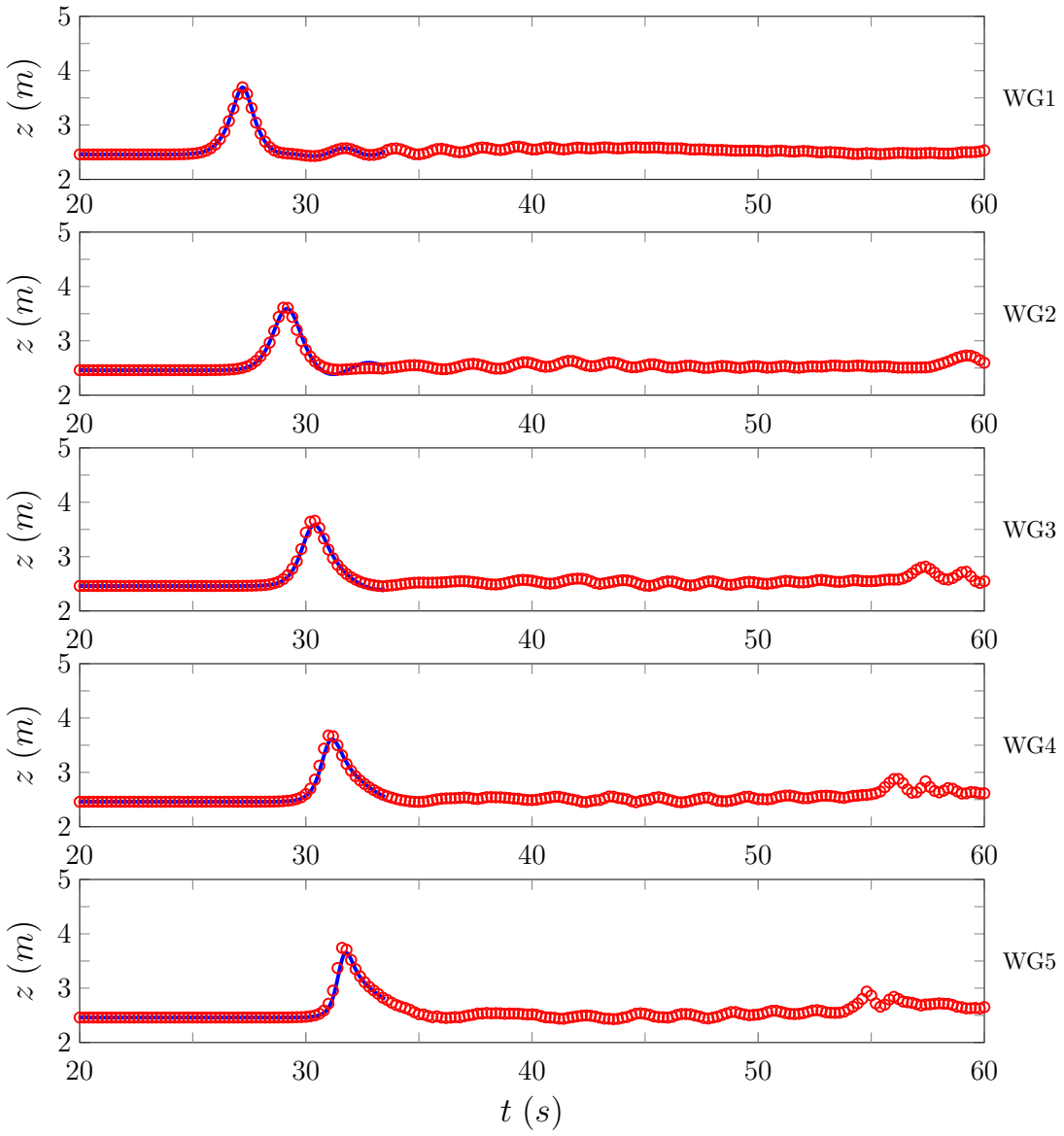


Figure 6.20: Comparison of the experimental (○) and numerical (–) wave gauge data produced by $FDVM_2$ for gauges 1 to 7.

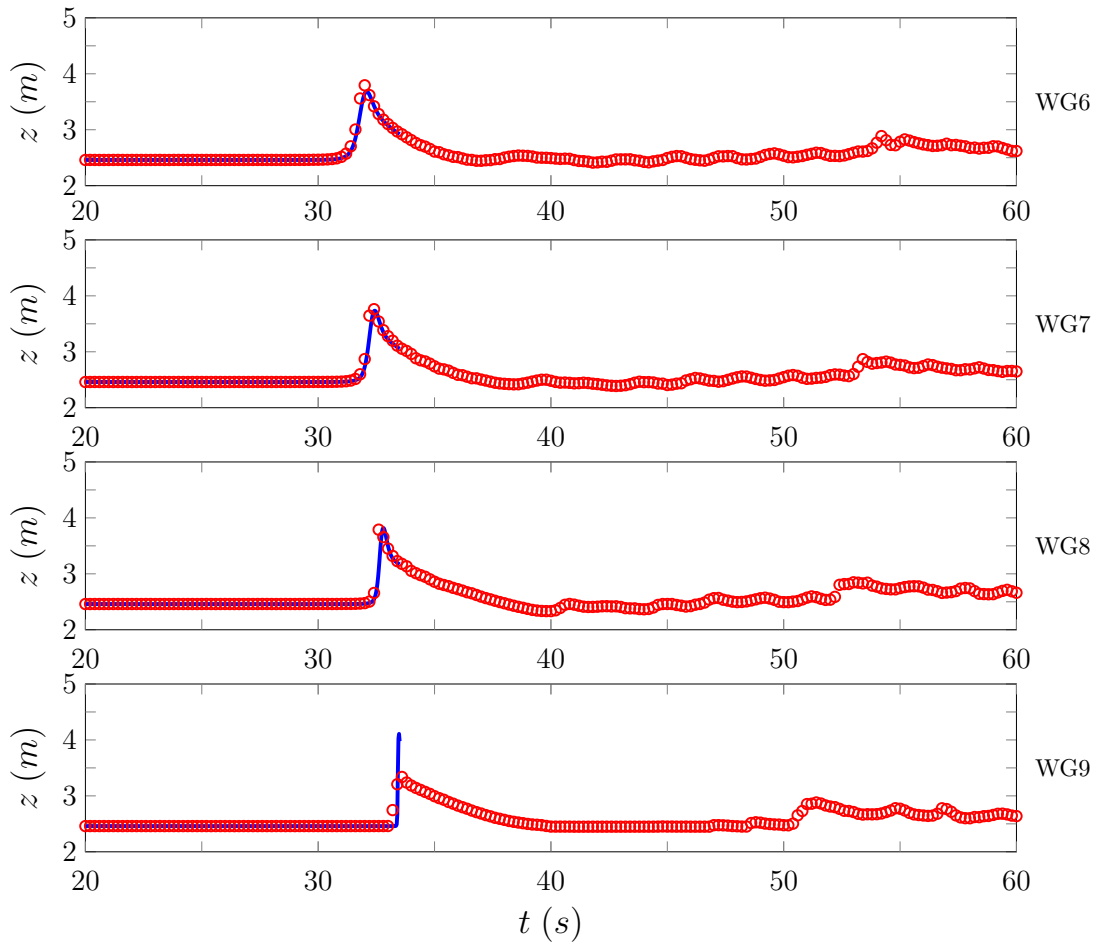


Figure 6.21: Comparison of the experimental (○) and numerical (–) wave gauge data produced by $FDVM_2$ for gauges 6 to 9.

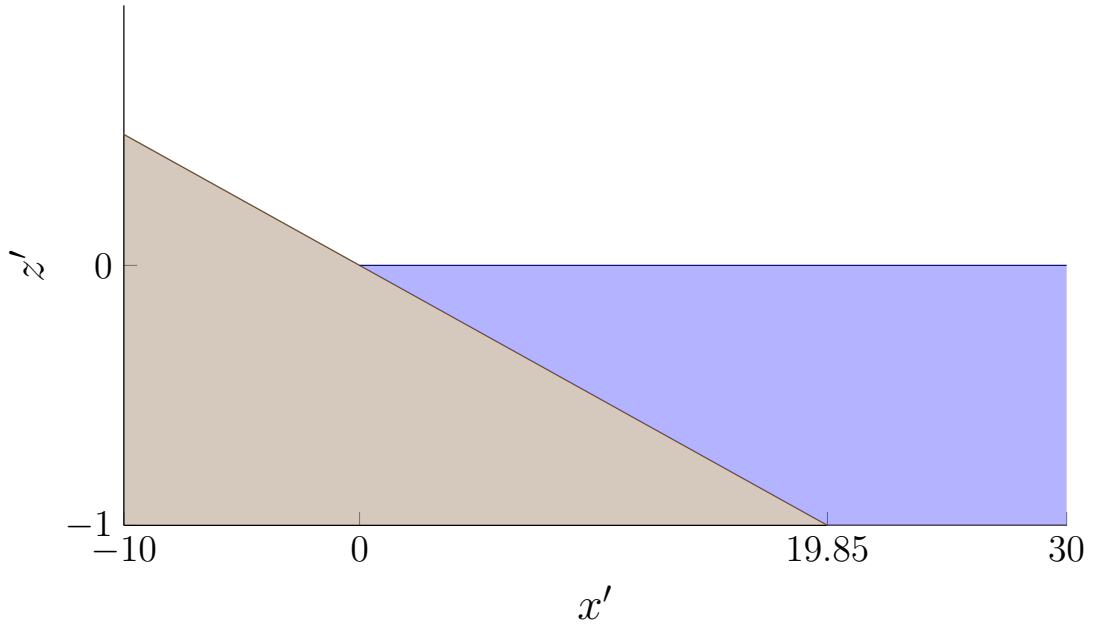


Figure 6.22: Diagram demonstrating the water (▫) and the ground (▫) for the Synolakis experiment with the normalised coordinates.

6.4 Runup of a Solitary Wave on a Linearly Sloped Beach

To study the run-up of incoming waves on linear beaches a series of experiments were conducted by Synolakis [57]. These experiments consisted of a number of runup events for a wide array of breaking and non-breaking waves where snapshots of the entire water surface were taken at certain times. These experiments were all performed on the beach profile depicted in Figure 6.22, where all the quantities are normalised [57]. To assess the computational models we recreated one of these experiments, which captured the runup of a non breaking solitary wave with a nonlinearity parameter of $\epsilon = 0.0185$.

This experiment allows us to compare the inundation behaviour of our numerical methods with experimental results. For this experiment the effect of dispersion on the run-up behaviour is minimal, and there is good agreement between numerical solutions of the SWWE and this particular experiment [58]. Therefore, the effect of the extra dispersive terms included by the Serre equations on the inundation process is not well tested by this experiment, but it does demonstrate robustness.

The numerical experiments used the normalised quantities reported by Syn-

olakis [57] to reproduce the experiment. The spatial domain was $x' \in [-30, 150]$ with a resolution of $\Delta x = 0.05$ and was run until $t' = 70$ with the CFL condition (3.28) satisfied by setting $\Delta t = 0.1\Delta x$. The spatial reconstruction used the input parameter $\theta = 1.2$ and gravity was normalised to match the coordinates and so $g = 1$.

6.4.1 Results

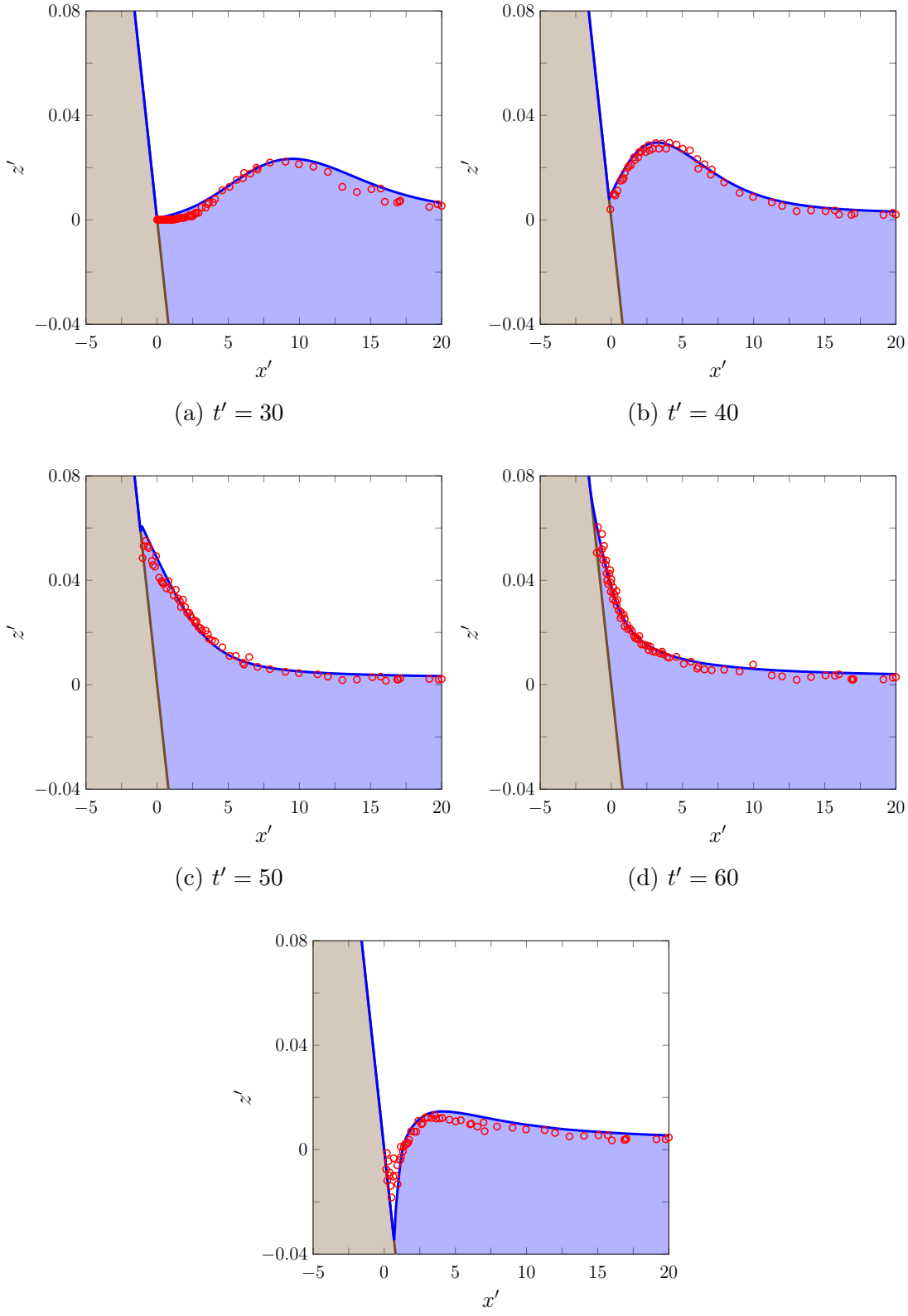
The normalised water surface data is given at the various times in Figure 6.24 for FDVM_2 and 6.23 for FEVM_2. The error in conservation of the conserved quantities are given in Tables 6.5 and 6.6 for FEVM_2 and FDVM_2 respectively.

Both methods reproduce the experimental results very well, replicating the incoming wave properties and the maximum runup very well. The experimental wave appears to be more skewed towards the shoreline, but this shape difference has all but disappeared as the wave begins to inundate the shore. The only other noticeable difference is that the numerical solutions appear to run down further than the experimental results. The observed larger rundown is likely caused by the lack of friction in the Serre equations.

The conserved quantities are well conserved by the method throughout the run up and rundown of the wave, particularly the mass. The total energy of the method is also well conserved, however the energy appears to have slightly increased in the method during the run-up process due to the handling of the dry bed problem. During this experiment kinetic energy is converted into gravitational potential energy and then back again as the wave is reflected, therefore uh and G will only be conserved in this experiment after the wave has completely reflected from the beach. Full reflection of the wave has not occurred by $t' = 70s$ and so the conservation results for uh and G were omitted from Tables 6.5 and 6.6.

The results for both FEVM_2 and FDVM_2 are identical in these Figures as these grids are quite fine and so these figures represent a good approximation to the true solution of the Serre equations. These numerical solutions demonstrate good agreement with experimental results and display the capability of the method to model the inundation of non-breaking waves.

In this chapter FEVM_2 and FDVM_2 were validated using experimental data. It was found that for most experiments the solutions of FEVM_2 and FDVM_2 were identical although FEVM_2 is the preferred method due to its greater robustness.

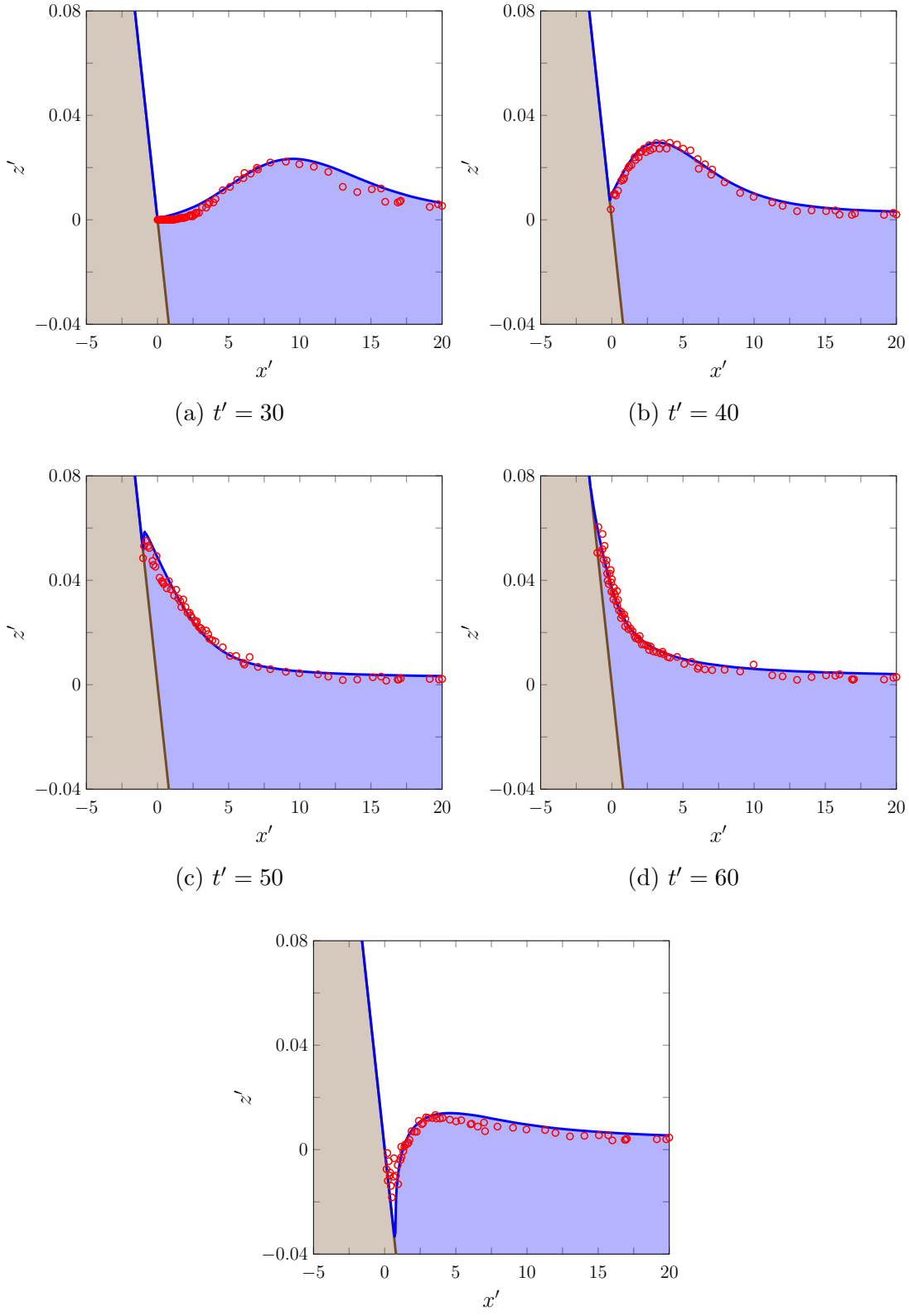


(a) $t' = 30$ (b) $t' = 40$

(c) $t' = 50$ (d) $t' = 60$

(e) $t' = 70$

Figure 6.23: A comparison of the water surface profiles for the experiment (○) and the numerical solution (–) produced by FEVM$_2$ at various times.



(a) $t' = 30$ (b) $t' = 40$

(c) $t' = 50$ (d) $t' = 60$

(e) $t' = 70$

Figure 6.24: A comparison of the water surface profiles for the experiment (○) and the numerical solution (–) produced by FDVM_2 at various times.

Quantity	$\mathcal{C}^*(\boldsymbol{q}^0)$	$\mathcal{C}^*(\boldsymbol{q}^*)$	$\mathcal{C}_1^*(\boldsymbol{q}^0, \boldsymbol{q}^*)$
h	140.4170	140.4170	7.65×10^{-12}
$\mathcal{H}$	68.3900	68.3914	2.16×10^{-5}

Table 6.5: Initial and final total amounts and the conservation error for all conserved quantities for FEVM_2's numerical solution of the runup experiment.

Quantity	$\mathcal{C}^*(\boldsymbol{q}^0)$	$\mathcal{C}^*(\boldsymbol{q}^*)$	$\mathcal{C}_1^*(\boldsymbol{q}^0, \boldsymbol{q}^*)$
h	140.4170	140.4170	1.11×10^{-7}
$\mathcal{H}$	68.3900	68.3914	2.16×10^{-5}

Table 6.6: Initial and final total amounts and the conservation error for all conserved quantities for FDVM_2's numerical solution of the runup experiment.

Chapter 7

Conclusion

The evolution of the dam-break problem for the Serre equations was comprehensively studied using various numerical methods resulting in the observation of new behaviours and the resolution of the differences previously reported in the literature.

A well balanced second-order FEVM was described for the one-dimensional Serre equations. This method makes use of a consistent polynomial representation of the quantities over the cells from which all necessary terms can be calculated locally over the cell; making it a readily parrallelisable computational method. The method uses a finite element and a finite volume method and thus is robust to steep gradients present in the conserved variables h and G.

A linear analysis of the convergence and dispersion properties of FEVM_2, all the FDM and all the FDVM [12] was performed. The analysis demonstrated that all FDVM, FEVM_2 and $\mathcal{D}$ are convergent methods, while $\mathcal{W}$ is only convergent when the mean background flow velocity is zero. The dispersion analysis demonstrated that all methods approximated the dispersion relation of the Serre equations with the expected order of accuracy. This analysis extended a previous analysis of the dispersion relationships of numerical methods [36] by allowing non-zero mean flow, combining the spatial and temporal analyses and comparing the real and imaginary parts of the dispersion error.

A comparison of the various numerical methods and the analytic solitary travelling wave solution of the Serre equations was performed. The expected order of accuracy and conservation properties of all the methods was observed. However, these results also demonstrated that the increase in accuracy achieved by a third-order method over a second-order method did not warrant the extra computational effort, justifying the further development of second-order methods

over third-order methods for future work. For this reason only the second-order FDVM and FEVM were developed further to allow varying bathymetry and dry beds.

The second-order FDVM and FEVM were then validated against the lake at rest steady state and the forced solutions. These results demonstrated that these methods are well balanced and accurately approximate all terms in the Serre equations in the presence of dry beds.

Finally the second-order FDVM and FEVM were compared to experimental data; demonstrating their modelling capabilities across a wide array of physical scenarios. These results established the greater robustness of the FEVM; as the FDVM was found to be unstable in the presence of large jumps in the water surface.

To summarise the major contributions of my research are

- Observation and justification of a new structure in the solution of the Serre equations to the dam-break problem;
- Development and description of the well balanced second-order finite element volume method that can handle dry beds and conserves h and G;
- Linear analysis of the convergence properties of the developed hybrid finite volume methods and the mentioned finite difference methods;
- Analysis of the dispersion properties of the numerical methods, allowing for non-zero mean flow velocity and accounting for the total dispersion error;
- Validation of the numerical method against analytic and forced solutions and experimental results.

7.1 Future Work

Following the work conducted in this thesis; some natural extensions are

- Inclusion of wave breaking in the model;
- Implementation of different boundary conditions;
- Incorporation of discontinuous bed profiles;
- Incorporation of bed friction;

- A complete analysis of the convergence properties of these methods;
- Extension of the FEVM to the two dimensional Serre equations on unstructured meshes.

Appendix A

Expressions for the Total Amount of Conserved Quantities for the Analytic Solutions

To calculate the conservation errors requires an analytic expression for the total amount of h, uh, G and $\mathcal{H}$ present in the initial conditions. Therefore to facilitate the validation tests against the analytic solutions described in Chapter 2 we present these analytic expressions for the initial conditions of solitary travelling wave and the lake at rest solutions. To allow for the simple calculation of the integrals in a concise way for any domains we present them in indefinite form.

A.1 Solitary Travelling Wave

For the solitary wave solution (2.13) with $t = 0$ the integrals of all the conserved quantities are

$$\int h(x,0)\,dx = a_0 x + \frac{a_1}{\kappa}\tanh(\kappa x) + \text{constant}, \tag{A.1a}$$

$$\int u(x,0)h(x,0)\,dx = \frac{a_1 c}{\kappa}\tanh(\kappa x) + \text{constant}, \tag{A.1b}$$

$$\begin{aligned}\int G(x,0)\,dx = \frac{ca_1}{3\kappa}\Bigg(&3 + 2a_0^2\kappa^2\text{sech}^2(\kappa x)\\ &+ 2a_0 a_1\kappa^2\text{sech}^4(\kappa x)\Bigg)\tanh(\kappa x) + \text{constant},\end{aligned} \tag{A.1c}$$

$$\int \mathcal{H}(x,0)\,dx = \frac{1}{2}\Bigg(\int g\left[h(x,0)\right]^2\,dx + \int h(x,0)\left[u(x,0)\right]^2\,dx + \int \left[h(x,0)\right]^3 \left[\frac{\partial u(x,0)}{\partial x}\right]^2\,dx \Bigg) \tag{A.1d}$$

where these integrals making up $\mathcal{H}$ are

$$\begin{aligned}\int g\left[h(x,0)\right]^2\,dx = {} & \frac{g}{12\kappa}\mathrm{sech}^3\left(\kappa x\right)\Big(9a_0^2\kappa x\cosh\left(\kappa x\right) + 3a_0^2\kappa x\cosh\left(3\kappa x\right) \\ & + 4a_1\left[3a_0 + 2a_1 + \left(3a_0 + a_1\right)\cosh\left(2\kappa x\right)\right]\sinh\left(\kappa x\right)\Big) \\ & + \text{constant},\end{aligned}$$

$$\begin{aligned}\int h(x,0)\left[u(x,0)\right]^2\,dx = {} & \frac{\sqrt{a_1}c^2}{\kappa}\Bigg(-\frac{a0}{\sqrt{a_0+a_1}}\mathrm{arctanh}\left(\frac{\sqrt{a1}\tanh\left(\kappa x\right)}{\sqrt{a_0+a1}}\right) \\ & + \frac{\sqrt{a_1}}{\kappa}\tanh\left(\kappa x\right)\Bigg) + \text{constant},\end{aligned}$$

$$\begin{aligned}\int \left[h(x,0)\right]^3\left[\frac{\partial u(x,0)}{\partial x}\right]^2\,dx = {} & \frac{2a_0^2c^2\kappa}{9\sqrt{a_1}\left(a_0 + a_1\mathrm{sech}^2\left(\kappa x\right)\right)} \\ & \times \left(a_0 + 2a_1 + a_0\cosh\left(2\kappa x\right)\right)\mathrm{sech}^2\left(\kappa x\right) \\ & \times \Bigg(-3a_0\sqrt{a_0+a_1}\mathrm{arctanh}\left(\frac{\sqrt{a1}\tanh\left(\kappa x\right)}{\sqrt{a_0+a1}}\right) \\ & + \sqrt{a_1}\left[3a_0 + a_1 - a_1\mathrm{sech}^2\left(\kappa x\right)\right]\tanh(\kappa x)\Bigg) + \text{constant}.\end{aligned}$$

Therefore, we have the analytic values of the total amounts of our conserved quantities for the solitary travelling wave solution (2.13) when $t = 0s$, as desired.

A.2 Lake At Rest

For the lake at rest solution (2.14) the total momentum and G in the system is straightforward to calculate as both are zero everywhere and so we have

$$\int u(x,0)h(x,0)\,dx = 0 + \text{constant}, \tag{A.2a}$$

$$\int G(x,0)\,dx = 0 + \text{constant}. \tag{A.2b}$$

To calculate the total mass and energy in the solution we must break up our domain into wet regions where $b(x) < a_0$ and dry regions where $b(x) \geq a_0$. For the dry regions the total amount of h and $\mathcal{H}$ are 0 and so we have

$$\int h(x,0)\,dx = 0, \tag{A.3a}$$

$$\int \mathcal{H}(x,0)\,dx = 0 \tag{A.3b}$$

whilst in a wet region we have

$$\int h(x,0)\,dx = a_0 x - \int b(x)\,dx, \tag{A.4a}$$

$$\int \mathcal{H}(x,0)\,dx = \frac{g}{2}\left(a_0^2 x - 2a_0 \int b(x)\,dx + \int b(x)^2\,dx\right). \tag{A.4b}$$

By summing all the wet regions in a given domain together we can calculate the total amount of h and $\mathcal{H}$ in the system from these expressions, in terms of the bed profile $b(x)$, as desired.

Appendix B

Basis Function and Function Space Definitions

For completeness we now provide the definitions of the basis functions of the FEM used by the FEVM described in Chapter 3 and the function spaces mentioned in Chapter 3. Beginning with the basis function definitions.

B.1 Basis Functions

Since all integrals of the basis functions are calculated with respect to the variable ξ, we give these basis functions in terms of ξ.

The basis functions ψ for h and G demonstrated in Figure 3.3 are

$$\psi^{+}_{j-1/2} = \begin{cases} \frac{1}{2}(1-\xi) & -1 \leq \xi \leq 1 \\ 0 & \text{otherwise,} \end{cases} \tag{B.1a}$$

$$\psi^{-}_{j+1/2} = \begin{cases} \frac{1}{2}(1+\xi) & -1 \leq \xi \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{B.1b}$$

While the basis functions ϕ for u and the test function v displayed in Figure 3.4 are given by

$$\phi_{j-1/2} = \begin{cases} 2\left(\xi+\frac{3}{2}\right)(\xi+2) & -2 \leq \xi \leq -1 \\ \frac{1}{2}\xi(\xi-1) & -1 \leq \xi \leq 1 \\ 0 & \text{otherwise,} \end{cases} \tag{B.2a}$$

$$\phi_{j} = \begin{cases} -(\xi-1)(\xi+1) & -1 \leq \xi \leq 1 \\ 0 & \text{otherwise,} \end{cases} \tag{B.2b}$$

$$\phi_{j+1/2} = \begin{cases} \frac{1}{2}\xi(\xi+1) & -1 \le \xi \le 1 \\ 2\left(\xi-2\right)\left(\xi-\frac{3}{2}\right) & 1 \le \xi \le 2 \\ 0 & \text{otherwise.} \end{cases} \tag{B.2c}$$

(B.2d)

Finally the basis functions γ for the bed profile b displayed in Figure 3.5 are given by

$$\gamma_{j-1/2} = \begin{cases} \frac{9}{2}\left(\xi+\frac{4}{3}\right)\left(\xi+\frac{5}{3}\right)(\xi+2) & -2 \le \xi \le -1 \\ \frac{9}{16}\left(\xi-1\right)\left(\xi-\frac{1}{3}\right)\left(\xi+\frac{1}{3}\right) & -1 \le \xi \le 1 \\ 0 & \text{otherwise,} \end{cases} \tag{B.3a}$$

$$\gamma_{j-1/6} = \begin{cases} \frac{27}{16}\left(\xi-1\right)\left(\xi-\frac{1}{3}\right)(\xi+1) & -1 \le \xi \le 1 \\ 0 & \text{otherwise,} \end{cases} \tag{B.3b}$$

$$\gamma_{j+1/6} = \begin{cases} -\frac{27}{16}\left(\xi-1\right)\left(\xi+\frac{1}{3}\right)(\xi+1) & -1 \le \xi \le 1 \\ 0 & \text{otherwise,} \end{cases} \tag{B.3c}$$

$$\gamma_{j-1/2} = \begin{cases} \frac{9}{16}\left(\xi+1\right)\left(\xi-\frac{1}{3}\right)\left(\xi+\frac{1}{3}\right) & -1 \le \xi \le 1 \\ -\frac{9}{2}\left(\xi-\frac{4}{3}\right)\left(\xi-\frac{5}{3}\right)(\xi-2) & 1 \le \xi \le 2 \\ 0 & \text{otherwise.} \end{cases} \tag{B.3d}$$

The calculation of the derivatives of these basis functions with respect to ξ are straightforward and hence omitted.

B.2 Function Spaces

The function spaces mentioned in Chapter 3 are $\mathbb{L}^2(\Omega)$ and $\mathbb{W}^{k,2}(\Omega)$. To be precise we now define these spaces here.

A function $f(x)$ is in $\mathbb{L}^2(\Omega)$ if

$$\left(\int_\Omega f(x)^2\, dx\right)^{\frac{1}{2}} < \infty.$$

While $f(x)$ is in $\mathbb{W}^{k,2}(\Omega)$ if

$$\left(\int_\Omega f(x)^2\, dx + \sum_{j=1}^{k}\int_\Omega \left[D^j f(x)\right]^2\, dx\right)^{\frac{1}{2}} < \infty.$$

where $D^j f(x)$ is the j^{th} weak derivative of $f(x)$.

Appendix C

Linear Analysis Results

In this appendix we present all the components to calculate the evolution matrix $\mathbf{E}$ for FDVM_1, FDVM_2, FDVM_3, $\mathcal{D}$ and $\mathcal{W}$. For the hybrid FDVM given the results in Chapter 4 for the FEVM_2 it is enough to provide only expressions for some of the operators. While for the FD methods we just present the evolution matrix.

C.1 Evolution Matrices for the Finite Difference Volume Methods

For the FDVM the evolution matrix can be constructed by taking the formulas for the elements of $\mathbf{F}$ from the FEVM and replacing the operators $\mathcal{R}^{+}_{j-1/2}$, $\mathcal{R}_j$, $\mathcal{R}^{-}_{j+1/2}$, $\mathcal{G}^{\eta}$ and $\mathcal{G}^{G}$ with the appropriate ones for the FDVM given in Tables C.1, C.2, C.3, C.4 and C.5. From $\mathbf{F}$ the evolution matrix $\mathbf{E}$ is then obtained by using the formulas given by the Runge-Kutta time stepping in Table C.6. Since $\mathcal{G}^{c}$ makes no contribution to the evolution matrix, it is omitted.

Method	Expression	Lowest Order Term of Error
FDVM_1	1	$-\frac{1}{24}k^2\Delta x^2$
FDVM_2 and FEVM_2	1	$-\frac{1}{24}k^2\Delta x^2$
FDVM_3	$\frac{26-2\cos(k\Delta x)}{24}$	$-\frac{3}{640}k^4\Delta x^4$

Table C.1: Factor $\mathcal{R}_j$ from reconstructing the nodal value at the midpoint for each method. Where the analytic value is $\mathcal{R}_j = \frac{k\Delta x}{2\sin\left(k\frac{\Delta x}{2}\right)}$.

Method	Formula	Lowest Order Term of Error
FDVM_1	1	$\frac{i}{2}k\Delta x$
FDVM_2 and FEVM_2	$\left(1-\frac{i\sin(k\Delta x)}{2}\right)$	$\frac{1}{12}k^2\Delta x^2$
FDVM_3	$\frac{1}{6}\left(5+2e^{-ik\Delta x}-e^{ik\Delta x}\right)$	$\frac{i}{12}k^3\Delta x^3$

Table C.2: Factor $\mathcal{R}^+_{j-1/2}$ from reconstruction of η and G at $x^+_{j+1/2}$ for each method. Where the analytic value is $\mathcal{R}^+_{j-1/2} = e^{-ik\Delta x/2}\frac{k\Delta x}{2\sin\left(\frac{k\Delta x}{2}\right)}$.

Method	Expression	Lowest Order Term of Error
FDVM_1	1	$-\frac{i}{2}k\Delta x$
FDVM_2 and FEVM_2	$1+\frac{i\sin(k\Delta x)}{2}$	$\frac{1}{12}k^2\Delta x^2$
FDVM_3	$\frac{1}{6}\left(5-e^{-ik\Delta x}+2e^{ik\Delta x}\right)$	$-\frac{i}{12}k^3\Delta x^3$

Table C.3: Factor $\mathcal{R}^-_{j+1/2}$ from reconstruction of η and G at $x^-_{j+1/2}$ for each method. Where the analytic value is $\mathcal{R}^-_{j+1/2}=e^{ik\Delta x/2}\dfrac{k\Delta x}{2\sin\left(\frac{k\Delta x}{2}\right)}$.

Method	Expression	Lowest Order Term of Error
FDVM_1	$\dfrac{-3U\Delta x^2\left(\dfrac{1+e^{ik\Delta x}}{2}\right)}{3\Delta x^2 H - H^3\left(2\cos\left(k\Delta x\right)-2\right)}$	$\dfrac{6+H^2k^2}{4H\left(3+H^2k^2\right)^2}Uk^2\Delta x^2$
FDVM_2	$\dfrac{-3U\Delta x^2\left(\dfrac{1+e^{ik\Delta x}}{2}\right)}{3\Delta x^2 H - H^3\left(2\cos\left(k\Delta x\right)-2\right)}$	$\dfrac{6+H^2k^2}{4H\left(3+H^2k^2\right)^2}Uk^2\Delta x^2$
FEVM_2	$\dfrac{-U\Delta x}{6}\left(1+\dfrac{i\sin\left(k\Delta x\right)}{2}+e^{ik\Delta x}\left(1-\dfrac{i\sin\left(k\Delta x\right)}{2}\right)\right)$ $\div\left(H\dfrac{\Delta x}{30}\left(4\cos\left(\dfrac{k\Delta x}{2}\right)-2\cos\left(k\Delta x\right)+8\right)\right.$ $\left.+\dfrac{H^3}{9\Delta x}\left(-16\cos\left(\frac{k\Delta x}{2}\right)+2\cos\left(k\Delta x\right)+14\right)\right)$	$-\dfrac{12+5H^2k^2}{40H\left(3+H^2k^2\right)^2}Uk^2\Delta x^2$
FDVM_3	$\dfrac{-36U\Delta x^2\left(\dfrac{-e^{-ik\Delta x}+9e^{ik\Delta x}-e^{2ik\Delta x}+9}{16}\right)}{36\Delta x^2 H - H^3\left(32\cos\left(k\Delta x\right)-2\cos\left(2k\Delta x\right)-30\right)}$	$\dfrac{243+49H^2k^2}{960H\left(3+H^2k^2\right)^2}Uk^4\Delta x^4$

Table C.4: Factor $\mathcal{G}^\eta$ from solving the elliptic equation (4.3c) for $v_{j+1/2}$ for each method. Where the analytic value is $\mathcal{G}^\eta = \dfrac{-3U}{3H+H^3k^2}\dfrac{1}{e^{-ik\Delta x/2}}\dfrac{k\Delta x}{2\sin\left(\frac{k\Delta x}{2}\right)}$.

Method	Expression	Lowest Order Term of Error
FDVM_1	$\dfrac{3\Delta x^2\left(\dfrac{1+e^{ik\Delta x}}{2}\right)}{3\Delta x^2 H - H^3\left(2\cos\left(k\Delta x\right)-2\right)}$	$-\dfrac{6+H^2k^2}{4H\left(3+H^2k^2\right)^2}k^2\Delta x^2$
FDVM_2	$\dfrac{3\Delta x^2\left(\dfrac{1+e^{ik\Delta x}}{2}\right)}{3\Delta x^2 H - H^3\left(2\cos\left(k\Delta x\right)-2\right)}$	$-\dfrac{6+H^2k^2}{4H\left(3+H^2k^2\right)^2}k^2\Delta x^2$
FEVM_2	$\dfrac{\Delta x}{6}\left(1+\dfrac{i\sin\left(k\Delta x\right)}{2}+e^{ik\Delta x}\left(1-\dfrac{i\sin\left(k\Delta x\right)}{2}\right)\right)$ $\div\left(H\dfrac{\Delta x}{30}\left(4\cos\left(\dfrac{k\Delta x}{2}\right)-2\cos\left(k\Delta x\right)+8\right)\right.$ $\left.+\dfrac{H^3}{9\Delta x}\left(-16\cos\left(\frac{k\Delta x}{2}\right)+2\cos\left(k\Delta x\right)+14\right)\right)$	$\dfrac{12+5H^2k^2}{40H\left(3+H^2k^2\right)^2}k^2\Delta x^2$
FDVM_3	$\dfrac{36\Delta x^2\left(\dfrac{-e^{-ik\Delta x}+9e^{ik\Delta x}-e^{2ik\Delta x}+9}{16}\right)}{36\Delta x^2 H - H^3\left(32\cos\left(k\Delta x\right)-2\cos\left(2k\Delta x\right)-30\right)}$	$-\dfrac{243+49H^2k^2}{960H\left(3+H^2k^2\right)^2}k^4\Delta x^4$

Table C.5: Factor $\mathcal{G}^G$ from solving the elliptic equation (4.3c) for $v_{j+1/2}$ for each method. Where the analytic value is $\mathcal{G}^G = \dfrac{3}{3H+H^3k^2}\dfrac{1}{e^{-ik\Delta x/2}}\dfrac{k\Delta x}{2\sin\left(\frac{k\Delta x}{2}\right)}$.

Method	Formula for $\mathbf{E}$
FDVM_1	$\mathbf{I} - \Delta t\mathbf{F}$
FDVM_2	$\mathbf{I} - \Delta t\mathbf{F} + \frac{1}{2}\Delta t^2\mathbf{F}^2$
FDVM_3	$\mathbf{I} - \Delta t\mathbf{F} + \frac{1}{2}\Delta t^2\mathbf{F}^2 - \frac{1}{6}\Delta t^3\mathbf{F}^3$

Table C.6: Formula for $\mathbf{E}$ given $\mathbf{F}$ determined by the SSP Runge-Kutta timestepping method.

C.2 Evolution Matrices for the Finite Difference Methods

By using (4.6) all the derivative approximations in the finite difference methods $\mathcal{D}$ and $\mathcal{W}$ can be written as operators that are constant in j and n as was done for the hybrid methods. Doing this we get that the evolution matrix for $\mathcal{D}$ is

$$\mathbf{E} = \begin{bmatrix} E_{0,0} & E_{0,1} & 1 & 0 \\ E_{1,0} & E_{1,1} & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{C.1}$$

with

$$\begin{aligned} E_{0,0} &= -\frac{2i\Delta t}{\Delta x} U \sin(k\Delta x), \\ E_{0,1} &= -\frac{2i\Delta t}{\Delta x} H \sin(k\Delta x), \\ E_{1,0} &= -\frac{6gi\Delta x\Delta t}{3\Delta x^2 - 2H^2(\cos(k\Delta x) - 1)} \sin(k\Delta x), \\ E_{1,1} &= -\frac{2i\Delta t}{\Delta x} U \sin(k\Delta x). \end{aligned}$$

While for $\mathcal{W}$ the evolution matrix is

$$\mathbf{E} = \begin{bmatrix} E_{0,0} & E_{0,1} & 0 & E_{0,3} \\ E_{1,0} & E_{1,1} & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{C.2}$$

with

$$E_{0,0} = 1 - \frac{\Delta t}{\Delta x}\left(-\frac{6gi\Delta x\Delta t}{3\Delta x^2 - 2H^2\left(\cos\left(k\Delta x\right)-1\right)}\sin\left(k\Delta x\right)\right)H\frac{i\sin\left(k\Delta x\right)}{2}$$
$$-\frac{\Delta t}{\Delta x}U\left(i\sin\left(k\Delta x\right) - \frac{\Delta t}{\Delta x}U\left(\cos\left(k\Delta x\right)-1\right)\right),$$
$$E_{0,1} = -\frac{\Delta t}{\Delta x}\left(H\frac{i\sin\left(k\Delta x\right)}{2}\left[1 - \frac{2i\Delta t}{\Delta x}U\sin\left(k\Delta x\right)\right] - U\left[\frac{\Delta t}{\Delta x}H\left(\cos\left(k\Delta x\right)-1\right)\right]\right),$$
$$E_{0,3} = -\frac{\Delta t}{\Delta x}H\frac{i\sin\left(k\Delta x\right)}{2},$$
$$E_{1,0} = -\frac{6gi\Delta x\Delta t}{3\Delta x^2 - 2H^2\left(\cos\left(k\Delta x\right)-1\right)}\sin\left(k\Delta x\right),$$
$$E_{1,1} = -\frac{2i\Delta t}{\Delta x}U\sin\left(k\Delta x\right).$$

Bibliography

[1] L. Euler. Principes generaux du mouvement des fluides. *Mémoires de l'académie des sciences de Berlin*, 11:274–315, 1757.

[2] A.J. Chorin. The numerical solution of the Navier-Stokes equations for an incompressible fluid. *Bulletin of the American Mathematical Society*, 73(6): 928–931, 1967.

[3] C. Taylor and P. Hood. numerical solution of the Navier-Stokes equations using the finite element technique. *Computers & Fluids*, 1:73–100, 1973.

[4] F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. *Journal of Computational Physics*, 131(2):267 – 279, 1997.

[5] D. Lannes and P. Bonneton. Derivation of asymptotic two-dimensional time-dependent equations for surface water wave propagation. *Physics of Fluids*, 21(1):16601–16610, 2009.

[6] The Clawpack Development Team. Clawpack documentation, 2018. URL `http://www.clawpack.org/`.

[7] Xiaoming Wang. Comcot, 2009. URL `http://223.4.213.26/archive/tsunami/cornell/comcot.htm`.

[8] Stephen Roberts. Anuga, 2018. URL `https://anuga.anu.edu.au/`.

[9] J. Grue, E.N. Pelinovsky, D. Fructus, T. Talipova, and C. Kharif. Formation of undular bores and solitary waves in the strait of Malacca caused by the 26 December 2004 Indian ocean tsunami. *Journal of Geophysical Research: Oceans*, 113, 2008.

[10] J.T. Kirby, F. Shi, B. Tehranirad, J.C. Harris, and S.T. Grilli. Dispersive tsunami waves in the ocean: Model equations and sensitivity to dispersion and coriolis effects. *Ocean Modelling*, 62:39–55, 2013.

[11] C. Zoppou. *Numerical Solution of the One-dimensional and Cylindrical Serre Equations for Rapidly Varying Free Surface Flows*. PhD thesis, Australian National University, Mathematical Sciences Institute, College of Physical and Mathematical Sciences, Australian National University, Canberra, ACT 2600, Australia, 2014.

[12] J.P.A. Pitt, C. Zoppou, and S.G. Roberts. Behaviour of the Serre equations in the presence of steep gradients revisited. *Wave Motion*, 76(1):61–77, 2018.

[13] C. Zoppou, J. Pitt, and S. Roberts. Numerical solution of the fully non-linear weakly dispersive Serre equations for steep gradient flows. *Applied Mathematical Modelling*, 48:70–95, 2017.

[14] R.M. Sorensen. *Basic Coastal Engineering*. Springer, 3 edition, 2006.

[15] F. Serre. Contribtion à l'étude des écoulements permanents et variables dans les canaux. *La Houille Blanche*, 6:830–872, 1953.

[16] C. H. Su and C. S. Gardner. Korteweg-de Vries equation and generalisations. III. Derivation of the Korteweg-de Vries equation and Burgers equation. *Journal of Mathematical Physics*, 10(3):536–539, 1969.

[17] 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.

[18] F. J. Seabra-Santos, D. P. Renouard, and A. 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, 1981.

[19] E. Barthélemy. Nonlinear shallow water theories for coastal waves. *Surveys in Geophysics*, 25(3):315–337, 2004.

[20] P. Bonneton, F. Chazel, D. Lannes, F. Marche, and M. Tissier. A splitting approach for the fully nonlinear and weakly dispersive Green-Naghdi model. *Journal of Computational Physics*, 230(4):1479–1498, 2011.

[21] J.A Liggett. *Fluid Mechanics*. McGraw-Hill Inc., 1994.

[22] O. Le Métayer, S. Gavrilyuk, and S. Hank. A numerical scheme for the Green-Naghdi model. *Journal of Computational Physics*, 229(6):2034–2045, 2010.

[23] M. Li, P. Guyenne, F. Li, and L. Xu. High order well-balanced CDG-FE methods for shallow water waves by a Green-Naghdi model. *Journal of Computational Physics*, 257(1):169–192, 2014.

[24] W. Choi and R. Camassa. Fully nonlinear internal waves in a two-fluid system. *Journal of Fluid Mechanics*, 396:1–36, 1999.

[25] J.D Carter and R. Cienfuegos. The kinematics and stability of solitary and cnoidal wave solutions of the Serre equations. *European Journal of Mechanics-B/Fluids*, 30(3):259–268, 2011.

[26] J. Pitt, C. Zoppou, and S.G Roberts. Importance of dispersion for shoaling waves. *Modelling and Simulation Society of Australia and New Zealand*, 22 (1):1725–1730, 2017.

[27] Y. A. Li. Hamiltonian structure and linear stability of solitary waves of the Green-Naghdi equations. *Journal of Nonlinear Mathematical Physics*, 9: 99–105, 2002.

[28] 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.

[29] D. Dutykh, D. Clamond, P. Milewski, and D. Mitsotakis. Finite volume and pseudo-spectral schemes for the fully nonlinear 1D Serre equations. *European Journal of Applied Mathematics*, 24(5):761–787, 2013.

[30] D. Mitsotakis, B. Ilan, and D. Dutykh. On the Galerkin/finite-element method for the Serre equations. *Journal of Scientific Computing*, 61(1): 166–195, 2014.

[31] D. Mitsotakis, D. Dutykh, and D. Carter. On the nonlinear dynamics of the traveling-wave solutions of the Serre system. *Wave Motion*, 70(1):166–182, 2017.

[32] J.S.A do Carmo, J.A Ferreira, L. Pinto, and G. Romanazzi. An improved Serre model: Efficient simulation and comparative evaluation. *Applied Mathematical Modelling*, 56:404–423, 2018.

[33] V.A Dougalis, A. Duran, M.A. Lopez-Marcos, and D.E. Mitsotakis. Numerical study of the stability of solitary waves of the Bona–Smith family of Boussinesq systems. *Journal of Nonlinear Science*, 17(6):569–607, 2007.

[34] R. Cienfuegos, E. Barthélemy, and P. Bonneton. A fourth-order compact finite volume scheme for fully nonlinear and weakly dispersive Boussinesq-type equations. part I: model development and analysis. *International Journal for Numerical Methods in Fluids*, 51(11):1217–1253, 2006.

[35] S. F. Bradford and B. F. Sanders. Finite volume schemes for the Boussinesq equations. In *Ocean Wave Measurement and Analysis (2001)*, pages 953–962. American Society of Civil Engineers, 2002.

[36] A. G. Filippini, M. Kazolea, and M. Ricchiuto. A flexible genuinely nonlinear approach for nonlinear wave propagation, breaking and run-up. *Journal of Computational Physics*, 310:381–417, 2016.

[37] J. Pitt. A second order well balanced hybrid finite volume and finite difference method for the Serre equations. Honour's thesis, Australian National University, Canberra, Australia, 2014.

[38] P.L Roe. Characteristic-based schemes for the Euler equations. *Annual Review of Fluid Mechanics*, 18(1):337–365, 1986.

[39] B. Van Leer. Towards the ultimate conservative difference scheme. IV. a new approach to numerical convection. *Journal of Computational Physics*, 23(3): 276–299, 1977.

[40] A. Harten. High resolution schemes for hyperbolic conservation laws. *Journal of Computational Physics*, 49(3):357–3935, 1983.

[41] P.J Davis and P. Rabinowitz. *Methods of Numerical Integration*, volume 2. Blaisdell Publishing Company, 1984.

[42] W.H Press, S.A Teukolsky, W.T Vetterling, and B.P Flannery. *Numerical Recipes in C*, volume 2. Cambridge University Press, 1996.

[43] A. Kurganov, S. Noelle, and G. Petrova. Semidiscrete central-upwind schemes for hyperbolic conservation laws and Hamilton-Jacobi equations. *Journal of Scientific Computing, Society for Industrial and Applied Mathematics*, 23(3):707–740, 2002.

[44] E. Audusse, F. Bouchut, M. Bristeau, R. Klein, and B. Perthame. A fast and stable well–balanced scheme with hydrostatic reconstruction for shallow water flows. *Journal of Scientific Computing, Society for Industrial and Applied Mathematics*, 25(6):2050–2065, 2004.

[45] S. Gottlieb, C. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *Review Society for Industrial and Applied Mathematics*, 43(1):89–112, 2001.

[46] R. Courant, K. Friedrichs, and H. Lewy. On the partial difference equations of mathematical physics. *IBM Journal of Research and Development*, 11(2): 215–234, 1967.

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

[48] S. D. Conte and C. De Boor. *Elementary numerical analysis: an algorithmic approach*, volume 3. McGraw-Hill Inc., 1980.

[49] P. Lax and R. Richtmyer. Survey of the stability of linear finite difference equations. *Communications on Pure and Applied Mathematics*, 9(2):267–293, 1956.

[50] A.T. Ippen and G. Kulin. The shoaling and breaking of the solitary wave. *Coastal Engineering Proceedings*, 1(5):4, 1954.

[51] J. L. Hammack and H. Segur. The Korteweg-de Vries equation and water waves. part 3. oscillatory waves. *Journal of Fluid Mechanics*, 84(2):337–358, 1978.

[52] S Beji and J.A. Battjes. Experimental investigation of wave propagation over a bar. *Coastal Engineering*, 19(1):151–162, 1993.

[53] S Beji and J.A. Battjes. Numerical simulation of nonlinear wave propagation over a bar. *Coastal Engineering*, 23(1):1–16, 1994.

[54] D. Lannes. *The Water Waves Problem: Mathematical Analysis and Asymptotics*, volume 1 of *American Mathematical Society*. Mathematical Surveys and Monographs, 2013.

[55] Y. Zhang, A.B. Kennedy, N. Panda, C. Dawson, and J.J. Westerink. Boussinesq-Green-Naghdi rotational water wave theory. *Coastal Engineering*, 73(1):13–27, 2013.

[56] V. Roeber. *Boussinesq-type mode for nearshore wave processes in fringing reef environment*. PhD thesis, University of Hawaii, Manoa, Honolulu, HI, U.S.A, 2010.

[57] C. E. Synolakis. The runup of solitary waves. *Journal of Fluid Mechanics*, 185:523–545, 1987.

[58] A. Bollermann, S. Noelle, and M. Lukáčová-Medvidová. Finite volume evolution Galerkin methods for the shallow water equations with dry beds. *Communications in Computational Physics*, 10(2):371–404, 2011.