

Refinement of the Hybrid Finite Volume Methods for the Serre equations for Rapidly Varying Flows and Dry Beds.

Jordan Pitt

October 2018

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



Contents

1	Introduction	1
1.1	Objectives of the Thesis	1
1.2	Original Contribution of the Thesis	1
1.3	Organisation of the Thesis	1
2	The Serre Equations	3
2.1	The Equations	3
2.1.1	Alternative form of the Serre Equations	5
2.2	Properties of the Serre Equations	6
2.2.1	Conservation Properties	6
2.2.2	Dispersion Properties	7
2.2.3	Analytic Solutions	8
2.2.4	Forced Solutions	10
2.2.5	Behaviour of Steep Gradients	11
3	Finite Element Volume Method	13
3.1	Notation for Numerical Grids	13
3.2	Structure Overview	14
(i)	Reconstruction	16
(ii)	Fluid Velocity	19
(iii)	Flux Across the Cell Interfaces	24
(iv)	Source Terms	27
(v)	Update Cell Averages	28
(vi)	Second-Order SSP Runge-Kutta Method	29
3.3	CFL condition	29
3.4	Boundary Conditions	29
3.5	Dry Beds	31

4 Linear Analysis of the Numerical Methods	33
4.1 Linearised Serre Equations with a Horizontal Bed	34
4.2 Evolution Matrix	35
(i) Reconstruction	36
(ii) Fluid Velocity	36
(iii) Flux Across the Cell Interfaces	39
(v) Update Cell Averages	42
(vi) Second-Order SSP Runge-Kutta Method	43
4.3 Convergence Analysis	43
4.3.1 Stability	44
4.3.2 Consistency	45
4.4 Dispersion Analysis	51
5 Numerical Validation	61
5.1 Measuring Convergence and Conservation	61
5.1.1 Measure of Convergence	61
5.1.2 Measures of Conservation	62
5.2 Analytic Solution for Horizontal Bed	62
5.2.1 Results for Solitary Travelling Wave Solution	63
5.3 Analytic Solution for Variable Bathymetry	65
5.3.1 Results for Lake at Rest	67
5.4 Forced Solutions	68
5.4.1 Results for Finite Water Depth	71
5.4.2 Results with Dry Beds	71
6 Experimental Validation	79
6.1 Evolution of Rectangular Depression	79
6.1.1 Results for $0.01m$ Rectangular Depression	80
6.1.2 Results for $0.03m$ Rectangular Depression	81
6.2 Periodic Waves Over A Submerged Bar	87
6.2.1 Low Frequency Results	88
6.2.2 High Frequency Results	89
6.3 Solitary Wave Over a Fringing Reef	89
6.3.1 Results	98
6.4 Runup of a Solitary Wave on a Linearly Sloped Beach	99
6.4.1 Results	104
7 Conclusion	109

A Finite Element Integrals	111
A.1 Basis Functions	111
B Linear Analysis Results	113
B.1 Evolution Matrices for the Finite Difference Volume Methods . . .	113
B.2 Evolution Matrices for the Finite Difference Methods	113
Bibliography	117

Chapter 1

Introduction

1.1 Objectives of the Thesis

1.2 Original Contribution of the Thesis

1.3 Organisation of the Thesis

Chapter 2

The Serre Equations

2.1 The Equations

There are three primary ways in which the Serre equations have been derived from the Euler equations in the literature; by asymptotic expansion [1, 2], directed fluid sheets [3] and depth integration [4, 5]. In this thesis the depth integration view of the equations is taken, although the derivation is omitted given the extent of literature already available.

From the depth-integration approach the Serre equations describe a free surface fluid defined by its height $h(x, t)$ above a stationary bed profile $b(x)$ with depth average of its horizontal velocity $u(x, t)$ as in Figure 2.1. The derivation is similar to that of the Shallow Water Wave Equations (SWWE) [], except for the Serre equations the vertical velocity $v(x, z, t)$ varies linearly with depth and is given by []

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

Because the vertical velocity of the Serre equations is not 0 throughout the depth of water as in the SWWE, the Serre equations possess a non-hydrostatic pressure distribution.

By depth integrating the Euler equations [] with a no-slip condition at the bed and a free surface condition at the free surface we obtain the depth integrated approximation of the conservation of mass and momentum equations

$$\frac{\partial h}{\partial t} + \frac{\partial(uh)}{\partial x} = 0, \quad (2.2a)$$

$$\frac{\partial(uh)}{\partial t} + \frac{\partial}{\partial x} \left(u^2 h + \frac{gh^2}{2} + \frac{h^2}{2} \Psi + \frac{h^3}{3} \Phi \right) + \frac{\partial b}{\partial x} \left(gh + h\Psi + \frac{h^2}{2} \Phi \right) = 0 \quad (2.2b)$$

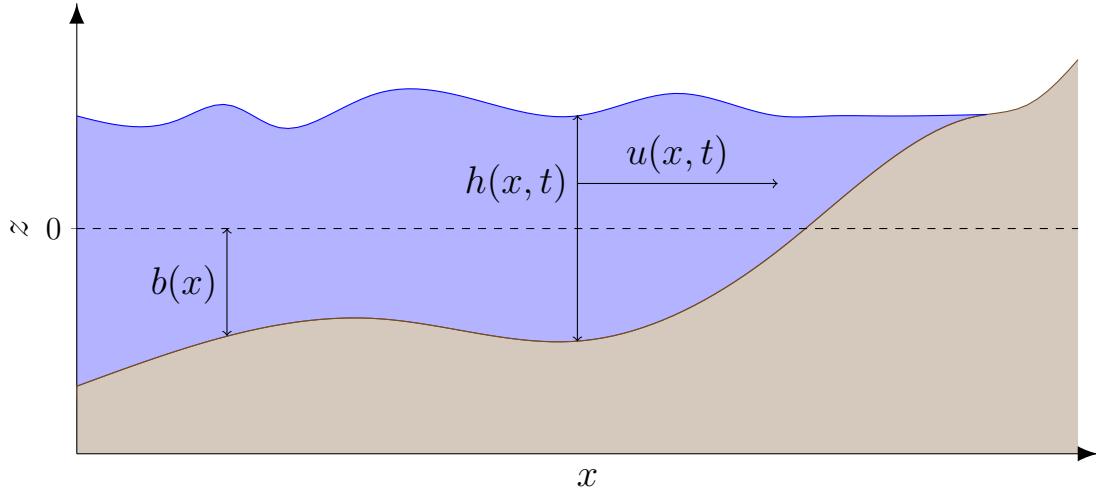


Figure 2.1: Diagram demonstrating the quantities used to describe the fluid (■) and the bed (■) for the Serre equations.

where the Φ and Ψ terms account for the non-hydrostatic part of the pressure and are

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

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

When $\Phi = \Psi = 0$ the Serre equations are equivalent to the SWWE where the vertical velocity is 0, only the hydrostatic pressure is present and there is no dispersion. Due to the presence of the Φ and Ψ terms the Serre equations are much more difficult to solve analytically and numerically than the SWWE. The primary reason for this is that whilst the SWWE are hyperbolic for finite water depth, 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

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

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

These equations are neither hyperbolic nor parabolic and are not in conservation law form as Φ contains a temporal derivative. As such even 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 mixed temporal and spatial derivative in Φ and Ψ (2.3). By rewriting the Serre equations and introducing a new conserved quantity G [6, 5, 7], the mixed temporal and spatial derivative can be removed and the Serre equations can be written in conservation law form.

Definition 2.1. The conserved quantity G is

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

The Serre equations (2.2) can then rewritten as conservation laws with a source term for the conserved variables h and G

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

$$\begin{aligned} \frac{\partial G}{\partial t} + \frac{\partial}{\partial x} \left(uG + \frac{gh^2}{2} - \frac{2}{3} h^3 \left[\frac{\partial u}{\partial x} \right]^2 + h^2 u \frac{\partial u}{\partial x} \frac{\partial b}{\partial x} \right) \\ + \frac{1}{2} h^2 u \frac{\partial u}{\partial x} \frac{\partial^2 b}{\partial x^2} - hu^2 \frac{\partial b}{\partial x} \frac{\partial^2 b}{\partial x^2} + gh \frac{\partial b}{\partial x} = 0. \end{aligned} \tag{2.5b}$$

The conserved quantity G resembles h multiplied by the irrotationality [8, 9].

This conservation law form makes the Serre equations well suited to be numerically solved using a finite volume method for the conservation of mass 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 equa-

tions is

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

$$\frac{\partial G}{\partial t} + \frac{\partial}{\partial x} \left(uG + \frac{gh^2}{2} - \frac{2}{3}h^3 \left[\frac{\partial u}{\partial x} \right]^2 \right) = 0, \quad (2.6b)$$

$$G = hu - \frac{\partial}{\partial x} \left(\frac{1}{3}h^3 \frac{\partial u}{\partial x} \right). \quad (2.6c)$$

2.2 Properties of the Serre Equations

The Serre equations possess a number of properties that can be used to assess the veracity of numerical methods. Because if a numerical method approximates the Serre equations accurately then the properties of the numerical method should approximate the properties of the Serre equations. In this thesis the conservation properties, dispersion properties, analytic solutions of the Serre are employed and so are presented here.

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

Finally the results of Pitt et al. [10] for the behaviour of the Serre equations in the presence of steep gradients are presented. These results satisfied one of the main objectives of the Thesis and contained behaviours that were not previously present in the literature.

2.2.1 Conservation Properties

Conservation of a quantity means that in a closed system the total amount of a quantity q remains constant in time.

Definition 2.2. The total amount of a quantity q in a system occurring on the interval $[a, b]$ at time t is

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

Conservation of a quantity q means that $\mathcal{C}_q(0) = \mathcal{C}_q(t)$ for all t . Given that the Serre equations (2.2) are conservation equations for mass and momentum and that the conservation of momentum equation can be rewritten as a conservation equation for G (2.5), the Serre equations conserve all these quantities. Additionally the Serre equations were derived by Green and Naghdi [3] by conserving the energy \mathcal{H}

$$\mathcal{H}(x, t) = \frac{1}{2} \left(g(h^2 + 2hb) + hu^2 + \frac{h^3}{3} \left(\frac{\partial u}{\partial x} \right)^2 + u^2 h \left[\frac{\partial b}{\partial x} \right]^2 - uh^2 \frac{\partial u}{\partial x} \frac{\partial b}{\partial x} \right). \quad (2.7)$$

The energy \mathcal{H} is the sum of the gravitational potential energy, the horizontal kinetic energy and the vertical kinetic energy which over the depth of water are

$$\int_b^{h+b} gz \, dx = g(h^2 + 2hb), \quad (2.8)$$

$$\int_b^{h+b} u^2 \, dx = hu^2, \quad (2.9)$$

$$\int_b^{h+b} v^2 \, dx = \frac{h^3}{3} \left(\frac{\partial u}{\partial x} \right)^2 + u^2 h \left[\frac{\partial b}{\partial x} \right]^2 - uh^2 \frac{\partial u}{\partial x} \frac{\partial b}{\partial x}, \quad (2.10)$$

respectively. For horizontal beds \mathcal{H} is also the Hamiltonian of the Serre equations [11].

For the system to be closed the flux terms of the conservation of mass and momentum equations at the boundaries must cancel and the integral of the source term over the domain must be zero.

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 wave number k of these solutions. For the Serre equations the dispersion relation [] is

$$\omega = U k \pm k \sqrt{gH} \sqrt{\frac{3}{(kH)^2 + 3}}. \quad (2.11)$$

Barthélemy [12] 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 water wave theory increases. The dispersion relation of the Serre

equations can be modified by introducing terms to reduce this difference [12], but such modifications are beyond the scope of this thesis.

From the dispersion relation (2.11) the phase velocity $v_p = \omega/k$ and the group velocity $v_g = \partial\omega/\partial k$ can be written in terms of wave number as

$$v_p = U \pm \sqrt{gH} \sqrt{\frac{3}{(kH)^2 + 3}}, \quad (2.12a)$$

$$v_g = U \pm \sqrt{gH} \left(\sqrt{\frac{3}{(kH)^2 + 3}} \mp (kH)^2 \sqrt{\frac{3}{((kH)^2 + 3)^3}} \right). \quad (2.12b)$$

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 $k \rightarrow 0$ then $v_p, v_g \rightarrow U \pm \sqrt{gH}$ and as $k \rightarrow \infty$ then $v_p, v_g \rightarrow U$. Therefore we have that

$$U - \sqrt{gH} \leq v_p \leq U + \sqrt{gH}, \quad (2.13a)$$

$$U - \sqrt{gH} \leq v_g \leq U + \sqrt{gH}. \quad (2.13b)$$

2.2.3 Analytic Solutions

Few analytic solutions have been discovered for the Serre equations. In particular there is a travelling wave solution for horizontal beds and a lake at rest solution for any 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 nonlinear and dispersive effects. Unlike the Euler equations this travelling wave solution has a closed form

$$h(x, t) = a_0 + a_1 \operatorname{sech}(\kappa(x - ct)), \quad (2.14a)$$

$$u(x, t) = c \left(1 - \frac{a_0}{h(x, t)} \right), \quad (2.14b)$$

$$b(x) = 0 \quad (2.14c)$$

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 straightforwardly derived, these are presented in Appendix [] 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 travelling periodic travelling wave solutions [13]. However, these solutions are not true solitons, due to their inelastic collisions with one another [14].

This analytic solution is a good test for the accuracy of numerical methods to solve the Serre equations with horizontal beds (2.6) for smooth solutions as all terms are smooth, vary in space and time and are non-zero inside the wave. Therefore, to accurately recreate this solitary wave the numerical method must have the appropriate accuracy for all terms in the equation. Additionally because this solution is a consequence of a balance between nonlinear and dispersive forces it can only be reproduced if the nonlinear and dispersive properties of the numerical scheme are properly balanced.

Lake at Rest

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

$$h(x, t) = \max \{a_0 - b(x), 0\}, \quad (2.15a)$$

$$u(x, t) = 0, \quad (2.15b)$$

$$G(x, t) = 0. \quad (2.15c)$$

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 the conserved quantities in the system are given in Appendix [] for reference.

For these quantities (2.15) the Serre equations (2.5) reduce to

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

$$\frac{\partial G}{\partial t} + \frac{\partial}{\partial x} \left(\frac{gh^2}{2} \right) + gh \frac{\partial b}{\partial x} = 0.$$

Since we have that $\partial h / \partial x = -\partial b / \partial x$ when $h \neq 0$, then G and h are constant in time and therefore so is u and thus we possess a stationary solution.

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 combat this modifications are made 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 of the numerical methods.

2.2.4 Forced Solutions

The analytic solutions of the Serre equations provide a stringent test when the bed is horizontal, as all terms in the equation are non-zero and vary in space and time inside the wave and therefore must be accurately approximated. However, for varying bathymetry there is only the lake at rest solution where all terms are constant in time and some are 0. 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.

To allow the verification of the accuracy of the numerical methods for solutions with varying bathymetry, which are not stationary solutions and have all terms non-zero in some region, forces us to use forced solutions. To do this we select some particular functions for all of our primitive quantities; h , u and b which we

denote h^* , u^* and b^* respectively. From these functions we calculate

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

$$\begin{aligned} S_G = & -\frac{\partial G^*}{\partial t} - \frac{\partial}{\partial x} \left(u^*G^* + \frac{g(h^*)^2}{2} - \frac{2}{3}(h^*)^3 \left[\frac{\partial u^*}{\partial x} \right]^2 + (h^*)^2 u^* \frac{\partial u^*}{\partial x} \frac{\partial b^*}{\partial x} \right) \\ & - \frac{1}{2}(h^*)^2 u^* \frac{\partial u^*}{\partial x} \frac{\partial^2 b^*}{\partial x^2} + (h^*)(u^*)^2 \frac{\partial b^*}{\partial x} \frac{\partial^2 b^*}{\partial x^2} - gh^* \frac{\partial b^*}{\partial x}. \end{aligned}$$

Now h^* , u^* and b^* will be solutions of the forced Serre equations in conservation law form with a source term

$$\frac{\partial h}{\partial t} + \frac{\partial(uh)}{\partial x} + S_{\text{mass}} = 0, \quad (2.16a)$$

$$\begin{aligned} \frac{\partial G}{\partial t} + \frac{\partial}{\partial x} \left(uG + \frac{gh^2}{2} - \frac{2}{3}h^3 \frac{\partial u^2}{\partial x} + h^2 u \frac{\partial u}{\partial x} \frac{\partial b}{\partial x} \right) \\ (2.16b) \end{aligned}$$

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

These forced Serre equations are then numerically solved by using operator splitting to split the Serre equations for which $S_{\text{mass}} = S_G = 0$ from the source term update. We then use the numerical methods to solve the Serre equation part and use a forward Euler step for the source term update, where S_{mass} and S_G are calculated analytically.

2.2.5 Behaviour of Steep Gradients - [Requires Modification]

To test the robustness of our numerical methods, we would like to perform validation tests on problems with steep gradients in the initial conditions. One particular group of these problems are the so called dam-break problems where water is initially still on a horizontal bed with a discontinuous jump in the water depth.

Unfortunately for the Serre equations the dam break problems have no known analytic solutions. Although some insight of the behaviour of steep gradients has been gained from asymptotic analysis, looking at the long term behaviour [] and from the analysis of the linearised Serre equations [].

Asymptotic Results

Beyond analytic solutions to the Serre equations there have also been studies of the long term behaviour of the Serre equations for situations that are difficult to treat analytically. One particular scenario of interest is the evolution of a moving discontinuous jump in h known as a bore, which can be observed naturally, for example tidal bores [1].

For the non-dispersive SWWE bores propagate at a fixed speed and have a constant shape [1]. For the Serre equations dispersion causes bores to break up into wave train, which are referred to as undular bores [1]. This process is more difficult to treat analytically particularly over short time spans and so we do not possess analytic solutions for the Serre equations for bores.

To gain some insight into the behaviour of bores for long time spans Whitham modulation techniques were applied to the Serre equations as $t \rightarrow \infty$ [13]. These techniques provided an estimate of the speed S^+ and amplitude A^+ of the front of a bore

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

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

where $\Delta = h_b/h_0$, h_b is the height of the bore and h_0 is the depth of still water in front of the bore. These estimates agreed well with numerical simulations provided that $\Delta < 1.43$ [13].

Results of the Paper

However, without analytic results there was previously some ambiguity in the short term behaviour of steep gradient problems with variety of observed behaviours for both actual discontinuous dam-break problems and their smoothed approximations; the so called smoothed dam-break problem.

To rectify this we conducted a comprehensive review of many numerical experiments for a variety of numerical methods and smoothings of the initial conditions, the results of which were published in [1].

Chapter 3

Finite Element Volume Method

3.1 Notation for Numerical Grids

For the purposes of the Finite Element Volume Method (FEVM) time and space will be discretised in different ways; time is broken up into time levels separated by constant durations and space is broken up into cells of constant width. 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

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

with $\Delta t \in \mathbb{R}$. Therefore, the duration between time levels is constant. The method can be extended to varying Δt and we restrict ourselves to constant Δt for simplicity. The goal of the FEVM is to update the quantities at the current time level t^n to the next time level t^{n+1} , allowing iteration through all time, 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

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

with $\Delta x \in \mathbb{R}$. This definition results in cells of constant width, these methods can be readily extended to cells of varying widths and we have restricted ourselves to the constant case for simplicity. Other points inside the j^{th} cell can be defined in relation to the midpoint so that

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

where $s \in [-\frac{1}{2}, \frac{1}{2}] \subset \mathbb{R}$, although for our purposes we restrict ourselves to rational s . Using this notation the j^{th} cell can be written $[x_{j-1/2}, x_{j+1/2}]$.

The 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 finite volume method the cell averages of quantities is also required. For each cell we define the average of a quantity at time level t^n

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

over the j^{th} cell.

In the FEVM we reconstruct quantities at various points inside the cell from the cell average values. Since neighbouring cells overlap at the cell edges $x_{j\pm 1/2}$, two reconstructions are possible there, 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. Since the particular time level is typically obvious from context and hence omitted the use of a superscript will not clutter the notation.

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

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

The evolution step proceeds as follows:

- (i) Reconstruction: The locations for the reconstruction of all the quantities in the j^{th} cell are displayed in Figure 3.1. For h , w and G they 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

$$\begin{aligned} h_{j-1/2}^+ &= \mathcal{R}_{j-1/2}^+ (\bar{\mathbf{h}}^n), & G_{j-1/2}^+ &= \mathcal{R}_{j-1/2}^+ (\bar{\mathbf{G}}^n), \\ h_j &= \mathcal{R}_j (\bar{\mathbf{h}}^n), & G_j &= \mathcal{R}_j (\bar{\mathbf{G}}^n), \\ h_{j+1/2}^- &= \mathcal{R}_{j+1/2}^- (\bar{\mathbf{h}}^n), & G_{j+1/2}^- &= \mathcal{R}_{j+1/2}^- (\bar{\mathbf{G}}^n), \end{aligned}$$

$$\begin{aligned} w_{j-1/2}^+ &= \mathcal{R}_{j-1/2}^+ (\bar{\mathbf{w}}^n), & b_{j-1/2} &= \mathcal{B}_{j-1/2} (\mathbf{b}), \\ w_j &= \mathcal{R}_j (\bar{\mathbf{w}}^n), & b_{j-1/6} &= \mathcal{B}_{j-1/6} (\mathbf{b}), \\ w_{j+1/2}^- &= \mathcal{R}_{j+1/2}^- (\bar{\mathbf{w}}^n), & b_{j+1/6} &= \mathcal{B}_{j+1/6} (\mathbf{b}), \\ && b_{j+1/2} &= \mathcal{B}_{j+1/2} (\mathbf{b}). \end{aligned}$$

These reconstruction operators are defined in subsection (i) below. To keep the notation simple the time superscript is omitted from the reconstructed quantities. This generates the vectors of these quantities reconstructed for every cell; $\hat{\mathbf{h}}$, $\hat{\mathbf{w}}$, $\hat{\mathbf{G}}$ and $\hat{\mathbf{b}}$ at time t^n which are

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

- (ii) Fluid Velocity: The remaining unknown quantity, the depth averaged fluid velocity, u is calculated at $x_{j-1/2}$, x_j and $x_{j+1/2}$ by solving the elliptic equation (2.1) with a second-order Finite Element Method (FEM) so that we possess $u_{j-1/2}$, u_j and $u_{j+1/2}$ for every cell. We denote the solution of the FEM by the map \mathcal{G} given in subsection (ii) below, which takes $\hat{\mathbf{h}}$, $\hat{\mathbf{G}}$ and $\hat{\mathbf{b}}$

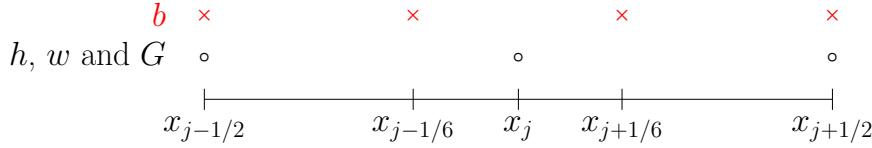


Figure 3.1: Location of the reconstructions for h , w , G and b inside the j^{th} cell.

as inputs. So that

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

- (iii) Flux Across Cell Interfaces: We calculate the average fluxes $F_{j-1/2}^n$ and $F_{j+1/2}^n$ across the cell boundaries $x_{j-1/2}$ and $x_{j+1/2}$ from time t^n to t^{n+1} using $\mathcal{F}_{j-1/2}$ and $\mathcal{F}_{j+1/2}$ defined in subsection (iii) below, so that

$$\begin{aligned} F_{j-1/2}^n &= \mathcal{F}_{j-1/2}(\hat{\mathbf{h}}, \hat{\mathbf{G}}, \hat{\mathbf{b}}, \hat{\mathbf{u}}), \\ F_{j+1/2}^n &= \mathcal{F}_{j+1/2}(\hat{\mathbf{h}}, \hat{\mathbf{G}}, \hat{\mathbf{b}}, \hat{\mathbf{u}}). \end{aligned}$$

- (iv) Source Terms: We calculate the source term contribution to the cell average of a quantity over a time step; S_j^n with the operator \mathcal{S} defined in subsection (iv) below

$$S_j^n = \mathcal{S}_j(\hat{\mathbf{h}}, \hat{\mathbf{w}}, \hat{\mathbf{b}}, \hat{\mathbf{u}}).$$

- (v) Update All the Cell Averages Using a Forward Euler Approximation: 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) Update All the Cell Averages Using a Second-Order SSP Runge-Kutta Method: We repeat steps (i)-(v) and use SSP Runge-Kutta time stepping to calculate $\bar{\mathbf{h}}$ and $\bar{\mathbf{G}}$ at t^{n+1} with second-order accuracy in space and time.

(i) 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 3.1. 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 functions that are linear over a cell and discontinuous at the cell edges. 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

$$q_{j-1/2}^+ = \mathcal{R}_{j-1/2}^+(\bar{q}) = \bar{q} - \frac{\Delta x}{2} d_j, \quad (3.1a)$$

$$q_j = \mathcal{R}_j(\bar{q}) = \bar{q}, \quad (3.1b)$$

$$q_{j+1/2}^- = \mathcal{R}_{j+1/2}^-(\bar{q}) = \bar{q} + \frac{\Delta x}{2} d_j \quad (3.1c)$$

where

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

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 [16]. When $\theta = 2$ the reconstruction introduces the least diffusion and is equivalent to the monotized central reconstruction [17].

Definition 3.1. The minmod function 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 0.

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

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), 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 first-order approximation which is necessarily 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 [] for which we have that

$$\bar{q} = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} q \, dx = q_j + \mathcal{O}(\Delta x^2). \quad (3.3)$$

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 $\partial^2 b / \partial x^2$ when they are calculated locally from the j^{th} cell. To accomplish this b is reconstructed with a cubic polynomial $C_j(x)$ centred around x_j

$$C_j(x) = c_0 (x - x_j)^3 + c_1 (x - x_j)^2 + c_2 (x - x_j) + c_3. \quad (3.4)$$

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 $C_j(x)$

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

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

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

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

Since a cubic is uniquely defined by its value at four locations, we associate this cubic with its value at the equally spaced locations $x_{j-1/2}$, $x_{j-1/6}$, $x_{j+1/6}$ and $x_{j+1/2}$ which are all inside the j^{th} cell. Furthermore we require a continuous bed profile across the cell edges and so we average the two reconstructions at the cell edges. Therefore, our reconstruction of the bed profile in the j^{th} cell is the cubic which takes these values

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

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

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

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

(ii) 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 Def 2.1. To form the FEM we take the weak form of Def 2.1 with test function τ which is

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

Integrating by parts with Dirichlet boundary conditions we get

$$\begin{aligned} \int_{\Omega} G\tau \, dx &= \int_{\Omega} uh \left(1 + \frac{\partial b}{\partial x} \frac{\partial b}{\partial x} \right) \tau \, dx + \int_{\Omega} \frac{1}{3} h^3 \frac{\partial u}{\partial x} \frac{\partial \tau}{\partial x} \, dx \\ &\quad - \int_{\Omega} \frac{1}{2} h^2 \frac{\partial b}{\partial x} u \frac{\partial \tau}{\partial x} \, dx - \int_{\Omega} \frac{1}{2} h^2 \frac{\partial b}{\partial x} \frac{\partial u}{\partial x} \tau \, dx. \end{aligned} \quad (3.6)$$

By assuming that t 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 the spatial domain 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 (2.5) 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 $u \in \mathbb{W}^{1,2}(\Omega)$.

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

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

which holds for all test functions τ . 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 defined in Appendix A are linear inside a cell and 0 elsewhere as shown in Figure 3.2. 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 our basis function approximations to h and G are in the appropriate function space.

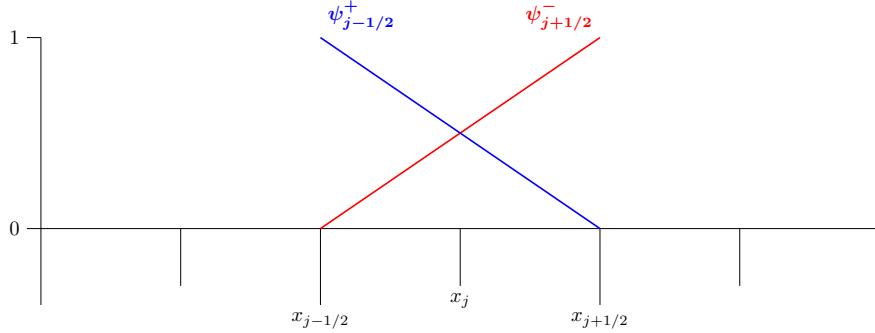


Figure 3.2: Discontinuous linear basis functions over a cell.

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

$$h = \sum_j \left(h_{j-1/2}^+ \psi_{j-1/2}^+ + h_{j+1/2}^- \psi_{j+1/2}^-, \right) \quad (3.8a)$$

$$G = \sum_j \left(G_{j-1/2}^+ \psi_{j-1/2}^+ + G_{j+1/2}^- \psi_{j+1/2}^-, \right). \quad (3.8b)$$

To calculate the flux and source terms of the evolution of G equation (2.5b) 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 $u \in \mathbb{W}^{1,2}$, this representation will be continuous across the cell edges $x_{j\pm 1/2}$. Therefore, we use the continuous quadratic basis functions $\phi_{j-1/2}$ depicted in Figure 3.3.

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

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

For the source term of the evolution of G equation (2.5b) 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 be a member of $\mathbb{W}^{2,2}$ which is smoother than $b \in \mathbb{W}^{1,2}$ as indicated by the elliptic equation (3.6). We choose the cubic basis functions γ which are continuous across the cell edges, as the bed profile will be continuous. These basis functions are shown in Figure 3.4 and from them we get our basis function approximation to b

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

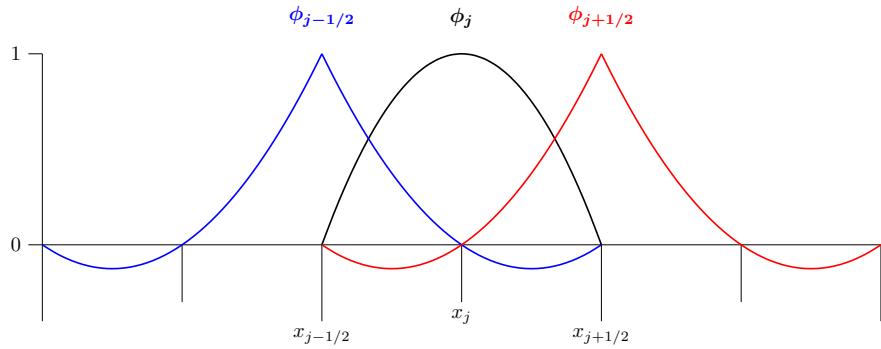


Figure 3.3: Continuous piecewise quadratic basis functions over a cell.

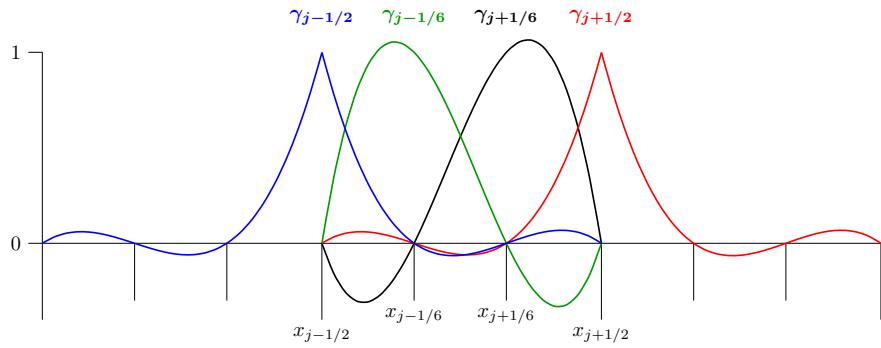


Figure 3.4: Continuous piecewise cubic basis functions over a cell.

With $b_{j-1/2}$ and $b_{j+1/2}$ being well defined as (3.5) imposes a unique reconstruction at the cell edges.

Calculation of Element-wise Matrices

The integral equation (3.7) holds for all τ . 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

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

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

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

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

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

We will demonstrate the rest of the process for the uh term as an example and provide the remaining integrals in Appendix A. 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 $[x_{j-1/2}, x_{j+1/2}]$, there are only a few non-zero contributions from the finite element approximations to h and u , so we have

$$\begin{aligned} & \frac{\Delta x}{2} \int_{-1}^1 (u_{j-1/2} \phi_{j-1/2} + u_j \phi_j + u_{j+1/2} \phi_{j+1/2}) \\ & \quad \left(h_{j-1/2}^+ \psi_{j-1/2}^+ + h_{j+1/2}^- \psi_{j+1/2}^- \right) \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} d\xi \\ &= \frac{\Delta x}{2} \left(h_{j-1/2}^+ \int_{-1}^1 \psi_{j-1/2}^+ \begin{bmatrix} \phi_{j-1/2} \phi_{j-1/2} & \phi_j \phi_{j-1/2} & \phi_{j+1/2} \phi_{j-1/2} \\ \phi_{j-1/2} \phi_j & \phi_j \phi_j & \phi_{j+1/2} \phi_j \\ \phi_{j+1/2} \phi_{j-1/2} & \phi_{j+1/2} \phi_j & \phi_{j+1/2} \phi_{j+1/2} \end{bmatrix} d\xi \right. \\ & \quad \left. + h_{j+1/2}^- \int_{-1}^1 \psi_{j+1/2}^- \begin{bmatrix} \phi_{j-1/2} \phi_{j-1/2} & \phi_j \phi_{j-1/2} & \phi_{j+1/2} \phi_{j-1/2} \\ \phi_{j-1/2} \phi_j & \phi_j \phi_j & \phi_{j+1/2} \phi_j \\ \phi_{j+1/2} \phi_{j-1/2} & \phi_{j+1/2} \phi_j & \phi_{j+1/2} \phi_{j+1/2} \end{bmatrix} d\xi \right) \\ & \quad \begin{bmatrix} u_{j-1/2} \\ u_j \\ u_{j+1/2} \end{bmatrix}. \end{aligned}$$

Calculating the integrals of all the basis function combinations we get

$$\begin{aligned} & \frac{\Delta x}{2} \int_{-1}^1 u h \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} d\xi = \\ & \frac{\Delta x}{2} \begin{bmatrix} \frac{7}{30} h_{j-1/2}^+ + \frac{1}{30} h_{j+1/2}^- & \frac{4}{30} h_{j-1/2}^+ & -\frac{1}{30} h_{j-1/2}^+ - \frac{1}{30} h_{j+1/2}^- \\ \frac{4}{30} h_{j-1/2}^+ & \frac{16}{30} h_{j-1/2}^+ + \frac{16}{30} h_{j+1/2}^- & \frac{4}{30} h_{j+1/2}^- \\ -\frac{1}{30} h_{j-1/2}^+ - \frac{1}{30} h_{j+1/2}^- & \frac{4}{30} h_{j+1/2}^- & \frac{1}{30} h_{j-1/2}^+ + \frac{7}{30} h_{j+1/2}^- \end{bmatrix} \\ & \quad \begin{bmatrix} u_{j-1/2} \\ u_j \\ u_{j+1/2} \end{bmatrix}. \quad (3.12) \end{aligned}$$

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 $G\tau$ generate the vector \mathbf{g}_j . Therefore, (3.7) can

be rewritten as

$$\sum_j \mathbf{A}_j \begin{bmatrix} u_{j-1/2} \\ u_j \\ u_{j+1/2} \end{bmatrix} = \sum_j \mathbf{g}_j. \quad (3.13)$$

This is a penta-diagonal matrix equation which can be solved by direct banded matrix solution techniques such as those of Press et al. [18] to obtain

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

as desired.

(iii) Flux Across the Cell Interfaces

We use the method of Kurganov et al. [19] 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 (2.13a).

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. [19] is

$$F_{j+\frac{1}{2}} = \frac{a_{j+\frac{1}{2}}^+ f(q_{j+\frac{1}{2}}^-) - a_{j+\frac{1}{2}}^- f(q_{j+\frac{1}{2}}^+)}{a_{j+\frac{1}{2}}^+ - a_{j+\frac{1}{2}}^-} + \frac{a_{j+\frac{1}{2}}^+ a_{j+\frac{1}{2}}^-}{a_{j+\frac{1}{2}}^+ - a_{j+\frac{1}{2}}^-} \left[q_{j+\frac{1}{2}}^+ - q_{j+\frac{1}{2}}^- \right] \quad (3.15)$$

where $a_{j+\frac{1}{2}}^+$ and $a_{j+\frac{1}{2}}^-$ are given by the wave speed bounds. Applying the wave speed bounds (2.13a) we obtain

$$a_{j+\frac{1}{2}}^- = \min \left\{ 0, u_{j+1/2}^- - \sqrt{gh_{j+1/2}^-}, u_{j+1/2}^+ - \sqrt{gh_{j+1/2}^+} \right\}, \quad (3.16)$$

$$a_{j+\frac{1}{2}}^+ = \max \left\{ 0, u_{j+1/2}^- + \sqrt{gh_{j+1/2}^-}, u_{j+1/2}^+ + \sqrt{gh_{j+1/2}^+} \right\}. \quad (3.17)$$

The flux functions $f(q_{j+\frac{1}{2}}^-)$ and $f(q_{j+\frac{1}{2}}^+)$ are evaluated using the reconstructed values of the j^{th} and $(j+1)^{th}$ cell respectively. From the continuity equation (2.5a) we have

$$\begin{aligned} f\left(h_{j+\frac{1}{2}}^-\right) &= u_{j+1/2}^- h_{j+1/2}^-, \\ f\left(h_{j+\frac{1}{2}}^+\right) &= u_{j+1/2}^+ h_{j+1/2}^+. \end{aligned}$$

For the evolution of G equation (2.5b) we have

$$\begin{aligned} f\left(G_{j+1/2}^-\right) &= u_{j+1/2}^- G_{j+1/2}^- + \frac{g}{2} \left(h_{j+1/2}^-\right)^2 - \frac{2}{3} \left(h_{j+1/2}^-\right)^3 \left[\left(\frac{\partial u}{\partial x}\right)_{j+1/2}^- \right]^2 \\ &\quad + \left(h_{j+1/2}^-\right)^2 u_{j+1/2}^- \left(\frac{\partial u}{\partial x}\right)_{j+1/2}^- \left(\frac{\partial b}{\partial x}\right)_{j+1/2}^-, \end{aligned} \quad (3.18a)$$

$$\begin{aligned} f\left(G_{j+1/2}^+\right) &= u_{j+1/2}^+ G_{j+1/2}^+ + \frac{g}{2} \left(h_{j+1/2}^+\right)^2 - \frac{2}{3} \left(h_{j+1/2}^+\right)^3 \left[\left(\frac{\partial u}{\partial x}\right)_{j+1/2}^+ \right]^2 \\ &\quad + \left(h_{j+1/2}^+\right)^2 u_{j+1/2}^+ \left(\frac{\partial u}{\partial x}\right)_{j+1/2}^+ \left(\frac{\partial b}{\partial x}\right)_{j+1/2}^+. \end{aligned} \quad (3.18b)$$

During the reconstruction process $h_{j-1/2}^+$, $h_{j+1/2}^-$, $G_{j-1/2}^+$, $G_{j+1/2}^-$ (3.1) were calculated, and the FEM provided $u_{j+1/2}^\pm = u_{j+1/2}$; because u is continuous across the cell boundaries. Approximations to $\left(\frac{\partial b}{\partial x}\right)_{j+1/2}^\pm$ and $\left(\frac{\partial u}{\partial x}\right)_{j+1/2}^\pm$ are 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. For u we have the quadratic $P_j^u(x)$ that passes through $u_{j-1/2}$, u_j and $u_{j+1/2}$ while for b we have the cubic $P_j^b(x)$ that passes through $b_{j-1/2}$, $b_{j-1/6}$, $b_{j+1/6}$ and $b_{j+1/2}$. So we have

$$P_j^u(x) = p_0^u (x - x_j)^2 + p_1^u (x - x_j) + p_2^u, \quad (3.19a)$$

$$P_j^b(x) = p_0^b (x - x_j)^3 + p_1^b (x - x_j)^2 + p_2^b (x - x_j) + p_3^b, \quad (3.19b)$$

By forcing the polynomials to pass through these reconstructed values we get that for $P_j^u(x)$

$$\begin{aligned} p_0^u &= \frac{u_{j-1/2} - 2u_j + u_{j+1/2}}{2\Delta x^2}, \\ p_1^u &= \frac{-u_{j-1/2} + u_{j+1/2}}{\Delta x}, \\ p_2^u &= u_j. \end{aligned}$$

While for $P_j^b(x)$ we get

$$p_0^b = \frac{-9b_{j-1/2} + 27b_{j-1/6} - 27b_{j+1/6} + 9b_{j+1/2}}{2\Delta x^3},$$

$$p_0^b = \frac{9b_{j-1/2} - 9b_{j-1/6} - 9b_{j+1/6} + 9b_{j+1/2}}{4\Delta x^2},$$

$$p_0^b = \frac{b_{j-1/2} - 27b_{j-1/6} + 27b_{j+1/6} - b_{j+1/2}}{8\Delta x},$$

$$p_0^b = \frac{-b_{j-1/2} + 9b_{j-1/6} + 9b_{j+1/6} - b_{j+1/2}}{16}.$$

Taking the derivative of the polynomials (3.19) we get

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

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

$$\left(\frac{\partial u}{\partial x} \right)_{j+1/2}^- = \frac{\partial}{\partial x} P_j^u(x_{j+1/2}), \quad (3.20a)$$

$$\left(\frac{\partial u}{\partial x} \right)_{j+1/2}^+ = \frac{\partial}{\partial x} P_{j+1}^u(x_{j+1/2}), \quad (3.20b)$$

$$\left(\frac{\partial b}{\partial x} \right)_{j+1/2}^- = \frac{\partial}{\partial x} P_j^b(x_{j+1/2}), \quad (3.20c)$$

$$\left(\frac{\partial b}{\partial x} \right)_{j+1/2}^+ = \frac{\partial}{\partial x} P_{j+1}^b(x_{j+1/2}). \quad (3.20d)$$

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. [20], who accomplished this for the SWWE. It was demonstrated that

this process could also be extended to the Serre equations [21]. To enforce well balancing the reconstruction of h is modified at the cell edges in the following way; first calculate

$$\dot{b}_{j+1/2}^- = w_{j+1/2}^- - h_{j+1/2}^- \quad \text{and} \quad \dot{b}_{j+1/2}^+ = w_{j+1/2}^+ - h_{j+1/2}^+. \quad (3.21)$$

Find the maximum

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

then define

$$\ddot{h}_{j+1/2}^- = \max \left\{ 0, w_{j+1/2}^- - \ddot{b}_{j+1/2} \right\}, \quad (3.22a)$$

$$\ddot{h}_{j+1/2}^+ = \max \left\{ 0, w_{j+1/2}^+ - \ddot{b}_{j+1/2} \right\}. \quad (3.22b)$$

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

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

which we use to calculate the flux term $F_{j+1/2}$ in (3.15) for the evolution equations for h (2.5a) and G (2.5b). Applying the same process but with different cells we obtain $F_{j-1/2}$ and we have

$$F_{j-1/2}^n = \mathcal{F}_{j-1/2} \left(\ddot{\mathbf{h}}, \hat{\mathbf{G}}, \hat{\mathbf{b}}, \hat{\mathbf{u}} \right), \quad (3.23a)$$

$$F_{j+1/2}^n = \mathcal{F}_{j+1/2} \left(\ddot{\mathbf{h}}, \hat{\mathbf{G}}, \hat{\mathbf{b}}, \hat{\mathbf{u}} \right) \quad (3.23b)$$

for the evolution of h and G equations as desired.

(iv) Source Terms

To evolve the Serre equations (2.5) 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 . The evolution equation for h (2.5a) has no source term, therefore we just present the calculation of the source term for the evolution equation for G (2.5b).

Following the work of Audusse et al. [20], we split our approximation to S_j into the centred source term S_{ci} and the corrective interface source terms $S_{j+\frac{1}{2}}^-$

and $S_{j+\frac{1}{2}}^+$. 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 steady state.

We calculate the centred source term using

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

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. However, to ensure that the terms cancel properly for a lake at rest we modify our approximation to $\frac{\partial b}{\partial x}$ 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}

$$\left(\frac{\partial u}{\partial x} \right)_j = \frac{\partial}{\partial x} P_j^u(x_j), \quad (3.24a)$$

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

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

To ensure well balancing the corrective interface source terms

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

$$S_{j-\frac{1}{2}}^+ = \frac{g}{2} \left(h_{j-\frac{1}{2}}^+ \right)^2 - \frac{g}{2} \left(\dot{h}_{j-\frac{1}{2}}^+ \right)^2$$

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 $\dot{h}_{j+\frac{1}{2}}^-$ and $\dot{h}_{j+\frac{1}{2}}^+$ from (3.22). Combining the centred and interface source terms our approximation to the source term for the evolution of G equation is

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

(v) Update Cell Averages

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

$$\bar{q}_j^{n+1} = \bar{q}_j^n + \frac{\Delta t}{\Delta x} \left(F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n + S_j^n \right) \quad (3.26)$$

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

(vi) Second-Order SSP Runge-Kutta Method

To increase the order of accuracy in time we employ the strong stability preserving Runge-Kutta method [22] which is a convex combination of first-order time steps (3.26) in the following way

$$\bar{q}_j^{(1)} = \bar{q}_j^n + \frac{\Delta t}{\Delta x} \left(F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n + S_j^n \right), \quad (3.27a)$$

$$\bar{q}_j^{(2)} = \bar{q}_j^{(1)} + \frac{\Delta t}{\Delta x} \left(F_{j+\frac{1}{2}}^{(1)} - F_{j-\frac{1}{2}}^{(1)} + S_j^{(1)} \right), \quad (3.27b)$$

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

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-Lowy (CFL) condition [23] 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

$$\Delta t \leq \frac{Cr}{\max_j \{ a_{j+1/2}^\pm \}} \Delta x \quad (3.28)$$

where $a_{j+1/2}^\pm$ are the wave-speed bounds used in Kurganovs 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. We therefore have ghost cells with the following associated quantities

$$\begin{aligned}\hat{\mathbf{h}}_{-1} &= \begin{bmatrix} h_{-3/2}^+ \\ h_{-1} \\ h_{-1/2}^- \end{bmatrix}, & \hat{\mathbf{h}}_{m+1} &= \begin{bmatrix} h_{m+1/2}^+ \\ h_{m+1} \\ h_{m+3/2}^- \end{bmatrix}, \\ \hat{\mathbf{w}}_{-1} &= \begin{bmatrix} w_{-3/2}^+ \\ w_{-1} \\ w_{-1/2}^- \end{bmatrix}, & \hat{\mathbf{w}}_{m+1} &= \begin{bmatrix} w_{m+1/2}^+ \\ w_{m+1} \\ w_{m+3/2}^- \end{bmatrix}, \\ \hat{\mathbf{G}}_{-1} &= \begin{bmatrix} G_{-3/2}^+ \\ G_{-1} \\ G_{-1/2}^- \end{bmatrix}, & \hat{\mathbf{G}}_{m+1} &= \begin{bmatrix} G_{m+1/2}^+ \\ G_{m+1} \\ G_{m+3/2}^- \end{bmatrix}, \\ \hat{\mathbf{u}}_{-1} &= \begin{bmatrix} u_{-3/2} \\ u_{-1} \\ u_{-1/2} \end{bmatrix}, & \hat{\mathbf{u}}_{m+1} &= \begin{bmatrix} u_{m+1/2} \\ u_{m+1} \\ u_{m+3/2} \end{bmatrix} \\ \hat{\mathbf{b}}_{-2} &= \begin{bmatrix} b_{-5/2} \\ b_{-13/6} \\ b_{-11/6} \\ b_{-3/2} \end{bmatrix}, & \hat{\mathbf{b}}_{-1} &= \begin{bmatrix} b_{-3/2} \\ b_{-7/6} \\ b_{-5/6} \\ b_{-1/2} \end{bmatrix}, & \hat{\mathbf{b}}_{m+1} &= \begin{bmatrix} b_{m+1/2} \\ b_{m+5/6} \\ b_{m+7/6} \\ b_{m+3/2} \end{bmatrix}, & \hat{\mathbf{b}}_{m+2} &= \begin{bmatrix} b_{m+3/2} \\ b_{m+11/6} \\ b_{m+13/6} \\ b_{m+5/2} \end{bmatrix}. \end{aligned}$$

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

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

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

These are then combined with the other element contributions in the global matrix (3.13).

3.5 Dry Beds - [Requires Modification]

Dry beds are handled adequately by all steps of our FEVM in their current form, except the FEM solution for u . For the FEM a dry bed presents two issues; when $h = 0$ the stiffness matrix will become singular, and when h is small u may become quite large leading to instabilities.

In previous work [24] direct banded matrix solvers such as the Thomas algorithm were employed to solve (3.14), however such methods rely on non-singular matrices and so are unsuitable for dry beds. To deal with this an LU decomposition algorithm by Press et al. [18] was used. This algorithm solves banded matrix problems using an LU decomposition with partial pivoting, which inserts small non-zero pivots when their 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. This results in a FEM that allows for $h = 0$, but still faces the problem of large u values when h is small.

To deal with large u values we restricted our solution of the FEM for u to cells where $\bar{h}_j > h_{tol}$ and modified our approximation to $h_{j-1/2}^+$ and $h_{j+1/2}^-$ in our stiffness matrix. To restrict the domain of our solution we search through the domain at the beginning of each first-order step and identify wet regions where $\bar{h}_j \geq h_{tol}$ and dry regions where $\bar{h}_j < h_{tol}$. In the dry regions we set h , G and u to be zero uniformly over the cell. In the wet regions we solve the FEM with the appropriate boundary conditions and the following modifications to $h_{j-1/2}^+$ and $h_{j+1/2}^-$

$$h_{j-1/2}^+ = h_{j-1/2}^+ \left(\frac{h_{j-1/2}^+ + h_{base}}{h_{j-1/2}^+ + h_{tol}} \right), \quad (3.31a)$$

$$h_{j+1/2}^- = h_{j+1/2}^- \left(\frac{h_{j+1/2}^- + h_{base}}{h_{j+1/2}^- + h_{tol}} \right). \quad (3.31b)$$

Where on the right hand side we mean $h_{j-1/2}^+$ and $h_{j+1/2}^-$ as calculated from the reconstruction (3.1). This modification ensures that as $h \rightarrow 0$ and correspondingly $G \rightarrow 0$ then $u \rightarrow 0$ by increasing the size of h in the stiffness matrix.

Typically we had $p_{tol} = 10^{-20}$, $h_{tol} = 10^{-12}$ and $h_{base} = 10^{-8}$ which allowed for a stable calculation of u in the presence of dry beds and small water depths. While the accuracy of the calculation of u by the FEM was maintained when the water depth is large compared to h_{tol} and h_{base} .

Chapter 4

Linear Analysis 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 [25]. 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 is increased. While a numerical method is stable if the errors from previous time steps are not amplified in 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 [12]. 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 our numerical methods for the linearised Serre equations with a horizontal bed. The effect of variations in the bed and nonlinear terms are important when studying the convergence 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 our 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 are derived from the linearised Serre equations with variable bathymetry having no effect [24], 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

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

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 the second-order FEVM and perform the convergence and dispersion analysis. We will then present the results of these analyses for all the other numerical methods in this thesis.

4.1 Linearised Serre Equations with a Horizontal Bed

The Serre equations with a horizontal bed (2.4) are linearised by considering waves as small perturbations $\delta\eta$ and δv on a flow with a mean height H and a mean velocity U respectively. So we have

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

$$u(x, t) = U + \delta v(x, t) + \mathcal{O}(\delta^2), \quad (4.2b)$$

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

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

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

and for G

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

These equations can be reformulated into conservation law form for the quantities η and G

$$\frac{\partial\eta}{\partial t} + \frac{\partial}{\partial x} (Hv + U\eta) = 0, \quad (4.4a)$$

$$\frac{\partial G}{\partial t} + \frac{\partial}{\partial x} (UG + UHv + gH\eta) = 0 \quad (4.4b)$$

where

$$G = UH + U\eta + Hv - \frac{H^3}{3} \frac{\partial^2 v}{\partial x^2}. \quad (4.4c)$$

We have absorbed the δ factor into the corresponding η and v terms to simplify the notation.

4.2 Evolution Matrix

To derive the evolution matrix, \mathbf{E} we study the behaviour of (4.4) when η and v are Fourier modes. For a general quantity q a Fourier mode is

$$q(x, t) = q(0, 0)e^{i(\omega t + kx)} \quad (4.5)$$

where k is the wavenumber, ω is the frequency and i is the imaginary number. 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

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

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

When all the quantities are Fourier modes 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 only vary with H , U , k , ω , Δx and Δt and hence are independent of j and n . By combining these operators the evolution matrix \mathbf{E} can be derived for the second-order FEVM for the linearised Serre equations

with horizontal bed. Since all of the constituent operators 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 laid out in Section 3.2. Since the linearised Serre equations with a horizontal bed has no Source Terms, step (iv) will be skipped.

(i) Reconstruction

Given $\bar{\eta}$ and $\bar{\mathbf{G}}$ 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 $d_j = (-\bar{q}_{j-1} + \bar{q}_{j+1}) / (2\Delta x)$ in the reconstruction. Applying (4.6) to the reconstructions (3.1) with the centred slope approximation we obtain

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

$$q_j = \bar{q}_j = \mathcal{R}_j \bar{q}_j, \quad (4.7b)$$

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

(ii) Fluid Velocity

To calculate $v_{j+1/2}$ 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 τ and integrating over the spatial domain Ω

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

For G we use the basis functions $\psi_{j-1/2}^+$ and $\psi_{j+1/2}^-$ defined in Appendix A, which means our approximation to G is linear inside a cell with discontinuous jumps at the cell edges. For τ and v we use the basis functions $\phi_{j-1/2}$, ϕ_j and $\phi_{j+1/2}$ defined in Appendix A so that τ and our approximation to v are quadratic polynomials inside a cell and are continuous across the cell edges. Substituting the finite element approximations to these quantities and only integrating over the j^{th} cell, we get

$$\begin{aligned}
& \sum_j \int_{x_{j-1/2}}^{x_{j+1/2}} \left(G_{j-1/2}^+ \psi_{j-1/2}^+ + G_{j+1/2}^- \psi_{j+1/2}^- \right) \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} dx = \\
& \sum_j UH \int_{x_{j-1/2}}^{x_{j+1/2}} \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} dx + \sum_j U \int_{x_{j-1/2}}^{x_{j+1/2}} \left(\eta_{j-1/2}^+ \psi_{j-1/2}^+ + \eta_{j+1/2}^- \psi_{j+1/2}^- \right) \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} dx \\
& + \sum_j H \int_{x_{j-1/2}}^{x_{j+1/2}} (v_{j-1/2} \phi_{j-1/2} + v_j \phi_j + v_{j+1/2} \phi_{j+1/2}) \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} dx \\
& + \sum_j \frac{H^3}{3} \int_{x_{j-1/2}}^{x_{j+1/2}} \left(v_{j-1/2} \frac{\partial \phi_{j-1/2}}{\partial x} + v_j \frac{\partial \phi_j}{\partial x} + v_{j+1/2} \frac{\partial \phi_{j+1/2}}{\partial x} \right) \begin{bmatrix} \frac{\partial \phi_{j-1/2}}{\partial x} \\ \frac{\partial \phi_j}{\partial x} \\ \frac{\partial \phi_{j+1/2}}{\partial x} \end{bmatrix} dx.
\end{aligned}$$

Calculating all the integrals of the appropriate basis function combinations we get

$$\begin{aligned}
\sum_j \frac{\Delta x}{6} \begin{bmatrix} G_{j-1/2}^+ \\ 2G_{j-1/2}^+ + 2G_{j+1/2}^- \\ G_{j+1/2}^- \end{bmatrix} &= \sum_j UH \frac{\Delta x}{6} \begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix} + \sum_j \frac{\Delta x}{6} U \begin{bmatrix} \eta_{j-1/2}^+ \\ 2\eta_{j-1/2}^+ + 2\eta_{j+1/2}^- \\ \eta_{j+1/2}^- \end{bmatrix} \\
&+ \sum_j \left(H \frac{\Delta x}{30} \begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix} + \frac{H^3}{9\Delta x} \begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix} \right) \begin{bmatrix} v_{j-1/2} \\ v_j \\ v_{j+1/2} \end{bmatrix}.
\end{aligned}$$

Using (4.7), we obtain

$$\begin{aligned}
\sum_j \frac{\Delta x}{6} \begin{bmatrix} \mathcal{R}_{j-1/2}^+ \bar{G}_j \\ 2\mathcal{R}_{j-1/2}^+ \bar{G}_j + 2\mathcal{R}_{j+1/2}^- \bar{G}_j \\ \mathcal{R}_{j+1/2}^- \bar{G}_j \end{bmatrix} &= \sum_j UH \frac{\Delta x}{6} \begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix} \\
&\quad + \sum_j \frac{\Delta x}{6} U \begin{bmatrix} \mathcal{R}_{j-1/2}^+ \bar{\eta}_j \\ 2\mathcal{R}_{j-1/2}^+ \bar{\eta}_j + 2\mathcal{R}_{j+1/2}^- \bar{\eta}_j \\ \mathcal{R}_{j+1/2}^- \bar{\eta}_j \end{bmatrix} \\
\sum_j \left(H \frac{\Delta x}{30} \begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix} + \frac{H^3}{9\Delta x} \begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix} \right) &\begin{bmatrix} e^{-ik\frac{\Delta x}{2}} v_j \\ v_j \\ e^{ik\frac{\Delta x}{2}} v_j \end{bmatrix}.
\end{aligned}$$

After simplifying

$$\begin{aligned}
\sum_j \frac{\Delta x}{6} \begin{bmatrix} \mathcal{R}_{j-1/2}^+ \\ 2\mathcal{R}_{j-1/2}^+ + 2\mathcal{R}_{j+1/2}^- \\ \mathcal{R}_{j+1/2}^- \end{bmatrix} \bar{G}_j &= \sum_j UH \frac{\Delta x}{6} \begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix} \\
&\quad + \sum_j \frac{\Delta x}{6} \begin{bmatrix} \mathcal{R}_{j-1/2}^+ \\ 2\mathcal{R}_{j-1/2}^+ + 2\mathcal{R}_{j+1/2}^- \\ \mathcal{R}_{j+1/2}^- \end{bmatrix} \bar{\eta}_j + \sum_j \left(H \frac{\Delta x}{30} \begin{bmatrix} 4e^{-ik\frac{\Delta x}{2}} + 2 - e^{ik\frac{\Delta x}{2}} \\ 2e^{-ik\frac{\Delta x}{2}} + 16 + 2e^{ik\frac{\Delta x}{2}} \\ -e^{-ik\frac{\Delta x}{2}} + 2 + 4e^{ik\frac{\Delta x}{2}} \end{bmatrix} \right. \\
&\quad \left. + \frac{H^3}{9\Delta x} \begin{bmatrix} 7e^{-ik\frac{\Delta x}{2}} - 8 + e^{ik\frac{\Delta x}{2}} \\ -8e^{-ik\frac{\Delta x}{2}} + 16 - 8e^{ik\frac{\Delta x}{2}} \\ e^{-ik\frac{\Delta x}{2}} - 8 + 7e^{ik\frac{\Delta x}{2}} \end{bmatrix} \right) v_j.
\end{aligned}$$

These vectors represent the contribution from the j^{th} cell to the three equations relating the quantities at $x_{j-1/2}$, x_j and $x_{j+1/2}$ respectively. Since the intercell flux only requires u at $x_{j+1/2}$ we only need to solve the third equation.

So far we have only given the contribution to the equation at $x_{j+1/2}$ from the j^{th} cell, but there is also a contribution from the adjacent $(j+1)^{th}$ cell because

$\phi_{j+1/2}$ is non-zero over both cells. Accounting for this we get

$$\begin{aligned}
& \frac{\Delta x}{6} \left(\mathcal{R}_{j+1/2}^- + \mathcal{R}_{j+1/2}^+ \right) \bar{G}_j = \\
& \quad \frac{\Delta x}{6} 2UH + \frac{\Delta x}{6} U \left(\mathcal{R}_{j+1/2}^- + \mathcal{R}_{j+1/2}^+ \right) \bar{\eta}_j \\
& \quad + \left(H \frac{\Delta x}{30} \left[-e^{-ik\frac{\Delta x}{2}} + 2 + 4e^{ik\frac{\Delta x}{2}} + e^{ik\Delta x} \left(4e^{-ik\frac{\Delta x}{2}} + 2 - e^{ik\frac{\Delta x}{2}} \right) \right] \right. \\
& \quad \left. + \frac{H^3}{9\Delta x} \left[e^{-ik\frac{\Delta x}{2}} - 8 + 7e^{ik\frac{\Delta x}{2}} + e^{ik\Delta x} \left(7e^{-ik\frac{\Delta x}{2}} - 8 + e^{ik\frac{\Delta x}{2}} \right) \right] \right) v_j \\
& \quad = \\
& \quad \frac{\Delta x}{3} UH + \frac{\Delta x}{6} U \left(\mathcal{R}_{j+1/2}^- + \mathcal{R}_{j+1/2}^+ \right) \bar{\eta}_j \\
& \quad + \left[H \frac{\Delta x}{30} \left(4 \cos \left(\frac{k\Delta x}{2} \right) - 2 \cos(k\Delta x) + 8 \right) + \frac{H^3}{9\Delta x} \left(-16 \cos \left(\frac{k\Delta x}{2} \right) + 2 \cos(k\Delta x) + 14 \right) \right] \\
& \quad \times e^{ik\frac{\Delta x}{2}} v_j.
\end{aligned}$$

From (4.6) $v_{j+1/2} = e^{ik\frac{\Delta x}{2}} v_j$ we have

$$\begin{aligned}
v_{j+1/2} &= \\
& \left[\frac{\Delta x}{6} \left(\mathcal{R}_{j+1/2}^- + \mathcal{R}_{j+1/2}^+ \right) \bar{G}_j - U \frac{\Delta x}{6} \left(\mathcal{R}_{j+1/2}^- + \mathcal{R}_{j+1/2}^+ \right) \bar{\eta}_j - \frac{\Delta x}{3} UH \right] \\
& \div \left[H \frac{\Delta x}{30} \left(4 \cos \left(\frac{k\Delta x}{2} \right) - 2 \cos(k\Delta x) + 8 \right) + \frac{H^3}{9\Delta x} \left(-16 \cos \left(\frac{k\Delta x}{2} \right) + 2 \cos(k\Delta x) + 14 \right) \right] \\
& = \mathcal{G}^C \bar{G}_j + \mathcal{G}^\eta \bar{\eta}_j + \mathcal{G}^c. \tag{4.8}
\end{aligned}$$

(iii) Flux Across the Cell Interfaces

The average intercell flux $F_{j+1/2}$ is approximated using the method of Kurganov et al. [19] (3.15). For the linearised Serre equations we have the wave speed bounds (2.13a), 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 $Fr = \frac{U}{\sqrt{gH}}$; (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 for a general quantity q (3.15) reduces to

$$F_{j+\frac{1}{2}} = f\left(q_{j+\frac{1}{2}}^+\right). \quad (4.10)$$

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

$$F_{j+\frac{1}{2}}^\eta = Hv_{j+1/2} + U\eta_{j+1/2}^+$$

since v is continuous $v_{j+1/2} = v_{j+1/2}^+ = v_{j+1/2}^-$. Using the FEM for $v_{j+1/2}$ (4.8) and the reconstruction (4.7) we have

$$\begin{aligned} F_{j+\frac{1}{2}}^\eta &= H(\mathcal{G}^G \bar{G}_j + \mathcal{G}^\eta \bar{\eta}_j + \mathcal{G}^c) + U\eta_{j+1/2}^+ \\ &= (H\mathcal{G}^\eta + U\mathcal{R}_{j+1/2}^+) \bar{\eta}_j + H\mathcal{G}^G \bar{G}_j + H\mathcal{G}^c \\ &= \mathcal{F}_{j+\frac{1}{2}}^{\eta,\eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,c} \end{aligned} \quad (4.11)$$

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

$$F_{j+\frac{1}{2}}^G = UG_{j+1/2}^+ + UHv_{j+1/2} + gH\eta_{j+1/2}^+$$

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

$$\begin{aligned} F_{j+\frac{1}{2}}^G &= UG_{j+1/2}^+ + UH(\mathcal{G}^G \bar{G}_j + \mathcal{G}^\eta \bar{\eta}_j + \mathcal{G}^c) + gH\eta_{j+1/2}^+ \\ &= (UH\mathcal{G}^\eta + gH\mathcal{R}_{j+1/2}^+) \bar{\eta}_j + (U\mathcal{R}_{j+1/2}^+ + UH\mathcal{G}^G) \bar{G}_j + UH\mathcal{G}^c \\ &= \mathcal{F}_{j+\frac{1}{2}}^{G,\eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,c} \end{aligned} \quad (4.12)$$

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 for a general quantity q (3.15) is

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

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

$$\begin{aligned} F_{j+\frac{1}{2}}^\eta = & \frac{U}{2\sqrt{gH}} \left[Hv_{j+1/2} + U\eta_{j+\frac{1}{2}}^- - Hv_{j+1/2} - U\eta_{j+\frac{1}{2}}^+ \right] \\ & + \frac{1}{2} \left[Hv_{j+1/2} + U\eta_{j+\frac{1}{2}}^- + Hv_{j+1/2} + U\eta_{j+\frac{1}{2}}^+ \right] \\ & + \frac{U^2 - gH}{2\sqrt{gH}} \left[\eta_{j+\frac{1}{2}}^+ - \eta_{j+\frac{1}{2}}^- \right]. \end{aligned} \quad (4.14)$$

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

$$\begin{aligned} F_{j+\frac{1}{2}}^\eta = & \left(H\mathcal{G}^\eta + \frac{U}{2} \left[\mathcal{R}_{j+1/2}^- + \mathcal{R}_{j+1/2}^+ \right] - \frac{\sqrt{gH}}{2} \left[\mathcal{R}_{j+1/2}^+ - \mathcal{R}_{j+1/2}^- \right] \right) \bar{\eta}_j \\ & + H\mathcal{G}^G \bar{G}_j + H\mathcal{G}^c \\ = & \mathcal{F}_{j+\frac{1}{2}}^{\eta,\eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,c}. \end{aligned} \quad (4.15)$$

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

$$\begin{aligned} F_{j+\frac{1}{2}}^G = & \frac{U}{2\sqrt{gH}} \left[UG_{j+\frac{1}{2}}^- + UHv_{j+1/2} + gH\eta_{j+\frac{1}{2}}^- - UG_{j+\frac{1}{2}}^+ - UHv_{j+1/2} - gH\eta_{j+\frac{1}{2}}^+ \right] \\ & + \frac{1}{2} \left[UG_{j+\frac{1}{2}}^- + UHv_{j+1/2} + gH\eta_{j+\frac{1}{2}}^- + UG_{j+\frac{1}{2}}^+ + UHv_{j+1/2} + gH\eta_{j+\frac{1}{2}}^+ \right] \\ & + \frac{U^2 - gH}{2\sqrt{gH}} \left[G_{j+\frac{1}{2}}^+ - G_{j+\frac{1}{2}}^- \right]. \end{aligned} \quad (4.16)$$

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

$$\begin{aligned} F_{j+\frac{1}{2}}^G = & \left(\frac{U\sqrt{gH}}{2} \left[\mathcal{R}_{j+1/2}^- - \mathcal{R}_{j+1/2}^+ \right] + UH\mathcal{G}^\eta + \frac{gH}{2} \left[\mathcal{R}_{j+1/2}^- + \mathcal{R}_{j+1/2}^+ \right] \right) \bar{\eta}_j \\ & + \left(UH\mathcal{G}^G + \frac{U}{2} \left[\mathcal{R}_{j+1/2}^- + \mathcal{R}_{j+1/2}^+ \right] - \frac{\sqrt{gH}}{2} \left[\mathcal{R}_{j+1/2}^+ - \mathcal{R}_{j+1/2}^- \right] \right) \bar{G}_j \\ & + UH\mathcal{G}^c \\ = & \mathcal{F}_{j+\frac{1}{2}}^{G,\eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,c}. \end{aligned} \quad (4.17)$$

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

$$F_{j+\frac{1}{2}} = f\left(q_{j+\frac{1}{2}}^-\right). \quad (4.18)$$

Substituting in the flux function for the continuity equation (4.4a) and the conservation of G equation (4.4b) we obtain

$$\begin{aligned} F_{j+\frac{1}{2}}^\eta &= \left(H\mathcal{G}^\eta + U\mathcal{R}_{j+1/2}^-\right)\bar{\eta}_j + H\mathcal{G}^G\bar{G}_j + H\mathcal{G}^c \\ &= \mathcal{F}_{j+\frac{1}{2}}^{\eta,\eta}\bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,G}\bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,c}, \end{aligned} \quad (4.19)$$

$$\begin{aligned} F_{j+\frac{1}{2}}^G &= \left(UH\mathcal{G}^\eta + gH\mathcal{R}_{j+1/2}^-\right)\bar{\eta}_j + \left(U\mathcal{R}_{j+1/2}^- + UH\mathcal{G}^G\right)\bar{G}_j + UH\mathcal{G}^c \\ &= \mathcal{F}_{j+\frac{1}{2}}^{G,\eta}\bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,G}\bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,c}. \end{aligned} \quad (4.20)$$

(v) Update Cell Averages

We have obtained the operators for the flux functions for all three cases, supercritical flow in both directions and subcritical flow. Substituting the appropriate flux approximation into the forward Euler step (3.26) we get

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

Since $\mathcal{F}_{j-\frac{1}{2}} = e^{-ik\Delta x}\mathcal{F}_{j+\frac{1}{2}}$ for all superscripts we have

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

This can be written in matrix form as

$$\begin{aligned} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^{n+1} &= \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n - (1 - e^{-ik\Delta x}) \frac{\Delta t}{\Delta x} \begin{bmatrix} \mathcal{F}^{\eta,\eta} & \mathcal{F}^{\eta,G} \\ \mathcal{F}^{G,\eta} & \mathcal{F}^{G,G} \end{bmatrix} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n \\ &= (\mathbf{I} - \Delta t \mathbf{F}) \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n \end{aligned} \quad (4.21)$$

for a single Euler step which is first-order in time and second-order in space. 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.

(vi) Second-Order SSP Runge-Kutta Method

The second-order SSP Runge Kutta time stepping in the FEVM (3.27) can be rewritten using (4.21) as

$$\left[\frac{\bar{\eta}}{G} \right]_j^{(1)} = (\mathbf{I} - \Delta t \mathbf{F}) \left[\frac{\bar{\eta}}{G} \right]_j^n, \quad (4.22a)$$

$$\left[\frac{\bar{\eta}}{G} \right]_j^{(2)} = (\mathbf{I} - \Delta t \mathbf{F}) \left[\frac{\bar{\eta}}{G} \right]_j^{(1)}, \quad (4.22b)$$

$$\left[\frac{\bar{\eta}}{G} \right]_j^{n+1} = \frac{1}{2} \left(\left[\frac{\bar{\eta}}{G} \right]_j^n + \left[\frac{\bar{\eta}}{G} \right]_j^{(2)} \right). \quad (4.22c)$$

Substituting (4.22a) and (4.22b) into (4.22c) we can write this in terms of the flux matrix \mathbf{F} and our cell averages at t^n as

$$\left[\frac{\bar{\eta}}{G} \right]_j^{n+1} = \frac{1}{2} \left(\left[\frac{\bar{\eta}}{G} \right]_j^n + (\mathbf{I} - \Delta t \mathbf{F})^2 \left[\frac{\bar{\eta}}{G} \right]_j^n \right).$$

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

$$\begin{aligned} \left[\frac{\bar{\eta}}{G} \right]_j^{n+1} &= \left(\mathbf{I} - \Delta t \mathbf{F} + \frac{1}{2} \Delta t^2 \mathbf{F}^2 \right) \left[\frac{\bar{\eta}}{G} \right]_j^n \\ &= \mathbf{E} \left[\frac{\bar{\eta}}{G} \right]_j^n \end{aligned} \quad (4.23)$$

which is in the desired form (4.1).

This is the evolution matrix \mathbf{E} for the second-order FEVM. 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 these flow scenarios individually.

Both the convergence and dispersion analysis then proceed by studying the properties of the evolution matrix \mathbf{E} . We begin with the convergence analysis.

4.3 Convergence Analysis

The linearised Serre equations are linear partial differential equations and therefore we can 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 of the linearised Serre equations. Together this stability and consistency condition 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

$$\rho(\mathbf{E}) \leq 1 \quad (4.24)$$

where $\rho(\mathbf{E})$ 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 $\rho(\mathbf{E})$ numerically for various values of Δx , Δt , k , H and U to check if (4.24) holds. We summarised our results in Figure 4.1 which is a plot of $\rho(\mathbf{E})$ against $\Delta x/\lambda$ for representative values of k , H and U ; where λ is the wavelength. We used $g = 9.81 \text{ m/s}^2$ and chose $\Delta t = 0.5 / (U + \sqrt{gH}) \Delta x$ to satisfy the CFL condition (3.28). This is the common choice of Δt in our numerical experiments.

The behaviour of $\rho(\mathbf{E})$ for $H = 1 \text{ m}$, $k = \frac{\pi}{10} \text{ m}^{-1}$ and $U = 0 \text{ m/s}$ and 1 m/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 our shallowness parameter $\sigma = \frac{1}{20}$ and so the Serre equations are applicable [].

In Figure 4.1 it can be seen that all methods have $\rho(\mathbf{E}) \leq 1$ for $U = 0 \text{ m/s}$ and are therefore stable. The two finite difference methods overlap and have $\rho(\mathbf{E}) = 1$ for all Δx values, while the second-order FDVM and the second-order FEVM also overlap with $\rho(\mathbf{E}) < 1$. However, when $U \neq 0 \text{ m/s}$ \mathcal{W} has $\rho(\mathbf{E}) > 1$ for all Δx and is therefore unstable. While for all other methods, $\rho(\mathbf{E}) \leq 1$ retain their stability when $U \neq 0 \text{ m/s}$.

The analytic value of $\rho(\mathbf{E})$ is given by using (4.6) to write

$$\left[\frac{\bar{\eta}}{G} \right]_j^{n+1} = e^{i\omega\Delta t} \left[\frac{\bar{\eta}}{G} \right]_j^n.$$

Therefore, the analytic growth factor is

$$\rho(\mathbf{E}) = |e^{i\omega\Delta t}| = \sqrt{\cos^2(\omega\Delta t) + \sin^2(\omega\Delta t)} = 1 \quad (4.25)$$

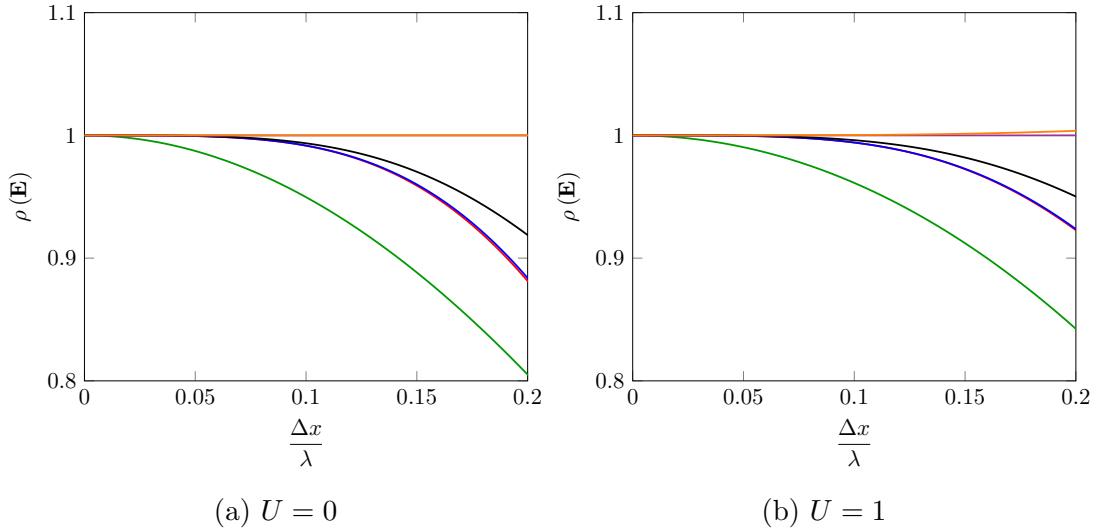


Figure 4.1: Spectral radius of \mathbf{E} for first-order FDVM (green), second-order FDVM (red), second-order FEVM (blue), third-order FDVM (black), \mathcal{D} (purple) and \mathcal{W} (orange). With $H = 1m$ and $k = \frac{\pi}{10}$.

since $\omega \in \mathbb{R}$ (2.11). Therefore numerical methods with $\rho(\mathbf{E})$ closer to 1 are closer to the analytic value. In this sense the two finite difference methods are best, although \mathcal{W} is unstable for $U \neq 0\text{m/s}$. As expected higher-order FDVM better approximate the analytic value of the growth factor and $\lim_{\Delta x \rightarrow 0} \rho(\mathbf{E}) = 1$ for all numerical methods.

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

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 the 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

$$\mathcal{T}^n = \mathbf{E} \left[\frac{\bar{\eta}}{\bar{G}} \right]_j^n - \left[\frac{\bar{\eta}}{\bar{G}} \right]_j^{n+1}. \quad (4.26)$$

To ensure consistency we must have that $\lim_{\Delta x, \Delta t \rightarrow 0} \|\mathcal{T}^n\| = 0$ for all n . Taking the norm of both sides of (4.26) we get

$$\|\mathcal{T}^n\| = \left\| \mathbf{E} \left[\frac{\bar{\eta}}{G} \right]_j^n - \left[\frac{\bar{\eta}}{G} \right]_j^{n+1} \right\|.$$

Making use of (4.6) we obtain

$$\|\mathcal{T}^n\| = \left\| \mathbf{E} \left[\frac{\bar{\eta}}{G} \right]_j^n - e^{i\omega \Delta t} \left[\frac{\bar{\eta}}{G} \right]_j^n \right\|.$$

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

$$\|\mathcal{T}^n\| \leq \|\mathbf{E} - e^{i\omega \Delta t} \mathbf{I}\| \left\| \left[\frac{\bar{\eta}}{G} \right]_j^n \right\|. \quad (4.27)$$

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

We calculated the Taylor series of $\mathbf{E} - e^{i\omega + \Delta t} \mathbf{I}$ for all the numerical methods for all flow scenarios; subcritical, critical and supercritical flows. We have reported the lowest order terms of the Taylor series in Tables 4.1 and 4.2 for the first-order FDVM, Table 4.3 for the second-order FDVM, Tables 4.4 and 4.5 for the third-order FDVM, Table 4.6 for the second-order FEVM, Table 4.7 for \mathcal{D} and Table 4.8 for \mathcal{W} . To be concise, the temporal and spatial errors for supercritical flow is only reported where it is different from the errors when $-1 \leq Fr \leq 1$, this only occurred for the spatial errors of the odd-order FDVM.

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

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

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

Element	Lowest Order Term of Error	
	Δx	Δt
$E_{0,0} - e^{i\omega_+ \Delta t}$	$-\frac{1}{2}\sqrt{gH}k^2\Delta t\Delta x$	$\frac{\sqrt{3gH\beta} + 3U}{\beta}ik\Delta t$
$E_{0,1}$	$\frac{3+\beta}{4\beta^2}ik^3\Delta t\Delta x^2$	$-\frac{3}{\beta}ik\Delta t$
$E_{1,0}$	$-\frac{1}{2}\sqrt{gH}k^2\Delta t\Delta x$	$\left(-gH + \frac{3U^2}{\beta}\right)ik\Delta t$
$E_{1,1} - e^{i\omega_+ \Delta t}$	$-\frac{1}{2}\sqrt{gH}k^2\Delta t\Delta x$	$\frac{\sqrt{3gH\beta} - 3U}{\beta}ik\Delta t$

Table 4.1: Table of the lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega_+ \Delta t} \mathbf{I}$ for the first-order FDVM with $-1 \leq Fr \leq 1$ and $\beta = 3 + k^2 H^2$.

Scheme	Lowest Order Δx Term of Error	
	$Fr < -1$	$Fr > 1$
$E_{0,0} - e^{i\omega_+ \Delta t}$	$\frac{1}{2}k^2U\Delta t\Delta x$	$-\frac{1}{2}k^2U\Delta t\Delta x$
$E_{0,1}$	$\frac{1}{2}gHk^2\Delta t\Delta x$	$\frac{1}{2}gHk^2\Delta t\Delta x$
$E_{1,1} - e^{i\omega_+ \Delta t}$	$\frac{1}{2}k^2U\Delta t\Delta x$	$-\frac{1}{2}k^2U\Delta t\Delta x$

Table 4.2: Table of the lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega_+ \Delta t} \mathbf{I}$ for the first-order FDVM which are different than those in Table 4.1.

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

Table 4.3: Table of the lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega_+ \Delta t} \mathbf{I}$ for the second-order FDVM with $-1 \leq Fr \leq 1$ and $\beta = 3 + k^2H^2$.

Element	Lowest Order Term of Error	
	Δx	Δt
$E_{0,0} - e^{i\omega_+ \Delta t}$	$-\frac{1}{12}\sqrt{gH}k^4\Delta t\Delta x^3$	$\frac{\sqrt{3gH\beta} + 3U}{\beta}ik\Delta t$
$E_{0,1}$	$\frac{\sqrt{gH}}{4\beta}ik^5\Delta t^2\Delta x^3$	$-\frac{3}{\beta}ik\Delta t$
$E_{1,0}$	$-\frac{1}{12}\sqrt{gH}k^4\Delta t\Delta x^3$	$\left(-gH + \frac{3U^2}{\beta}\right)ik\Delta t$
$E_{1,1} - e^{i\omega_+ \Delta t}$	$-\frac{1}{12}\sqrt{gH}k^4\Delta t\Delta x^3$	$\frac{\sqrt{3gH\beta} - 3U}{\beta}ik\Delta t$

Table 4.4: Table of the lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega_+ \Delta t} \mathbf{I}$ for the third-order FDVM with $-1 \leq Fr \leq 1$ and $\beta = 3 + k^2H^2$.

Scheme	Lowest Order Δx Term of Error	
	$Fr < -1$	$Fr > 1$
$E_{0,0} - e^{i\omega_+ \Delta t}$	$\frac{1}{12}k^4U\Delta t\Delta x^3$	$-\frac{1}{12}k^4U\Delta t\Delta x^3$
$E_{0,1}$	$\frac{1}{4\beta}iUk^5\Delta t^2\Delta x^3$	$-\frac{1}{4\beta}iUk^5\Delta t^2\Delta x^3$
$E_{1,0}$	$\frac{1}{12}gHk^4\Delta t^2\Delta x^3$	$-\frac{1}{12}gHk^4\Delta t^2\Delta x^3$
$E_{1,1} - e^{i\omega_+ \Delta t}$	$\frac{1}{12}k^4U\Delta t\Delta x^3$	$-\frac{1}{12}k^4U\Delta t\Delta x^3$

Table 4.5: Table of the lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega_+ \Delta t} \mathbf{I}$ for the third-order FDVM which are different than those in Table 4.1 with $\beta = 3 + k^2 H^2$.

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

Table 4.6: Table of the lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega_+ \Delta t} \mathbf{I}$ for the second-order FEVM with $-1 \leq Fr \leq 1$ and $\beta = 3 + k^2 H^2$.

Element	Lowest Order Term of Error	
	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: Table of the lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega \Delta t} \mathbf{I}$ for \mathcal{D} with $\beta = 3 + k^2 H^2$.

Element	Lowest Order Term of Error	
	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: Table of the lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega_+ \Delta t} \mathbf{I}$ for \mathcal{W} with $\beta = 3 + k^2 H^2$.

4.4 Dispersion Analysis

To study the dispersion of the numerical methods ω for the numerical methods must be calculated. Making use of (4.6) in (4.23) we get

$$e^{i\omega\Delta t} \left[\frac{\bar{\eta}}{G} \right]_j^n = \mathbf{E} \left[\frac{\bar{\eta}}{G} \right]_j^n. \quad (4.28)$$

Assuming that \mathbf{E} has an eigenvalue decomposition $\mathbf{E} = \mathbf{P}^{-1}\Lambda\mathbf{P}$ and substituting it into (4.28) we get

$$e^{i\omega\Delta t} \left[\frac{\bar{\eta}}{G} \right]_j^n = \mathbf{P}^{-1}\Lambda\mathbf{P} \left[\frac{\bar{\eta}}{G} \right]_j^n. \quad (4.29)$$

Left multiplying (4.29) by \mathbf{P} we obtain

$$e^{i\omega\Delta t} \mathbf{P} \left[\frac{\bar{\eta}}{G} \right]_j^n = \Lambda \mathbf{P} \left[\frac{\bar{\eta}}{G} \right]_j^n. \quad (4.30)$$

Since Λ is a diagonal matrix we must have that $e^{i\omega_+\Delta t} = \lambda_+$ and $e^{i\omega_-\Delta t} = \lambda_-$ where λ_\pm are the eigenvalues of \mathbf{E} and ω_\pm are the positive and negative branches of the dispersion relation. Therefore, the dispersion relation of a numerical method is

$$\tilde{\omega}_\pm = \frac{1}{i\Delta t} \log [\lambda_\pm]. \quad (4.31)$$

By comparing $\tilde{\omega}_\pm$ with the analytic ω_\pm given by the linearised Serre equations (2.11) we can determine the error in the dispersion relation for the numerical method. The real part of $\tilde{\omega}_\pm$ determines the speed of a wave, while the imaginary part determines the change in amplitude. For ω_\pm the imaginary part is zero and so the amplitude of waves of the linearised Serre equations are constant in time. We only present the results for the positive branch of the dispersion relation as the results for the negative and positive branches are very similar.

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

In Figures 4.2 and 4.3 we present the plots for $kH = \pi/10$ so that the water is shallow as $\sigma = 1/20$ and therefore, the Serre equations are appropriate. We present the real and imaginary errors separately as they account for different physical phenomenon and also present the total relative error as a measure of the overall difference of behaviour between waves in the numerical method and the waves of the linearised Serre equations.

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 $\Delta x \rightarrow 0$, 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 dilation of waves the roles are reversed with the two finite difference methods either dilating the waves very little (\mathcal{W} for $U > 0$) or not at all. When taking both effects into account with the complete error we see that the first-order FDVM has the largest dispersion error followed by \mathcal{W} , \mathcal{D} , the second-order FEVM, the second-order FDVM and finally the third-order FDVM 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. These results justify choosing the hybrid methods over these finite difference methods for the Serre equations.

Figures 4.2 and 4.3 furthermore demonstrate that the second-order FDVM is superior to the FEVM not just for the complete dispersion error, but its real and imaginary parts individually as well. Therefore, the second-order FDVM will do a better job in accurately modelling the speed and amplitude of waves than the second-order FEVM. Interestingly, for the two finite difference methods and the first-order FDVM there seems to be some trade-off between predicting the speed or amplitude of the waves very well.

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

These kH values are the same as those by Filippini et al. [26], 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. [26]. 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 . In particular the non-zero value of U changes the real part of the dispersion error for the first-order FDVM, most significantly 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 ω determines not only the speed of waves but also their amplitude. In particular it is possible that a method like the first-order FDVM performs very well for the real part of the dispersion error and poorly for the imaginary part, leading to false conclusions about the accuracy of the method.

The Taylor series expansion of $\tilde{\omega}$ was also derived for all the numerical methods. We have compiled the lowest order terms of the Taylor series for $\tilde{\omega}_+ - \omega_+$ 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.

We also present the lowest order terms of the Taylor series for $\tilde{\omega}_+ - \omega_+$ 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 odd-order numerical methods. We can see that for all the flow scenarios that our FDVM and the FEVM have the correct order of accuracy when approximating ω_+ .

Finally we present the lowest order terms of the Taylor series for $\tilde{\omega}_+ - \omega_+$ 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 both have the correct order of accuracy in space and time.

Because all methods were demonstrated to have the expected order of accuracy in approximating ω_+ this implies that for small Δx values the order of accuracy will be the primary driver of the dispersion error.

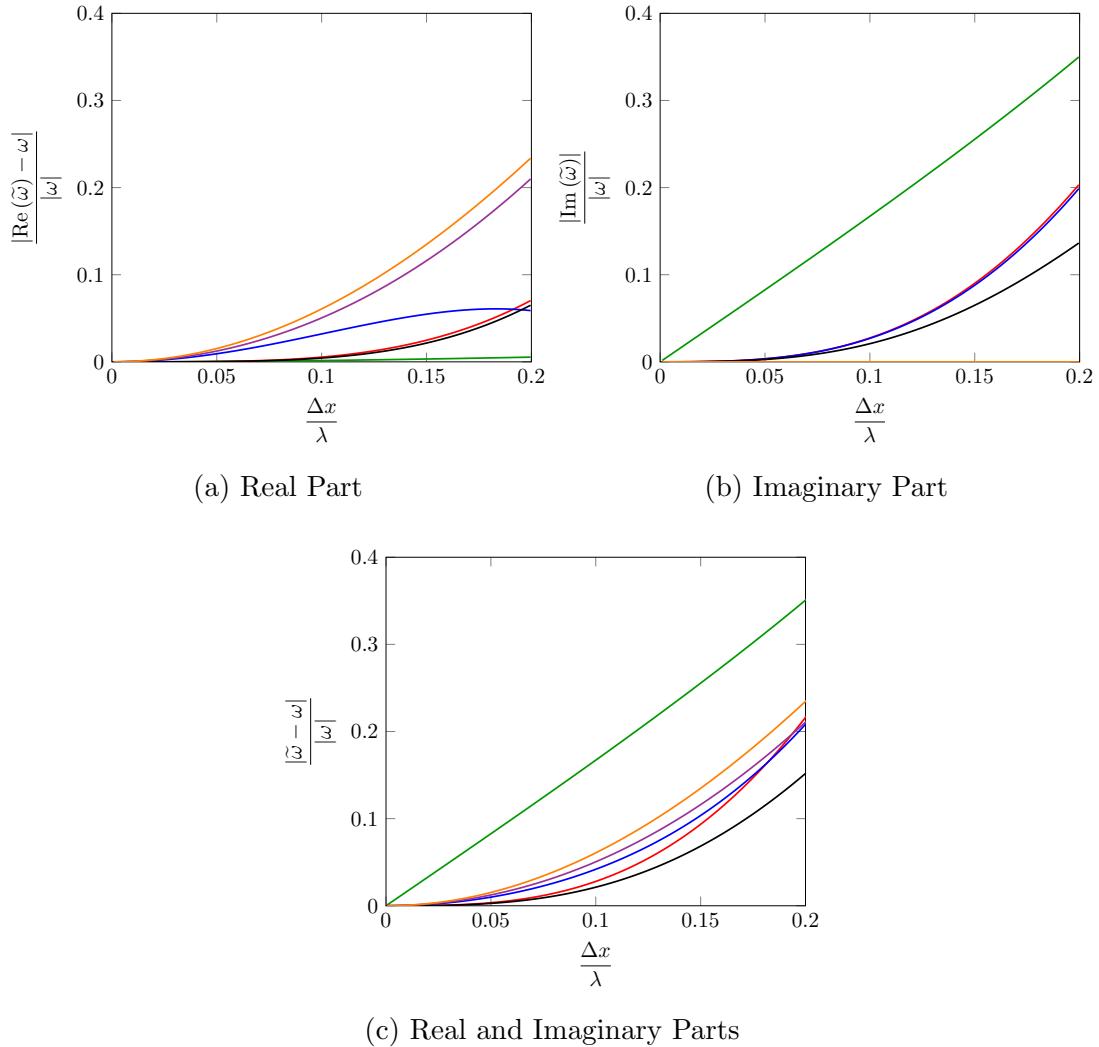


Figure 4.2: Relative dispersion error for first-order FDVM (—), second-order FDVM (—), second-order FEVM (—), third-order FDVM (—), \mathcal{D} (—) and \mathcal{W} (—). With $H = 1m$, $k = \frac{\pi}{10}$ and $U = 0m/s$.

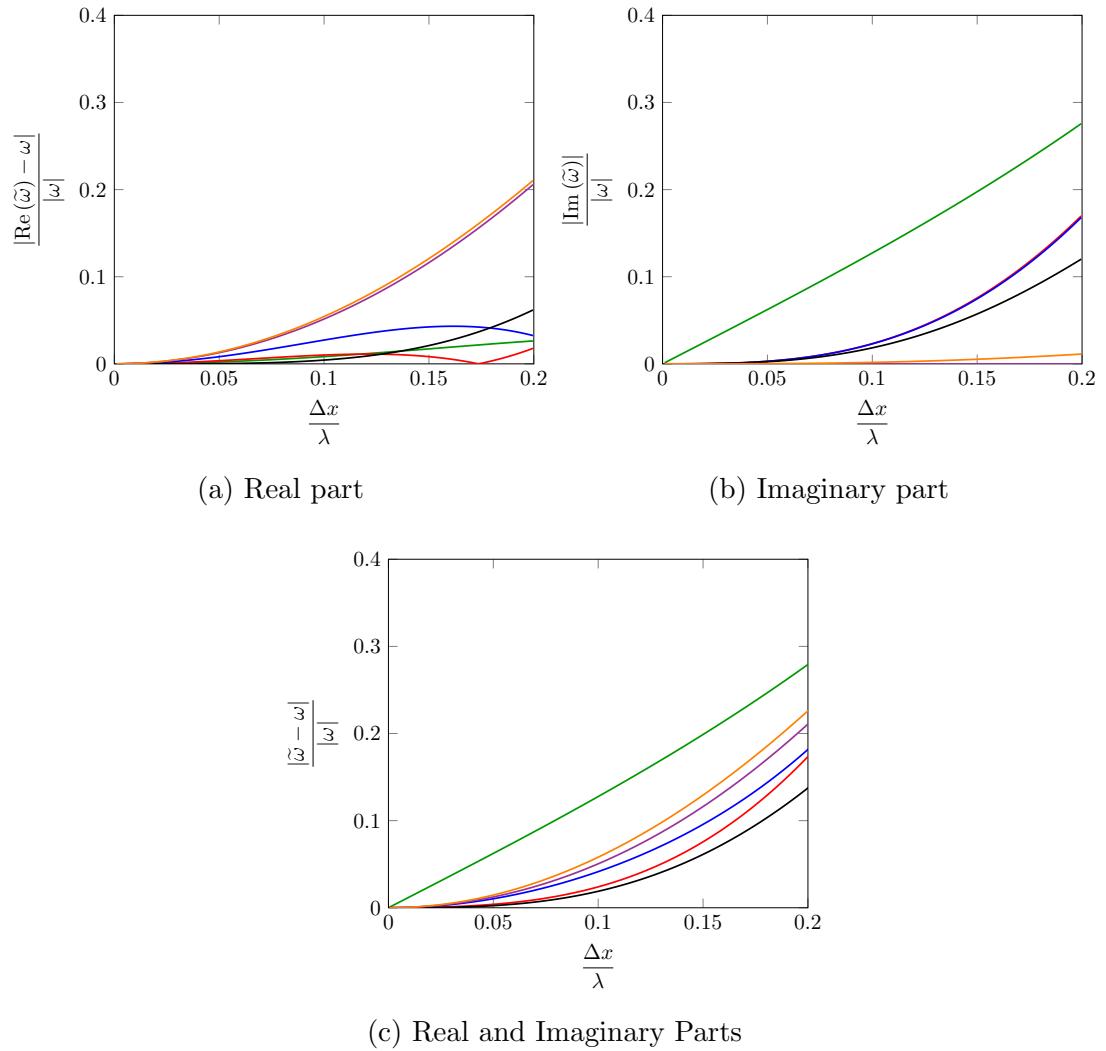


Figure 4.3: Relative dispersion error for first-order FDVM (—), second-order FDVM (—), second-order FEVM (—), third-order FDVM (—), \mathcal{D} (—) and \mathcal{W} (—). With $H = 1m$, $k = \frac{\pi}{10}$ and $U = 1m/s$.

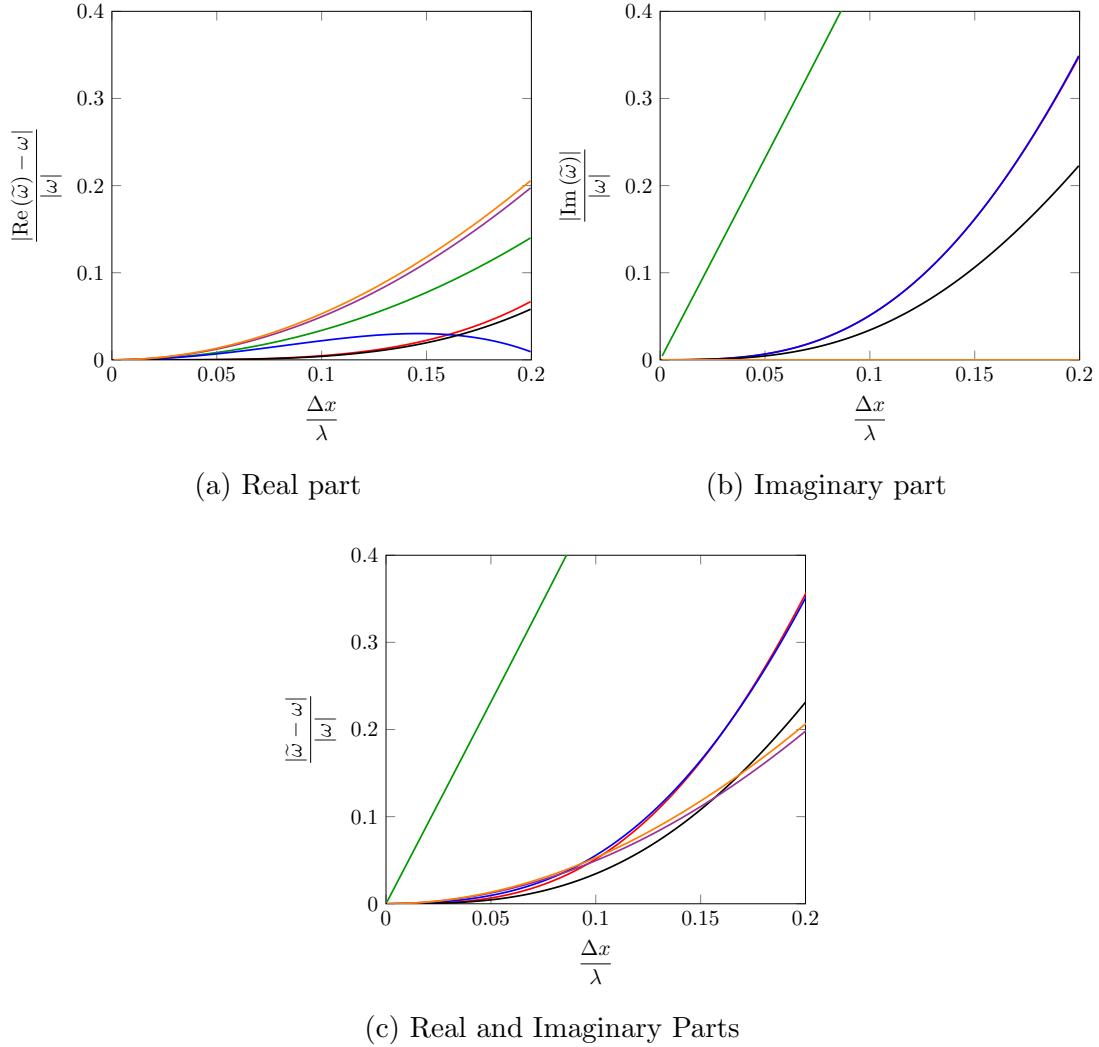


Figure 4.4: Relative dispersion error for first-order FDVM (—), second-order FDVM (—), second-order FEVM (—), third-order FDVM (—), \mathcal{D} (—) and \mathcal{W} (—). With $H = 1m$, $k = 2.5$ and $U = 0m/s$.

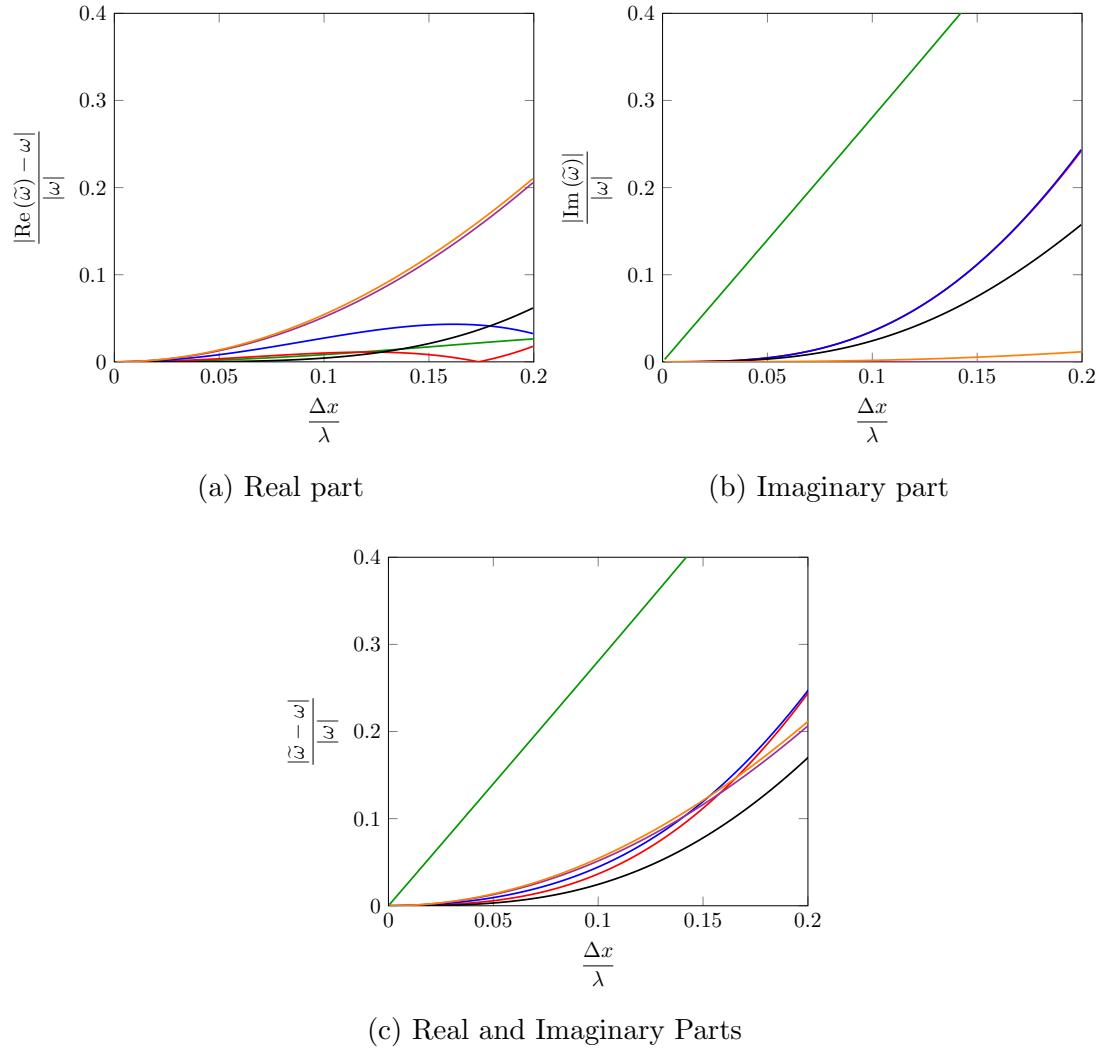


Figure 4.5: Relative dispersion error for first-order FDVM (—), second-order FDVM (—), second-order FEVM (—), third-order FDVM (—), \mathcal{D} (—) and \mathcal{W} (—). With $H = 1m$, $k = 2.5$ and $U = 1m/s$.

Scheme	Lowest Order Term of Error	
	Δx	Δt
FDVM ₁	$-\left(2\sqrt{gH} - \sqrt{\frac{3U}{\beta}}\right) \frac{ik^2}{4} \Delta x$	$\frac{i\omega_+^2}{2} \Delta t$
FDVM ₂	$\frac{2\beta U - 3\sqrt{3gH}\beta}{\beta^2} \frac{k^3}{24} \Delta x^2$	$-\frac{\omega_+^3}{6} \Delta t^2$
FEVM ₂	$\left(U + \frac{(42 + 15k^2 H^2)\sqrt{3gH}\beta}{20\beta^2}\right) \frac{k^3}{12} \Delta x^2$	$-\frac{\omega_+^3}{6} \Delta t^2$
FDVM ₃	$-(2\sqrt{gH} - \sqrt{3\beta}U) \frac{ik^4}{24} \Delta x^3$	$-\frac{i\omega_+^4}{24} \Delta t^3$

Table 4.9: Table showing lowest order error term for approximating ω_+ for all FDVM and the FEVM. With $-1 \leq Fr \leq 1$ and $\beta = 3 + H^2 k^2$.

Scheme	Lowest Order Δx Term of Error	
	$Fr < -1$	$Fr > 1$
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: Table showing the lowest order spatial error term for approximating ω_+ for all FDVM and the FEVM for supercritical Froude numbers where different from Table 4.9. With $\beta = 3 + H^2 k^2$.

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

Table 4.11: Table showing lowest order error term for approximating ω_+ for \mathcal{D} and \mathcal{W} .

Chapter 5

Numerical Validation

5.1 Measuring Convergence and Conservation

The numerical methods are assessed in this chapter by investigating their convergence and conservation properties. The convergence of these numerical methods is studied using analytic solutions to the governing equations and forced solutions to the forced Serre equations. While conservation is investigated by comparing the total of a conserved quantity in a numerical solution with the total 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 \mathbf{q} at the grid locations x_j and the corresponding numerical solution at those locations \mathbf{q}^* ; the L_1 norm is

$$L_1(\mathbf{q}, \mathbf{q}^*) = \begin{cases} \frac{\|\mathbf{q}^* - \mathbf{q}\|_1}{\|\mathbf{q}\|_1} & \|\mathbf{q}\|_1 > 0 \\ \|\mathbf{q}^*\|_1 & \|\mathbf{q}\|_1 = 0 \end{cases}. \quad (5.1)$$

5.1.2 Measures of Conservation

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

$$C_1(q, \mathbf{q}^*) = \begin{cases} \frac{|\mathcal{C}^*(\mathbf{q}^*) - \mathcal{C}(q(x, 0))|}{|\mathcal{C}(q(x, 0))|} & |\mathcal{C}(q(x, 0))| > 0 \\ |\mathcal{C}^*(\mathbf{q}^*)| & |\mathcal{C}(q(x, 0))| = 0 \end{cases}. \quad (5.2)$$

$\mathcal{C}^*(\mathbf{q}^*)$ 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 $\mathcal{C}^*(\mathbf{q}^*)$ 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 $\mathcal{C}^*(\mathbf{q}^*)$ with $\mathcal{C}^*(\mathbf{q}^0)$; where \mathbf{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

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

5.2 Analytic Solution for Horizontal Bed

To assess the ability of our numerical methods to solve the Serre equations with a horizontal bed (2.6) we use the solitary travelling wave solution (2.14) described in Chapter 2. This is a particular member of the family of periodic travelling wave solutions [13], but all these solutions except the trivial 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 (2.6) must be adequately approximated by the numerical method to properly reproduce the analytic solution. Therefore this analytic solution serves as a very good 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 (2.14) with $a_0 = 1m$, $a_0 = 0.7m$ 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 $k \in [6, \dots, 19]$. The CFL condition was satisfied with CFL number $Cr = 0.5$ by setting $\Delta t = Cr/\sqrt{g(a_0 + a_1)}$. For FDVM₂ and FEVM₂ $\theta = 1.2$ was used as the limiting parameter in the generalised minmod limiter (3.2).

For the parameters $a_0 = 1m$ and $a_0 = 0.7m$ the non-linearity parameter is $\epsilon = a_1/a_0 = 0.7$; this is large but beneath most of the well known breaking thresholds for water waves [27]. Because ϵ is large the nonlinear effects are large and therefore so are the 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 and the boundary, therefore the used Dirichlet boundary conditions were appropriate.

5.2.1 Results for Solitary Travelling Wave Solution

An example numerical solution with $\Delta x = 100/2^{11}m$ 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 FDVM₁ 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

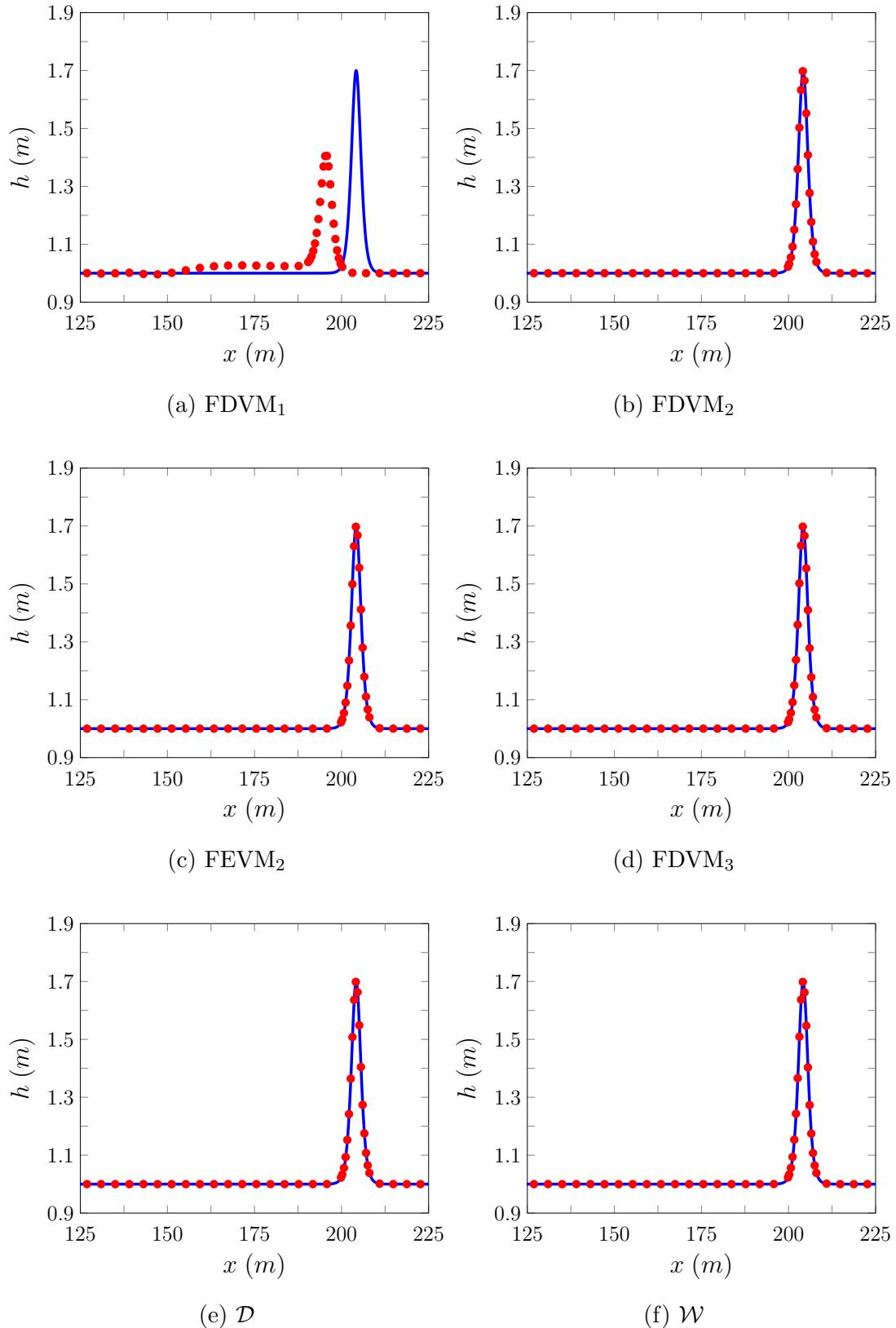


Figure 5.1: Comparison of the analytic solution (—) and numerical solution with $\Delta x = 100/2^{11}m$ (●) for the soliton problem at $t = 50s$ for all methods.

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 (2.5a) as compared to the evolution equation of G (2.5b); 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, FDVM₁ had the poorest reproduction of the analytic solution. However, the benefit of increasing the order of accuracy is significantly diminished for the the third-order accurate FDVM₃.

For the second-order methods we find that FDVM₂ consistently produces the smallest L_1 error followed by FEVM₂, \mathcal{W} and \mathcal{D} . The difference between the FDVM₂ and FEVM₂ is significant with errors of FEVM₂ being 2 to 4 times larger than FDVM₂. Therefore, FDVM₂ is reproducing the solitary wave solution more accurately than FEVM₂.

The finite difference methods produce very similar errors which are twice as large as the errors from FEVM₂. Additionally, the round-off effects which increase the L_1 error of the finite difference methods occur for larger Δx values than the hybrid finite volume methods. This demonstrates the superiority of the higher-order hybrid finite volume methods over the finite difference methods.

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. Given these results, only FEVM₂ and FDVM₂ 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 (2.2) and assess the employed well balancing method we compare various numerical solutions to the lake at rest analytic stationary solution (2.15).

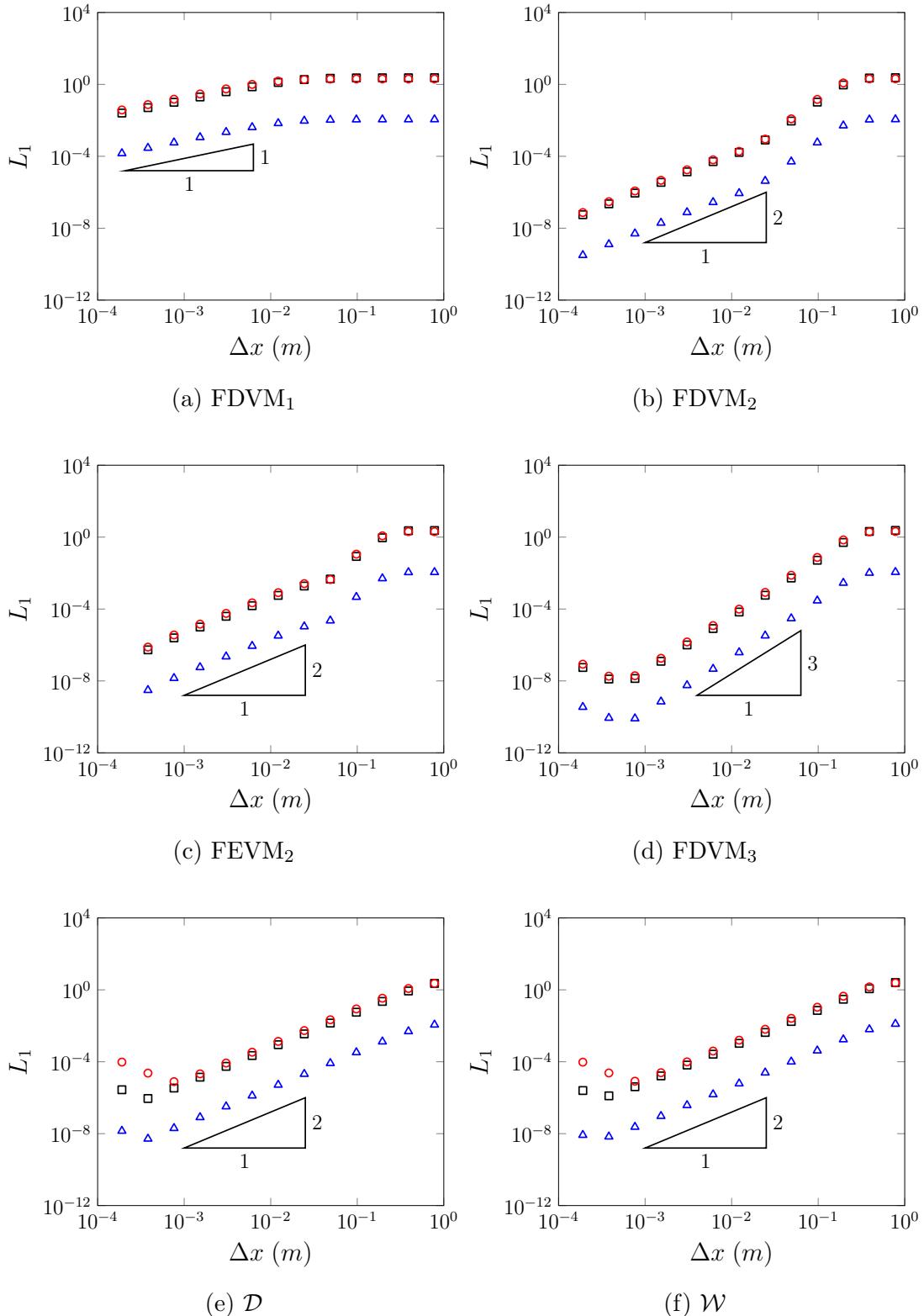


Figure 5.2: Convergence plots as measured by the L_1 norm for h (Δ), u (\square) and G (\circ) for the soliton problem for all methods.

The particular lake at rest solution (2.15) associated with the bed profile

$$b(x) = a_1 \sin(a_2 x) \quad (5.4)$$

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 $a_2 = 2\pi/50$ were chosen. These parameter values result in a periodic bed where lakes with flat surfaces submerge the troughs and the peaks are exposed.

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 $k \in [8, \dots, 17]$ and the CFL condition (3.28) was satisfied by having $\Delta t = Cr/\sqrt{g}$ with condition number $Cr = 0.5$. While the standard limiting parameter $\theta = 1.2$ was used in the generalised minmod limiter (3.2) for both FEVM₂ and FDVM₂. 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 FEVM₂ and FDVM₂ at some later time to the analytic solution, which are the initial conditions. To demonstrate the utility of the well balancing method the numerical solutions of two versions of FEVM₂ and FDVM₂ are presented; one where the well balancing method described in Chapter 3 is employed and one where no well balancing method is used.

5.3.1 Results for Lake at Rest

Example numerical solutions with $\Delta x = 100/2^{10}m$ at $t = 10s$ for all versions of FEVM₂ and FDVM₂ are given in Figure 5.3. The analytic solution and the numerical solution in these figures are indistinguishable at this scale as both methods have reproduced the analytic solution.

Examination of the L_1 errors depicted in Figure 5.4 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 the 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 in each cell for each time step; hence the second-order of the slopes. For methods without well balancing; the errors

are significantly larger although they are converging to the analytic solution. The methods without well balancing have 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 .

These results demonstrate the need for the well-balancing to be included in 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 comparisons 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 (2.16) can be determined and accounted for analytically, the only origin of error in the numerical solutions for the forced solutions are the numerical methods themselves and thus the theoretical second-order accuracy of FEVM₂ and FDVM₂ 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 solutions in both situations. To ensure that all terms of the Serre equations were accurately approximated in the numerical method the functions

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

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

$$b^*(x) = a_6 \sin(a_7 x) \quad (5.5c)$$

for the primitive variables were chosen. These functions produce a 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 $a_1 = 0.5$, $a_2 = 2\pi/(10a_7)$, $a_3 = -3\pi/(2a_7)$, $a_4 = \pi/(16a_7)$, $a_5 = 0.5$, $a_6 = 1.0$ and $a_7 = \pi/25$ with $a_0 = 1$ for the finite

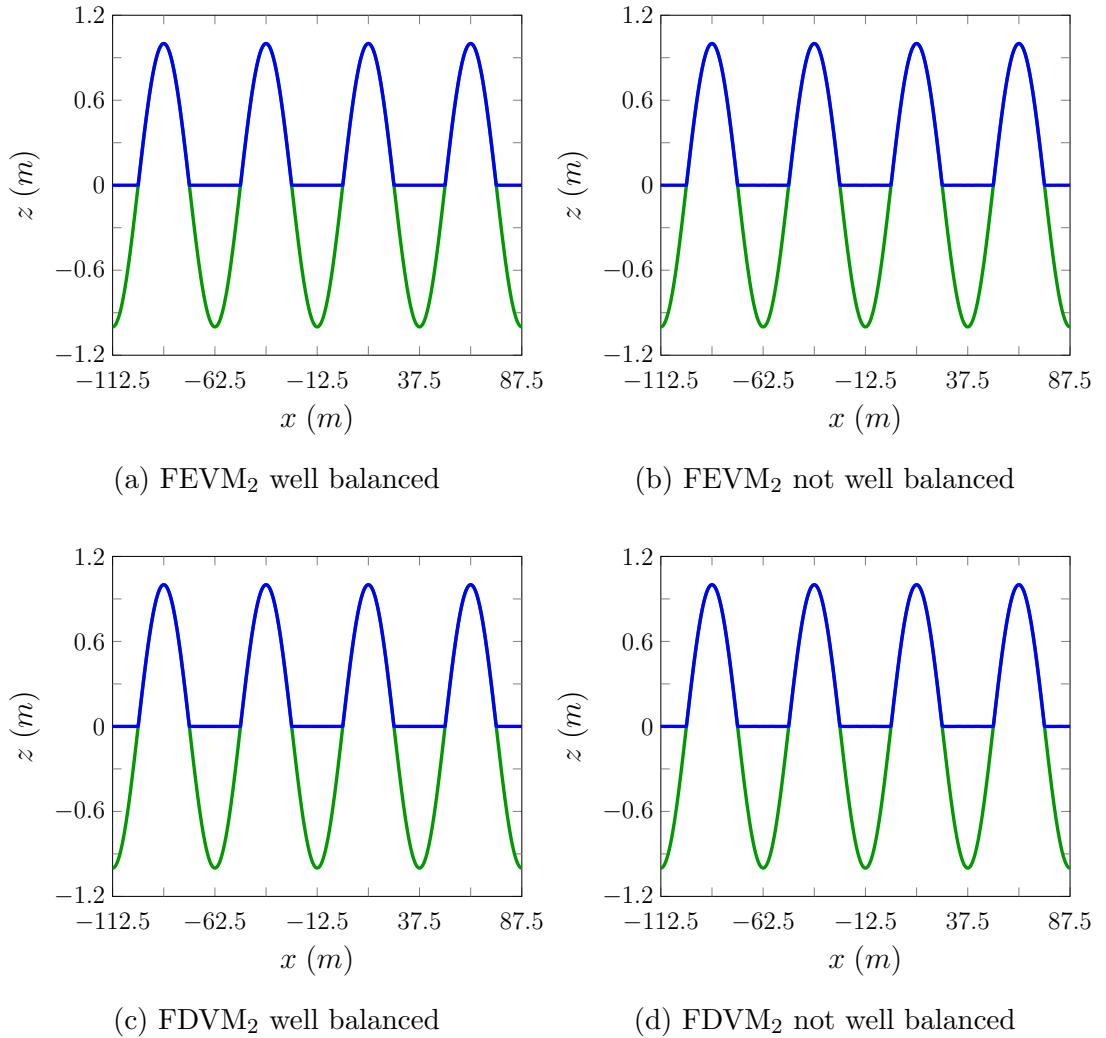


Figure 5.3: Comparison of the analytic solution (—) and numerical solution with $\Delta x = 100/2^{10} \text{ m}$ (—) for the lake at rest problem at $t = 10\text{s}$ for all methods.

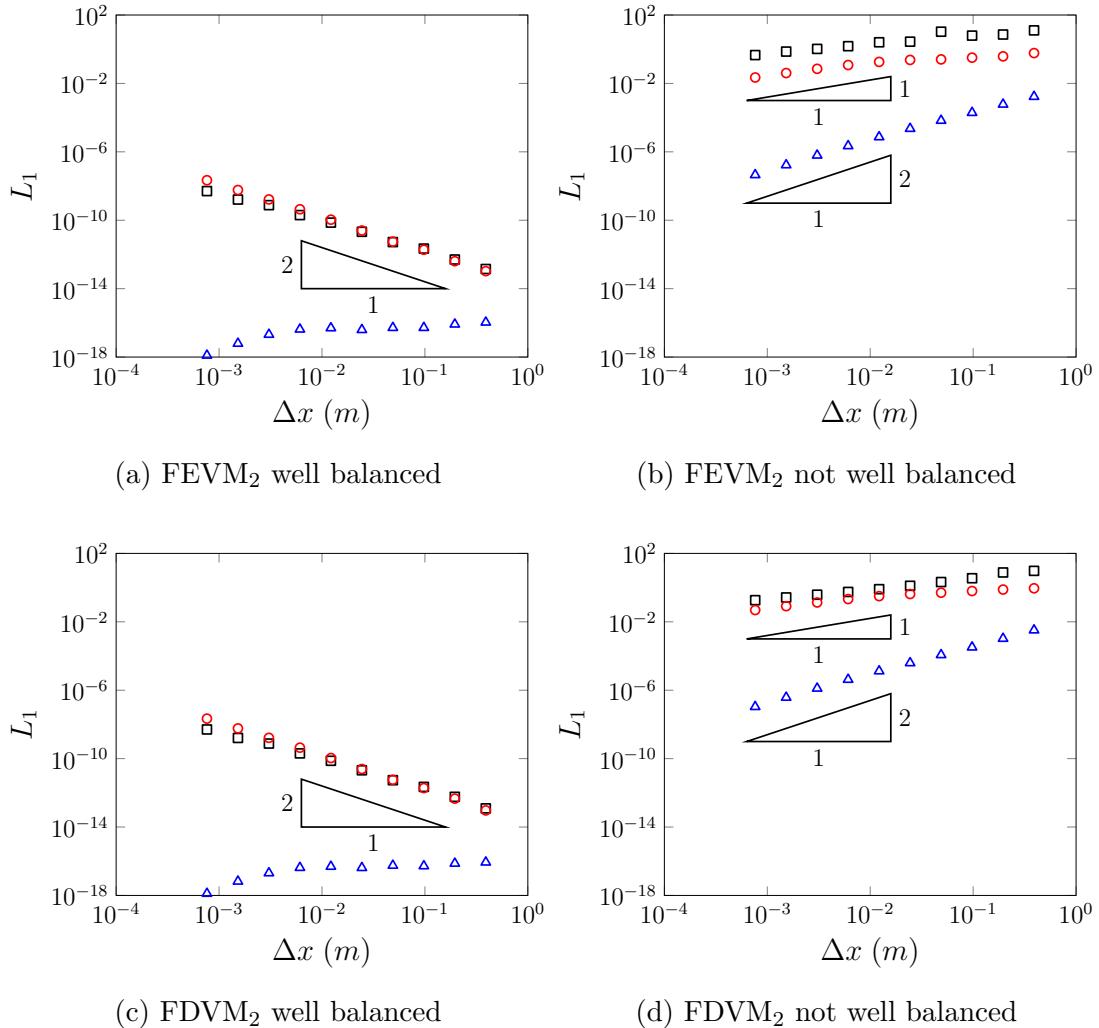


Figure 5.4: Convergence plots as measured by the L_1 norm for h (Δ), u (\square) and G (\diamond) for the lake at rest problem at $t = 10s$ for all methods.

water depth forced solution and $a_0 = 0$ 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 $k \in [8, \dots, 16]$. To satisfy the CFL condition (3.28) the temporal resolution $\Delta t = Cr / (a_2 + a_5 + \sqrt{g(a_0 + a_1)})$ was chosen with condition number $Cr = 0.5$. The value $\theta = 1.2$ was used in the generalised minmod limiter (3.2) for both FEVM₂ and FDVM₂ and Dirichlet boundary conditions were applied at the domain edges.

5.4.1 Results for Finite Water Depth

For the finite water depth case $a_0 = 1m$ was chosen in equation (5.5a). An example of the numerical solutions of FEVM₂ and FDVM₂ are given in Figures 5.5 and 5.6 for $\Delta x = 100/2^{10}m$ 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 , uh and G for the FEVM₂ and FDVM₂ are given in Figure 5.7. Both methods recover the expected second-order accuracy. Since the source term of the modified Serre equations is added analytically and all terms must be accurately approximated by the methods 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 FEVM₂ and FDVM₂.

Example numerical solutions demonstrating the evolution of the wave is given in Figures 5.8 and 5.9 for FEVM₂ and FDVM₂ with $\Delta x = 100/2^{10}m$ 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

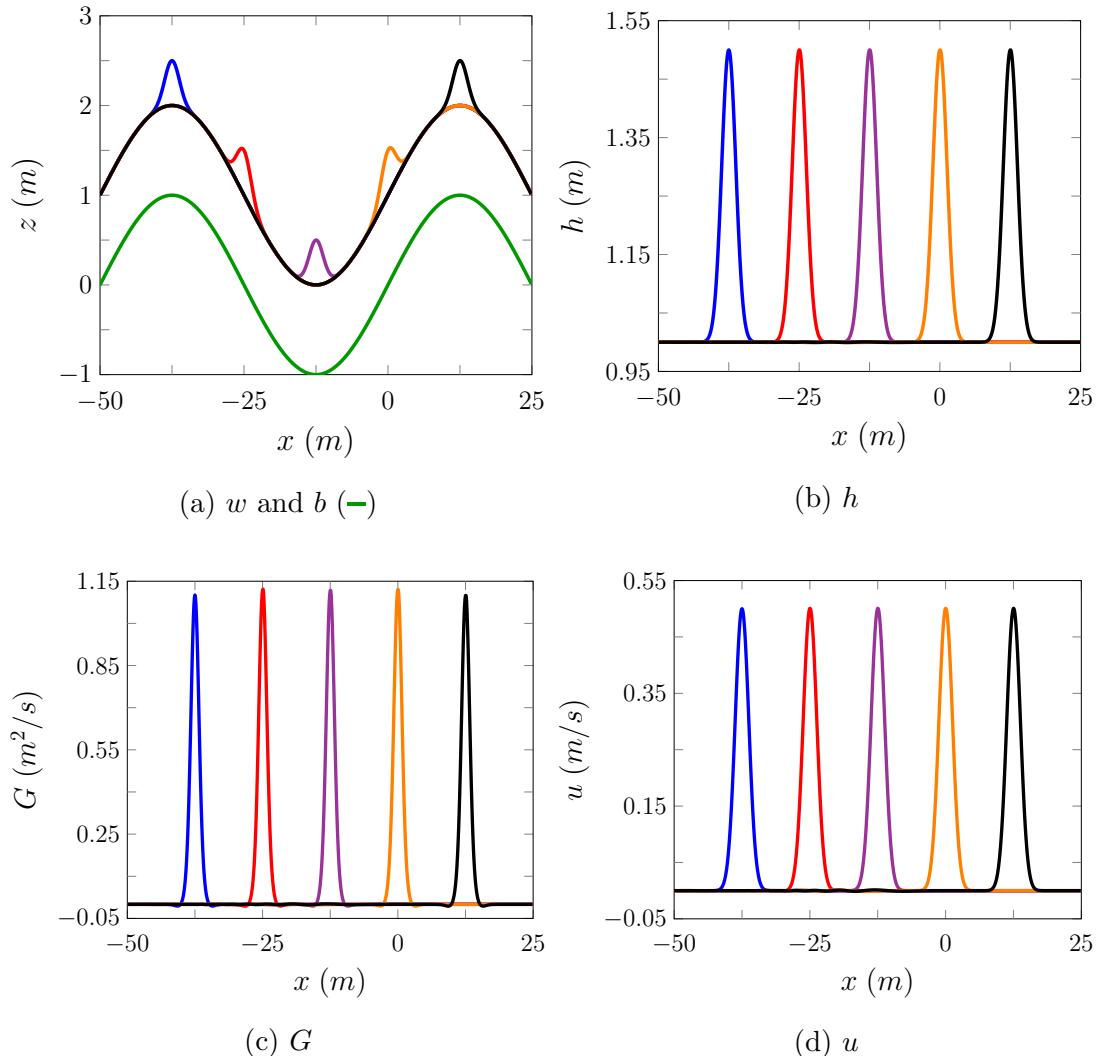


Figure 5.5: Plots of various quantities for the FEVM numerical solution with $\Delta x = 100/2^{10}m$ at $0s$ (---), $2.5s$ (---), $5.0s$ (---), $7.5s$ (---), $10.0s$ (---) of the forced Serre equations for the finite water test.

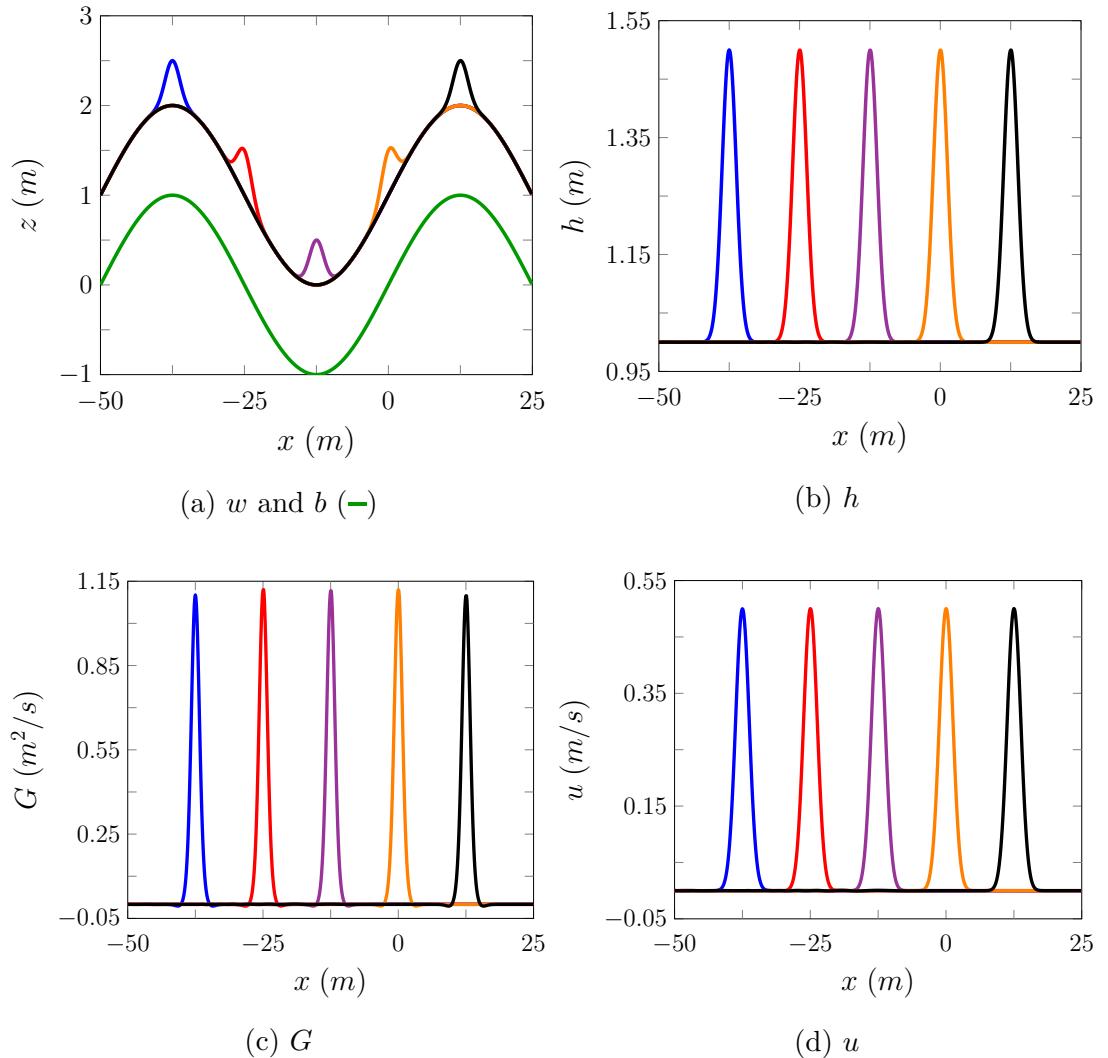


Figure 5.6: Plots of various quantities for the FDVM numerical solution with $\Delta x = 100/2^{10}m$ at 0s (—), 2.5s (—), 5.0s (—), 7.5s (—), 10.0s (—) of the forced Serre equations for the finite water test.

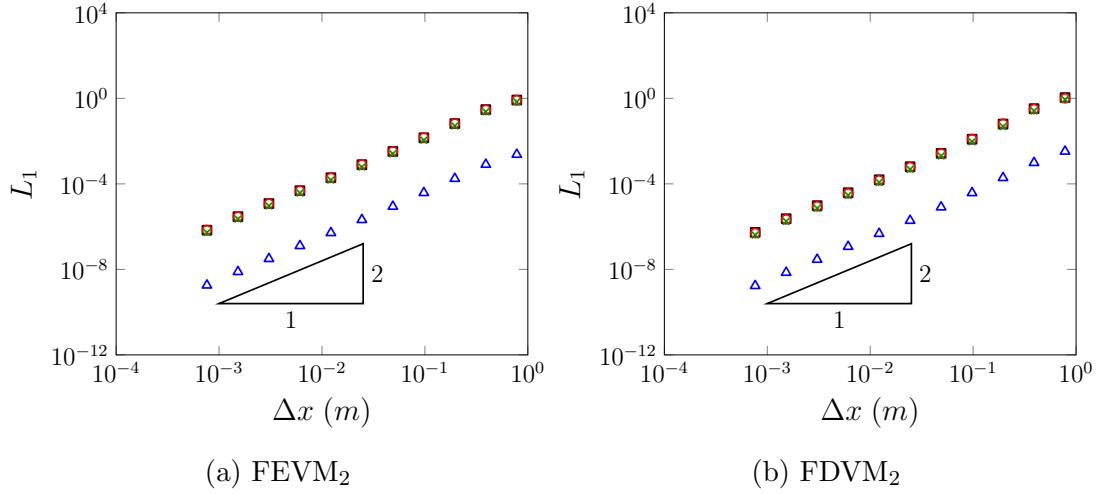


Figure 5.7: Convergence plots as measured by the L_1 norm for h (\triangle), u (\square), uh (\times) and G (\circ) for the finite water forced solution problem for FEVM and FDVM at $t = 10s$.

to the elliptic solver (3.14). By choosing larger values of these quantities the errors in u can be significantly damped. However, this leads to convergence plots that are second-order accurate in all the terms but stop converging when the L_1 errors become dominated by the artificial h_{base} and h_{tol} terms. For our purposes smaller h_{base} and h_{tol} are preferable when demonstrating convergence of the conserved quantities for small grids, resulting in large errors in u when h is small.

The L_1 errors for both methods are given in Figure 5.10. Both methods possess second-order convergence in all the quantities except u . Because in conservation law form (2.5) all the flux and source terms only depend on u multiplied by some power of h ; 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 $h > 10^{-3}$ as in Figure 5.11, we recover the expected second-order accuracy in all quantities.

Therefore these methods 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.

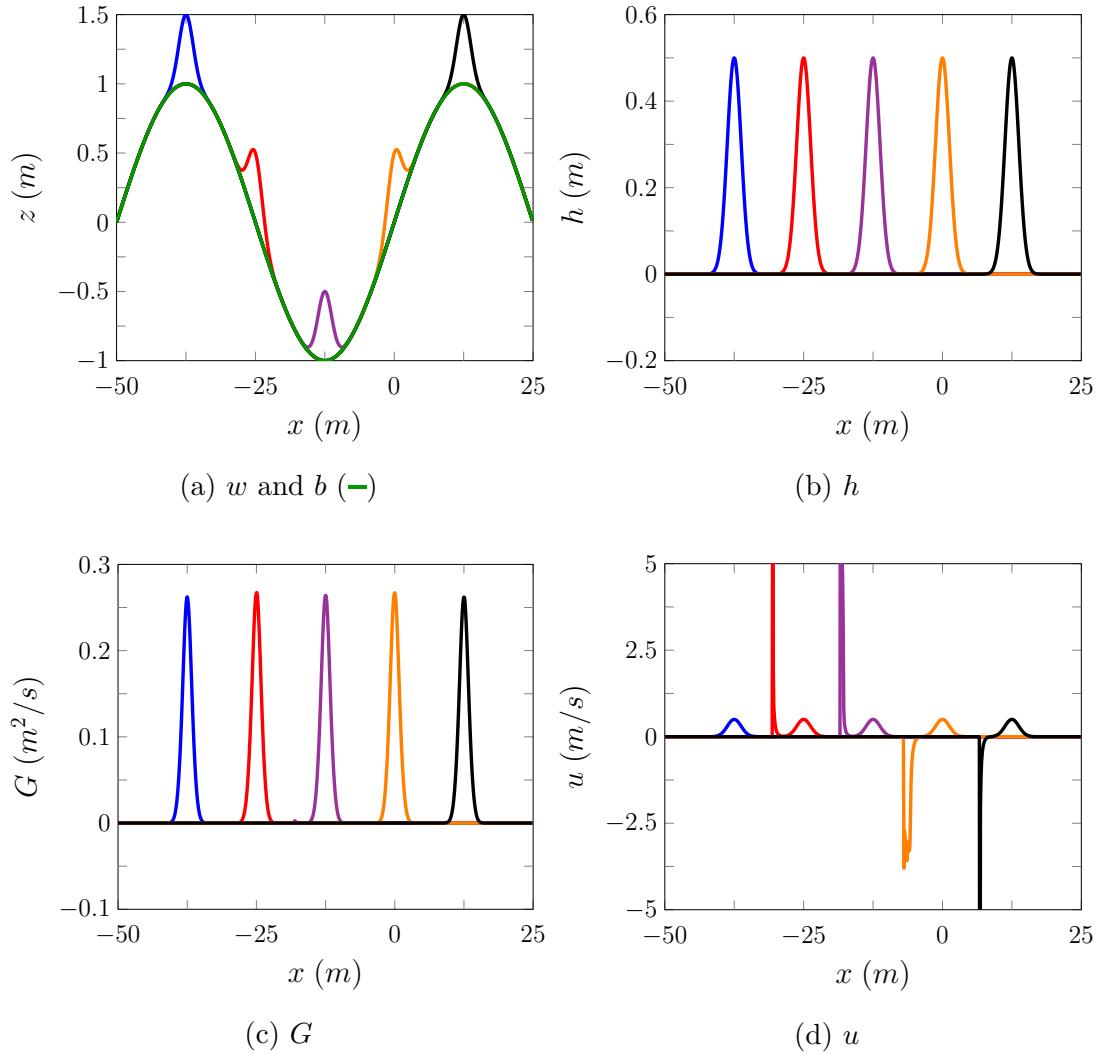


Figure 5.8: Plots of various quantities for the FEVM numerical solution at 0s (—), 2.5s (—), 5.0s (—), 7.5s (—), 10.0s (—) of the dry bed forced solution problem.

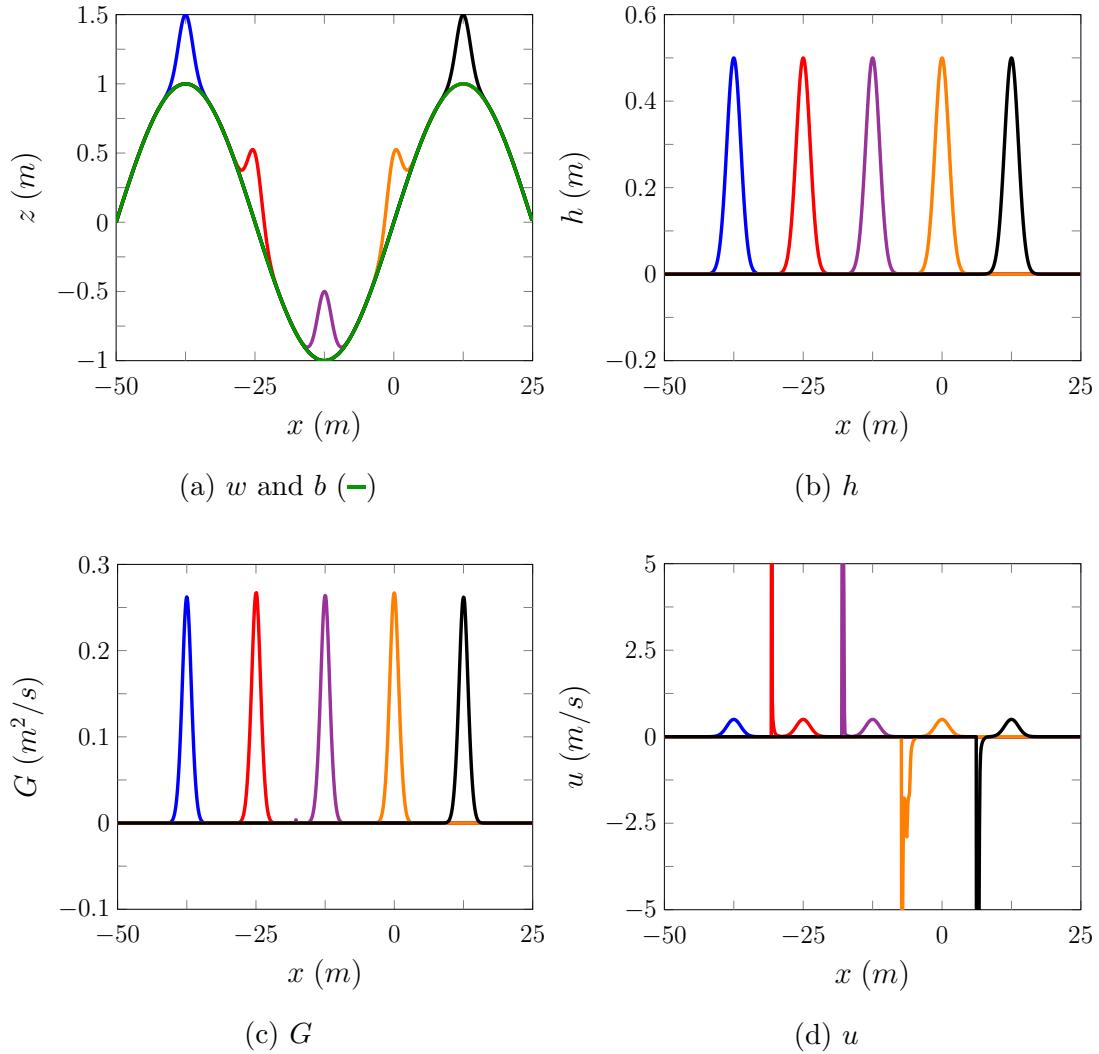


Figure 5.9: Plots of various quantities for the FDVM numerical solution at 0s (---), 2.5s (---), 5.0s (---), 7.5s (---), 10.0s (---) of the dry bed forced solution problem.

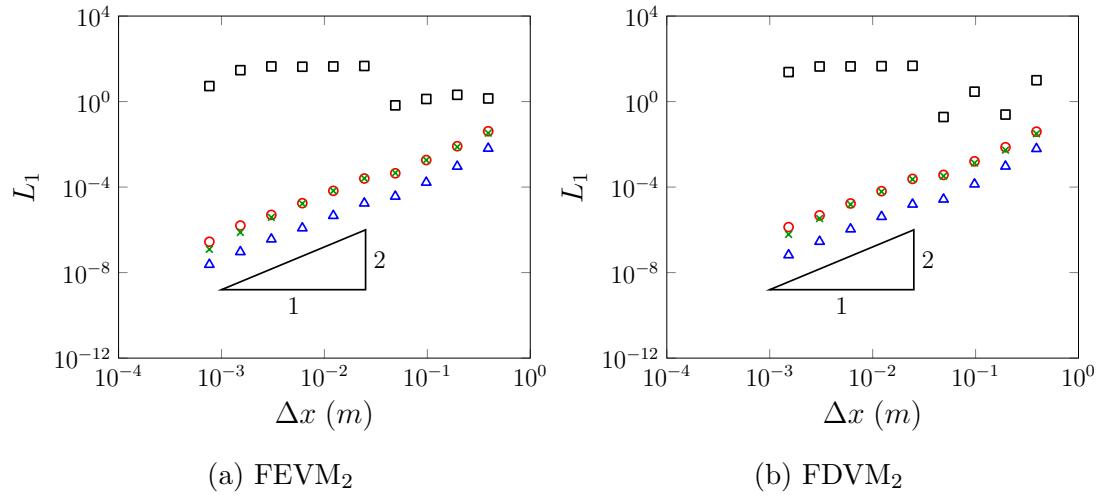


Figure 5.10: Convergence plots as measured by the L_1 norm for h (Δ), u (\square), uh (\times) and G (\circ) for the dry bed forced solution problem for FEVM and FDVM at $t = 10s$.

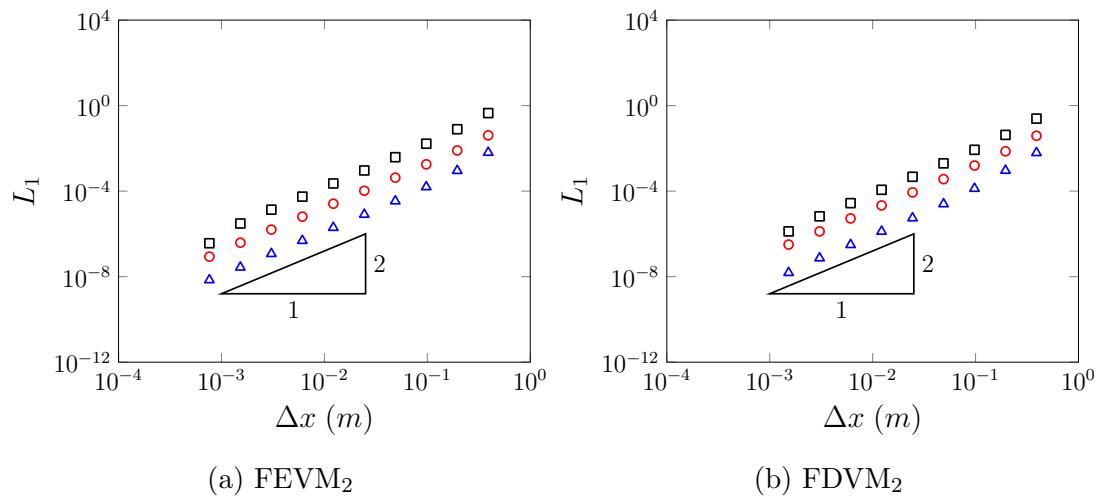


Figure 5.11: Convergence plots for regions where $h > 10^{-3}m$ as measured by the L_1 norm for h (Δ), u (\square) and G (\circ) for the dry bed forced solution problem for FEVM and FDVM at $t = 10s$.

Chapter 6

Experimental Validation

6.1 Evolution of Rectangular Depression

A series of experiments studying the evolution of rectangular depressions was conducted by Hammack and Segur [28]. 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 water in the wave tank is $0.1m$ deep and initially still with the piston in the up position; the experiment begins when the piston suddenly moves down. This creates a sudden depression, generating waves that are recorded at wave gauges located $0m$, $5m$, $10m$, $15m$ and $20m$ from the right edge of the piston. A diagram of the wavetank with the wave gauge locations marked is given in Figure 6.1.

These experiments provide a good benchmark for the capability of the numerical method to accurately model problems with steep gradients in the free surface. However, since the high frequency waves that are supposed to be present in the oscillatory wave trains are very sensitive to smoothing of the initial conditions and diffusion [15] it is expected that solutions of the Serre equations should produce many more oscillations in the wave train than could be observed experimentally. Hammack and Segur [28] 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. When the nonlinearity is higher more of these high frequency waves in the dispersive wave train are observed, however, still not as many as are produced in the solutions of the Serre equations.

This experiment was modelled numerically using the reflected problem, with the wall as the axis of symmetry. In the numerical experiments the domain is

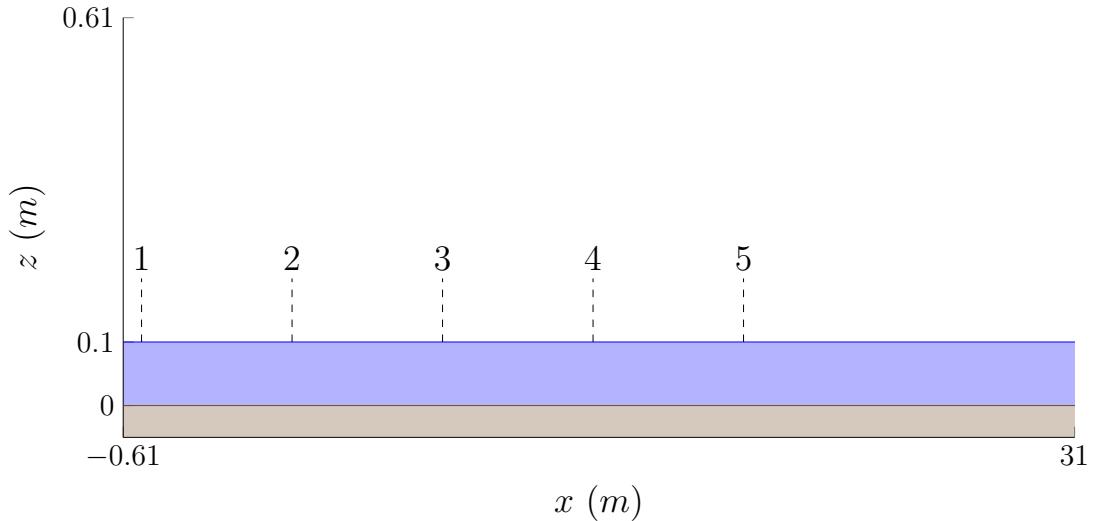


Figure 6.1: Diagram demonstrating the water (■) and the bed (□) for the Segur experiments, with the wave gauge locations marked.

$[-60m, 60m]$ and the experiment is run for $50s$ with $g = 9.81m/s^2$. For the spatial resolution we set $\Delta x = 0.01m$ and satisfy the CFL condition (3.28) by setting $\Delta t = 0.5/\sqrt{g \cdot 0.1}$. The limiting parameter $\theta = 1.2$ was used in the reconstruction for FEVM₂ and FDVM₂.

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 6.3 for FEVM₂ and FDVM₂ respectively. 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 the initial triangular wave and 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 friction in the Serre equations (2.5). 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; which significantly attenuates the higher frequency waves in the generated dis-

Quantity	$\mathcal{C}^*(\mathbf{q}^0)$	$\mathcal{C}^*(\mathbf{q}^*)$	$\mathcal{C}_1^*(\mathbf{q}^0, \mathbf{q}^*)$
Mass	11.9888	11.9888	0
Momentum	0	7.44×10^{-18}	7.44×10^{-18}
G	0	1.56×10^{-18}	1.56×10^{-18}
Energy	5.8751204081	5.8750869556	5.70×10^{-6}

Table 6.1: Total amounts and error in conservation for all quantities for FEVM₂ numerical solution of the 0.01m rectangular depression.

Quantity	$\mathcal{C}^*(\mathbf{q}^0)$	$\mathcal{C}^*(\mathbf{q}^*)$	$\mathcal{C}_1^*(\mathbf{q}^0, \mathbf{q}^*)$
Mass	11.9888	11.9888	0
Momentum	0	-1.19×10^{-17}	-1.19×10^{-17}
G	0	-8.05×10^{-18}	-8.05×10^{-18}
Energy	5.8751204081	5.87508358202	6.27×10^{-6}

Table 6.2: Total amounts and error in conservation for all quantities for FDVM₂ numerical solution of the 0.01m rectangular depression.

persive wavetrain [15]. Given these limitations the numerical methods do a very good job of replicating the experimental behaviour.

Both FEVM₂ and FDVM₂ have produced visually identical results at this scale and have demonstrated very good conservation of all the quantities in Tables 6.1 and 6.2. Given the extensive review of these methods [15] for steep gradient problems, this indicates that these solutions are indicative of true solutions of the Serre equations which agree well with the 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 6.4 and 6.5 for FEVM₂ and FDVM₂ respectively. The conservation of all the conserved quantities are given in Tables 6.3 and 6.3 for FEVM₂ and FDVM₂ 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, because there is more energy and the piston has to move further to generate the wave the effects that damp

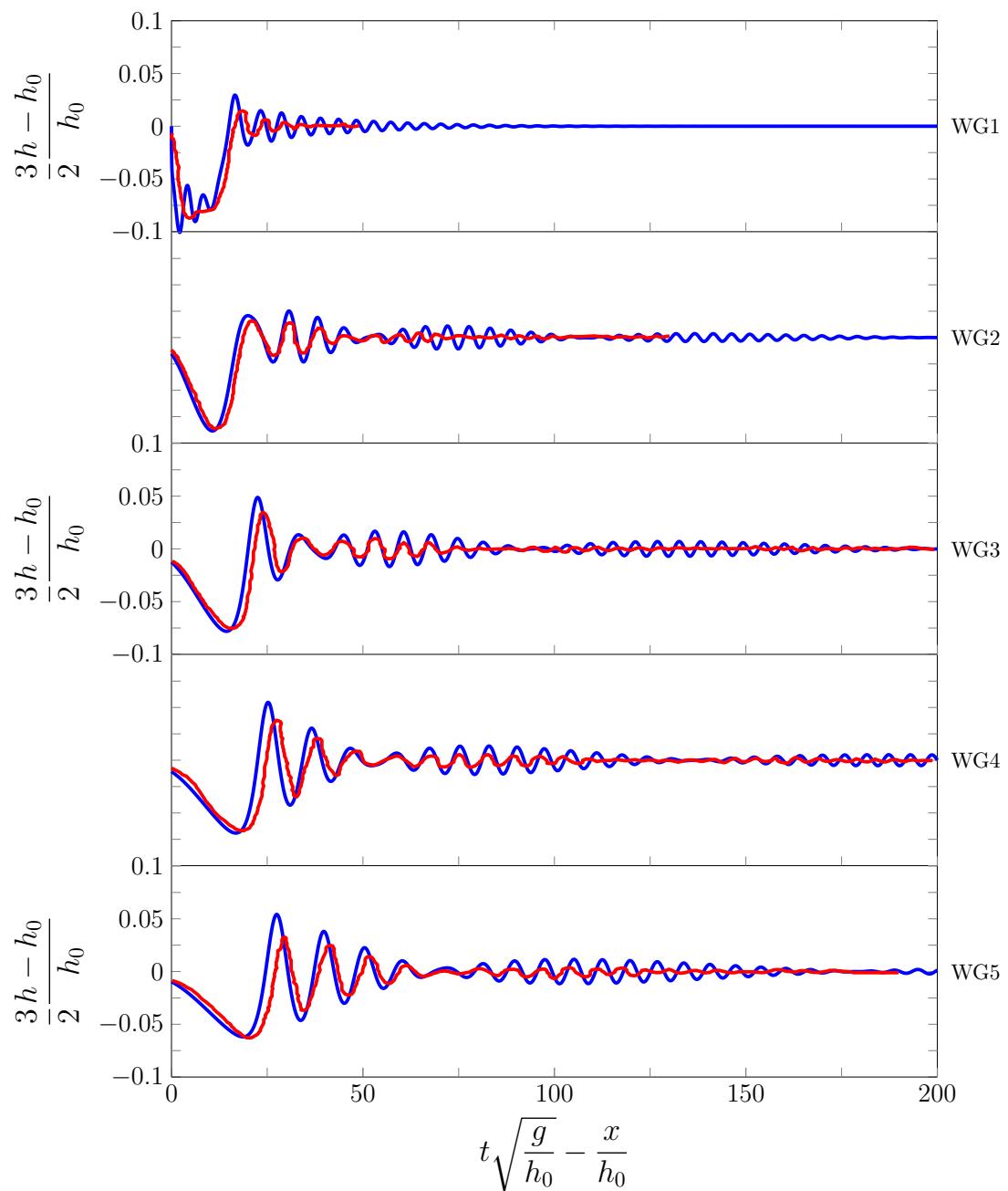


Figure 6.2: FEVM

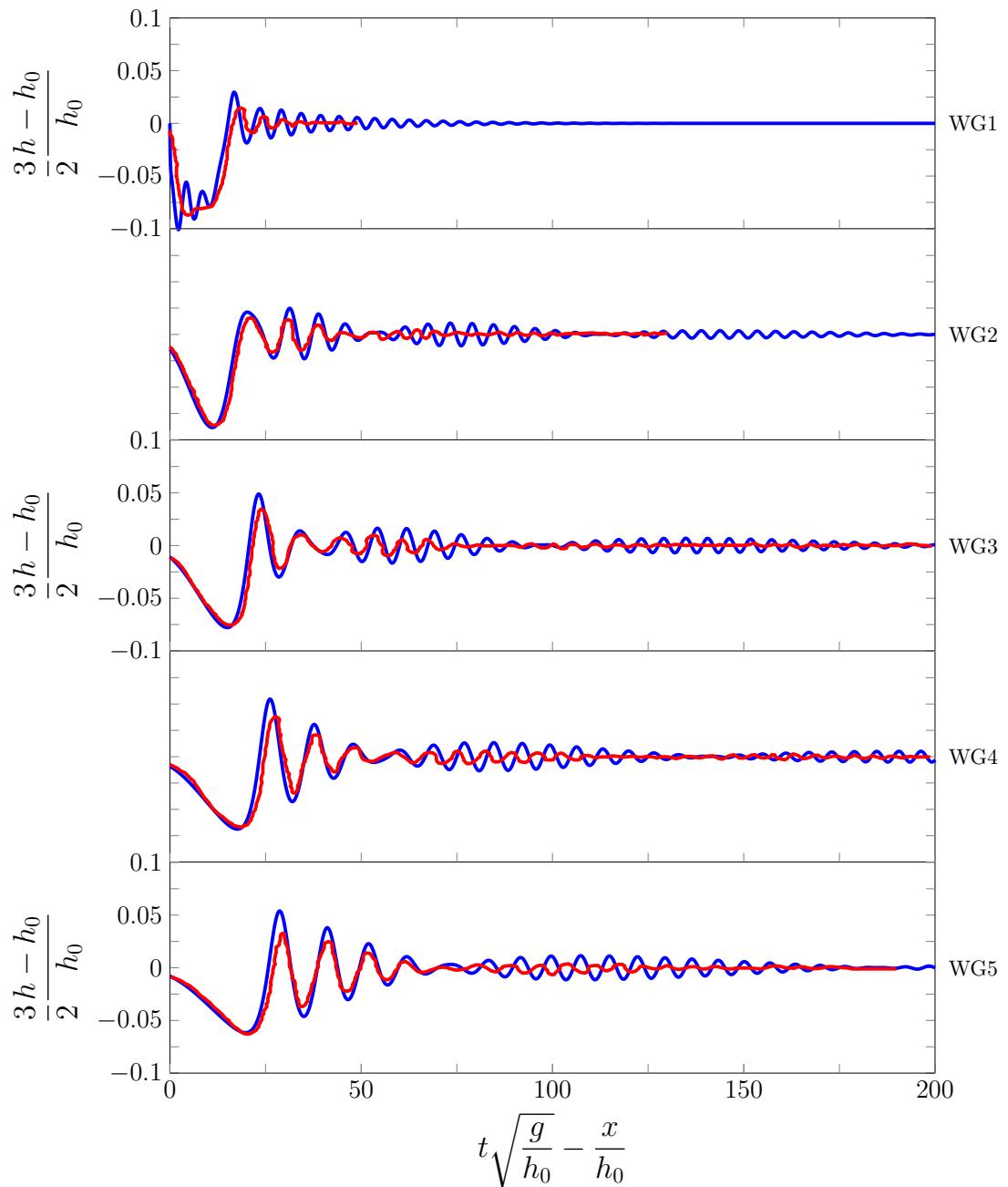


Figure 6.3: FDVM

Quantity	$\mathcal{C}^*(\mathbf{q}^0)$	$\mathcal{C}^*(\mathbf{q}^*)$	$\mathcal{C}_1^*(\mathbf{q}^0, \mathbf{q}^*)$
Mass	11.9644	11.9644	0
Momentum	0	-7.75×10^{-17}	-7.75×10^{-17}
G	0	-3.33×10^{-16}	-3.33×10^{-16}
Energy	5.85596887293	5.85524264766	1.24×10^{-4}

Table 6.3: Total amounts and error in conservation for all quantities for FEVM₂ numerical solution of the 0.03m rectangular depression.

Quantity	$\mathcal{C}^*(\mathbf{q}^0)$	$\mathcal{C}^*(\mathbf{q}^*)$	$\mathcal{C}_1^*(\mathbf{q}^0, \mathbf{q}^*)$
Mass	11.9644	11.9644	0
Momentum	0	-9.09×10^{-17}	-9.09×10^{-17}
G	0	-1.16×10^{-16}	-1.16×10^{-16}
Energy	5.85596887293	5.85520859829	1.30×10^{-4}

Table 6.4: Total amounts and error in conservation for all quantities for FDVM₂ numerical solution of the 0.03m rectangular depression.

the oscillations in the experiments are more significant leading to a much larger difference between the numerical and experimental results; particularly for the amplitude and speed of the 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 represent 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 that the accuracy of the numerical methods in the presence of steep gradients in the free surface.

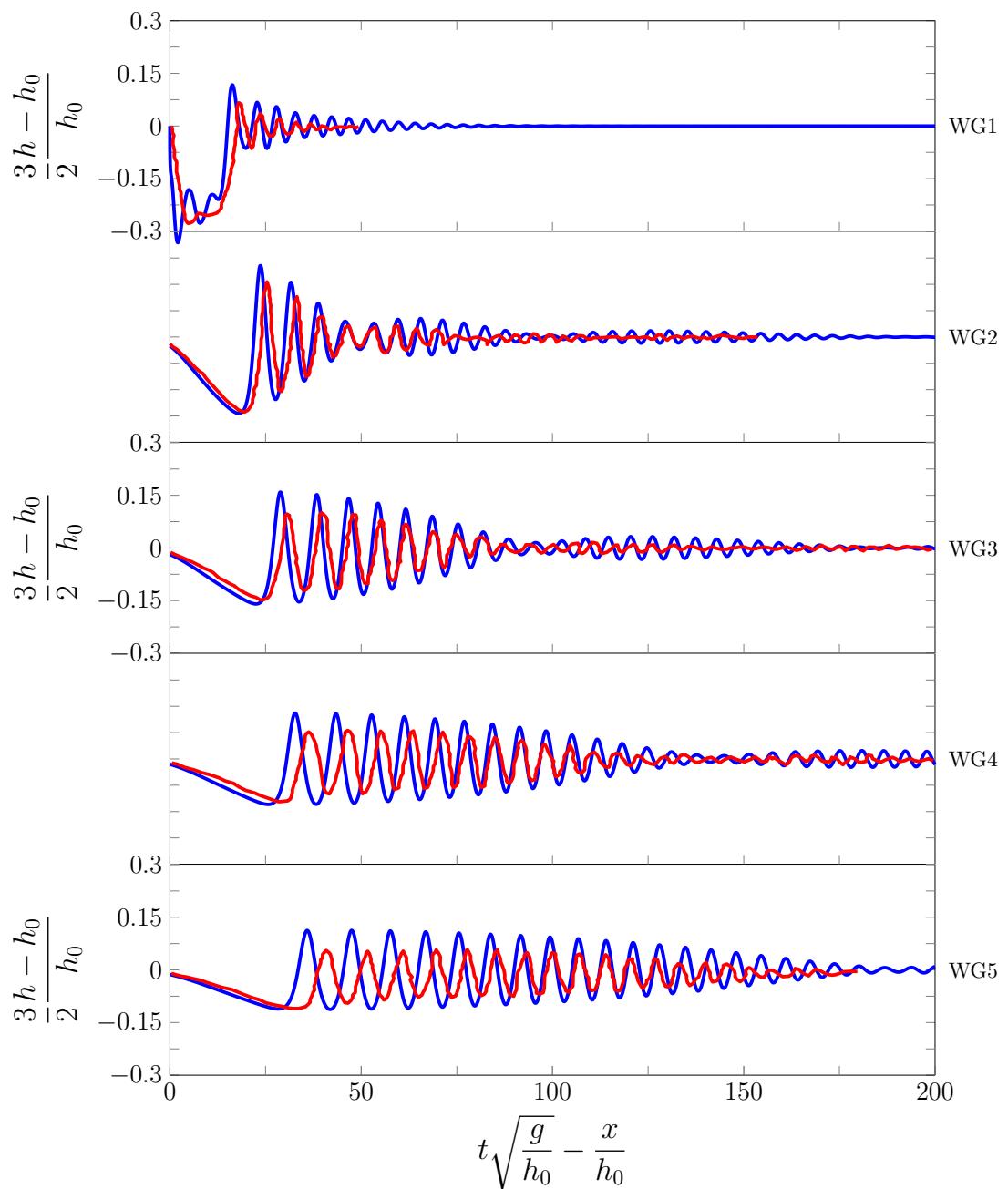


Figure 6.4: FEVM

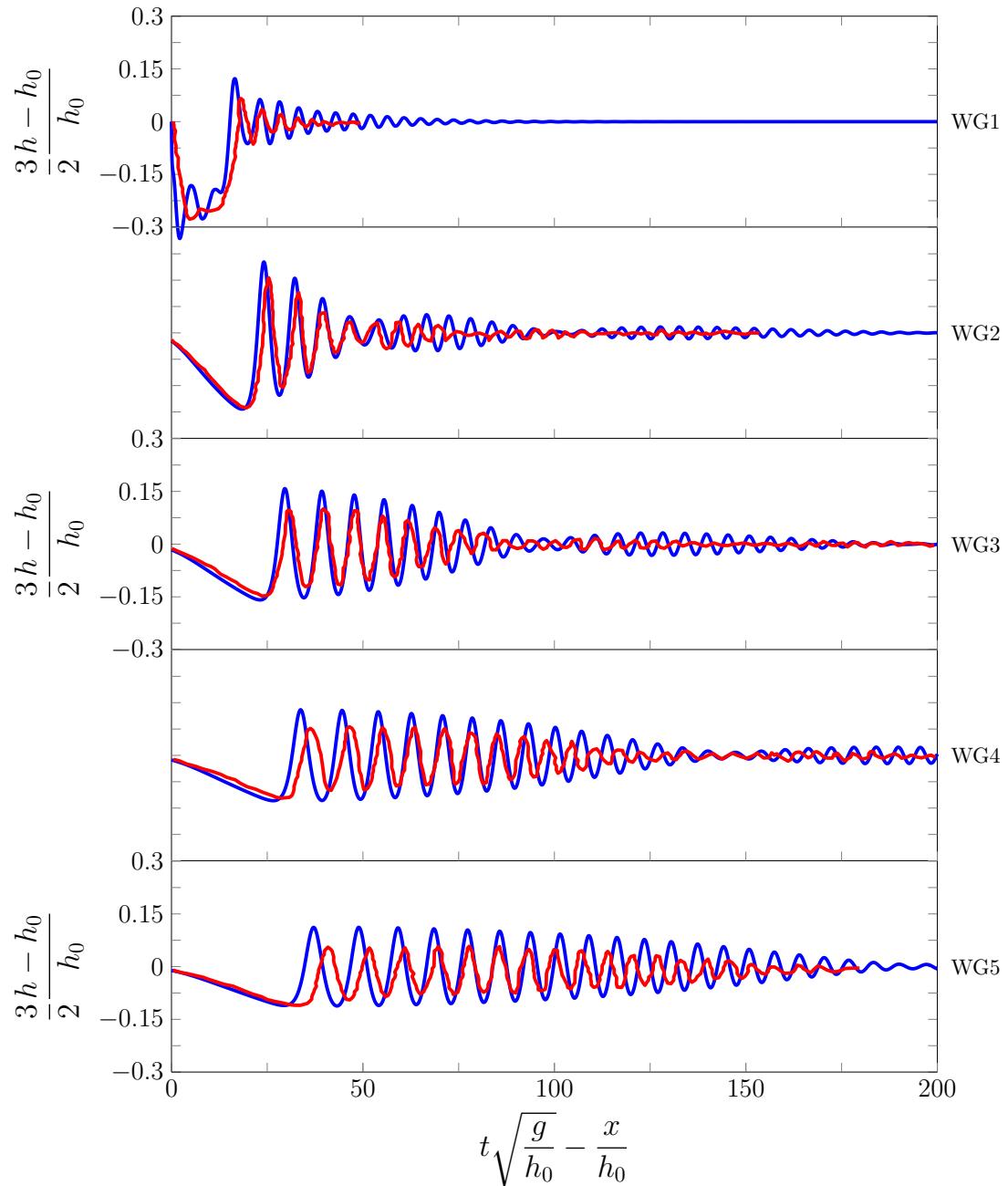


Figure 6.5: FDVM

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 [29, 30]. 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 good benchmark for our numerical schemes abilities to accurately model both the interaction of variable bathymetry and dispersive waves. For our purposes we will focus on the monochromatic wave experiments of Beji and Battjes [30].

The experiments of Beji and Battjes [30] were conducted in a wave tank 37.7m long, 0.8m wide and 0.75m high. A diagram of the longitudinal section of the wave tank is given in Figure 6.6. 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 to the right, over the submerged trapezoidal bar towards a wave absorbing sloped 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 non-linearity parameter $\epsilon = 0.01/0.4 = 0.025$.

We numerically simulated these experiments over the spatial domain [5.7m, 150m] with $\Delta x = 0.1/2^4 m$ and $\Delta t = Sp/2^5$ where $Sp = 0.039$ is the experimental sampling period. These Δx and Δt values satisfy the CFL condition (3.28) for these experiments. 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 boundary effects in our results.

To simulate the incoming waves at the upstream boundary we used the first wave gauge as our left boundary condition and used linear extrapolation to calculate the other required h values in the left ghost cell. The velocity boundary conditions were calculated from the height values by solving the continuity equation (2.2a) assuming u and h are travelling wave solutions

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

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

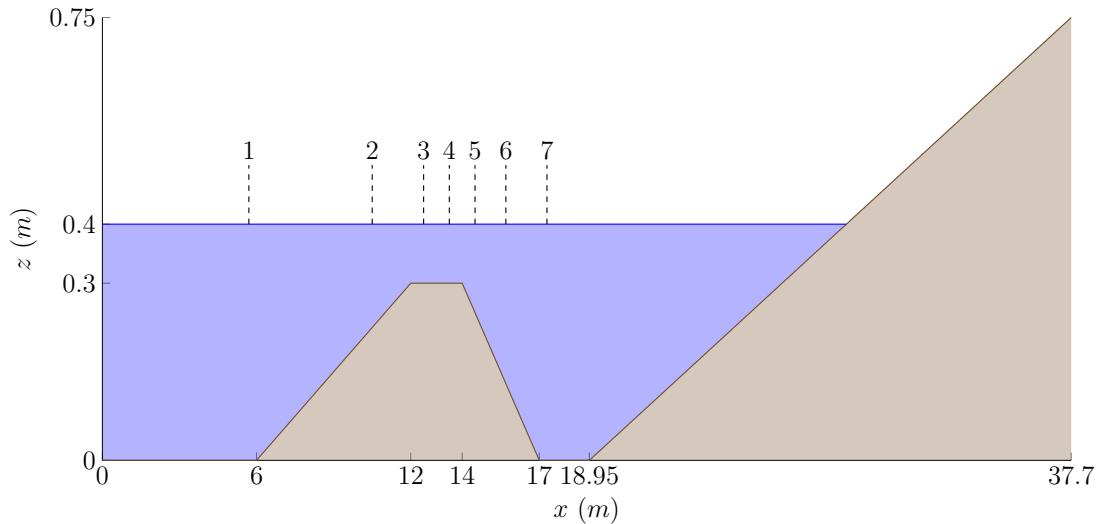


Figure 6.6: Diagram demonstrating the water (■) and the ground (■) for the Beji experiments, with the wave gauge locations marked.

high frequency experiments.

6.2.1 Low Frequency Results

A comparison of wave heights of the experimental and numerical results are located in Figures 6.7 and 6.8 for FEVM₂ and Figures 6.9 and 6.10 for FDVM₂. 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. This validates the numerical schemes for simulating shoaling of dispersion 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 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 water waves, as the numerical solutions of improved dispersion equations [30, 31] accurately reproduce the experimental results on the leeward side.

The dispersion of the Serre equations is vital to recreating the experimental results for wave gauges 2 to 5, as non-dispersive equations such as the SWWE do a

very poor job at simulating this experiment [10]. However, to properly reproduce the experimental results on the leeward side of the slope at wave gauges 6 and 7 would require improving the dispersion characteristics of the underlying Serre equations as done by Barthélemy [12]. Such an improvement can be incorporated into the hybrid FDVM and FEVM numerical methods [5] but is beyond the scope of this thesis.

6.2.2 High Frequency Results

The wave heights of the experimental and numerical results are given in Figures 6.11 and 6.12 for FEVM_2 . While the results for FDVM_2 are given in Figures 6.13 and 6.14. As for the low frequency experiment, these numerical schemes 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 improving the dispersion properties of the governing partial differential equations lead to a much better agreement between the numerical and experimental results [30, 31]. Because the difference between the dispersion relation of the Serre equations and water waves is largest for higher frequency and therefore for shorter waves [12] the earlier divergence between experimental and numerical results is expected.

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 [30, 31, 7, 32]. 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 [30].

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 [33]. These experiments were performed in a wave tank

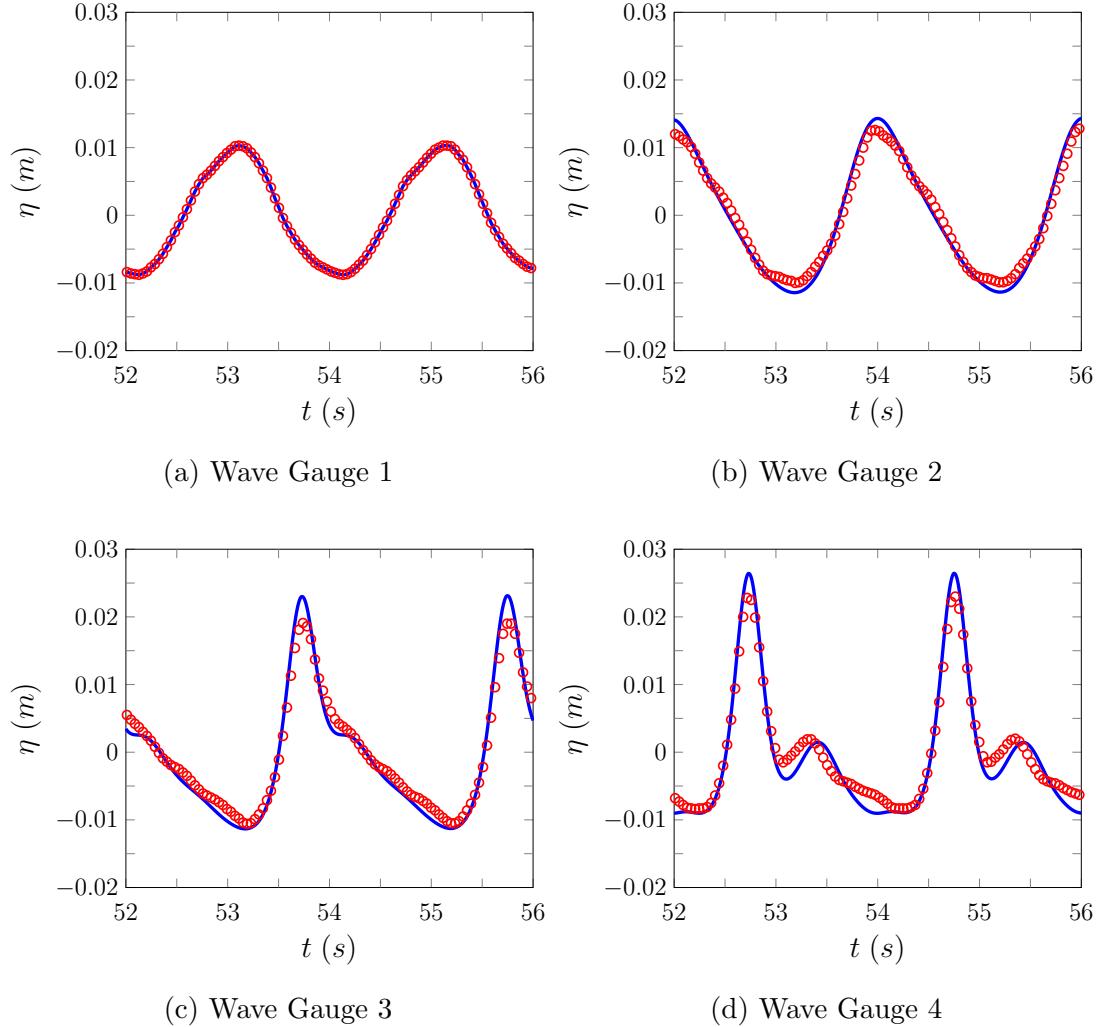


Figure 6.7: Comparison of the wave heights η of the numerical results for the FEVM₂ (—) and the experimental results (○) for wave gauges 1 - 4 for the low frequency experiment.

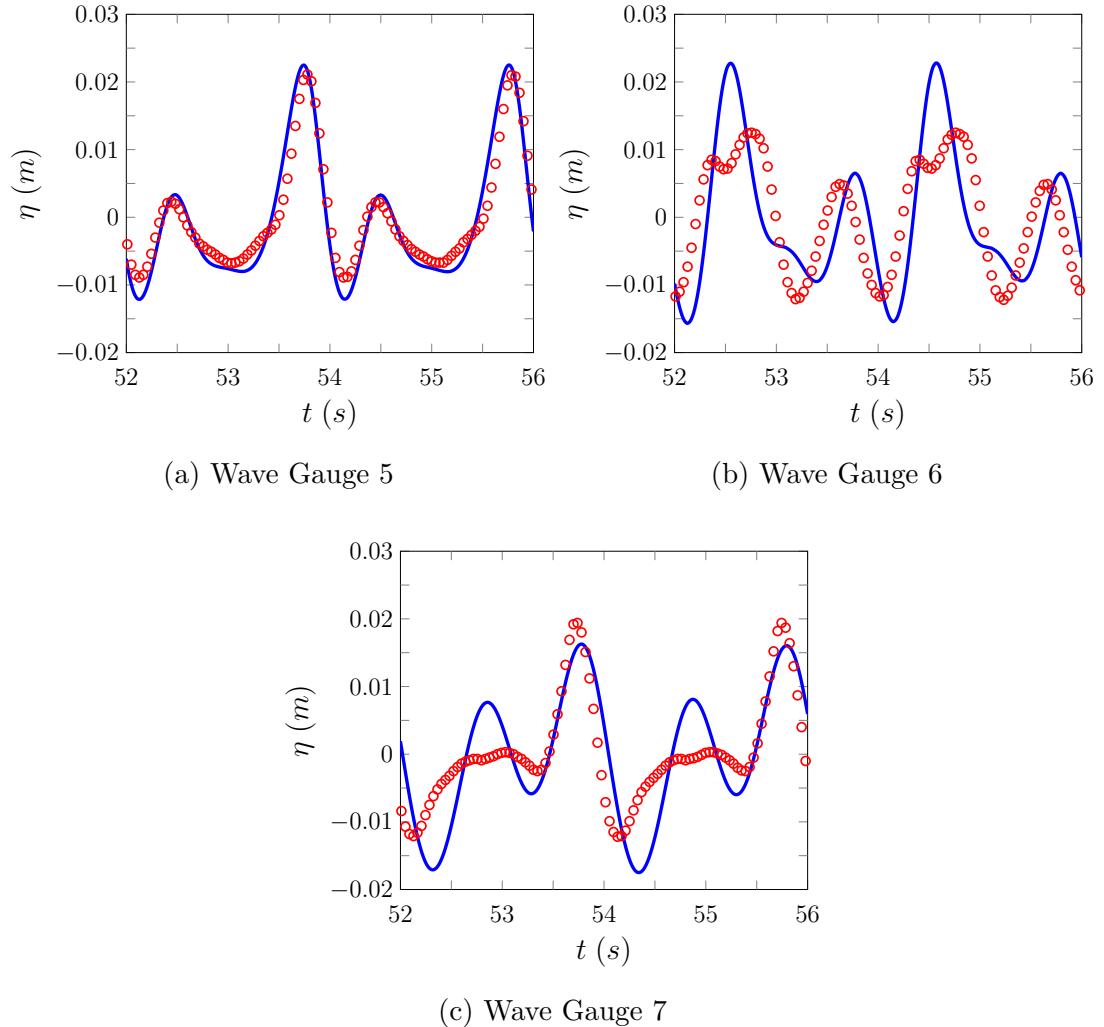


Figure 6.8: Comparison of the wave heights η of the numerical results for the FEVM₂ (—) and the experimental results (○) for wave gauges 5 - 7 for the low frequency experiment.

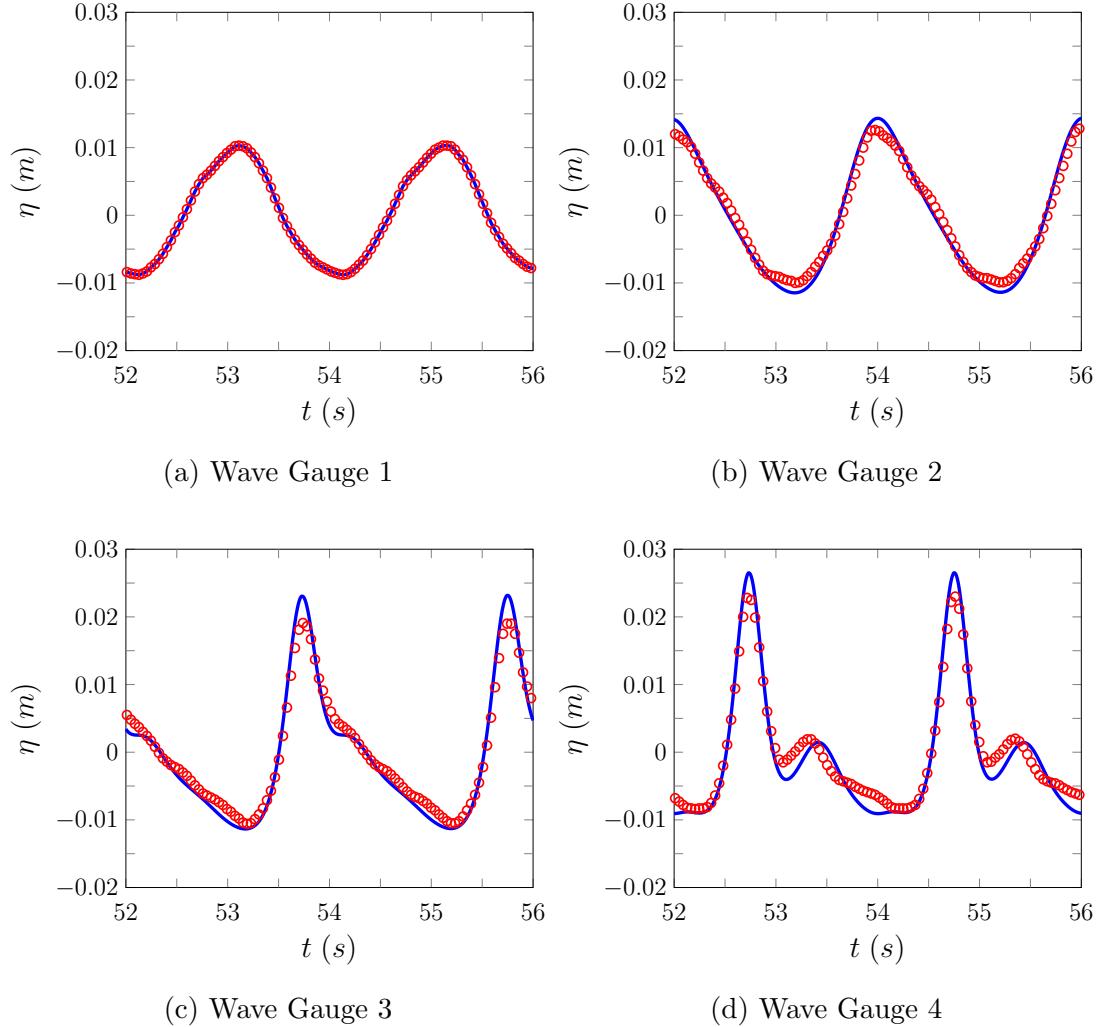


Figure 6.9: Comparison of the wave heights η of the numerical results for the FDVM₂ (—) and the experimental results (○) for wave gauges 1 - 4 for the low frequency experiment.

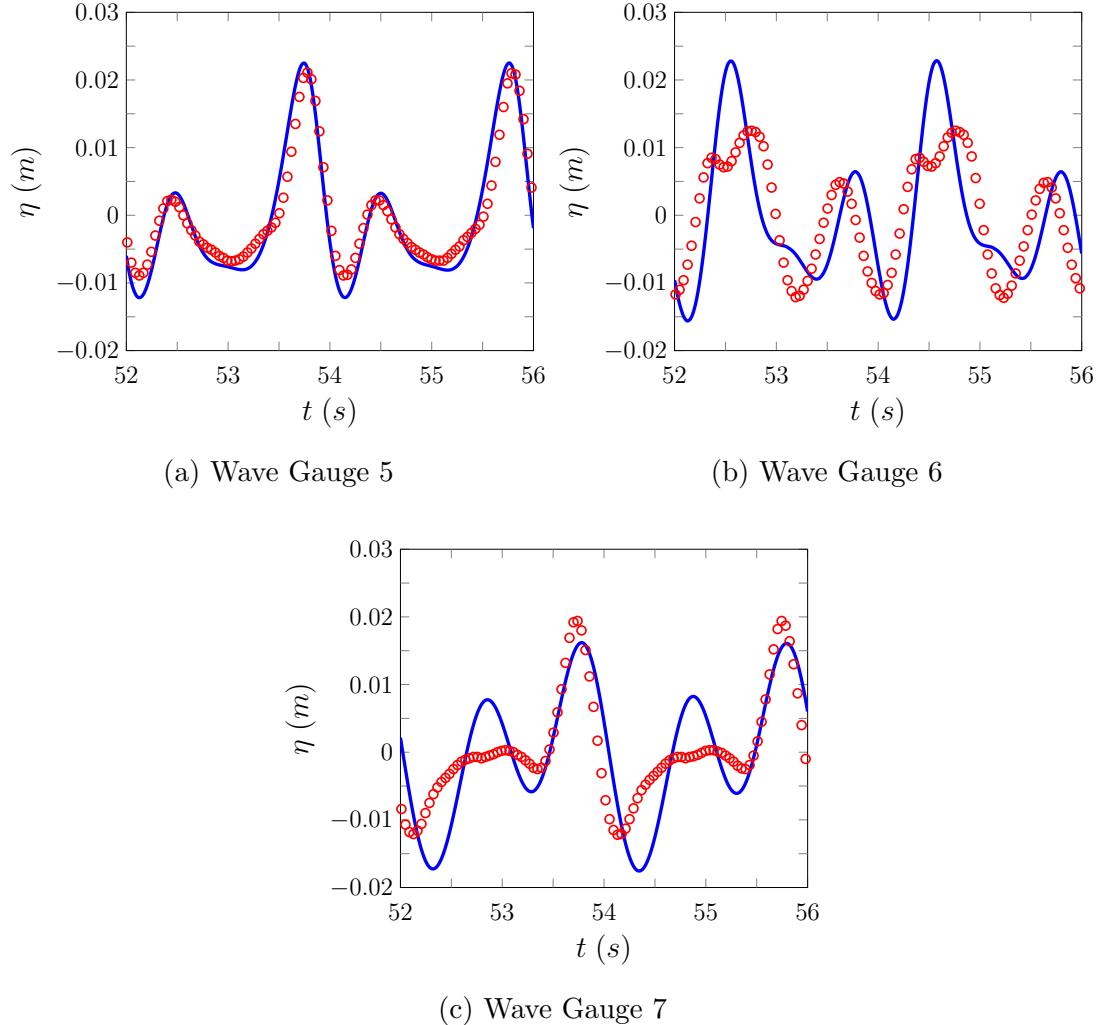


Figure 6.10: Comparison of the wave heights η of the numerical results for the FDVM₂ (—) and the experimental results (○) for wave gauges 5 - 7 for the low frequency experiment.

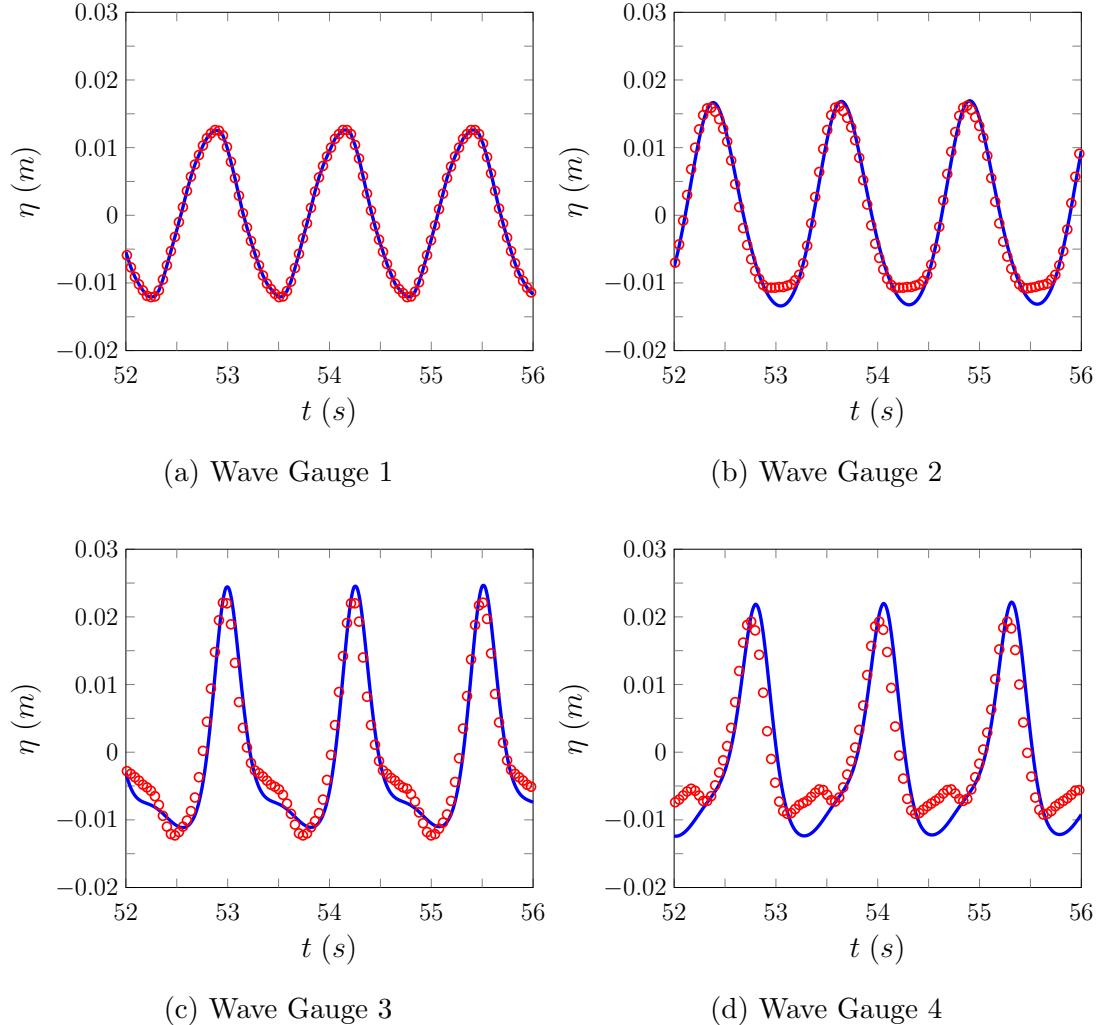


Figure 6.11: Comparison of the wave heights η of the numerical results for the FEVM₂ (—) and the experimental results (○) for wave gauges 1 - 4 for the low frequency experiment.

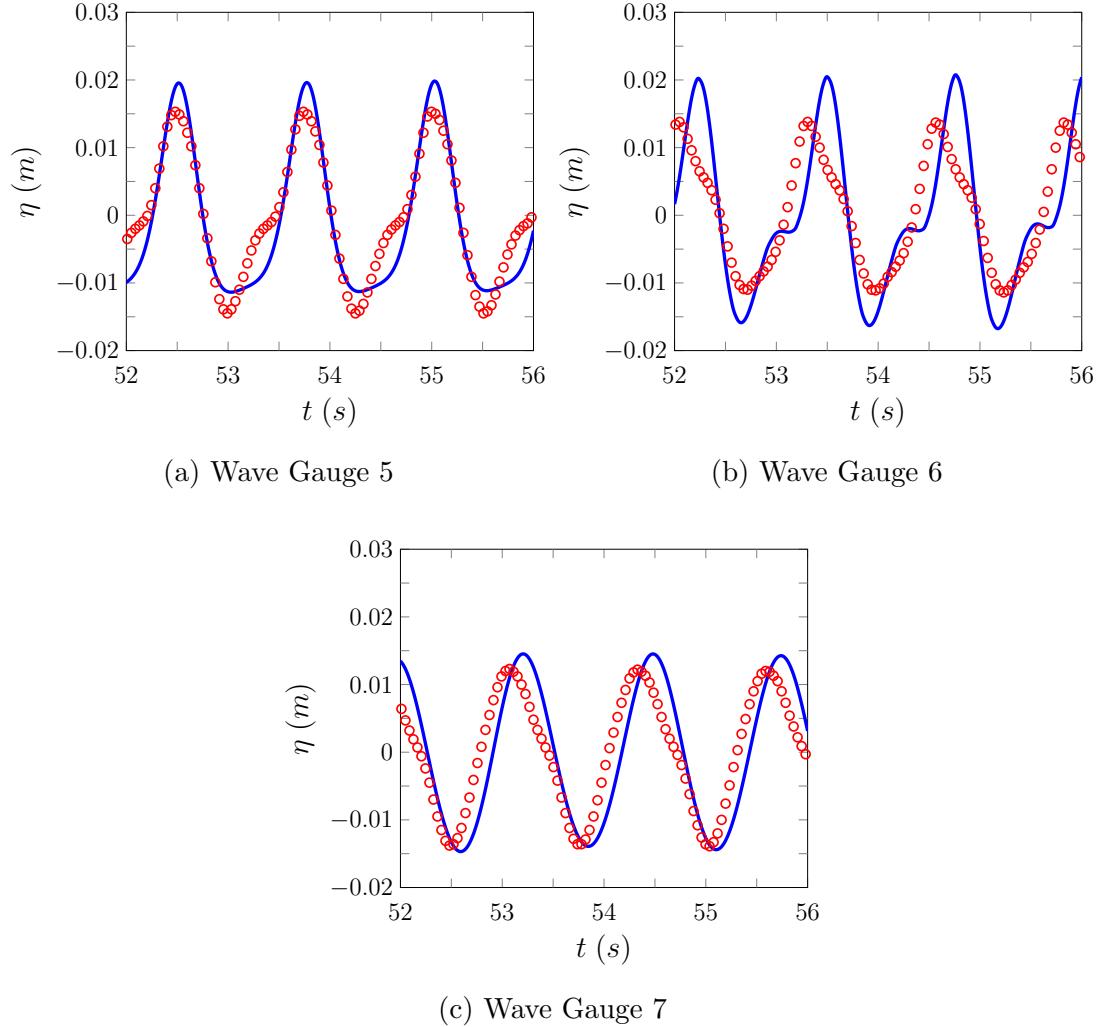


Figure 6.12: Comparison of the wave heights η of the numerical results for the FEVM₂ (—) and the experimental results (○) for wave gauges 5 - 7 for the high frequency experiment.

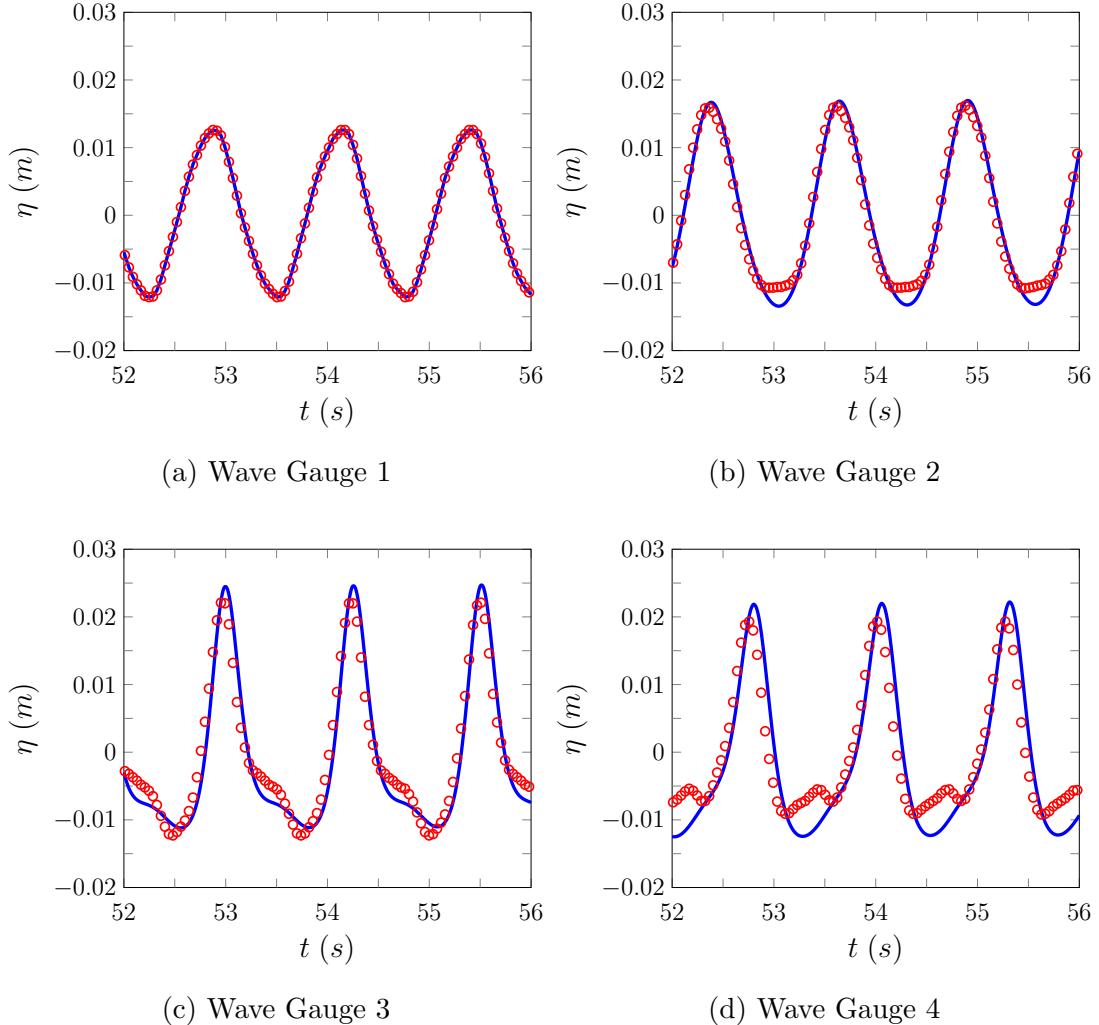


Figure 6.13: Comparison of the wave heights η of the numerical results for the FDVM₂ (—) and the experimental results (○) for wave gauges 1 - 4 for the high frequency experiment.

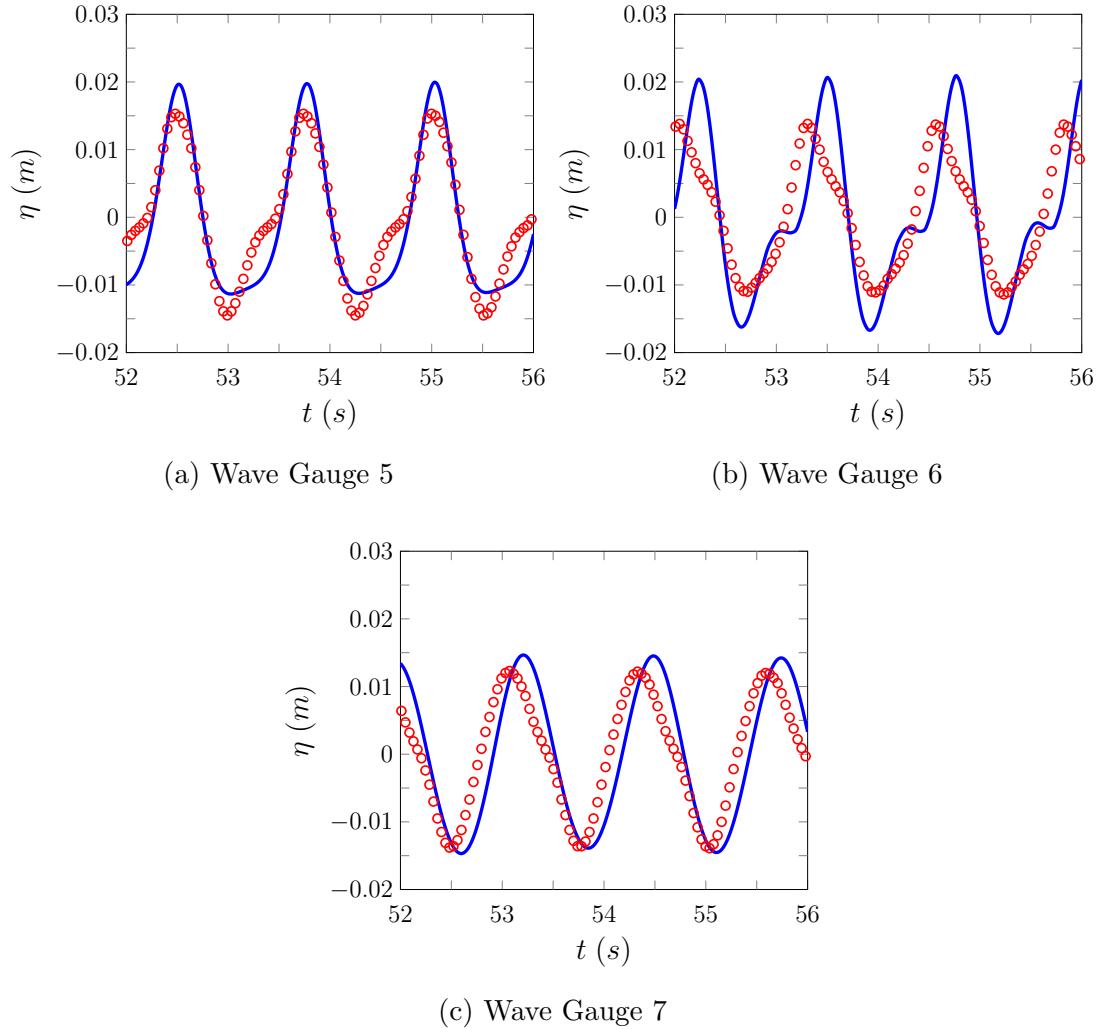


Figure 6.14: Comparison of the wave heights η of the numerical results for the FDVM₂ (—) and the experimental results (○) for wave gauges 5 - 7 for the high frequency experiment.

$3.66m$ wide, $83.7m$ long and $4.57m$ high with removable beds that allowed for the wide range of experiments reported by Roeber [33]. We have computationally modelled the experiment with the bathymetry displayed in Figure 6.15, where a solitary wave is generated from the wavemaker at $0m$ and is recorded at the wave gauges at $17.6m$, $28.6m$, $35.9m$, $40.6m$, $44.3m$, $46.1m$, $48.2m$, $50.4m$, $54.4m$, $58.0m$, $61.7m$, $65.4m$, $72.7m$ and $80.0m$.

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 thin 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 $t \approx 33s$. 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 $\Delta t = Sp/8s$ where $Sp = 0.02$ 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 and no effects from the boundary conditions were observed in the numerical results.

6.3.1 Results

The wave gauge results comparing the numerical and experimental data are displayed in Figures 6.18 and 6.19 for FDVM₂ and 6.16 and 6.17 for FEVM₂.

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 FDVM₂ becomes unstable and the solution blows up. Be-

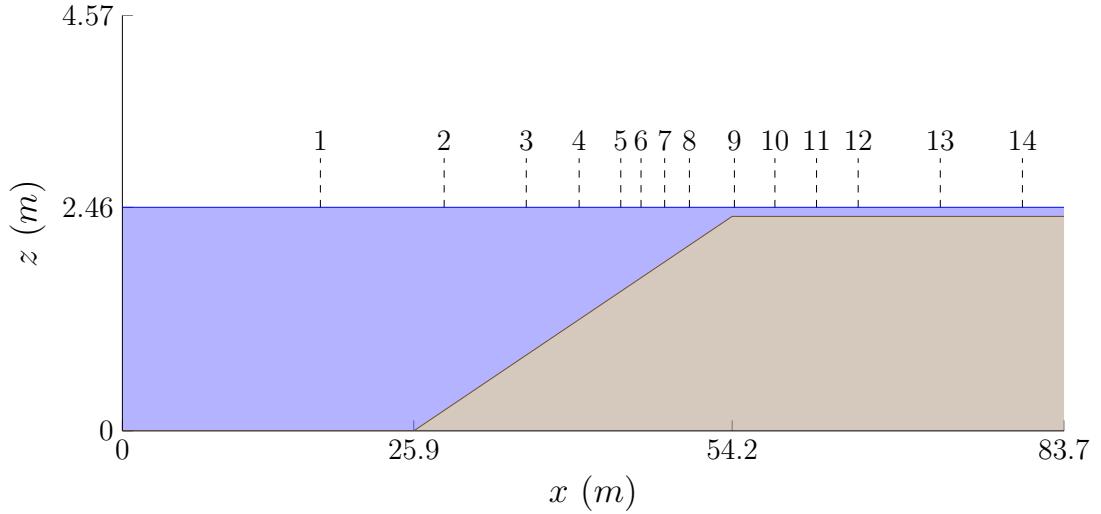


Figure 6.15: Diagram demonstrating the water (■) and the ground (■) for the Beji experiments, with the wave gauge locations marked.

cause of this the numerical solution of FDVM₂ 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. This is a much higher relative jump in the water heights than previously observed for these methods [15] and so their instability for these problems was not previously apparent. The FEVM₂ 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 FEVM₂ are quite good and the numerical methods 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.

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 [34]. 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 runs were all

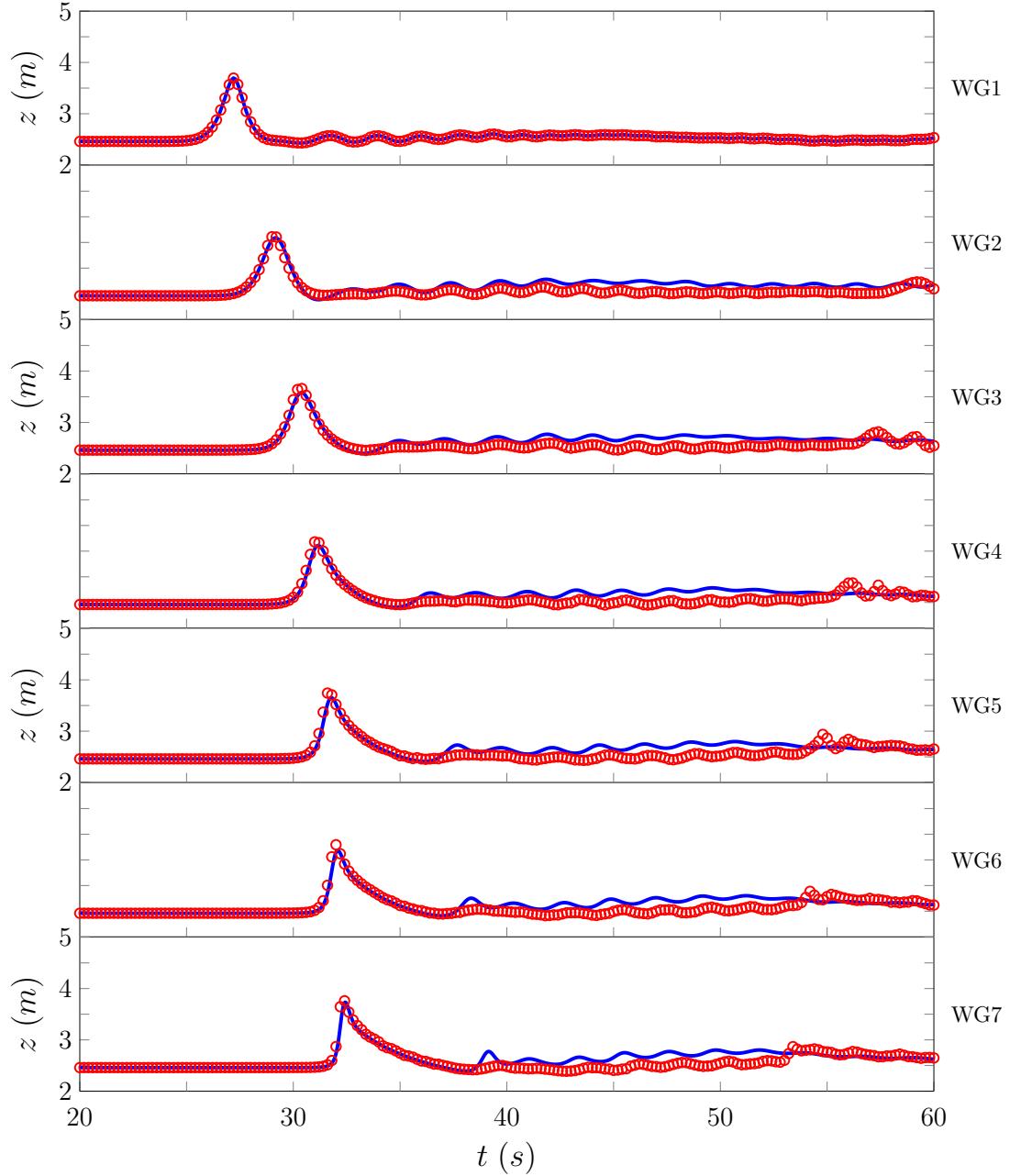


Figure 6.16: Comparison of the experimental (○) and numerical (—) wave gauge data produced by FEVM₂ for gauges 1 to 7.

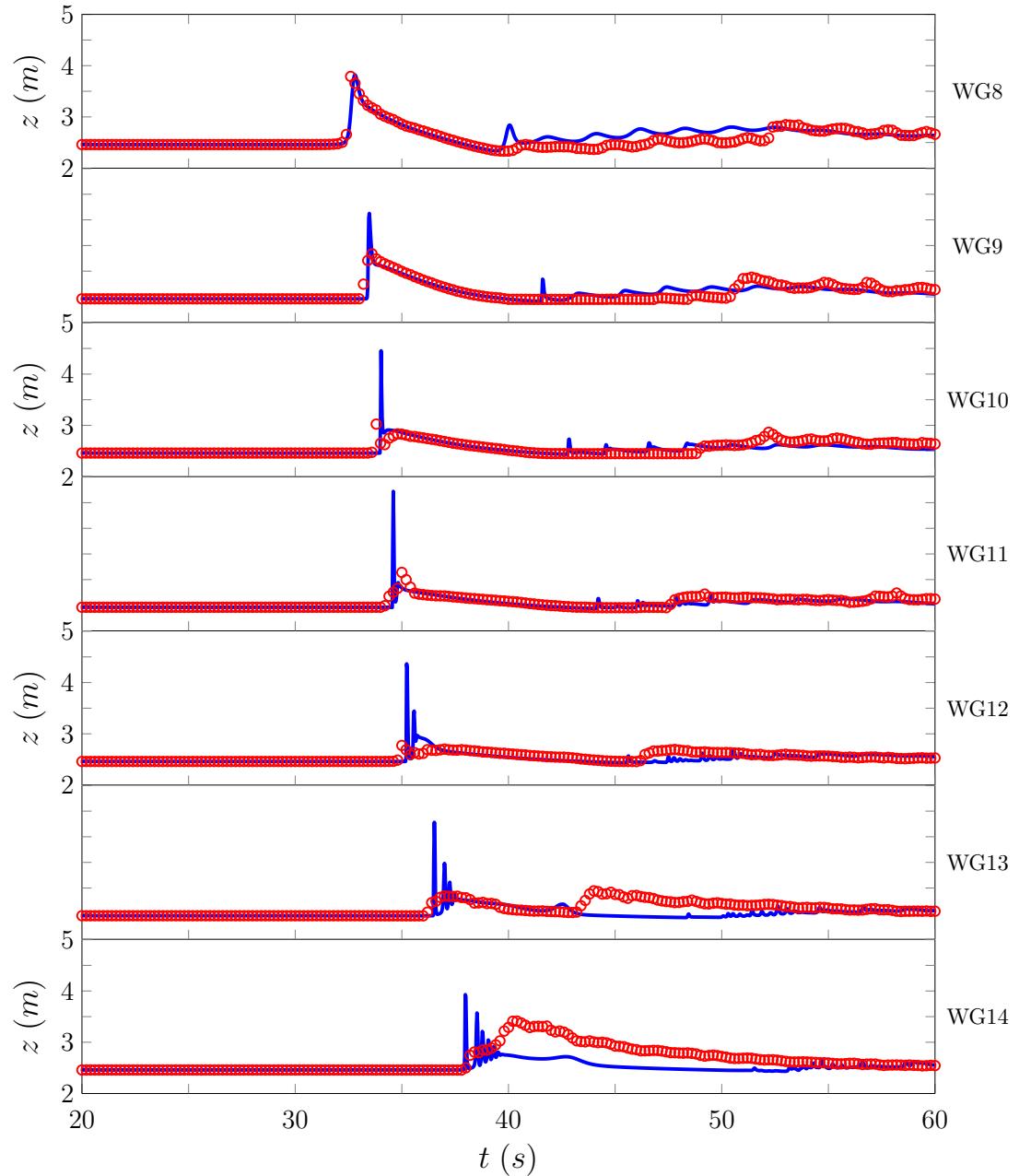


Figure 6.17: Comparison of the experimental (○) and numerical (—) wave gauge data produced by FEVM₂ for gauges 8 to 14.

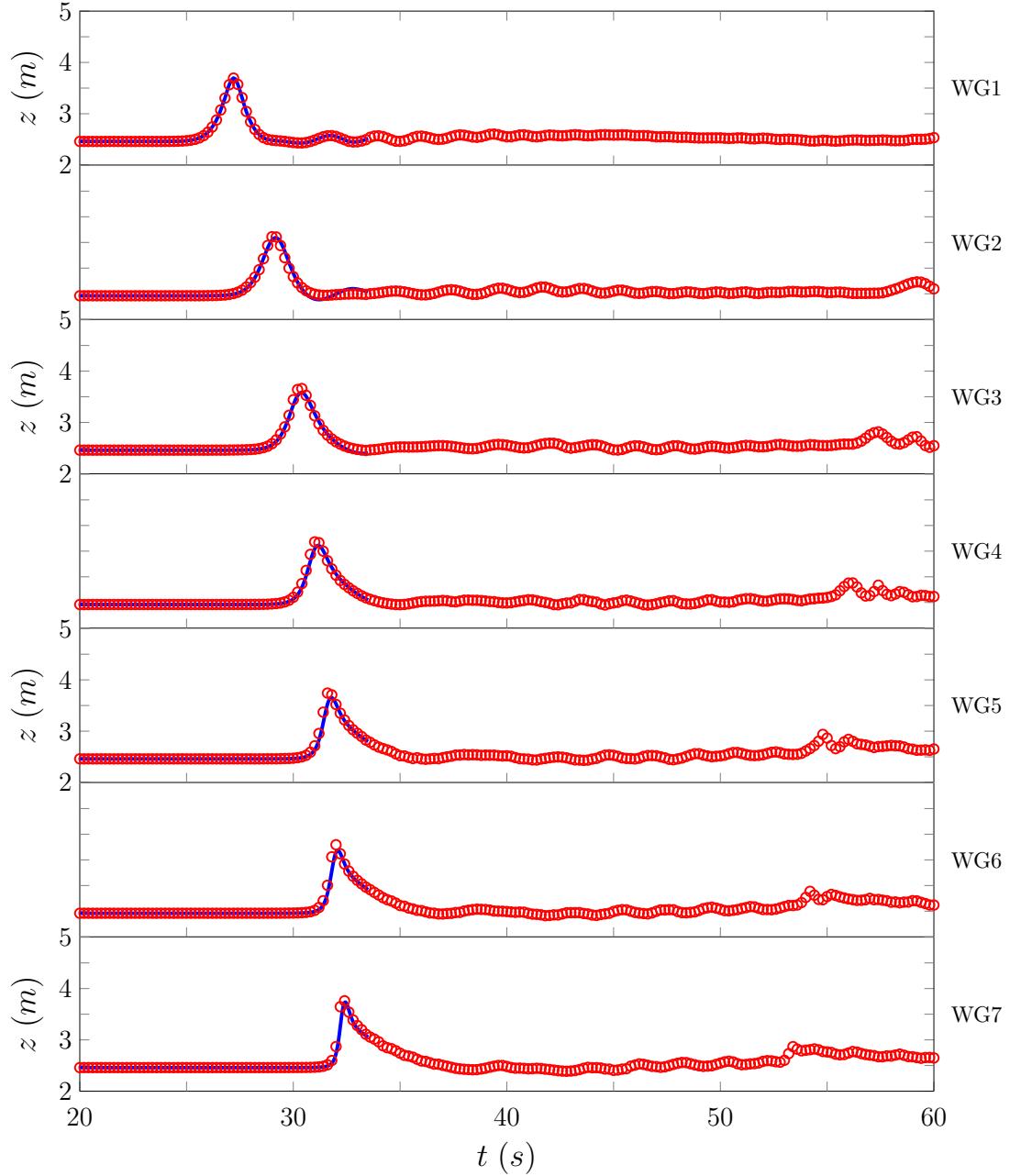


Figure 6.18: Comparison of the experimental (○) and numerical (—) wave gauge data produced by FDVM₂ for gauges 1 to 7.

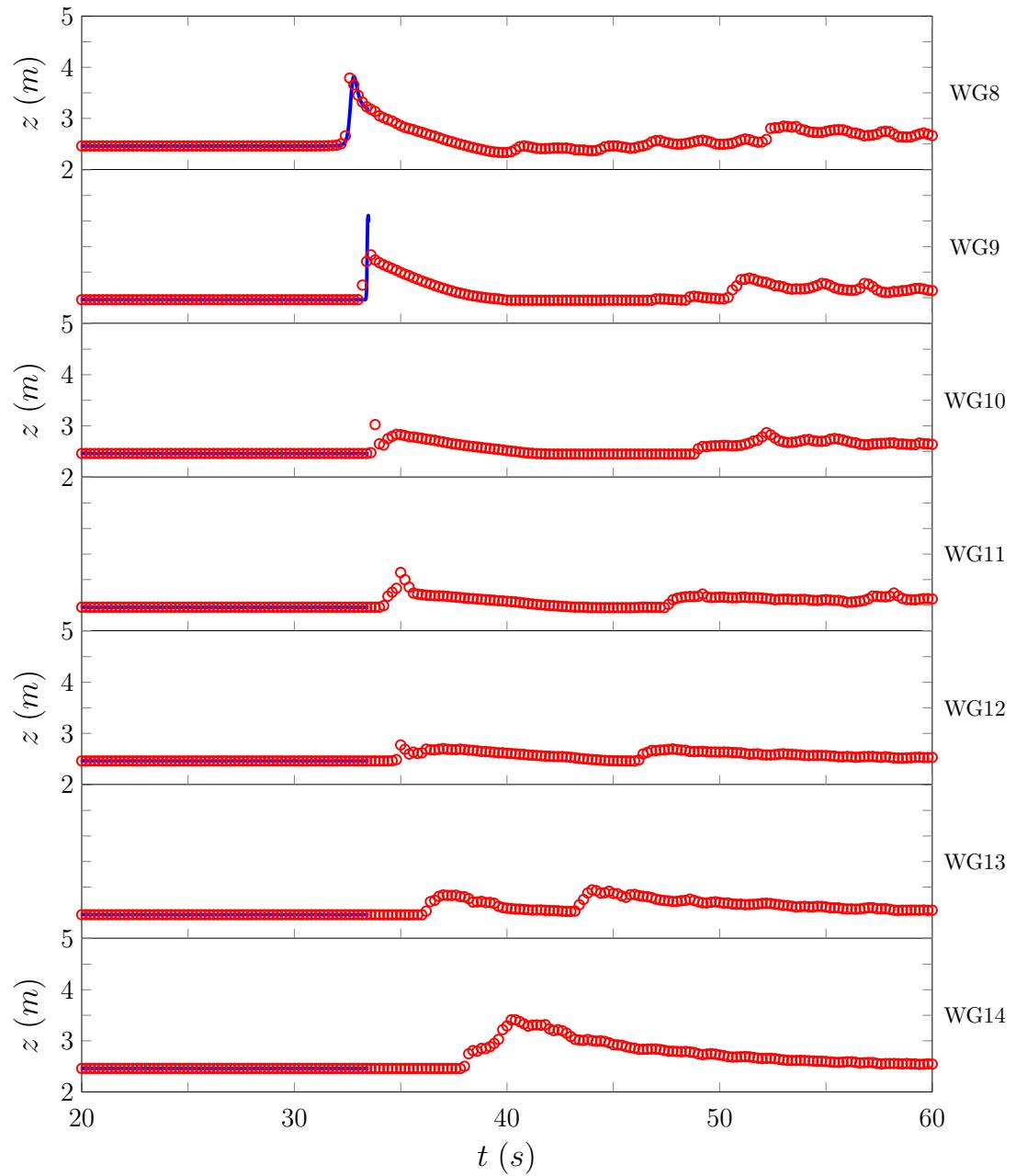


Figure 6.19: Comparison of the experimental (○) and numerical (—) wave gauge data produced by FDVM₂ for gauges 8 to 14.

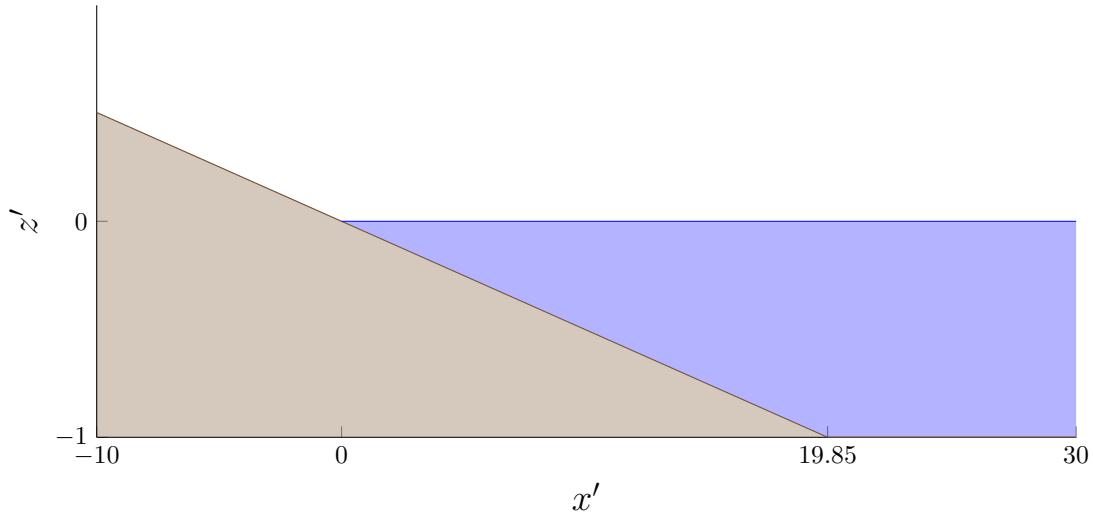


Figure 6.20: Diagram demonstrating the water (■) and the ground (□) for the Synolakis experiments with the normalised coordinates.

performed on the beach profile depicted in Figure 6.20, where all the quantities are normalised [34]. To assess the computational models we recreated one of these runs, 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 [1]. Therefore, the effect of the extra dispersive terms included by the Serre equations on the inundation process is not well tested by this experiment.

The numerical experiments used the normalised quantities reported by Synolakis [34] 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 6.22 for FDVM₂ and 6.21 for FEVM₂. The error in conservation of the conserved quantities are given in Tables 6.5 and 6.6 for FEVM₂ and FDVM₂ respectively.

Quantity	$\mathcal{C}^*(\mathbf{q}^0)$	$\mathcal{C}^*(\mathbf{q}^*)$	$\mathcal{C}_1^*(\mathbf{q}^0, \mathbf{q}^*)$
Mass	140.41696534392645	140.416965345	-7.65×10^{-12}
Momentum	-0.31905013851476244	0.320317183357	0.0040
G	-0.31907372312696913	0.320448924368	0.0043
Energy	-68.38995818739492	-68.3914362911	2.16×10^{-5}

Table 6.5: Total amounts and error in conservation for all quantities for FEVM₂ numerical solution of the runup experiment.

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 caused by the lack of friction in the numerical method.

The conserved quantities are well conserved by the method; in particular the mass which even with the wetting and drying of cells was very well conserved. The total energy of the method is also very 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 momentum and G are not necessarily conserved by this experiment. However, at the end of the experiment the total momentum and G of the numerical solution has been well reflected from the beach, affirming the capabilities of the method.

The results for both FEVM₂ and FDVM₂ 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 for the inundation of non-breaking waves.

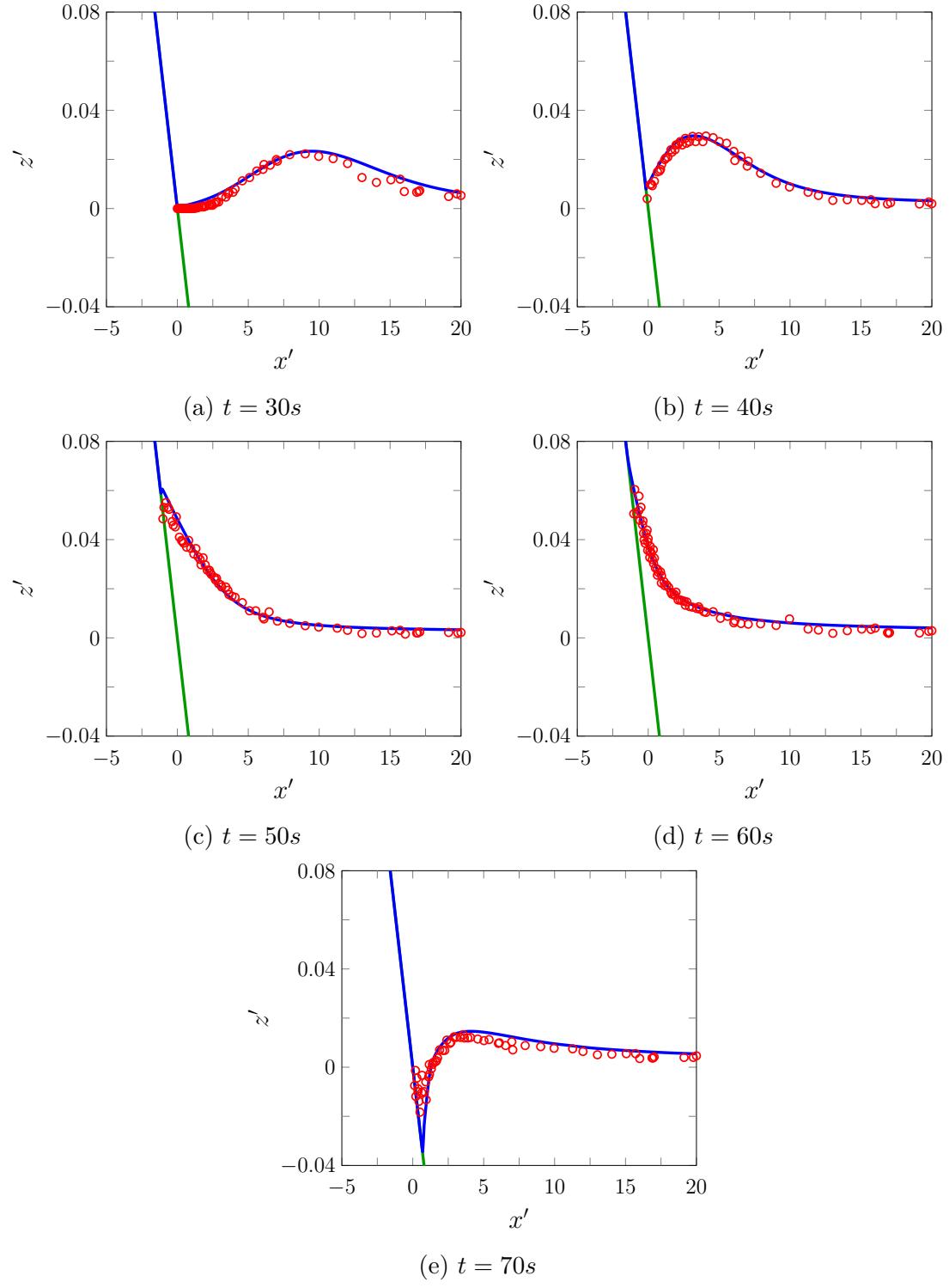


Figure 6.21: A comparison of the water surface profiles for the experiment (○) and the numerical solution (—) produced by FEVM₂ at various times.

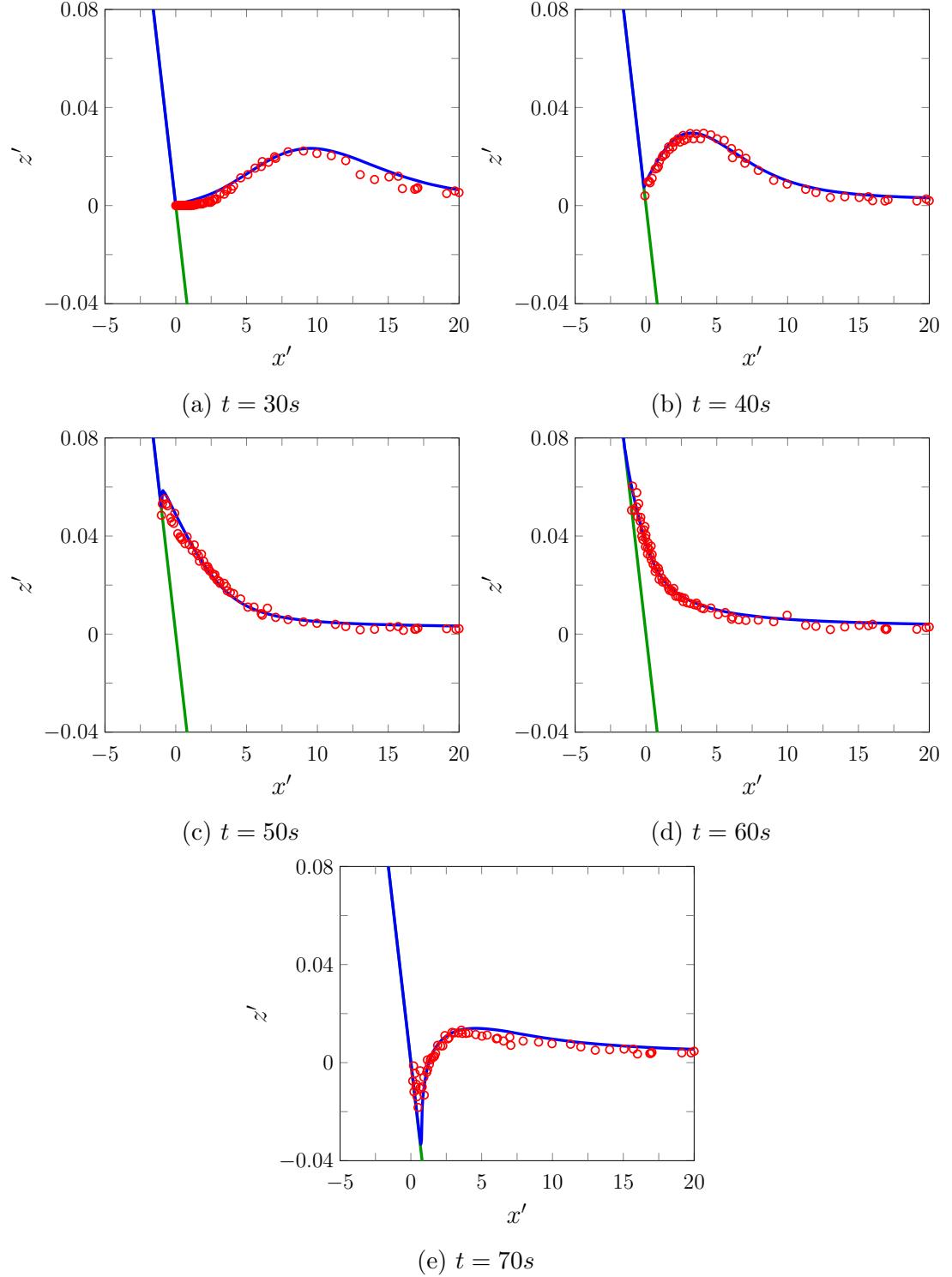


Figure 6.22: A comparison of the water surface profiles for the experiment (○) and the numerical solution (—) produced by FDVM₂ at various times.

Quantity	$\mathcal{C}^*(\mathbf{q}^0)$	$\mathcal{C}^*(\mathbf{q}^*)$	$\mathcal{C}_1^*(\mathbf{q}^0, \mathbf{q}^*)$
Mass	140.41696534392645	140.416949821	1.11×10^{-7}
Momentum	-0.31905013851476244	0.319579326905	0.0017
G	-0.31907372312696913	0.320668702008	0.0050
Energy	-68.38995818739492	-68.3914356777	2.16×10^{-5}

Table 6.6: Total amounts and error in conservation for all quantities for FDVM₂ numerical solution of the runup experiment.

Chapter 7

Conclusion

Appendix A

Finite Element Integrals

We now provide some of the details left out of Chapter 3; the definitions for the basis functions in the ξ space and the other finite element matrices.

A.1 Basis Functions

ψ

$$\psi_{j-1/2}^+ = \begin{cases} \frac{1}{2}(1-\xi) & -1 \leq \xi \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.1})$$

$$\psi_{j+1/2}^- = \begin{cases} \frac{1}{2}(1+\xi) & -1 \leq \xi \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.2})$$

ϕ

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

$$\phi_j = \begin{cases} -(\xi - 1)(\xi + 1) & -1 \leq \xi \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.4})$$

$$\phi_{j+1/2} = \begin{cases} \frac{1}{2}\xi(\xi + 1) & -1 \leq \xi \leq 1 \\ 2(\xi - 2)(\xi - \frac{3}{2}) & 1 \leq \xi \leq 2 \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.5})$$

$$(\text{A.6})$$

γ

$$\gamma_{j-1/2} = \begin{cases} \frac{9}{2} \left(\xi + \frac{4}{3} \right) \left(\xi + \frac{5}{3} \right) (\xi + 2) & -2 \leq \xi \leq -1 \\ \frac{9}{16} \left(\xi - 1 \right) \left(\xi - \frac{1}{3} \right) \left(\xi + \frac{1}{3} \right) & -1 \leq \xi \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.7})$$

$$\gamma_{j-1/6} = \begin{cases} \frac{27}{16} \left(\xi - 1 \right) \left(\xi - \frac{1}{3} \right) (\xi + 1) & -1 \leq \xi \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.8})$$

$$\gamma_{j+1/6} = \begin{cases} -\frac{27}{16} \left(\xi - 1 \right) \left(\xi + \frac{1}{3} \right) (\xi + 1) & -1 \leq \xi \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.9})$$

$$\gamma_{j-1/2} = \begin{cases} \frac{9}{16} \left(\xi + 1 \right) \left(\xi - \frac{1}{3} \right) \left(\xi + \frac{1}{3} \right) & -1 \leq \xi \leq 1 \\ -\frac{9}{2} \left(\xi - \frac{4}{3} \right) \left(\xi - \frac{5}{3} \right) (\xi - 2) & 1 \leq \xi \leq 2 \\ 0 & \text{otherwise} \end{cases} \quad (\text{A.10})$$

Appendix B

Linear Analysis Results

In this appendix we present all the pieces one would need to calculate the Evolution matrix \mathbf{E} for all the methods. For the hybrid FDVM given the results in Chapter 4 for the FEVM it is enough to provide only expressions for some of the operators. While for the FD methods we just present the evolution matrix as its calculation is straightforward given their description by Zoppou et al. [24].

B.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 , $\mathcal{R}_{j+1/2}^-$ and \mathcal{G} with the appropriate ones for the FDVM given in Tables B.1, B.2, B.3 and B.4. From \mathbf{F} the appropriate \mathbf{E} is then obtained by using the appropriate formulas given by the Runge-Kutta time stepping in Table B.5.

B.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

Scheme	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 B.1: Factor \mathcal{R}_j from transformation between nodal and cell average values.
Where the analytic value is $\mathcal{R}_j = \frac{k\Delta x}{2 \sin(k\frac{\Delta x}{2})}$.

Scheme	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	$\frac{1}{6}(5 + 2e^{-ik\Delta x} - e^{ik\Delta x})$	$\frac{i}{12}k^3\Delta x^3$

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

Scheme	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	$\frac{1}{6}(5 - e^{-ik\Delta x} + 2e^{ik\Delta x})$	$-\frac{i}{12}k^3\Delta x^3$

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

Scheme	Expression	Lowest Order Term of Error
FDVM ₁	$\frac{3\Delta x^2 \left(\frac{1+e^{ik\Delta x}}{2} \right)}{3\Delta x^2 H - H^3 (2 \cos(k\Delta x) - 2)}$	$-\frac{6 + H^2 k^2}{4H (3 + H^2 k^2)^2} k^2 \Delta x^2$
FDVM ₂	$\frac{3\Delta x^2 \left(\frac{1+e^{ik\Delta x}}{2} \right)}{3\Delta x^2 H - H^3 (2 \cos(k\Delta x) - 2)}$	$-\frac{6 + H^2 k^2}{4H (3 + H^2 k^2)^2} k^2 \Delta x^2$
FEVM ₂	$\begin{aligned} & \left(\frac{\Delta x}{6} \left(1 + \frac{i \sin(k\Delta x)}{2} + e^{ik\Delta x} \left(1 - \frac{i \sin(k\Delta x)}{2} \right) \right) \right) \\ & \div \left(H \frac{\Delta x}{30} \left(2 \left(2 \cos\left(\frac{k\Delta x}{2}\right) - \cos(k\Delta x) + 4 \right) \right. \right. \\ & \left. \left. + \frac{H^3}{9\Delta x} \left(-16 \cos\left(\frac{k\Delta x}{2}\right) + 2 \cos(k\Delta x) + 14 \right) \right) \right) \end{aligned}$	$\frac{12 + 5H^2 k^2}{40H (3 + H^2 k^2)^2} k^2 \Delta x^2$
FDVM ₃	$\frac{36\Delta x^2 \left(\frac{-e^{-ik\Delta x} + 9e^{ik\Delta x} - e^{2ik\Delta x} + 9}{16} \right)}{36\Delta x^2 H - H^3 (32 \cos(k\Delta x) - 2 \cos(2k\Delta x) - 30)}$	$-\frac{243 + 49H^2 k^2}{960H (3 + H^2 k^2)^2} k^4 \Delta x^4$

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

Order	Formula for \mathbf{E}
First	$\mathbf{I} - \Delta t \mathbf{F}$
Second	$\mathbf{I} - \Delta t \mathbf{F} + \frac{1}{2} \Delta t^2 \mathbf{F}^2$
Third	$\mathbf{I} - \Delta t \mathbf{F} + \frac{1}{2} \Delta t^2 \mathbf{F}^2 - \frac{1}{6} \Delta t^3 \mathbf{F}^3$

Table B.5: Formula for \mathbf{E} given \mathbf{F} determined by the SSP Runge-Kutta timestepping method.

for the hybrid methods. Doing this we get that the evolution matrix for \mathcal{D} is

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

While for \mathcal{W} we get that evolution matrix is

$$\mathbf{E} = \begin{bmatrix} E_{0,0} & E_{0,1} & 0 & -\frac{\Delta t}{\Delta x} H \frac{i \sin(k\Delta x)}{2} \\ E_{1,0} & -\frac{2i\Delta t}{\Delta x} U \sin(k\Delta x) & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad (\text{B.1})$$

with

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

Bibliography

- [1] F. Serre. Contribution à l'étude des écoulements permanents et variables dans les canaux. *La Houille Blanche*, 6:830–872, 1953.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] W. Choi and R. Camassa. Fully nonlinear internal waves in a two-fluid system. *Journal of Fluid Mechanics*, 396:1–36, 1999.

- [9] 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.
- [10] 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.
- [11] 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.
- [12] E. Barthélemy. Nonlinear shallow water theories for coastal waves. *Surveys in Geophysics*, 25(3):315–337, 2004.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] P.L Roe. Characteristic-based schemes for the euler equations. *Annual Review of Fluid Mechanics*, 18(1):337–365, 1986.
- [17] 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.
- [18] W.H Press, S.A Teukolsky, W.T Vetterling, and B.P Flannery. *Numerical Recipes in C*, volume 2. Cambridge University Press, 1996.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] C. Zoppou, J. Pitt, and S. Roberts. Numerical solution of the fully nonlinear weakly dispersive Serre equations for steep gradient flows. *Applied Mathematical Modelling*, 48:70–95, 2017.
- [25] 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.
- [26] 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.
- [27] A.T. Ippen and G. Kulin. The shoaling and breaking of the solitary wave. *Coastal Engineering Proceedings*, 1(5):4, 1954.
- [28] 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.
- [29] S Beji and J.A. Battjes. Experimental investigation of wave propagation over a bar. *Coastal Engineering*, 19(1):151–162, 1993.
- [30] S Beji and J.A. Battjes. Numerical simulation of nonlinear wave propagation over a bar. *Coastal Engineering*, 23(1):1–16, 1994.

- [31] D. Lannes. *The Water Waves Problem: Mathematical Analysis and Asymptotic*, volume 1 of *American Mathematical Society. Mathematical Surveys and Monographs*, 2013.
- [32] 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.
- [33] 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.
- [34] C. E. Synolakis. The runup of solitary waves. *Journal of Fluid Mechanics*, 185:523–545, 1987.