

Simulation of Rapidly Varying and Dry  
Bed Flow using the Serre equations  
solved by a Finite Element Volume  
Method.

Jordan Peter Anthony Pitt

January 2019

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





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



# **Declaration**

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

Jordan Peter Anthony Pitt



# Acknowledgements

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

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

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

Finally, I would like to thank my dear friend Joseph Gibson who donated a computer that hosted many numerical experiments.



# Abstract

Recent research in numerical wave modelling has focused on developing computational methods for solving non-linear, dispersive wave equations as an extension to methods solving the non-linear shallow water wave equations. These equations contain extra terms that allow for dispersion; improving their modelling capabilities for water waves. An interesting example of these non-linear dispersive equations for modelling water waves are the Serre equations.

In this work an efficient and robust numerical method for the one-dimensional Serre equations was developed. This method uses a finite element method to solve an elliptic equation and a finite volume method to solve the remaining conservation equations and is hence termed the Finite Element Volume Method (FEVM). The use of a finite element method and a finite volume method makes the FEVM adaptable to unstructured meshes and parallelisable. The FEVM allows for the recovery of the lake at rest steady state and the simulation of flow over dry beds.

The convergence and dispersion properties of the FEVM was determined using a linear analysis. The FEVM was validated against analytic and forced solutions of the Serre equations, demonstrating its convergence and conservation properties. Finally, it was validated against experimental data for a wide array of physical scenarios, establishing its utility as a realistic model.

All these analyses and validations were also conducted on other methods allowing comparisons between them and the FEVM. Overall the FEVM was found to be the most robust whilst being adequately accurate and is therefore, the recommended approach for solving the Serre equations.



# Contents

<b>Acknowledgements</b>	vii
<b>Abstract</b>	ix
<b>1 Introduction</b>	1
1.1 Objectives of the Thesis . . . . .	2
1.2 Original Contribution of the Thesis . . . . .	3
1.2.1 Publications . . . . .	3
1.3 Organisation of the Thesis . . . . .	8
<b>2 The Serre Equations</b>	9
2.1 The One-Dimensional Serre Equations . . . . .	10
2.1.1 Alternative Form . . . . .	12
2.2 Properties . . . . .	13
2.2.1 Conservation Properties . . . . .	14
2.2.2 Dispersion Properties . . . . .	15
2.2.3 Analytic Solutions . . . . .	16
2.3 Forced Solutions . . . . .	18
2.4 Behaviour in the Presence of Steep Gradients . . . . .	19
<b>3 Finite Element Volume Method</b>	25
3.1 Notation for Numerical Grids . . . . .	26
3.2 Structure Overview . . . . .	27
3.2.1 Reconstruction . . . . .	30
3.2.2 Fluid Velocity . . . . .	32
3.2.3 Flux Across the Cell Interfaces . . . . .	37
3.2.4 Source Terms . . . . .	40
3.2.5 Update Cell Averages . . . . .	41
3.2.6 Second-Order SSP Runge-Kutta Method . . . . .	42

3.3	CFL condition . . . . .	42
3.4	Boundary Conditions . . . . .	42
3.5	Dry Beds . . . . .	43
<b>4</b>	<b>Linear Analysis</b>	<b>47</b>
4.1	Linearised Serre Equations . . . . .	49
4.2	Evolution Matrix . . . . .	49
4.2.1	Reconstruction . . . . .	50
4.2.2	Fluid Velocity . . . . .	51
4.2.3	Flux Across the Cell Interfaces . . . . .	53
4.2.4	Update Cell Averages . . . . .	56
4.2.5	Second-Order SSP Runge-Kutta Method . . . . .	57
4.3	Convergence Analysis . . . . .	58
4.3.1	Stability . . . . .	58
4.3.2	Consistency . . . . .	59
4.4	Dispersion Analysis . . . . .	62
<b>5</b>	<b>Numerical Validation</b>	<b>71</b>
5.1	Measuring Convergence and Conservation . . . . .	72
5.1.1	Measure of Convergence . . . . .	72
5.1.2	Measures of Conservation . . . . .	72
5.2	Solitary Travelling Wave Solution . . . . .	73
5.3	Lake at Rest Solution . . . . .	79
5.4	Forced Solutions . . . . .	83
5.4.1	Results for a Wet Bed . . . . .	86
5.4.2	Results with a Dry Bed . . . . .	87
<b>6</b>	<b>Experimental Validation</b>	<b>95</b>
6.1	Evolution of a Negative Rectangular Wave . . . . .	95
6.1.1	Results for the $0.01m$ Negative Rectangular Wave . . . . .	97
6.1.2	Results for the $0.03m$ Negative Rectangular Wave . . . . .	100
6.2	Periodic Waves Over A Submerged Bar . . . . .	103
6.2.1	Low Frequency Results . . . . .	106
6.2.2	High Frequency Results . . . . .	111
6.3	Solitary Wave Over a Fringing Reef . . . . .	111
6.4	Run-up Experiment . . . . .	117

<i>CONTENTS</i>	xiii
<b>7 Conclusion</b>	<b>129</b>
7.1 Future Work . . . . .	130
<b>A Additional Conservation Information</b>	<b>133</b>
A.1 Solitary Travelling Wave Solution . . . . .	133
A.2 Lake At Rest Solution . . . . .	135
<b>B Finite Element Method Definitions</b>	<b>137</b>
B.1 Basis Function Definitions . . . . .	137
B.2 Function Space Definitions . . . . .	139
<b>C Linear Analysis Results</b>	<b>141</b>
C.1 Finite Difference Volume Methods . . . . .	143
C.2 Finite Difference Methods . . . . .	150
C.3 Consistency Results . . . . .	152
<b>D Publications</b>	<b>159</b>
<b>Bibliography</b>	<b>166</b>



# Chapter 1

## Introduction

A significant portion of the world's people and critical infrastructure is located near the coast. While the ocean provides many opportunities it also poses significant hazards from tsunamis and storm surges. Furthermore, the dynamics of ocean waves drives other physical phenomena; such as the break-up of sea-ice and the erosion of beaches. Therefore, accurate modelling of ocean waves is important to our society.

The physics of water can be described using Newton's second law. From which the partial differential equations initially presented by Euler in 1757 [1] can be derived. The Euler equations were then extended to include viscosity, producing the full Navier-Stokes equations [2, 3]. Numerical methods [4, 5, 6] have been developed to solve the Euler equations; however due to their complexity these methods can only accurately resolve fluid behaviour over small scales and not the scales required to model tsunamis along a coastline.

For this reason the central focus of water wave modelling has been simplified water wave theories that approximate the behaviour of the free surface of water governed by the Euler equations. The most popular class of these approximate water wave theories are the shallow water wave theories where the characteristic water depth  $h_0$  is far smaller than the characteristic wave length  $\lambda_0$ , so that  $\sigma = h_0/\lambda_0 \ll 1$ . For tsunamis and storm surges  $h_0$  is typically  $4\text{km}$  far from the coastline and  $\lambda_0$  can be  $100\text{km}$  and so  $\sigma \ll 1$ .

Neglecting all terms of order  $\mathcal{O}(\sigma^2)$  the full Euler equations reduce to the Shallow Water Wave Equations (SWWE) [7] which describe fully non-linear non-dispersive waves. Retaining higher powers of  $\sigma$  leads to a class of equations known as 'Boussinesq-type' equations. Boussinesq-type equations are then classified by the powers of  $\sigma$  they retain and their retained non-linearity. The non-linearity

$\epsilon = a_0/h_0$  compares the characteristic amplitude of the waves  $a_0$  to the water depth  $h_0$ . These wave models form a spectrum with the SWWE being the simplest and most restrictive model and the Boussinesq-type models retaining the highest powers of  $\sigma$  and largest  $\epsilon$  being the most complex and least restrictive. The Serre equations are one particular Boussinesq-type equation that retains all terms of order  $\mathcal{O}(\sigma^4)$  and makes no assumption on the size of  $\epsilon$  [7]. This allows the Serre equations to model water waves better than the SWWE in intermediate water depths where  $\sigma^2$  terms can be significant but  $\sigma^4$  terms are not. These intermediate water depths tend to occur as tsunamis and storm surges approach the coastline and interact with the varying bathymetry. Since the Serre equations allow arbitrary wave height they are the most appropriate model for water waves for the  $\mathcal{O}(\sigma^4)$  class of Boussinesq-type equations.

There has previously been a significant amount of research into developing large scale, efficient and robust computational methods for the SWWE [8, 9, 10]. The SWWE neglect all terms of order  $\mathcal{O}(\sigma^2)$  in the Euler equations and so do not capture all water wave behaviour; in particular dispersion. Recent research has highlighted the need for dispersive wave models for the evolution of tsunamis [11, 12]. For the purposes of ocean wave modelling the Serre equations are the best placed [7]; retaining high-order  $\sigma$  terms and allowing arbitrary wave amplitude. Hence the overarching goal of our research is the development of large-scale, efficient and robust computational methods for the Serre equations for the purposes of wave modelling.

## 1.1 Objectives of the Thesis

In view of the overarching goal, the primary motivation of this thesis was the development of a numerical method for solving the one-dimensional Serre equations. This method should be robust to steep gradients in the free surface, handle dry beds and be extendable to the two-dimensional Serre equations using unstructured meshes.

Some of these goals were achieved through the development of the Finite Element Volume Method (FEVM). The FEVM is an improvement of the Finite Difference Volume Methods (FDVM) described by Zoppou [13]. The FEVM can adequately handle dry beds and uses a finite element method instead of a finite difference method, making it suitable for unstructured meshes.

The FEVM was assessed with a linear analysis, a validation against analytic

and forced solutions and experimental results. At all stages of this assessment the method is compared to at least one other method to demonstrate its strengths and weaknesses. Overall, the FEVM was found to be superior to other methods and satisfies all the objectives of the thesis.

## 1.2 Original Contribution of the Thesis

My research made the following original contributions to the field:

- Implementation of the third-order FDVM.
- Observation and justification of a new structure in the solution of the Serre equations in the presence of steep gradients in the free surface.
- Extension of the second-order FDVM to allow for dry beds.
- Development and description of the well-balanced second-order FEVM that can handle dry beds.
- A linear analysis of convergence for all developed finite volume based methods as well as some finite difference methods.
- A complete linear analysis of the dispersion properties for all developed finite volume based methods as well as some finite difference methods.
- Validation of FEVM and the second-order FDVM using forced solutions where all terms of the Serre equations are present for both wet and dry beds.
- Comparison of the numerical solutions of FEVM and the second-order FDVM with experimental results in the presence of dry beds and with wave breaking.

### 1.2.1 Publications

The publications generated from my research are compiled here in chronological order. A brief summary, my contribution to the paper and the relevance of the paper to this thesis are also provided. The publication list is reproduced in Appendix D where the abstracts of the publications are also provided.

## A Solution of the Conservation Law Form of the Serre Equations

*Australia and New Zealand Industrial and Applied Mathematics Journal (2016)*

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

### **Summary:**

A second-order FDVM for the one-dimensional Serre equations with a horizontal bed is described and validated against an analytic solution.

### **My Contribution:**

I reproduced the results of my coauthors with my own implementation of the method. This method was consistent with a SWWE solver allowing the computational cost of solving the Serre equations and the SWWE to be compared.

### **Relevance to Thesis:**

The method described in this paper was extended during my research to allow for varying bathymetry [14] and dry beds, producing the second-order FDVM. The results of a linear analysis and numerical experiments for the second-order FDVM are reported in the thesis. The linear analysis of the convergence and dispersion properties performed in this thesis assume a completely wet horizontal bed, and therefore these results apply to the method described in this paper. These linear results can be found in Chapter 4 and Appendix C. The extended version of the method was validated against analytic and forced solutions in Chapter 5 and compared to experimental results in Chapter 6.

## Numerical Solution of the Fully Non-Linear Weakly Dispersive Serre Equations for Steep Gradient Flows

*Applied Mathematical Modelling (2017)*

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

### **Summary:**

The first-, second-, and third-order FDVM for the one-dimensional Serre equations with variable bathymetry were described. The results of a linear analysis of the dispersion properties of the methods for waves on quiescent water are provided. These methods were then validated against an analytic solution. A well-balanced version of the second-order FDVM is described and further validated against experimental results.

### **My Contribution:**

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

### **Relevance to Thesis:**

The first-, second- and third-order FDVM described in this paper are the methods whose results are reported in this thesis. The results of a linear analysis of the dispersion and convergence properties of all these methods are provided in Chapter 4 and Appendix C. The linear analysis of dispersion presented in this paper was extended to allow for a mean background flow in the water on which the waves occur. The results of the analytic solution validation in this paper are reproduced in Chapter 5. This validation was extended to study the convergence and conservation properties of more quantities. The second-order FDVM was then validated against forced solutions in Chapter 5 and experimental results in Chapter 6. The results of the second-order FDVM for the negative rectangular wave experiment [15] and the periodic waves over a submerged bar experiment [16] in this paper are reproduced in this thesis.

## Importance of Dispersion for Shoaling Waves

*22nd International Congress on Modelling and Simulation (2017)*

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

### **Summary:**

The numerical solutions of the second-order FDVM [14] for the Serre equations were compared to the numerical solutions of ANUGA [10] for the SWWE. This comparison investigated the influence of dispersion on shoaling waves for an artificial representative example and the experiments of Beji and Battjes [16]. The artificial representative example simulates the shoaling of a solitary wave over a fringing reef. While the experiments of Beji and Battjes [16] investigate the behaviour of periodic waves travelling over a submerged bar.

### **My Contribution:**

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

### **Relevance to Thesis:**

The experimental results of the second-order FDVM for the periodic waves over a submerged bar experiments of Beji and Battjes [16] are reproduced in Chapter 6.

## Behaviour of the Serre Equations in the Presence of Steep Gradients Revisited

*Wave Motion (2018)*

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

### **Summary:**

The first-, second- and third-order FDVM as well as two second-order finite difference methods for the Serre equations were used to numerically study the behaviour of steep gradients in the free surface. The convergence and conservation properties of many numerical solutions of all these methods were used to justify the observed behaviours. One such behaviour was not previously reported by the literature. This behaviour best represents the solution of the Serre equations in the presence of steep gradients over short time spans. The effect of the used numerical method, the grid resolution and the smoothing of the initial conditions was studied. Allowing the source of the differences in the observed behaviours reported in the literature to be understood.

### **My Contribution:**

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

### **Relevance to Thesis:**

The behaviours observed and justified in this paper are summarised in Chapter 2. The results of this paper demonstrate the utility of using a finite volume based method to solve the Serre equations in the presence of steep gradients. Hence, the further development of these methods in this thesis.

## 1.3 Organisation of the Thesis

Chapter 2 proceeds by presenting the one-dimensional Serre equations in conservation law form with a source term. The dispersion and conservation properties and known analytic solutions of the Serre equations are also presented. The forced Serre equations and the concept of forced solutions are introduced. Followed by a summary of the main results of my investigation into the behaviour of the Serre equations in the presence of steep gradients in the free surface [17].

This is followed by Chapter 3 which describes in detail the FEVM. In this thesis the results of other numerical methods are also provided. Descriptions of these methods can be found in the literature [14, 17].

Chapter 4 provides a linear analysis of the convergence and dispersion properties of the FEVM in detail. The analysis begins with the linearised Serre equations over a horizontal bed and then derives the evolution matrix; through which the convergence and dispersion properties can be studied. The results of the linear analysis are also provided for all the methods used by Pitt et al. [17] as a comparison.

The convergence and conservation properties of the FEVM and all the numerical methods used by Pitt et al. [17] are then assessed in Chapter 5 using analytic and forced solutions of the Serre equations. While in Chapter 6 the FEVM and the second-order FDVM are validated against experimental results.

Chapter 7 summarises the major contributions and findings of this thesis and provides ideas for future work.

Appendix A provides expressions for the total amount of the conserved quantities over any domain for all of the analytic solutions described in this thesis.

The basis functions used by the FEVM and the precise definition of the function spaces mentioned in this thesis are provided in Appendix B.

Appendix C provides the evolution matrix produced by the linear analysis for the first-, second- and third-order FDVM and two second-order finite difference methods. It also includes the consistency results for these methods, as they were omitted from Chapter 4.

Finally the publications I contributed to throughout my research and their abstracts are provided in chronological order in Appendix D.

# Chapter 2

## The Serre Equations

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

The Serre equations are a system of partial differential equations that describe the free-surface waves of fluids whose motion is dominated by gravitational forces. They are an approximation to the Euler equations [1]; describing waves in shallow water when the characteristic depth of the water  $h_0$  is much smaller than the characteristic wavelength of the waves  $\lambda_0$  so that the shallowness parameter  $\sigma = h_0/\lambda_0 \ll 1$ . Typically, water is considered to be shallow when  $\sigma \leq 1/20$  [18]. Tsunamis and storm surges are shallow water waves as the typical ocean depth in which they occur is  $4\text{km}$  and both can have wavelengths up to  $100\text{km}$  long.

The Serre equations for one-dimensional flows over horizontal beds were first derived by Serre [19] using asymptotic expansion. They were later derived using depth integration by Su and Gardner [20]. They are equivalent to the Green-Naghdi equations [21] derived using the theory of directed fluid sheets. The Serre equations were then extended to spatially varying bathymetry by Seabra-Santos et al. [22].

The Serre equations are fully non-linear and thus applicable across the entire range of wave amplitudes  $a_0$  which are usually characterised using the non-linearity parameter  $\epsilon = a_0/h_0$ . The fluid described by the Serre equations possesses a non-hydrostatic pressure distribution and is dispersive in nature, as are real fluids. Furthermore, the dispersion relationship of linear waves of the Serre equations well approximates the linear wave theory for the Euler equations [23]. For these reasons the Serre equations are considered one of the best approximate water wave models up to wave-breaking [7, 24].

In this chapter we present the one-dimensional Serre equations and a refor-

mulation of these equations into conservation law form with a source term. We then present the relevant properties of the Serre equations and introduce the forced solutions which are used to assess the validity of the numerical methods. Finally, the contribution of my research [17] to understanding the behaviour of steep gradients in the free-surface for the Serre equations is summarised.

## 2.1 The One-Dimensional Serre Equations

In this thesis we take the Serre equations as derived from the depth-integration approach [20, 22]. Given the extent of the literature already available for the derivation of these equations we will only introduce the relevant quantities and present the equations. The one-dimensional Serre equations describe the behaviour of unsteady free surface fluid flow for an inviscid fluid with a constant density  $\rho$  neglecting wave-breaking. In this thesis we will neglect bottom friction effects.

The primitive variables of the Serre equations are the height  $h(x, t)$  of a column of fluid above the stationary bed profile given by  $b(x)$  and the average horizontal velocity  $u(x, t)$  of the column of fluid which are all shown in Figure 2.1. The stage  $w(x, t) = h(x, t) + b(x)$  gives the absolute location of the free surface. Additionally, to relate the primitive variables of the Serre equations to the intuitive variables of the Euler equations we introduce the horizontal velocity  $u'(x, z, t)$  and the vertical velocity  $v'(x, z, t)$  of a fluid particle.

The derivation is similar to that of the Shallow Water Wave Equations (SWWE) [25]. In particular, we assume that the horizontal velocity of a fluid particle  $u'(x, z, t)$  inside a column of fluid equals the horizontal velocity of the column of fluid  $u(x, t)$ . By integrating the conservation of mass equation for an inviscid fluid with a no-slip condition at the bed we obtain the distribution of the vertical velocity of a fluid particle throughout a column of fluid [13]

$$v'(x, z, t) = u \frac{\partial b}{\partial x} - (z - b) \frac{\partial u}{\partial x}. \quad (2.1)$$

Unlike the SWWE the vertical velocity of a fluid particle in the Serre equations is not zero throughout the depth of water. This leads to a non-hydrostatic pressure distribution on the fluid particles which is derived by integrating the conservation of vertical momentum equation of the Euler equations with a no-slip condition

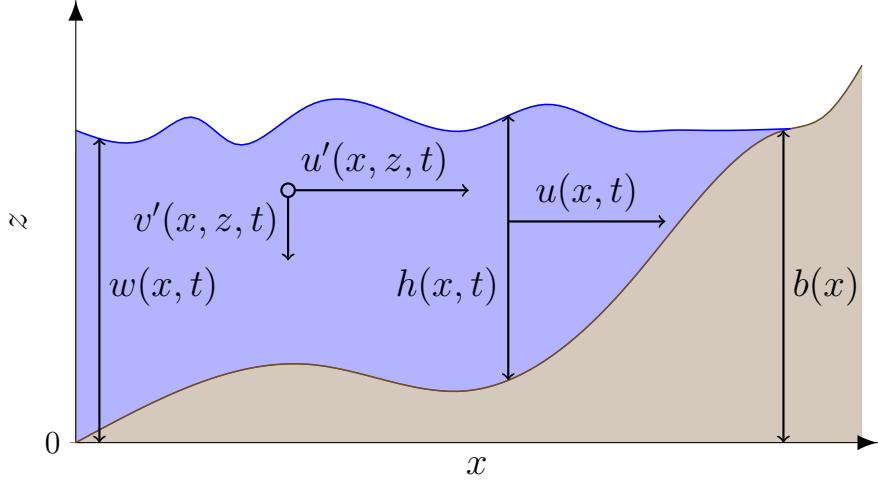


Figure 2.1: Diagram demonstrating a free surface flow (□) over a bed (■) where  $w(x, t)$  is the absolute location of the free surface,  $v'(x, z, t)$  is the vertical velocity of a fluid particle,  $u'(x, z, t)$  is the horizontal velocity of a fluid particle,  $h(x, t)$  is the height of a column of fluid,  $u(x, t)$  is the average horizontal velocity of a column of fluid and  $b(x)$  is the stationary bed profile.

at the bed [13]

$$p'(x, z, t) = \underbrace{\rho g (h + b - z)}_{\text{hydrostatic pressure}} + \rho (h + b - z) \Psi + \frac{1}{2} \rho (h^2 - [z - b]^2) \Phi \quad (2.2)$$

where

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

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

The hydrostatic pressure term  $\rho g (h + b - z)$  is the pressure due to the weight of water above the fluid particle. The non-hydrostatic pressure terms are a consequence of the non-zero vertical velocity which modifies the underlying hydrostatic pressure distribution causing dispersion of waves on the free surface.

Integrating the Euler equations [13, 20] over the entire depth with a no-slip condition at the bed, a free surface condition at the free surface, the vertical velocity relation (2.1) and the pressure distribution (2.2) we obtain the Serre

equations

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

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

which are depth integrated approximations to the conservation of mass and horizontal momentum equations of the Euler equations. When  $\Phi = \Psi = 0$  the Serre equations reduce to the SWWE where the vertical velocity is zero over the depth, the pressure distribution is hydrostatic and there is no dispersion.

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

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

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

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

### 2.1.1 Alternative Form

A major hurdle for developing numerical methods for the Serre equations is the presence of the temporal derivative in  $\Phi$  and  $\Psi$  (2.3). Using the quantity  $G$  [13, 26, 27] the Serre equations can be rewritten in conservation law form with a

source term

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

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

where

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

The introduced quantity  $G$  resembles  $h$  multiplied by the irrotationality [28, 29].

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

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

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

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

with

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

## 2.2 Properties

The Serre equations possess a number of desirable properties for the modelling of water waves; in particular their conservation of fundamental quantities and their dispersion relation. If a numerical method accurately approximates the

Serre equations then the numerical method should reproduce the conservation and dispersion properties of the Serre equations. Furthermore, the numerical method should accurately reproduce the analytic solutions of the Serre equations. In this thesis the conservation and dispersion properties and the analytic solutions of the Serre equations are used to assess the veracity of numerical methods.

### 2.2.1 Conservation Properties

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

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

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

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

Integrating the Serre equations in both non-conservation law form (2.4) and conservation law form (2.6) for a closed system we get that  $h$ ,  $uh$  and  $G$  are conserved by the Serre equations. Additionally, the Green-Naghdi equations [21] which are equivalent to the Serre equations for one-dimensional flows were derived by conserving the energy

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

Therefore, the one-dimensional Serre equations should also conserve  $\mathcal{H}$ . The energy  $\mathcal{H}$  is the sum of the gravitational potential energy

$$\frac{1}{2} \int_b^{h+b} gz dx = \frac{1}{2} gh(h + 2b),$$

the horizontal kinetic energy

$$\frac{1}{2} \int_b^{h+b} (u')^2 dx = \frac{1}{2} hu^2$$

and the vertical kinetic energy

$$\frac{1}{2} \int_b^{h+b} (v')^2 dx = \frac{1}{2} \left( \frac{h^3}{3} \left[ \frac{\partial u}{\partial x} \right]^2 + u^2 h \left[ \frac{\partial b}{\partial x} \right]^2 - uh^2 \frac{\partial u}{\partial x} \frac{\partial b}{\partial x} \right)$$

for a column of fluid. Where the vertical velocity of a fluid particle  $v'$  for the Serre equations is given by (2.1). For horizontal beds  $\mathcal{H}$  is the Hamiltonian of the Serre equations [30].

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

### 2.2.2 Dispersion Properties

The dispersion properties are studied by linearising the Serre equations with a horizontal bed, assuming periodic wave solutions and then deriving a relationship between the frequency  $\omega$  and the wave number  $k$  of these solutions. For the Serre equations the dispersion relation [27] is

$$\omega^\pm = U k \pm k \sqrt{gH} \sqrt{\frac{3}{(kH)^2 + 3}} \quad (2.9)$$

where  $U$  and  $H$  are the mean velocity and height of the fluid respectively. For the dispersion relation two wave frequencies  $\omega^+$  and  $\omega^-$  are possible for each wavenumber allowing for downwind and upwind travelling waves respectively. Barthélemy [23] compared this dispersion relation to the dispersion relation given by the linear theory for water waves and demonstrated its utility when  $k$  is small. However, when  $k$  is large the difference between the dispersion relation of the Serre equations and that of the linear water wave theory increases.

From the dispersion relation (2.9) the phase velocity  $v_p^\pm = \omega^\pm/k$  and the group velocity  $v_g^\pm = \partial\omega^\pm/\partial k$  can be written in terms of the wave number as

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

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

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

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

Therefore, we have that

$$U - \sqrt{gH} \leq v_p^- \leq U \leq v_p^+ \leq U + \sqrt{gH}, \quad (2.10a)$$

$$U - \sqrt{gH} \leq v_g^- \leq U \leq v_g^+ \leq U + \sqrt{gH} \quad (2.10b)$$

so that the speed of waves in the Serre equations are bounded by the speed of waves in the SWWE.

### 2.2.3 Analytic Solutions

Currently few analytic solutions have been discovered for the Serre equations. There is a travelling wave solution for a horizontal bed [31] and the trivial stationary lake at rest solution for arbitrary bathymetry.

#### Solitary Travelling Wave Solution

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

$$h(x, t) = a_0 + a_1 \operatorname{sech}^2(\kappa [x - ct]), \quad (2.11a)$$

$$u(x, t) = c \left( 1 - \frac{a_0}{h(x, t)} \right), \quad (2.11b)$$

$$b(x) = 0 \quad (2.11c)$$

with

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

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

From these equations the total amounts of  $h$  (A.1a),  $uh$  (A.1b),  $G$  (A.1c) and  $\mathcal{H}$  (A.1d) at  $t = 0s$  can be derived. These expressions are provided in Appendix A.

This solitary wave solution has an amplitude of  $a_1$ , an infinite wavelength and propagates on water  $a_0$  deep. It is one member of a family of smooth periodic travelling wave solutions [31]. Note that these solitary wave solutions are not true solitons, due to their inelastic collisions with one another [32].

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

### Lake at Rest

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

$$h(x, t) = \max \{a_0 - b(x), 0\}, \quad (2.12a)$$

$$u(x, t) = 0, \quad (2.12b)$$

$$G(x, t) = 0. \quad (2.12c)$$

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. With the maximum function included in the water depth to allow for dry regions of the bed when  $b(x) > a_0$ . We write these quantities in terms of  $b(x)$  as this solution holds for all bed profiles. The corresponding total amounts of  $h$  and  $\mathcal{H}$  can be calculated by summing their total amounts in wet regions which given by the integrals (A.2). While the total amounts of  $u$  and  $G$  are zero. The expressions to calculate the total amount of the conserved quantity over any domain are provided in Appendix A.

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

$$\frac{\partial h}{\partial t} = 0, \quad \frac{\partial G}{\partial t} = 0.$$

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

For naive numerical methods for the Serre equations the hydrostatic pressure terms do not balance the bed slope terms. This causes the numerical solutions of an initially still lake to produce non-physical velocities, degrading their convergence to the solution. To combat this, modifications are made to the flux

and source term approximations so that these terms do balance, leading to a so called ‘well-balanced’ method. This analytic solution then provides a test for the effectiveness of these well-balancing modifications to the numerical methods.

## 2.3 Forced Solutions

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

Currently the verification of the order of accuracy of the numerical methods for transient solutions with varying bathymetry requires the use of forced solutions. To do this we select some particular functions for all of the primitive quantities;  $h$ ,  $u$  and  $b$  which we denote  $h^*$ ,  $u^*$  and  $b^*$  respectively. To force these functions  $h^*$ ,  $u^*$  and  $b^*$  to be solutions of the Serre equations (2.6) we add the terms  $S_h$  and  $S_G$  to obtain the forced Serre equations

$$\frac{\partial h}{\partial t} + \frac{\partial(uh)}{\partial x} + S_h = 0, \quad (2.13a)$$

$$\frac{\partial G}{\partial t} + \frac{\partial}{\partial x} \left( uG + \frac{gh^2}{2} - \frac{2}{3}h^3 \left[ \frac{\partial u}{\partial x} \right]^2 + h^2u \frac{\partial u}{\partial x} \frac{\partial b}{\partial x} \right) \\ + \frac{1}{2}h^2u \frac{\partial u}{\partial x} \frac{\partial^2 b}{\partial x^2} - hu^2 \frac{\partial b}{\partial x} \frac{\partial^2 b}{\partial x^2} + gh \frac{\partial b}{\partial x} + S_G = 0 \quad (2.13b)$$

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

where

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

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

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

Note that since the choice of the forced solutions  $h^*$ ,  $u^*$  and  $b^*$  is arbitrary the solutions of the forced Serre equations need not be conservative or retain any properties of the underlying Serre equations.

## 2.4 Behaviour in the Presence of Steep Gradients

To ensure that the developed numerical methods are robust, their capability to handle initial condition value problems with quantities possessing discontinuities must be tested. One group of these initial condition value problems that has been of particular interest to the water wave community is the dam-break problem [26, 31, 33, 34, 35]. The dam-break problem consists of an initially still body of water with a discontinuous jump in its surface between two depth values. So that

$$h(x, 0) = \begin{cases} h_l & x < x_0 \\ h_r & x \geq x_0, \end{cases} \quad (2.14a)$$

$$u(x, 0) = 0, \quad (2.14b)$$

$$G(x, 0) = 0, \quad (2.14c)$$

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

where  $h_l$  and  $h_r$  are the water depth to the left and right of  $x_0$  respectively.

Currently, these dam-break problems (2.14) have no known analytic solutions for the Serre equations. However, some insight into the behaviour of the evolution of these initial condition problems has been gained from asymptotic [31] and linear [36] analyses.

There have been a number of numerical solutions to dam-break problems presented in the literature [26, 31, 33, 34, 35] which have used a variety of numerical methods. Some of these numerical methods could not handle discontinuous initial conditions [31, 33, 34, 35] and so smooth approximations to the initial conditions (2.14) were employed. The variety of numerical approaches has lead to different conclusions about the behaviour of the evolution of dam-break problems for the Serre equations in the literature. To resolve these differences a comprehensive

review of a particular dam-break problem with a variety of numerical methods and smoothing of the initial conditions was performed [17].

The convergence and conservation properties of many numerical solutions of the first-, second- and third-order finite difference volume methods as well as two second-order finite difference methods were investigated [17]. All the numerical solutions affirmed the existence of a previously unreported behaviour of the Serre equations in the presence of steep gradients over short time spans. This new behaviour was justified using the convergence of the numerical solutions to one another as the resolution of the method increased. Since this convergence was observed for all high-order accurate methods; the likelihood that this new behaviour was a result of a particular quirk of the method was small. Indicating that this behaviour is representative of the true solution of the Serre equations in the presence of steep gradients.

The analytic solution of the Serre equations to the dam-break problem is not currently known. These numerical solutions [17] currently provide the best approximation to the solution of the Serre equations to the dam-break problem over short time spans. Furthermore, these results demonstrate the effect of smoothing the initial conditions and low grid resolutions on the observed behaviours. This allows others to appropriately choose the smoothness of the initial conditions and the resolution of their numerical method.

The relevant results garnered from the asymptotic [31] and linear [36] analyses for the evolution of the dam-break problem are presented here, followed by a summary of the numerical results [17], which constituted a significant portion of my research. Since the rest of this thesis is focused around the development and validation of the second-order finite element volume method and the paper [17] is focused on other methods, we provide only a brief summary of this work in this thesis.

## Asymptotic and Linear Results

The asymptotic analysis of El et al. [31] used Whitham modulation to study the evolution of dispersive shock waves of the Serre equations as  $t \rightarrow \infty$ . Because, a dispersive shock wave is generated in the evolution of the dam-break problem in the Serre equations; these results are very useful. In particular, they provide a relationship between the initial heights of the dam-break problem  $h_l$  and  $h_r$  and the amplitude  $A^+$  and speed  $S^+$  of the leading wave in the resulting dispersive

wave train. These quantities are obtained by solving

$$\frac{\Delta}{(A^+ + 1)^{1/4}} - \left( \frac{3}{4 - \sqrt{A^+ + 1}} \right)^{21/10} \left( \frac{2}{1 + \sqrt{A^+ + 1}} \right)^{2/5} = 0, \quad (2.15a)$$

$$S^+ = \sqrt{g(A^+ + 1)} \quad (2.15b)$$

where

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

These expressions were found to agree well with numerical simulations provided that  $\Delta < 1.43$  [31].

For the linearised Serre equations we obtain the dispersion relation (2.9) from which the phase and group velocities can be determined (2.10). The negative and positive branches of the phase and group velocities are separated, implying that there should be a separation of the upwind and downwind parts of the dispersive wave-train [36]. Therefore, the structure of dispersive shock waves of the Serre equations should be two separate dispersive wave trains.

### Numerical Solutions for the Smoothed Dam-break Problem

To resolve the differences present in the literature a variety of numerical methods were used to solve the most common class of smoothed versions of the dam-break problem initial conditions (2.14). This was termed the smoothed dam-break problem and is given by

$$h(x, 0) = h_r + \frac{h_l - h_r}{2} \left( 1 + \tanh \left( \frac{x_0 - x}{\alpha} \right) \right), \quad (2.16a)$$

$$u(x, 0) = 0, \quad (2.16b)$$

$$G(x, 0) = 0, \quad (2.16c)$$

$$b(x) = 0 \quad (2.16d)$$

where  $\alpha$  controls the width of the transition from  $h_l$  to  $h_r$  and thus the steepness of the initial gradient in the water surface. Most of the numerical simulations were focused on solutions for  $h_l = 1.8m$ ,  $h_r = 1m$  and  $x_0 = 500m$  with a final time of  $t = 30s$ . The smoothing parameter  $\alpha$  and the resolution of the methods were varied to investigate their influence on the observed behaviour of the numerical solution. Four structures were observed in the numerical solutions; the

non-oscillatory structure, the flat structure, the node structure and the growth structure. Numerical solutions at  $t = 30s$  for the mentioned  $h_l$ ,  $h_r$  and  $x_0$  values demonstrating examples of the observed structures are shown in Figure 2.2.

The growth structure was previously not reported in the literature. The growth structure was consistently observed in numerical solutions of the smoothed dam-break problem for high-order accurate methods as  $\alpha$  and the grid resolution were increased. Therefore, the growth structure represents the true structure of the solution of the Serre equations for the dam-break problem with the corresponding  $h_l$  and  $h_r$  values at  $t = 30s$ . This conclusion was supported by measuring the convergence of the numerical solutions to one another and observing the conservation of the conserved quantities as the resolution was increased.

The finite difference volume methods introduced diffusive errors caused by their flux approximation, so that the amplitude and number of oscillations in the dispersive wave train increased as the resolution increased. While the finite difference methods introduce dispersive errors decreasing the amplitude and number of oscillations in the dispersive wave train as the resolution was increased. Therefore, the numerical solutions of both classes of methods serve as upper and lower bounds on the amplitude and number of oscillations in the dispersive wave train. These bounds further demonstrated that the observed growth structure best represented the true solution of the Serre equations to this dam-break problem.

The observation of other behaviours at  $t = 30s$  is caused by; small  $\alpha$  values which overly smooth the initial conditions, low-order numerical methods introducing large diffusive errors and low numerical resolutions which cannot resolve the high frequency waves observed in the growth structure. These structures exist on a spectrum where the severity of these effects determine the observed behaviour. So that, the most severe damping effects produced the non-oscillatory structure and the least severe effects produced the growth structure. These effects explained the observations of different structures previously present in the literature [26, 31, 33, 34].

The differences in the observed structures are primarily driven by the different internal structures of the dispersive shock wave, so that for the flat, node and growth structure in Figure 2.2 the front of the dispersive wave trains are indistinguishable. Therefore, the results of numerical solutions that have not resolved all the internal structure present in the growth structure still agree well with the Whitham modulation results (2.15) of El et al. [31].

The amplitude of waves at the centre of the growth structure decays over time, resulting in the observation of the flat structure when  $t$  is large. These

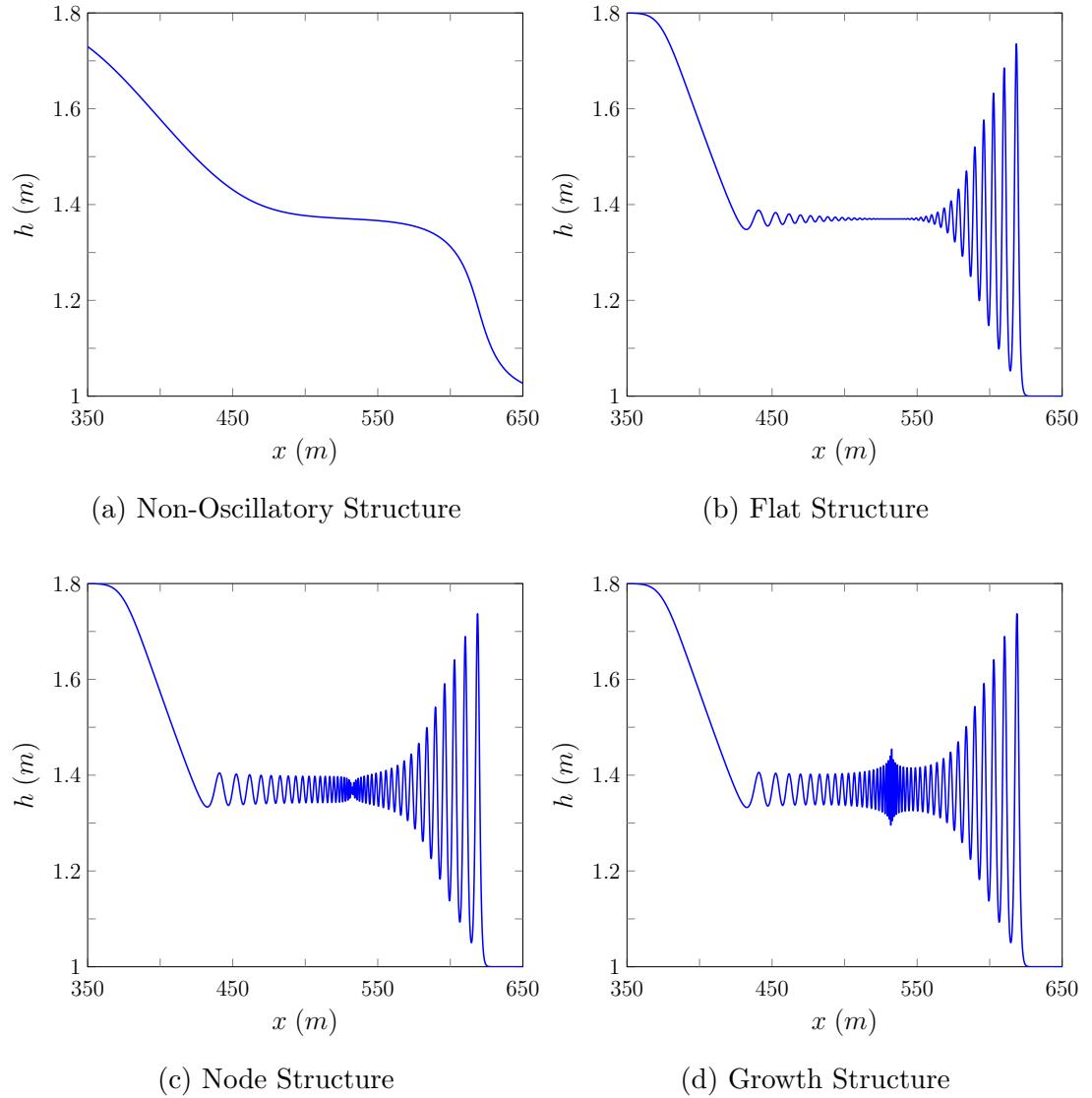


Figure 2.2: Examples of the different structures observed in numerical solutions to the smoothed dam-break problem (2.16) displayed by Pitt et al. [17].

results agree with the linear argument put forth by Dougalis et al. [36]. This indicates that for smaller times the non-linear terms of the Serre equations play a significant role in the evolution of steep gradients. While for long times the linear terms are dominant and thus a separation of the dispersive wave trains is observed.

In this chapter the Serre equations and their relevant properties were given. The forced Serre equations were introduced and a summary of the main results for the evolution of steep gradients in the free surface was provided. We now will describe the finite element volume method, whose development was a primary objective of this thesis.

# Chapter 3

## Finite Element Volume Method

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

A variety of numerical methods have been used to solve the Serre equations; from complete finite difference methods [31, 37] and finite element methods [27, 33, 34] to combinations of finite difference and finite volume methods [14, 26]. Splitting techniques have also been employed, most commonly to split the Serre equations into their non-linear and dispersive parts; resulting in an elliptic operator for the dispersive part and the SWWE for the non-linear part [32, 38, 39].

Numerical methods that make use of the conservation law form of the Serre equations (2.6) [14, 26, 27] are the most promising for the two dimensional Serre equations with variable bathymetry. The primary reason for this is that these methods are robust and extend well to unstructured meshes with complex geometries which are the meshes most commonly used for modelling physical scenarios. Secondly, to properly handle the elliptic operator produced by the non-linear and dispersive splitting requires overly restrictive assumptions about the smoothness of the physical quantities, particularly the water depth.

I have developed an extension of the Finite Difference Volume Methods (FDVM) [14, 26] that uses a finite element method in place of the finite difference method. This second-order FEVM which will be referred to as FEVM<sub>2</sub> was a main objective of this thesis; it consists of two main parts a Finite Element Method (FEM) to solve (2.7) and a Finite Volume Method (FVM) to solve (2.6) hence its name. Making use of these two methods results in a numerical method with a number of desirable properties. It is robust in the presence of steep gradients in the free surface [17], robust during the wetting and drying of beds and the FEM provides a complete profile of the horizontal velocity inside a cell. This last point indicates

that this method is the ideal variant of the finite volume based methods [14] to be extended to solve the two-dimensional Serre equations on unstructured meshes with parallelised code.

In addition to the FEVM<sub>2</sub>, the first- and second-order FDVM of Le Métayer et al. [26] and Zoppou et al. [14] were reproduced. These methods will be referred to as FDVM<sub>1</sub> and FDVM<sub>2</sub> respectively. Furthermore, the third-order FDVM<sub>3</sub> was implemented during my research. I have also reproduced the second-order naive finite difference method [40] and the finite difference method of El et al. [31]; which I refer to as  $\mathcal{D}$  and  $\mathcal{W}$  respectively. Descriptions of these methods have already been published [14, 17] and therefore, are omitted from this thesis.

### 3.1 Notation for Numerical Grids

In the FEVM<sub>2</sub>, time and space will be discretised in different ways. Time is broken up into time levels separated by a constant duration  $\Delta t$  and space is broken up into cells of constant width  $\Delta x$ . The FEVM can be extended to allow for varying  $\Delta t$  and  $\Delta x$  values, with this description restricted to the constant case for simplicity. The notation for time is quite simple; from an initial time  $t^0$  we define the  $n^{th}$  time level where  $n \in \mathbb{N}$  to be

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

The goal of FEVM<sub>2</sub> is to update the quantities at the current time level  $t^n$  to the next time level  $t^{n+1}$  by solving the Serre equations.

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

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

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

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

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

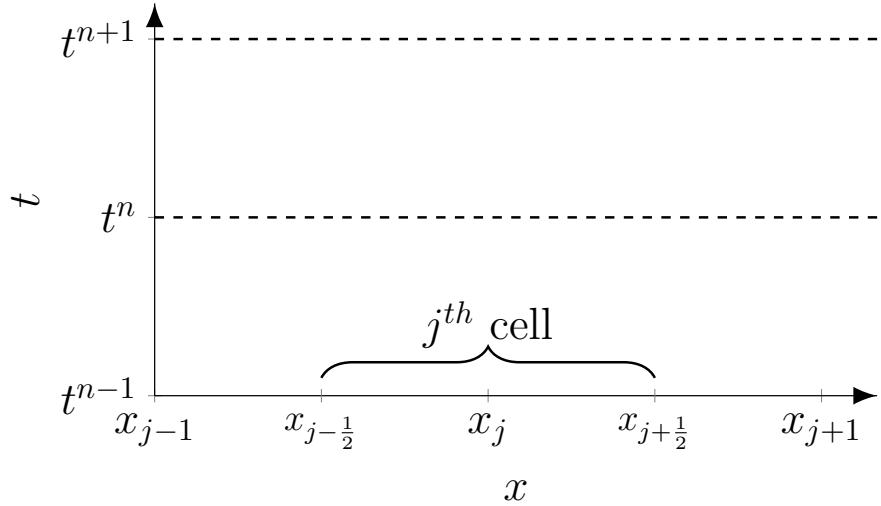


Figure 3.1: Diagram of the time levels  $t^{n-1}$ ,  $t^n$  and  $t^{n+1}$  at which the numerical solution of the Serre equations will be calculated. Also shown is the  $j^{th}$  cell with midpoint  $x_j$  spanning  $x_{j-1/2}$  to  $x_{j+1/2}$  which is a volume of the FVM and an element of the FEM.

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

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

These are the nodal values of  $q$ . Since the FEVM uses a FVM the cell averages of the quantities are also required. We define the average of a quantity over the  $j^{th}$  cell at time level  $t^n$  as

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

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

## 3.2 Structure Overview

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

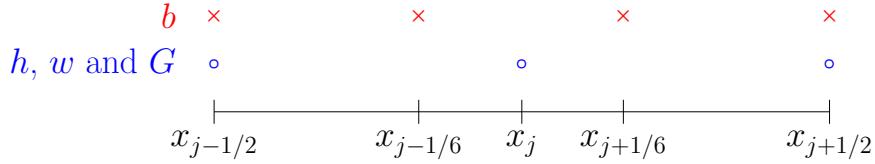


Figure 3.2: The locations of the reconstructions for  $h$ ,  $w$ ,  $G$  (○) and  $b$  (✖) inside the  $j^{th}$  cell.

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

$$\bar{\mathbf{q}}^n = \begin{bmatrix} \bar{q}_0^n \\ \bar{q}_1^n \\ \vdots \\ \bar{q}_m^n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{bmatrix}$$

where  $q$  is a generic quantity representing the vectors for  $h$ ,  $G$  and  $w$ . The evolution step proceeds by (i) reconstructing the quantities over the cell, (ii) calculating the fluid velocity, (iii) approximating the flux, (iv) approximating the source term, (v) updating the cell averages and then (vi) applying second-order Strong Stability Preserving (SSP) Runge-Kutta time stepping.

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

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

$$\begin{aligned} q_{j\pm 1/2}^\pm &= \mathcal{R}_{j\pm 1/2}^\pm(\bar{\mathbf{q}}^n), & b_{j\pm 1/2} &= \mathcal{B}_{j\pm 1/2}(\mathbf{b}), \\ q_j &= \mathcal{R}_j(\bar{\mathbf{q}}^n), & b_{j\pm 1/6} &= \mathcal{B}_{j\pm 1/6}(\mathbf{b}). \end{aligned}$$

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

$$\hat{\mathbf{q}} = \begin{bmatrix} q_{-1/2}^+ \\ q_0 \\ q_{1/2}^- \\ \vdots \\ q_{m+1/2}^- \end{bmatrix}, \quad \hat{\mathbf{b}} = \begin{bmatrix} b_{-1/2} \\ b_{-1/6} \\ b_{1/6} \\ b_{1/2} \\ \vdots \\ b_{m+1/2} \end{bmatrix}$$

where  $q$  is a generic quantity demonstrating these vectors for  $h$ ,  $w$  and  $G$ .

- (ii) Fluid Velocity: The remaining unknown quantity, the horizontal velocity of the fluid column,  $u$  is calculated at  $x_{j-1/2}$ ,  $x_j$  and  $x_{j+1/2}$  in each cell by solving (2.7) with a second-order FEM. We denote the solution map of the FEM by  $\mathcal{G}$ , which takes  $\hat{\mathbf{h}}$ ,  $\hat{\mathbf{G}}$  and  $\hat{\mathbf{b}}$  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 temporally averaged fluxes  $F_{j-1/2}^n$  and  $F_{j+1/2}^n$  across the cell boundaries  $x_{j-1/2}$  and  $x_{j+1/2}$  using  $\mathcal{F}_{j-1/2}$  and  $\mathcal{F}_{j+1/2}$ , so that

$$F_{j\pm 1/2}^n = \mathcal{F}_{j\pm 1/2}(\hat{\mathbf{h}}, \hat{\mathbf{G}}, \hat{\mathbf{b}}, \hat{\mathbf{u}}).$$

- (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  $\mathcal{S}$

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

- (v) Update Cell Averages: We update the cell average values from time  $t^n$  to  $t^{n+1}$  with a forward Euler approximation, resulting in a method that is second-order accurate in space and first-order in time.
- (vi) Second-Order SSP Runge-Kutta Method: We repeat steps (i)-(v) and use SSP Runge-Kutta time stepping to obtain  $\bar{\mathbf{h}}$  and  $\bar{\mathbf{G}}$  at  $t^{n+1}$  with second-order accuracy in space and time.

### 3.2.1 Reconstruction

We now provide details for the reconstruction of  $h$ ,  $w$ ,  $G$  and  $b$  in the  $j^{th}$  cell at the locations shown in Figure 3.2. For the purposes of the reconstruction we will assume that the mesh is structured. The reconstruction methods described here can be extended to unstructured meshes through generalisations of the interpolation techniques employed here [41, 42]. For  $h$ ,  $w$  and  $G$  the reconstructions are performed from the cell averages. While  $b$  is reconstructed from the nodal values.

#### Reconstruction of the $h$ , $w$ and $G$

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

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

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

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

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  $\theta$  parameter changes the diffusion introduced by the reconstruction. When  $\theta = 1$  the reconstruction introduces the most diffusion and is equivalent to the minmod reconstruction [43]. When  $\theta = 2$  the reconstruction introduces the least diffusion and is equivalent to the monotized central reconstruction [44].

**Definition 3.1.** The minmod function

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

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

The non-linear limiting used to calculate  $d_j$  ensures that the reconstruction of  $h$ ,  $w$  and  $G$  inside the cell is Total Variation Diminishing (TVD) [45], hence it does

not introduce non-physical oscillations. The TVD property of this reconstruction is achieved by constraining the slope  $d_j$  to zero near local extrema, resulting in a piecewise constant reconstruction which is TVD. Away from local extrema  $d_j$  will be the gradient with the smallest absolute value, making our reconstruction second-order accurate.

The reconstruction operator  $\mathcal{R}_j$  is second-order accurate regardless of the presence of local extrema. This can be seen through the error analysis of the midpoint quadrature rule [46] for which we have that

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

### Reconstruction of the Bed Profile

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

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

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

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

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

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

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

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

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

We require a continuous bed profile and so we average the two reconstructions at the cell edge from the adjacent cells. Therefore, our reconstruction of the bed profile in the  $j^{th}$  cell is the cubic which takes these values

$$b_{j-1/2} = \mathcal{B}_{j-1/2}(\mathbf{b}) = \frac{1}{2} (C_j(x_{j-1/2}) + C_{j-1}(x_{j-1/2})) , \quad (3.4a)$$

$$b_{j-1/6} = \mathcal{B}_{j-1/6}(\mathbf{b}) = C_j(x_{j-1/6}), \quad (3.4b)$$

$$b_{j+1/6} = \mathcal{B}_{j+1/6}(\mathbf{b}) = C_j(x_{j+1/6}), \quad (3.4c)$$

$$b_{j+1/2} = \mathcal{B}_{j+1/2}(\mathbf{b}) = \frac{1}{2} (C_j(x_{j+1/2}) + C_{j+1}(x_{j+1/2})) . \quad (3.4d)$$

### 3.2.2 Fluid Velocity

In the FEVM we solve for the primitive variable  $u$  given  $h$ ,  $G$  and  $b$  using a finite element approximation to (2.7). For the FEM we begin with the weak form of (2.7) using test function  $v$  over the spatial domain  $\Omega$  which is

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

Integrating this equation by parts with zero Dirichlet boundary conditions we get

$$\begin{aligned} \int_{\Omega} Gv \, dx &= \int_{\Omega} uh \left( 1 + \left[ \frac{\partial b}{\partial x} \right]^2 \right) v \, dx + \int_{\Omega} \frac{1}{3} h^3 \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} \, dx \\ &\quad - \int_{\Omega} \frac{1}{2} uh^2 \frac{\partial b}{\partial x} \frac{\partial v}{\partial x} \, dx - \int_{\Omega} \frac{1}{2} h^2 \frac{\partial b}{\partial x} \frac{\partial u}{\partial x} v \, dx. \end{aligned} \quad (3.5)$$

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

We simplify (3.5) by performing the integration over the cells and then sum-

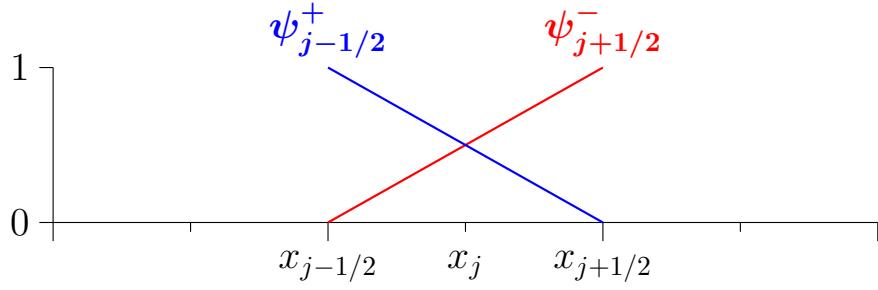


Figure 3.3: Support of the discontinuous linear basis functions  $\psi$  which are non-zero over the  $j^{th}$  cell.

ming the integrals together to get the equation for the entire domain

$$\sum_{j=0}^m \left( \int_{x_{j-1/2}}^{x_{j+1/2}} \left[ \left( uh \left( 1 + \left[ \frac{\partial b}{\partial x} \right]^2 \right) - \frac{1}{2} h^2 \frac{\partial b}{\partial x} \frac{\partial u}{\partial x} - G \right) v + \left( \frac{1}{3} h^3 \frac{\partial u}{\partial x} - \frac{1}{2} uh^2 \frac{\partial b}{\partial x} \right) \frac{\partial v}{\partial x} \right] dx \right) = 0 \quad (3.6)$$

which holds for all test functions  $v$ . The next step is to replace the functions for  $h$ ,  $G$ ,  $b$ ,  $v$  and  $u$  with their corresponding basis function approximations. Expressions for these basis functions are provided in Appendix B.

### Basis Function Approximations

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

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

$$q = \sum_{j=0}^m \left( q_{j-1/2}^+ \psi_{j-1/2}^+ + q_{j+1/2}^- \psi_{j+1/2}^- \right). \quad (3.7)$$

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

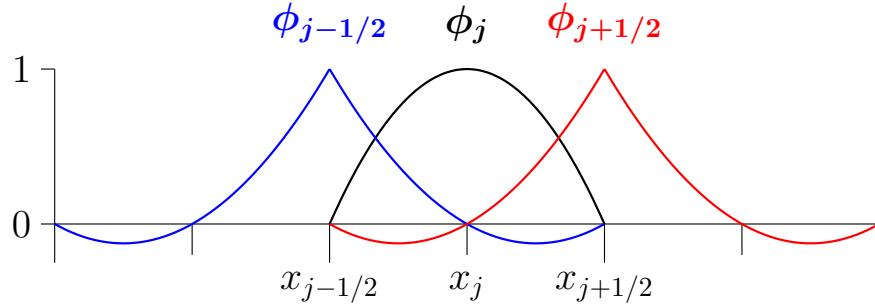


Figure 3.4: Support of the continuous piecewise quadratic basis functions  $\phi$  which are non-zero over the  $j^{th}$  cell.

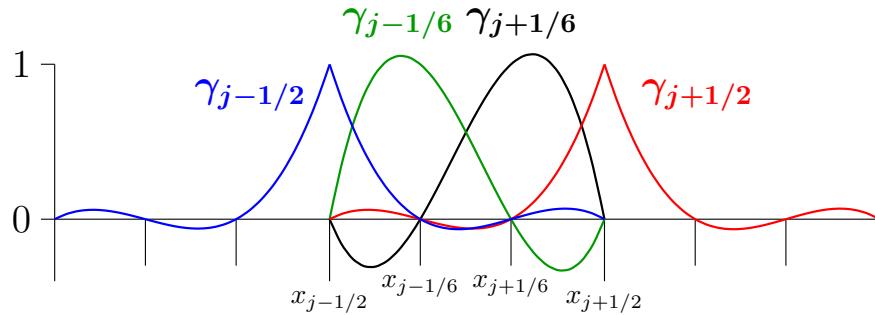


Figure 3.5: Support of the continuous piecewise cubic basis functions  $\gamma$  which are non-zero over the  $j^{th}$  cell.

Therefore, we use the continuous quadratic basis functions  $\phi_{j\pm 1/2}$  and  $\phi_j$  (B.2) depicted in Figure 3.4.

From the basis functions  $\phi$  our basis function approximation to  $u$  is

$$u = u_{-1/2}\phi_{-1/2} + \sum_{j=0}^m (u_j\phi_j + u_{j+1/2}\phi_{j+1/2}). \quad (3.8)$$

For the source term of the evolution of  $G$  equation (2.6b) we require a local approximation to the second derivative of the bed that is also second-order accurate. To allow for an appropriate second derivative of the bed profile,  $b$  must be a member of  $\mathbb{W}^{2,2}(\Omega)$  which is smoother than required by (3.5). We choose the cubic basis functions  $\gamma$  (B.3) which are continuous across the cell edges, as the bed profile will be continuous. These basis functions are shown in Figure 3.5 and from them we get our basis function approximation to  $b$

$$b = b_{-1/2}\gamma_{-1/2} + \sum_{j=0}^m (b_{j-1/6}\gamma_{j-1/6} + b_{j+1/6}\gamma_{j+1/6} + b_{j+1/2}\gamma_{j+1/2}). \quad (3.9)$$

### Calculation of Element-wise Matrices

The integral equation (3.6) holds for all  $v$ . However, since our solution space has the basis functions  $\phi$  it is sufficient to satisfy (3.6) for all  $\phi$  to generate the solution. Since only the basis functions  $\phi_{j-1/2}$ ,  $\phi_j$  and  $\phi_{j+1/2}$  are non-zero over the  $j^{th}$  cell we can calculate the  $j^{th}$  term in the sum (3.6) like so

$$\begin{aligned} & \int_{x_{j-1/2}}^{x_{j+1/2}} \left( \left[ uh \left( 1 + \left[ \frac{\partial b}{\partial x} \right]^2 \right) - \frac{1}{2} h^2 \frac{\partial b}{\partial x} \frac{\partial u}{\partial x} - G \right] \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} \right. \\ & \quad \left. + \left[ \frac{1}{3} h^3 \frac{\partial u}{\partial x} - \frac{1}{2} h^2 \frac{\partial b}{\partial x} u \right] \frac{\partial}{\partial x} \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} \right) dx \quad (3.10) \end{aligned}$$

where we use our finite element approximations for  $h$  (3.7),  $G$  (3.7),  $u$  (3.8) and  $b$  (3.9). This integral can be generalised by moving to the natural reference  $\xi$ -space, as the basis functions which are non-zero in one element are just translations of the non-zero basis functions in another element. The mapping from the  $x$ -space to the  $\xi$ -space is

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

Therefore, the  $j^{th}$  cell  $[x_{j-1/2}, x_{j+1/2}]$  gets mapped to  $[-1, 1]$  in the  $\xi$ -space. Making the change of variables from  $x$  to  $\xi$  in (3.10) we get

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

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

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

Since the integral is computed over  $[-1, 1]$ , there are only a few non-zero contributions from the finite element approximations to  $h$  and  $u$ , so the  $uh$  term becomes

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

Calculating the integrals of all the basis function combinations we get

$$\begin{aligned} & \frac{\Delta x}{2} \int_{-1}^1 uh \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} d\xi = \\ & \frac{\Delta x}{60} \begin{bmatrix} 7h_{j-1/2}^+ + h_{j+1/2}^- & 4h_{j-1/2}^+ & -h_{j-1/2}^+ - h_{j+1/2}^- \\ 4h_{j-1/2}^+ & 16h_{j-1/2}^+ + 16h_{j+1/2}^- & 4h_{j+1/2}^- \\ -h_{j-1/2}^+ - h_{j+1/2}^- & 4h_{j+1/2}^- & h_{j-1/2}^+ + 7h_{j+1/2}^- \end{bmatrix} \begin{bmatrix} u_{j-1/2} \\ u_j \\ u_{j+1/2} \end{bmatrix}. \end{aligned}$$

### 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  $Gv$  in (3.6) generate the element wise vector  $\mathbf{g}_j$ . These element wise matrices and vectors are then assembled into the global stiffness matrix  $\mathbf{A}$  and the global right hand side term  $\mathbf{g}$  thus (3.6) is rewritten as

$$\mathbf{A} \hat{\mathbf{u}} = \mathbf{g}. \tag{3.11}$$

This is a penta-diagonal matrix equation which can be solved by direct banded matrix solution techniques such as those of Press et al. [47] to obtain

$$\hat{\mathbf{u}} = \mathcal{G}(\hat{\mathbf{h}}, \hat{\mathbf{G}}, \hat{\mathbf{b}}) = \mathbf{A}^{-1}\mathbf{g} \quad (3.12)$$

as desired.

### 3.2.3 Flux Across the Cell Interfaces

We use the method of Kurganov et al. [48] 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 (2.10).

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 approximation of the flux term given by Kurganov et al. [48] is

$$F_{j+\frac{1}{2}} = \frac{a_{j+\frac{1}{2}}^+ f(q_{j+\frac{1}{2}}^-) - a_{j+\frac{1}{2}}^- f(q_{j+\frac{1}{2}}^+)}{a_{j+\frac{1}{2}}^+ - a_{j+\frac{1}{2}}^-} + \frac{a_{j+\frac{1}{2}}^+ a_{j+\frac{1}{2}}^-}{a_{j+\frac{1}{2}}^+ - a_{j+\frac{1}{2}}^-} (q_{j+\frac{1}{2}}^+ - q_{j+\frac{1}{2}}^-) \quad (3.13)$$

where  $a_{j+\frac{1}{2}}^+$  and  $a_{j+\frac{1}{2}}^-$  are given by bounds on the wave speed. Applying the wave speed bounds (2.10) we obtain

$$a_{j+\frac{1}{2}}^- = \min \left\{ 0, u_{j+1/2}^- - \sqrt{gh_{j+1/2}^-}, u_{j+1/2}^+ - \sqrt{gh_{j+1/2}^+} \right\}, \quad (3.14)$$

$$a_{j+\frac{1}{2}}^+ = \max \left\{ 0, u_{j+1/2}^- + \sqrt{gh_{j+1/2}^-}, u_{j+1/2}^+ + \sqrt{gh_{j+1/2}^+} \right\}. \quad (3.15)$$

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

$$f\left(h_{j+\frac{1}{2}}^\pm\right) = u_{j+1/2}^\pm h_{j+1/2}^\pm.$$

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

$$\begin{aligned} f\left(G_{j+\frac{1}{2}}^\pm\right) &= u_{j+1/2}^\pm G_{j+1/2}^\pm + \frac{g}{2} \left(h_{j+1/2}^\pm\right)^2 - \frac{2}{3} \left(h_{j+1/2}^\pm\right)^3 \left[ \left(\frac{\partial u}{\partial x}\right)_{j+1/2}^\pm \right]^2 \\ &\quad + \left(h_{j+1/2}^\pm\right)^2 u_{j+1/2}^\pm \left(\frac{\partial u}{\partial x}\right)_{j+1/2}^\pm \left(\frac{\partial b}{\partial x}\right)_{j+1/2}^\pm. \end{aligned} \quad (3.16)$$

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

### Calculation of Derivatives

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 now required to calculate the flux (3.16). To calculate these derivatives in  $u$  and  $b$  we use the basis function approximation to these quantities in the FEM to define the reconstruction polynomial of these quantities over a cell. For  $u$  we have the quadratic

$$P_j^u(x) = p_0^u (x - x_j)^2 + p_1^u (x - x_j) + p_2^u \quad (3.17)$$

that passes through  $u_{j-1/2}$ ,  $u_j$  and  $u_{j+1/2}$  (3.12). While for  $b$  we have the cubic

$$P_j^b(x) = p_0^b (x - x_j)^3 + p_1^b (x - x_j)^2 + p_2^b (x - x_j) + p_3^b \quad (3.18)$$

that passes through  $b_{j-1/2}$ ,  $b_{j-1/6}$ ,  $b_{j+1/6}$  and  $b_{j+1/2}$  (3.4). Because the cell edge values were averaged during the reconstruction of the bed,  $P_j^b(x)$  will be different from  $C_j(x)$ .

For  $P_j^u(x)$  we obtain the coefficients

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

While for  $P_j^b(x)$  the coefficients are

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

$$p_1^b = \frac{9b_{j-1/2} - 9b_{j-1/6} - 9b_{j+1/6} + 9b_{j+1/2}}{4\Delta x^2},$$

$$p_2^b = \frac{b_{j-1/2} - 27b_{j-1/6} + 27b_{j+1/6} - b_{j+1/2}}{8\Delta x},$$

$$p_3^b = \frac{-b_{j-1/2} + 9b_{j-1/6} + 9b_{j+1/6} - b_{j+1/2}}{16}.$$

Taking the derivative of the polynomials (3.17) and (3.18) 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

$$\begin{aligned}\left(\frac{\partial u}{\partial x}\right)_{j+1/2}^- &= \frac{\partial}{\partial x} P_j^u(x_{j+1/2}), \\ \left(\frac{\partial u}{\partial x}\right)_{j+1/2}^+ &= \frac{\partial}{\partial x} P_{j+1}^u(x_{j+1/2}), \\ \left(\frac{\partial b}{\partial x}\right)_{j+1/2}^- &= \frac{\partial}{\partial x} P_j^b(x_{j+1/2}), \\ \left(\frac{\partial b}{\partial x}\right)_{j+1/2}^+ &= \frac{\partial}{\partial x} P_{j+1}^b(x_{j+1/2}).\end{aligned}$$

Therefore, we possess all the terms needed to calculate the approximation to the flux (3.13) for  $h$  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 Modification to Flux Approximation

To recover the lake at rest steady state solution we follow the work of Audusse et al. [49], who accomplished this for the SWWE. Previously, we demonstrated that this process can be extended to the Serre equations [40]. To enforce well-balancing the reconstruction of  $h$  is modified at the cell edges in the following way.

First calculate

$$\dot{b}_{j+1/2}^- = w_{j+1/2}^- - h_{j+1/2}^-, \quad \dot{b}_{j+1/2}^+ = w_{j+1/2}^+ - h_{j+1/2}^+. \quad (3.19)$$

Find the maximum

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

then define

$$\ddot{h}_{j+1/2}^- = \max \left\{ 0, w_{j+1/2}^- - \ddot{b}_{j+1/2} \right\}, \quad (3.20a)$$

$$\ddot{h}_{j+1/2}^+ = \max \left\{ 0, w_{j+1/2}^+ - \ddot{b}_{j+1/2} \right\}. \quad (3.20b)$$

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 (3.13) for  $h$  and  $G$  instead of  $\hat{\mathbf{h}}$ . Applying the same process but with different cells we obtain  $F_{j-1/2}$  and we have

$$F_{j\pm 1/2}^n = \mathcal{F}_{j\pm 1/2}(\ddot{\mathbf{h}}, \hat{\mathbf{G}}, \hat{\mathbf{b}}, \hat{\mathbf{u}}).$$

for the evolution of  $h$  and  $G$  equations as desired.

### 3.2.4 Source Terms

To evolve the Serre equations (2.6), we require an approximation to the source term at the cell centre  $x_j$  from time  $t^n$  to  $t^{n+1}$  which we denote as  $S_j^n$ . Equation (2.6a) has no source term, therefore we only present the calculation of the source term for equation (2.6b).

Following the work of Audusse et al. [49] to produce a well-balanced method, we split our approximation to  $S_j^n$  into the centred source term  $S_{ci}$  and the corrective interface source terms  $S_{j+\frac{1}{2}}^-$  and  $S_{j+\frac{1}{2}}^+$

$$S_j^n = S_{j+\frac{1}{2}}^- + \Delta x S_{ci} + S_{j-\frac{1}{2}}^+.$$

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

We calculate the centred source term using

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

Where we use  $h_j$  from the reconstruction process (3.1) and  $u_j$  from the solution of (3.12). To calculate the derivatives we employ our polynomial representations of  $u$  (3.17) and  $b$  (3.18) 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 (3.19). Therefore, the following approximations are used to calculate  $S_{ci}$

$$\begin{aligned}\left(\frac{\partial u}{\partial x}\right)_j &= \frac{\partial}{\partial x} P_j^u(x_j), \\ \left(\frac{\partial b}{\partial x}\right)_j &= \frac{\dot{b}_{j+1/2}^- - \dot{b}_{j-1/2}^+}{\Delta x}, \\ \left(\frac{\partial^2 b}{\partial x^2}\right)_j &= \frac{\partial^2}{\partial x^2} P_j^b(x_j).\end{aligned}$$

To ensure well-balancing the corrective interface source terms

$$\begin{aligned}S_{j+\frac{1}{2}}^- &= \frac{g}{2} \left( \ddot{h}_{j+\frac{1}{2}}^- \right)^2 - \frac{g}{2} \left( h_{j+\frac{1}{2}}^- \right)^2, \\ S_{j-\frac{1}{2}}^+ &= \frac{g}{2} \left( h_{j-\frac{1}{2}}^+ \right)^2 - \frac{g}{2} \left( \ddot{h}_{j-\frac{1}{2}}^+ \right)^2\end{aligned}$$

are also added. These corrective terms make use of  $h_{j+\frac{1}{2}}^-$  and  $h_{j+\frac{1}{2}}^+$  obtained from the reconstruction (3.1) and the modified values  $\ddot{h}_{j+\frac{1}{2}}^-$  and  $\ddot{h}_{j+\frac{1}{2}}^+$  from (3.20). Combining the centred and interface source terms our approximation to the source term for  $G$  is

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

This operator  $\mathcal{S}_j$  is slightly different to the example given in the overview of the evolution step as it takes  $\hat{\mathbf{h}}$  and  $\ddot{\mathbf{h}}$  as inputs. This change was made to make clear that  $\ddot{\mathbf{h}}$  is required to calculate  $S_j^n$ . This would be obscured by a more consistent notation that is possible since  $\ddot{\mathbf{h}}$  only depends on  $\hat{\mathbf{h}}$ ,  $\hat{\mathbf{w}}$  and  $\hat{\mathbf{b}}$ .

### 3.2.5 Update Cell Averages

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

$$\bar{q}_j^{n+1} = \bar{q}_j^n + \frac{\Delta t}{\Delta x} \left( F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n + S_j^n \right) \quad (3.21)$$

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 first-order in time.

### 3.2.6 Second-Order SSP Runge-Kutta Method

To increase the order of accuracy in time we employ the Strong Stability Preserving (SSP) Runge-Kutta method [50] which is a convex combination of the first-order time steps (3.21) in the following way

$$\bar{q}_j^{(1)} = \bar{q}_j^n + \frac{\Delta t}{\Delta x} \left( F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n + S_j^n \right), \quad (3.22a)$$

$$\bar{q}_j^{(2)} = \bar{q}_j^{(1)} + \frac{\Delta t}{\Delta x} \left( F_{j+\frac{1}{2}}^{(1)} - F_{j-\frac{1}{2}}^{(1)} + S_j^{(1)} \right), \quad (3.22b)$$

$$\bar{q}_j^{n+1} = \frac{1}{2} \left( \bar{q}_j^n + \bar{q}_j^{(2)} \right). \quad (3.22c)$$

This results in a time stepping method that preserves the stability of the first-order method (3.21) 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 [51] which is necessary for stability. The CFL condition ensures that time steps are small enough so that information is only transferred between neighbouring cells. For the Serre equations the CFL condition is

$$\Delta t \leq \frac{Cr}{\max_j \left\{ a_{j+1/2}^\pm \right\}} \Delta x \quad (3.23)$$

where  $a_{j+1/2}^\pm$  are the wave-speed bounds used in the flux approximation (3.15) and  $0 \leq Cr \leq 1$  is the Courant number. Typically, we use the conservative  $Cr = 0.5$  for our numerical experiments.

## 3.4 Boundary Conditions

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

is required, while for  $b$  we require two ghost cells at each boundary. The ghost cells for  $h$ ,  $w$  and  $G$  written for a generic quantity  $q$  are

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

For  $u$  and  $b$  the ghost cells are

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

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

To ensure that the solution of  $u$  using (3.12) agrees with the boundary conditions  $\hat{\mathbf{u}}_{-1}$  and  $\hat{\mathbf{u}}_m$  the element-wise stiffness matrices  $\mathbf{A}_0$  and  $\mathbf{A}_m$  and vectors  $\mathbf{g}_0$  and  $\mathbf{g}_m$  must be modified in the following way

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

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

These are then assembled with the other element contributions in the global stiffness matrix  $\mathbf{A}$  and right hand side vector  $\mathbf{g}$  in (3.11).

## 3.5 Dry Beds

Dry beds are handled adequately by all steps of the FEVM in their current form, except the FEM for  $u$ . A dry bed presents two issues; when  $h$  and  $G$  are small

then small errors in  $h$  and  $G$  can produce large errors in  $u$  leading to instabilities and when  $h = 0$  the stiffness matrix  $\mathbf{A}$  (3.12) becomes singular.

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

$$G = uh + \mathcal{O}(h^3). \quad (3.24)$$

Since  $h \ll 1$  we neglect the  $\mathcal{O}(h^3)$  terms, and thus when  $h$  is small  $G$  is equal to the momentum  $uh$ , and the challenges posed by  $h \rightarrow 0$  for the SWWE and the Serre equations are equivalent. Therefore, we can apply the dry bed handling techniques from the SWWE to the Serre equations; in particular a desingularisation transformation [52].

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

$$u = \frac{(uh)h}{h(h + h_{base})} \quad (3.25)$$

where  $h_{base}$  is a small chosen parameter. The analytical error introduced by this transformation decreases as  $h_{base}$  decreases. However, as noted by Kurganov and Petrova [52] small values of  $h_{base}$  lead to large numerical errors in the calculation of  $u$ . To avoid such errors  $h_{base}$  can be made larger or following Kurganov and Petrova [52] different desingularisation transformations can be employed. For the main purpose of this thesis; the validation tests reported in Chapter 5 we found the simpler transformation with small values of  $h_{base}$  more useful, keeping in mind that large numerical errors in  $u$  were possible for small values of  $h$ .

To adapt the calculation of  $u$  in (3.25) to (2.7) we view it as a transformation of the quantity  $h$  which is equivalent to

$$h \rightarrow h \left( \frac{h + h_{base}}{h} \right). \quad (3.26)$$

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

values of  $h$  in the finite element method is

$$h_{j-1/2}^+ = h_{j-1/2}^+ \left( \frac{h_{j-1/2}^+ + h_{base}}{h_{j-1/2}^+ + h_{tol}} \right), \quad (3.27a)$$

$$h_{j+1/2}^- = h_{j+1/2}^- \left( \frac{h_{j+1/2}^- + h_{base}}{h_{j+1/2}^- + h_{tol}} \right) \quad (3.27b)$$

where on the right hand side are the reconstructed values of  $h$  from (3.1) and the left hand side are the values of  $h$  used to defined the basis functions of the FEM (3.7). This transformation is applied to all terms in the FEM avoiding the singularity as  $h \rightarrow 0$ ; and in the case where  $G = uh$  the transformation is equivalent to (3.25) for the SWWE.

Even with the transform (3.27), the matrix  $\mathbf{A}$  can become singular. The methods of Zoppou et al. [14] made use of direct banded matrix solvers such as the Thomas algorithm [53] to solve (3.12) which rely on non-singular matrices making them unsuitable when  $h = 0$ . This was resolved by employing an LU decomposition algorithm described by Press et al. [47]. This algorithm solves banded matrix problems using an LU decomposition with partial pivoting. When the value of a pivot is below some tolerance value  $p_{tol}$  the algorithm replaces its value with  $p_{tol}$ , avoiding the errors associated with small pivot values during an LU decomposition. It does this while also keeping the banded matrix structure, and so is not as memory intensive as a standard LU decomposition. Typically we set  $p_{tol} = 10^{-20}$  allowing the matrix solver to accurately invert  $\mathbf{A}$  and thus solve (3.12) when  $h = 0$ .

Solving (3.12) using the LU decomposition algorithm of Press et al. [47] where the transformation (3.27) has been applied to the reconstructed values of  $h$  an approximation to  $u$  is obtained that is valid in the presence of dry beds. Additionally, to avoid numerical errors becoming dominant when  $h$  is very small we place a cut-off on  $h$  beyond which  $h = G = u = 0$  and the cells are dry; this is given by  $h_{tol}$ . This drying of the cells is performed for the whole cell based on the cell average value of  $h$  so that if  $\bar{h}_j \leq h_{tol}$  then

$$\begin{aligned} h_{j-1/2}^+ &= 0 & G_{j-1/2}^+ &= 0 & w_{j-1/2}^+ &= b_{j-1/2} \\ h_j &= 0 & G_j &= 0 & w_j &= b_j, \\ h_{j+1/2}^- &= 0 & G_{j+1/2}^- &= 0 & w_{j+1/2}^- &= b_{j+1/2} \end{aligned}$$

and

$$\begin{aligned} u_{j-1/2} &= 0 & \text{if} & \quad \bar{h}_{j-1} \leq h_{tol} \\ u_j &= 0 & \text{if} & \\ u_{j+1/2} &= 0 & \quad \bar{h}_{j+1} \leq h_{tol}. \end{aligned}$$

The drying procedure occurs after the solution of (3.12). In the numerical experiments the typical values used were  $h_{tol} = 10^{-12}$  and  $h_{base} = 10^{-8}$ .

In this chapter FEVM<sub>2</sub> was described, including the details for the well balancing and dry bed handling procedures. A linear analysis of the convergence and dispersion properties of FEVM<sub>2</sub> will now be performed.

# Chapter 4

## Linear Analysis

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

- FEVM<sub>2</sub> which is a second-order Finite Element Volume Method (FEVM) described in Chapter 3.
- The Finite Difference Volume Methods (FDVM) of first- (FDVM<sub>1</sub>), second- (FDVM<sub>2</sub>) and third-order (FDVM<sub>3</sub>) described by Zoppou et al. [14].
- The second-order finite difference methods  $\mathcal{D}$  and  $\mathcal{W}$  described by Pitt et al. [17].

We provide the full description of the linear analysis for FEVM<sub>2</sub> as an example and then just present the results for FDVM<sub>1</sub>, FDVM<sub>2</sub>, FDVM<sub>3</sub>,  $\mathcal{D}$  and  $\mathcal{W}$ . Supplementary details for FDVM<sub>1</sub>, FDVM<sub>2</sub>, FDVM<sub>3</sub>,  $\mathcal{D}$  and  $\mathcal{W}$  can be found in Appendix C.

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 it approximates. 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 [54]. A numerical scheme is consistent if the error introduced by the numerical method over a time step approaches zero as the spatial and temporal resolution increases. A numerical method is stable if the errors from previous time steps are not amplified over 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 [23]. Therefore, how well the dispersion relation of a numerical method approximates the dispersion relation of the Serre equations is of particular interest.

We analyse the convergence and dispersion properties of the whole numerical method applied to the solution of the linearised Serre equations with a horizontal bed. The whole scheme is considered with the spatial and temporal approximations analysed simultaneously. The effect of variations in the bed and non-linear 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, estimate the convergence properties of the non-linear Serre equations with varying bathymetry by investigating the linearised Serre equations with a horizontal bed. 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 non-linear terms are included.

The dispersion properties of the Serre equations are derived from the linearised Serre equations with a horizontal bed [14]. Because the dispersion analysis includes the spatial and temporal approximations simultaneously the presented analysis of the dispersion properties of the numerical method is a complete analysis extending the work of Filippini et al. [39].

The linear analyses of convergence and dispersion properties for the finite volume based methods rely on establishing a relationship of the form

$$\left[ \frac{\bar{h}}{\bar{G}} \right]_j^{n+1} = \mathbf{E} \left[ \frac{\bar{h}}{\bar{G}} \right]_j^n \quad (4.1)$$

where  $\mathbf{E}$  is the  $2 \times 2$  evolution matrix relating the cell average conserved quantities  $h$  and  $G$  at time level  $t^n$  with the cell average conserved quantities at time level  $t^{n+1}$ . For a linear analysis the evolution matrix  $\mathbf{E}$  will be independent of  $n$  and  $j$ . The evolution matrix  $\mathbf{E}$  is obtained in the analyses by representing the numerical solution in terms of Fourier modes. By analysing the properties of  $\mathbf{E}$  and comparing it with the exact evolution matrix we can determine the convergence and dispersion properties of its associated numerical method.

We derive  $\mathbf{E}$  in (4.1) for FEVM<sub>2</sub> and perform the convergence and dispersion analysis. We then present the results of the analyses for FDVM<sub>1</sub>, FDVM<sub>2</sub>, FDVM<sub>3</sub>,  $\mathcal{D}$  and  $\mathcal{W}$ . In Appendix C a process that produces the evolution matrix  $\mathbf{E}$  for FDVM<sub>1</sub>, FDVM<sub>2</sub>, FDVM<sub>3</sub> is provided while for  $\mathcal{D}$  and  $\mathcal{W}$  the evolution ma-

trix itself is provided. These results extend those published by Zoppou et al. [14] by including more methods, analysing the convergence properties and allowing non-zero background mean velocities.

## 4.1 Linearised Serre Equations

The Serre equations with a horizontal bed (2.5) are linearised by considering waves as small perturbations  $\delta \times \eta(x, t)$  and  $\delta \times \mu(x, t)$  on a flow with a mean height  $H$  and a mean velocity  $U$  respectively, where  $\delta \ll 1$ . So we have

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

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

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

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

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

and for  $G$

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

Absorbing the  $\delta$  factor into the corresponding  $\eta$  and  $\mu$  terms and rewriting these equations in conservation law form for  $\eta$  and  $G$  we obtain

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

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

where

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

## 4.2 Evolution Matrix

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

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

where  $k$  is the wavenumber,  $\omega^\pm$  is the frequency (2.9) and  $i$  is the imaginary number. The Fourier modes are the eigenfunctions of these linearised Serre equations (4.4). Since the eigenfunctions form a basis of the solution space, their dispersion and convergence properties are inherited by all solutions of (4.4). Therefore, it is sufficient to study only the convergence and dispersion properties for Fourier mode solutions captured by the evolution matrix  $\mathbf{E}$ .

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

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

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

For 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 act as factors on their inputs. These factors vary with  $H$ ,  $U$ ,  $k$ ,  $\omega^\pm$ ,  $\Delta x$  and  $\Delta t$  and are independent of  $j$  and  $n$ . By combining the factors associated with these operators the evolution matrix  $\mathbf{E}$  can be derived for FEVM<sub>2</sub> for the linearised Serre equations with a horizontal bed. Since all the constituent factors that produce  $\mathbf{E}$  are independent of  $j$  and  $n$  then  $\mathbf{E}$  will also be independent of  $j$  and  $n$ , as desired. We will now derive expressions for all these factors, following the structure of the method laid out in Section 3.2. Since the linearised Serre equations with a horizontal bed have no source terms step (iv), which approximates the source terms is not necessary.

### 4.2.1 Reconstruction

Given the cell averages  $\bar{\eta}$  and  $\bar{G}$  at  $t^n$ , the first step of our numerical method is to reconstruct  $\eta$  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  $\eta$  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

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

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

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

Note that the factors given by the reconstructions operators  $\mathcal{R}_{j-1/2}^+$ ,  $\mathcal{R}_j$  and  $\mathcal{R}_{j+1/2}^-$  are independent of  $j$ . Furthermore, from (4.6) we have that reconstructions of  $\eta$  and  $G$  at other locations are  $\mathcal{R}_{j-1/2}^+$ ,  $\mathcal{R}_j$  and  $\mathcal{R}_{j+1/2}^-$  multiplied by a shift factor. In particular, we have that the reconstruction operator  $\mathcal{R}_{j+1/2}^+$  for  $q_{j+\frac{1}{2}}^+$  is given by

$$q_{j+\frac{1}{2}}^+ = e^{ik\Delta x} q_{j-\frac{1}{2}}^+ = e^{ik\Delta x} \mathcal{R}_{j-1/2}^+ \bar{q}_j = \mathcal{R}_{j+1/2}^+ \bar{q}_j. \quad (4.7d)$$

### 4.2.2 Fluid Velocity

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

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

The FEM then proceeds for (4.4c) as in Chapter 3 for the non-linear Serre equation using the basis functions defined in Appendix B. So that  $G$  has the basis functions  $\psi_{j-1/2}^+$  and  $\psi_{j+1/2}^-$  (B.1), which means our approximation to  $G$  is linear inside a cell with discontinuous jumps at the cell edges. For  $v$  and  $\mu$  the basis functions  $\phi_{j-1/2}$ ,  $\phi_j$  and  $\phi_{j+1/2}$  (B.2) are used so that  $v$  and our approximation to  $\mu$  are quadratic polynomials inside a cell and are continuous across the cell edges.

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

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

To calculate the intercell flux we require  $\mu$  at  $x_{j+1/2}$ . From the element wise matrices and vectors for the  $j^{th}$  and  $(j+1)^{th}$  cells the equation that relates all

the quantities at  $x_{j+1/2}$  is

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

Using (4.7) and (4.6), we obtain

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

Rearranging this equation we have that

$$\mu_{j+1/2} = \mathcal{G}^\eta \bar{\eta}_j + \mathcal{G}^G \bar{G}_j + \mathcal{G}^c \quad (4.8)$$

where

$$\mathcal{G}^\eta = -U\mathcal{G}^G,$$

$$\mathcal{G}^G = \frac{\Delta x}{6\mathcal{G}_D} \left( \mathcal{R}_{j+1/2}^- + \mathcal{R}_{j+1/2}^+ \right),$$

$$\mathcal{G}^c = -2UH \frac{\Delta x}{6\mathcal{G}_D}$$

and the common divisor  $\mathcal{G}_D$  is

$$\begin{aligned} \mathcal{G}_D & = H \frac{\Delta x}{30} \left( 4 \cos \left( \frac{k\Delta x}{2} \right) - 2 \cos(k\Delta x) + 8 \right) \\ & \quad + \frac{H^3}{3} \frac{1}{3\Delta x} \left( -16 \cos \left( \frac{k\Delta x}{2} \right) + 2 \cos(k\Delta x) + 14 \right). \end{aligned}$$

So that the factors  $\mathcal{G}^\eta$ ,  $\mathcal{G}^G$ ,  $\mathcal{G}^c$  that make up the operator  $\mathcal{G}$  do not depend on  $n$  or  $j$  as desired.

### 4.2.3 Flux Across the Cell Interfaces

The average intercell flux  $F_{j+1/2}$  is approximated using (3.13). For the linearised Serre equations we have the wave speed bounds (2.10), so that

$$a_{j+1/2}^- = \min \left\{ 0, U - \sqrt{gH} \right\}, \quad a_{j+1/2}^+ = \max \left\{ 0, U + \sqrt{gH} \right\}. \quad (4.9)$$

This method has three different approximations to  $F_{j+1/2}$  depending on the Froude number  $Fr = U/\sqrt{gH}$ ; (i) supercritical flow to the left where  $Fr < -1$ , (ii) critical and subcritical flow in both directions where  $-1 \leq Fr \leq 1$  and (iii) supercritical flow to the right where  $Fr > 1$ . We will derive the flux operators for each of these cases separately.

#### Left Supercritical Flow $Fr < -1$ :

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

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

for a generic quantity  $q$ .

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

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

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

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

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

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

$$F_{j+\frac{1}{2}}^\eta = \mathcal{F}_{j+\frac{1}{2}}^{\eta,\eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,c}$$

where

$$\begin{aligned} \mathcal{F}_{j+\frac{1}{2}}^{\eta,\eta} &= H\mathcal{G}^\eta + U\mathcal{R}_{j+1/2}^+, \\ \mathcal{F}_{j+\frac{1}{2}}^{\eta,G} &= H\mathcal{G}^G, \\ \mathcal{F}_{j+\frac{1}{2}}^{\eta,c} &= H\mathcal{G}^c. \end{aligned}$$

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

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

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

$$F_{j+\frac{1}{2}}^G = UG_{j+1/2}^+ + UH(\mathcal{G}^\eta \bar{\eta}_j + \mathcal{G}^G \bar{G}_j + \mathcal{G}^c) + gH\eta_{j+1/2}^+$$

which can be rewritten as

$$F_{j+\frac{1}{2}}^G = \mathcal{F}_{j+\frac{1}{2}}^{G,\eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,c}$$

where

$$\begin{aligned}\mathcal{F}_{j+\frac{1}{2}}^{G,\eta} &= UH\mathcal{G}^\eta + gH\mathcal{R}_{j+1/2}^+, \\ \mathcal{F}_{j+\frac{1}{2}}^{G,G} &= U\mathcal{R}_{j+1/2}^+ + UH\mathcal{G}^G, \\ \mathcal{F}_{j+\frac{1}{2}}^{G,c} &= UH\mathcal{G}^c.\end{aligned}$$

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

When the flow is subcritical we have  $-1 \leq Fr \leq 1$ , which means that  $a_{j+1/2}^- = U - \sqrt{gH}$  and  $a_{j+1/2}^+ = U + \sqrt{gH}$ . Therefore, the flux approximation (3.13) becomes

$$\begin{aligned}F_{j+\frac{1}{2}} &= \frac{U}{2\sqrt{gH}} \left[ f(q_{j+\frac{1}{2}}^-) - f(q_{j+\frac{1}{2}}^+) \right] + \frac{1}{2} \left[ f(q_{j+\frac{1}{2}}^-) + f(q_{j+\frac{1}{2}}^+) \right] \\ &\quad + \frac{U^2 - gH}{2\sqrt{gH}} \left[ q_{j+\frac{1}{2}}^+ - q_{j+\frac{1}{2}}^- \right].\end{aligned}\tag{4.11}$$

Substituting in the flux function for  $\eta$  given by (4.4a) we get

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

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

$$F_{j+\frac{1}{2}}^\eta = \mathcal{F}_{j+\frac{1}{2}}^{\eta,\eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,c}$$

where

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

For the flux function of  $G$  (4.4b) the flux approximation (4.11) becomes

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

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

$$F_{j+\frac{1}{2}}^G = \mathcal{F}_{j+\frac{1}{2}}^{G,\eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,c}$$

where

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

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

For a rightward supercritical flow 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 (3.13) which reduces to

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

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

$$F_{j+\frac{1}{2}}^\eta = \mathcal{F}_{j+\frac{1}{2}}^{\eta,\eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,c}$$

where

$$\begin{aligned}\mathcal{F}_{j+\frac{1}{2}}^{\eta,\eta} &= H\mathcal{G}^\eta + U\mathcal{R}_{j+1/2}^-, \\ \mathcal{F}_{j+\frac{1}{2}}^{\eta,G} &= H\mathcal{G}^G, \\ \mathcal{F}_{j+\frac{1}{2}}^{\eta,c} &= H\mathcal{G}^c\end{aligned}$$

and

$$F_{j+\frac{1}{2}}^G = \mathcal{F}_{j+\frac{1}{2}}^{G,\eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,c}$$

where

$$\begin{aligned}\mathcal{F}_{j+\frac{1}{2}}^{G,\eta} &= UH\mathcal{G}^\eta + gH\mathcal{R}_{j+1/2}^-, \\ \mathcal{F}_{j+\frac{1}{2}}^{G,G} &= U\mathcal{R}_{j+1/2}^+ + UH\mathcal{G}^G, \\ \mathcal{F}_{j+\frac{1}{2}}^{G,c} &= UH\mathcal{G}^c\end{aligned}$$

respectively.

#### 4.2.4 Update Cell Averages

We have obtained the operators for the flux functions for supercritical, critical and subcritical flow. Substituting the appropriate flux approximation into the forward Euler step, (3.21) 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  $\mathcal{F}_{j-\frac{1}{2}}^{\eta,\eta} = e^{-ik\Delta x} \mathcal{F}_{j+\frac{1}{2}}^{\eta,\eta}$ ,  $\mathcal{F}_{j-\frac{1}{2}}^{\eta,G} = e^{-ik\Delta x} \mathcal{F}_{j+\frac{1}{2}}^{\eta,G}$ ,  $\mathcal{F}_{j-\frac{1}{2}}^{G,\eta} = e^{-ik\Delta x} \mathcal{F}_{j+\frac{1}{2}}^{G,\eta}$  and  $\mathcal{F}_{j-\frac{1}{2}}^{G,G} = e^{-ik\Delta x} \mathcal{F}_{j+\frac{1}{2}}^{G,G}$  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 using the identity matrix  $\mathbf{I}$  as

$$\begin{aligned}\begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^{n+1} &= \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n - (1 - e^{-ik\Delta x}) \frac{\Delta t}{\Delta x} \begin{bmatrix} \mathcal{F}^{\eta,\eta} & \mathcal{F}^{\eta,G} \\ \mathcal{F}^{G,\eta} & \mathcal{F}^{G,G} \end{bmatrix} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n \\ &= (\mathbf{I} - \Delta t \mathbf{F}) \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n\end{aligned}\tag{4.12}$$

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

### 4.2.5 Second-Order SSP Runge-Kutta Method

To achieve second-order accurate time stepping, the second-order SSP Runge-Kutta scheme (3.22) is used. This scheme uses the following convex combination of the Euler steps (4.12)

$$\left[ \frac{\bar{\eta}}{G} \right]_j^{(1)} = (\mathbf{I} - \Delta t \mathbf{F}) \left[ \frac{\bar{\eta}}{G} \right]_j^n, \quad (4.13a)$$

$$\left[ \frac{\bar{\eta}}{G} \right]_j^{(2)} = (\mathbf{I} - \Delta t \mathbf{F}) \left[ \frac{\bar{\eta}}{G} \right]_j^{(1)}, \quad (4.13b)$$

$$\left[ \frac{\bar{\eta}}{G} \right]_j^{n+1} = \frac{1}{2} \left( \left[ \frac{\bar{\eta}}{G} \right]_j^n + \left[ \frac{\bar{\eta}}{G} \right]_j^{(2)} \right). \quad (4.13c)$$

Substituting (4.13a) and (4.13b) into (4.13c) 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

$$\begin{aligned} \left[ \frac{\bar{\eta}}{G} \right]_j^{n+1} &= \left( \mathbf{I} - \Delta t \mathbf{F} + \frac{1}{2} \Delta t^2 \mathbf{F}^2 \right) \left[ \frac{\bar{\eta}}{G} \right]_j^n \\ &= \mathbf{E} \left[ \frac{\bar{\eta}}{G} \right]_j^n \end{aligned} \quad (4.14)$$

which is in the desired form (4.1).

This is the evolution matrix  $\mathbf{E}$  for FEVM<sub>2</sub>. The matrix  $\mathbf{E}$  is dependent on the flux matrix  $\mathbf{F}$  and therefore will depend on the Froude number. The Froude number is constant over time in this analysis and so we can investigate supercritical, subcritical and critical flow individually.

The convergence and dispersion analysis then proceed by studying the properties of the evolution matrix  $\mathbf{E}$  for FEVM<sub>2</sub>. As a comparison we also provide the results for the finite difference volume methods FDVM<sub>1</sub>, FDVM<sub>2</sub> and FDVM<sub>3</sub> described by Zoppou et al. [14] and the finite difference methods  $\mathcal{D}$  and  $\mathcal{W}$  described by Pitt et al. [17].

The evolution matrices for FDVM<sub>1</sub>, FDVM<sub>2</sub>, FDVM<sub>3</sub> can be derived following the derivation of the evolution matrix of FEVM<sub>2</sub> using the expressions for its constituent operators provided in Appendix C. For  $\mathcal{D}$  and  $\mathcal{W}$  the evolution matrices are (C.10) and (C.11) respectively. We now present the results of the convergence analysis.

## 4.3 Convergence Analysis

We apply the Lax-equivalence theorem to demonstrate the convergence of our numerical methods by establishing their stability and consistency. Stability is established for the numerical methods using a Von Neumann stability analysis. While consistency is established for the numerical methods by representing the numerical solution as Fourier modes (4.5) and demonstrating the consistency of the numerical method for these Fourier mode solutions. Since the Fourier modes form a basis of the solution space of the linearised Serre equations, this demonstrates consistency for all numerical solutions. The presented stability and consistency analyses 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 over the current time step. To accomplish this we must ensure

$$\rho(\mathbf{E}) \leq 1 \quad (4.15)$$

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  $\Delta x$ ,  $\Delta t$ ,  $k$ ,  $H$  and  $U$  to check if (4.15) 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  $\lambda = 2\pi/k$  is the wavelength. We used  $g = 9.81m/s^2$  and chose  $\Delta t = 0.5 / (U + \sqrt{gH}) \Delta x$  to satisfy the CFL condition (3.23). This is the common choice of  $\Delta t$  in our numerical experiments.

The behaviour of  $\rho(\mathbf{E})$  for  $H = 1m$ ,  $k = \pi/10m^{-1}$  and  $U = 0m/s$  and  $1m/s$  is shown in Figure 4.1 and is representative of the behaviour for all other values of  $H$ ,  $k$  and  $U$ . For these  $k$  and  $H$  values the shallowness parameter  $\sigma = 1/20$  and so the Serre equations are applicable [23].

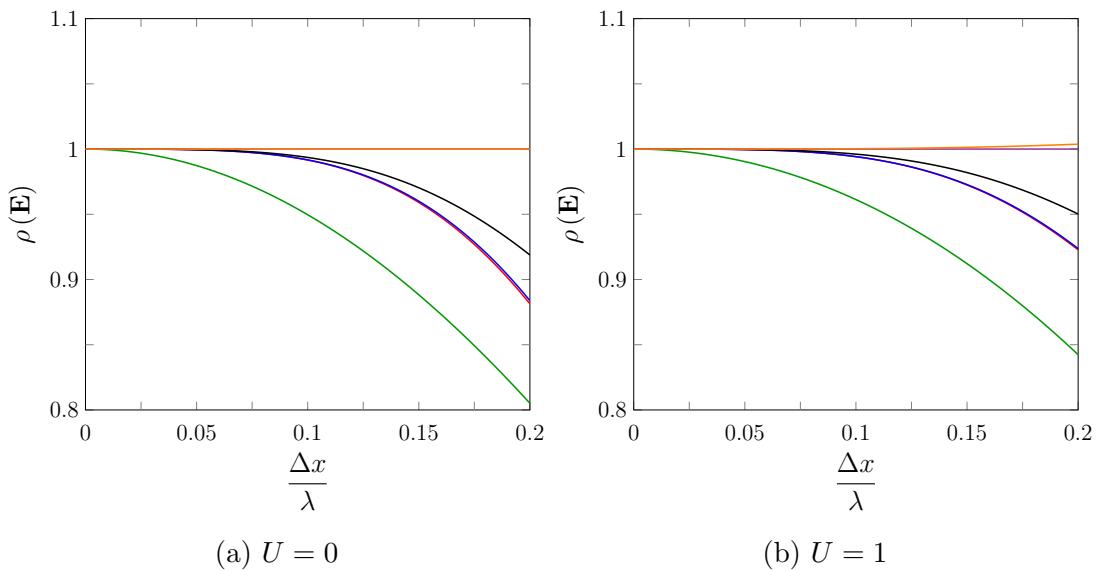


Figure 4.1: Spectral radius of  $\mathbf{E}$  against  $\Delta x/\lambda$  for FDVM<sub>1</sub> (green line), FDVM<sub>2</sub> (red line), FEVM<sub>2</sub> (blue line), FDVM<sub>3</sub> (black line),  $\mathcal{D}$  (purple line) and  $\mathcal{W}$  (orange line). With  $H = 1m$  and  $k = \pi/10m^{-1}$ .

In Figure 4.1 it can be seen that all methods have  $\rho(\mathbf{E}) \leq 1$  for  $U = 0m/s$  and are therefore stable. The two finite difference methods overlap and have  $\rho(\mathbf{E}) = 1$  for all  $\Delta x$  values, while FDVM<sub>2</sub> and FEVM<sub>2</sub> also overlap with  $\rho(\mathbf{E}) < 1$ . However, when  $U \neq 0m/s$  the method  $\mathcal{W}$  has  $\rho(\mathbf{E}) > 1$  for all  $\Delta x$  values and is therefore unstable. All other methods have  $\rho(\mathbf{E}) \leq 1$ , retaining their stability when  $U \neq 0m/s$ .

We observed the same results for a wide range of  $k$ ,  $H$  and  $U$  values and Froude numbers. All methods except  $\mathcal{W}$  were found to be stable for any combination of these variables. While  $\mathcal{W}$  was only stable when  $U = 0\text{m/s}$ . This differs from the stability result for  $\mathcal{D}$  and  $\mathcal{W}$  reported by Pitt et al. [17] as that analysis assumed  $U = 0\text{m/s}$ .

### 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 convergence, it is sufficient 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  $t^{n+1}$ ,  $\mathcal{T}_j^n$  is

$$\mathcal{T}_j^n = \mathbf{E} \left[ \frac{\bar{\eta}}{G} \right]_j^n - \left[ \frac{\bar{\eta}}{G} \right]_j^{n+1}. \quad (4.16)$$

To ensure consistency we must have that  $\lim_{\Delta x, \Delta t \rightarrow 0} \|\mathcal{T}_j^n\|_2 = 0$  for all  $n$  and  $j$ , where  $\|\cdot\|_2$  is the  $L_2$  vector norm. Importantly the location  $x_j$  and time  $t^n$  will be fixed as  $\Delta x$  and  $\Delta t$  vary, so that the measurement of the error  $\mathcal{T}_j^n$  is consistent. Taking the  $L_2$  norm of both sides of (4.16) and using (4.6) we obtain

$$\|\mathcal{T}_j^n\|_2 = \left\| \mathbf{E} \left[ \frac{\bar{\eta}}{G} \right]_j^n - e^{i\omega^\pm \Delta t} \left[ \frac{\bar{\eta}}{G} \right]_j^n \right\|_2.$$

Using the matrix norm induced by  $L_2$ , the Frobenius norm  $\|\cdot\|_F$  we have that

$$\|\mathcal{T}_j^n\|_2 \leq \left\| \mathbf{E} - e^{i\omega^\pm \Delta t} \mathbf{I} \right\|_F \left\| \left[ \frac{\bar{\eta}}{G} \right]_j^n \right\|_2.$$

Both  $\bar{\eta}_j^n$  and  $\bar{G}_j^n$  approach their respective midpoint values  $\eta_j^n$  and  $G_j^n$  as  $\Delta x$  goes to zero due to the shrinking of the  $j^{th}$  cell. Even though  $\bar{\eta}_j^n$  and  $\bar{G}_j^n$  vary they are always finite. Therefore, if  $\lim_{\Delta x, \Delta t \rightarrow 0} \left\| \mathbf{E} - e^{i\omega^\pm \Delta t} \mathbf{I} \right\|_F = 0$  then  $\lim_{\Delta x, \Delta t \rightarrow 0} \|\mathcal{T}_j^n\|_2 = 0$  as desired.

We calculated the Taylor series of  $\mathbf{E} - e^{i\omega^\pm \Delta t} \mathbf{I}$  for all the numerical methods for all flow scenarios; subcritical, critical and supercritical flows. Since the results are the same for  $\omega^+$  and  $\omega^-$  we only report the results for  $\omega^+$ . For FEVM<sub>2</sub> the lowest order  $\Delta x$  and  $\Delta t$  terms of the Taylor series of  $\mathbf{E} - e^{i\omega^+ \Delta t} \mathbf{I}$  can be found in Table 4.1. From this table it can be seen that the Taylor series of all the elements of  $\mathbf{E} - e^{i\omega^+ \Delta t} \mathbf{I}$  have a factor of  $\Delta t$ . So that

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

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

From Tables 4.1 we have that  $\mathbf{M}_0$  is independent of  $\Delta t$  and finite so that as  $\Delta t \rightarrow 0$  then  $|\Delta t| (\|\mathbf{M}_0\|_F + \|\mathcal{O}(\Delta t)\|_F) \rightarrow 0$  and therefore  $\left\| \mathbf{E} - e^{i\omega^+ \Delta t} \mathbf{I} \right\|_F \rightarrow 0$ . Therefore, for FEVM<sub>2</sub> we have  $\lim_{\Delta x, \Delta t \rightarrow 0} \|\mathcal{T}_j^n\|_2 = 0$  for all  $n$  and  $j$  and

---

Element	Lowest Order Terms of $\mathbf{E} - e^{i\omega^+\Delta t}\mathbf{I}$ for FEVM <sub>2</sub>	
	$\Delta x$	$\Delta t$
$E_{0,0} - e^{i\omega^+\Delta t}$	$-\frac{i(54 + 45H^2k^2 + 10H^4k^4)}{120\beta^2}Uk^3\Delta t\Delta x^2$	$\frac{\sqrt{3gH\beta} + 3U}{\beta}ik\Delta t$
$E_{0,1}$	$\frac{\beta - 3}{\beta^2}\frac{ik^3}{40}\Delta t\Delta x^2$	$-\frac{3}{\beta}ik\Delta t$
$E_{1,0}$	$-\left(gH - \frac{15U^2}{\beta} + \frac{9U^2}{\beta}\right)\frac{k^3}{120}\Delta t\Delta x^2$	$\left(-gH + \frac{3U^2}{\beta}\right)ik\Delta t$
$E_{1,1} - e^{i\omega^+\Delta t}$	$\frac{126 + 75H^2k^2 + 10H^4k^4}{\beta^2}\frac{k^3}{120}iU\Delta t\Delta x^2$	$\frac{\sqrt{3gH\beta} - 3U}{\beta}ik\Delta t$

---

Table 4.1: Lowest order terms of the Taylor series for the elements of  $\mathbf{E} - e^{i\omega^+\Delta t}\mathbf{I}$  for FEVM<sub>2</sub> for all values of  $Fr$ . Here  $\beta = 3 + k^2H^2$ .

so FEVM<sub>2</sub> is consistent for Fourier mode solutions implying consistency for all solutions as desired.

All methods were found to be consistent using the same reasoning. Their Tables can be found in Appendix C. In particular we have Tables C.9 and C.10 for FDVM<sub>1</sub>, Table C.11 for FDVM<sub>2</sub>, Tables C.12 and C.13 for FDVM<sub>3</sub>, Table C.14 for  $\mathcal{D}$  and Table C.15 for  $\mathcal{W}$ .

## 4.4 Dispersion Analysis

To study the dispersion properties of the numerical method, we must calculate the dispersion relation of the numerical method that relates the frequency  $\tilde{\omega}^\pm$  of the numerical solution to the wavenumber  $k$ . Making use of (4.6) in (4.14) we get

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

Therefore, the evolution matrix  $\mathbf{E}$  of an exact method has the eigenvalues  $e^{i\omega^+ \Delta t}$  and  $e^{i\omega^- \Delta t}$  where  $\omega^\pm$  are the positive and negative branches of the dispersion relation of the linearised Serre equations (2.9). For approximate numerical methods the dispersion relation for  $\tilde{\omega}^\pm$  can be calculated by taking the eigenvalues of its evolution matrix  $\lambda^\pm$  like so

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

By comparing  $\tilde{\omega}^\pm$  with the analytic  $\omega^\pm$  given by the linearised Serre equations (2.9) we can determine the error in the dispersion relation for the numerical method. The real part of the frequency determines the speed of a wave, while the imaginary part determines the change in amplitude. For the linearised Serre equations the imaginary part of  $\omega^\pm$  is zero and so the amplitude of waves are constant in time. We only present the results for the positive branch of the dispersion relation comparing  $\tilde{\omega}^+$  and  $\omega^+$  as the behaviour of the negative branch is very similar.

The relative error in the dispersion relation was plotted against  $\Delta x/\lambda$  for representative values of  $H$ ,  $U$  and  $k$ . Where  $\lambda = 2\pi/k$  is the wavelength of the wave with wave number  $k$ . We used  $g = 9.81 m/s^2$  and chose  $\Delta t = 0.5 / (U + \sqrt{gH}) \Delta x$  to satisfy the CFL condition (3.23).

In Figures 4.2 and 4.3 we present the plots for  $kH = \pi/10$  where  $\sigma = kH/2\pi = 1/20$  and so the water is shallow and thus the Serre equations are appropriate.

We present the real and imaginary errors separately to isolate the errors in the speed and amplitude of the wave for the numerical method. The total error is also reported as a measure of the overall error in the dispersion relation of the numerical method.

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

For the real part of the dispersion error all the FEVM and the FDVM outperform the two finite difference methods and therefore will better approximate the speed of waves of the linearised Serre equations. However, for the amplitude of waves the roles are reversed with the two finite difference methods either scaling the waves very little or not at all. When taking both effects into account with the total error we see that the FDVM<sub>1</sub> has the largest dispersion error followed by  $\mathcal{W}$ ,  $\mathcal{D}$ , FEVM<sub>2</sub>, FDVM<sub>2</sub> and finally FDVM<sub>3</sub> 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. Overall the methods built around a FVM perform better than the finite difference methods of the same order.

Furthermore, Figures 4.2 and 4.3 demonstrate that FDVM<sub>2</sub> is superior to FEVM<sub>2</sub> not just for the complete dispersion error, but its real and imaginary parts individually as well. Therefore, FDVM<sub>2</sub> should more accurately model the speed and amplitude of waves.

We observed similar results across a wide array of  $k$ ,  $H$  and  $U$  values. However, as  $kH$  is increased the distinction between FDVM<sub>2</sub> and FEVM<sub>2</sub> becomes less pronounced. This can be seen in Figure 4.4 where  $kH = 2.5$  and  $\sigma = 5/4\pi > 1/20$  where the water is no longer shallow.

These  $kH$  values are the same as those reported by Filippini et al. [39], 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. [39]. Furthermore, we extended their work by allowing for non-zero values of  $U$ , combining the spatial and temporal approximations and examining the imaginary and total error in the dispersion relation. This work also extended the dispersion analysis presented by Zoppou et al. [14] for FDVM<sub>1</sub>, FDVM<sub>2</sub> and FDVM<sub>3</sub> by including non-zero values of  $U$ .

Figure 4.5 demonstrates that the results of the real part of the dispersion error is slightly different if we allow for non-zero values of  $U$ . For example the non-zero value of  $U$  significantly changes the real part of the dispersion error for FDVM<sub>1</sub> when  $kH = 2.5$ . Therefore, for some methods allowing for non-zero values of

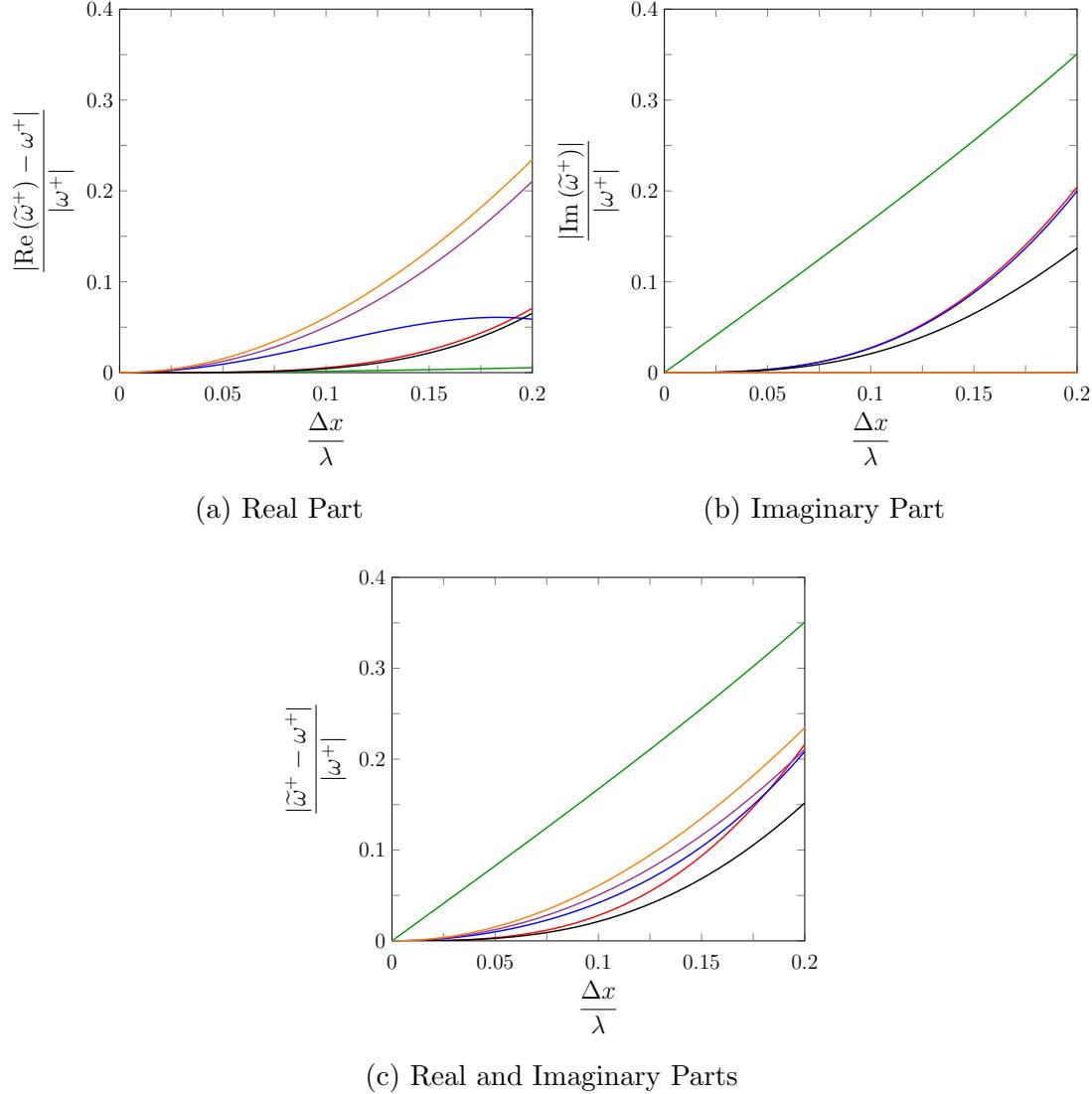


Figure 4.2: Relative dispersion error against  $\Delta x/\lambda$  when  $H = 1m$ ,  $k = \pi/10m^{-1}$  and  $U = 0m/s$  for FDVM<sub>1</sub> (—), FDVM<sub>2</sub> (—), FEVM<sub>2</sub> (—), FDVM<sub>3</sub> (—),  $\mathcal{D}$  (—) and  $\mathcal{W}$  (—).

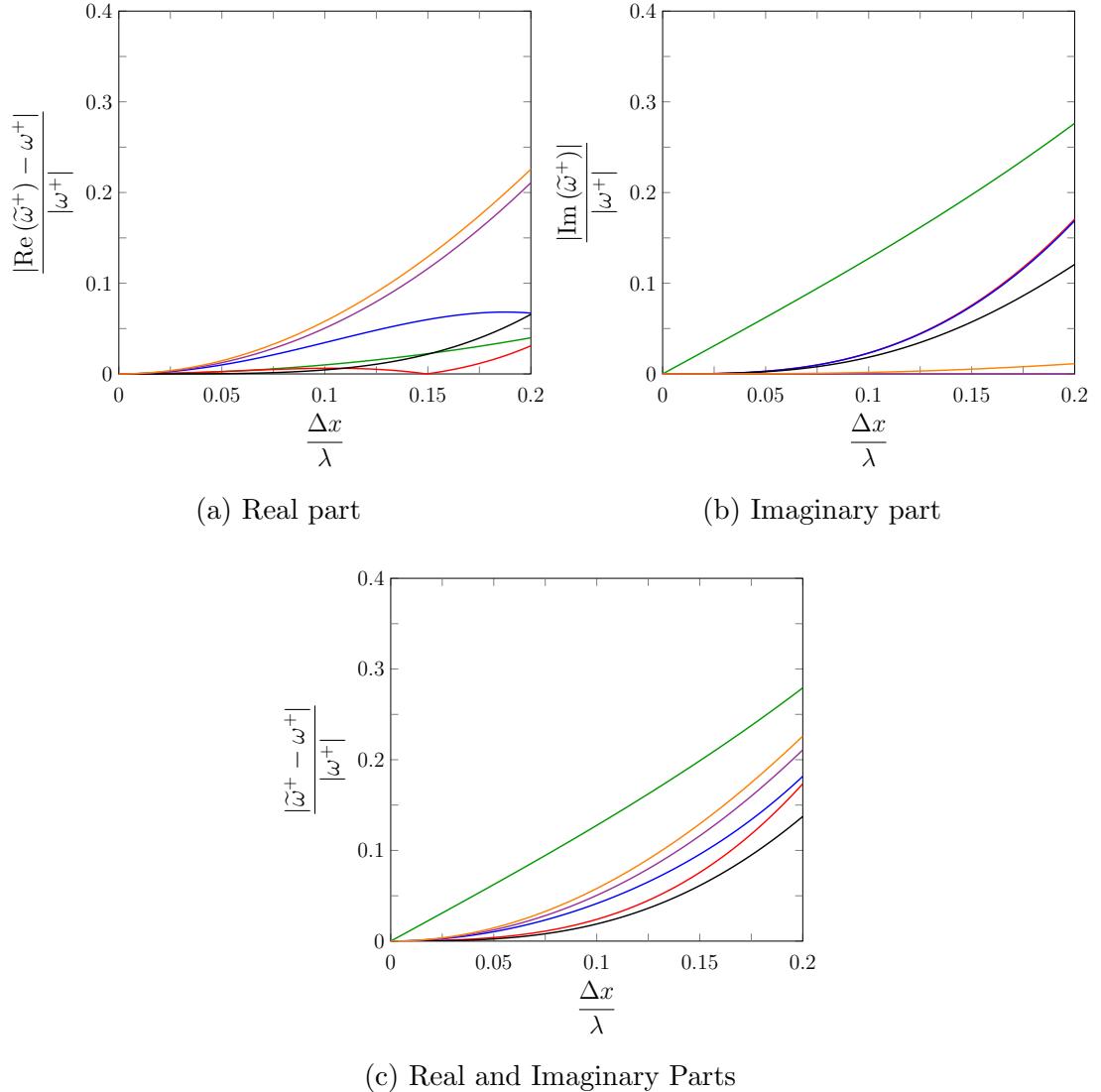


Figure 4.3: Relative dispersion error against  $\Delta x/\lambda$  when  $H = 1m$ ,  $k = \pi/10m^{-1}$  and  $U = 1m/s$  for FDVM<sub>1</sub> (—), FDVM<sub>2</sub> (—), FEVM<sub>2</sub> (—), FDVM<sub>3</sub> (—),  $\mathcal{D}$  (—) and  $\mathcal{W}$  (—).

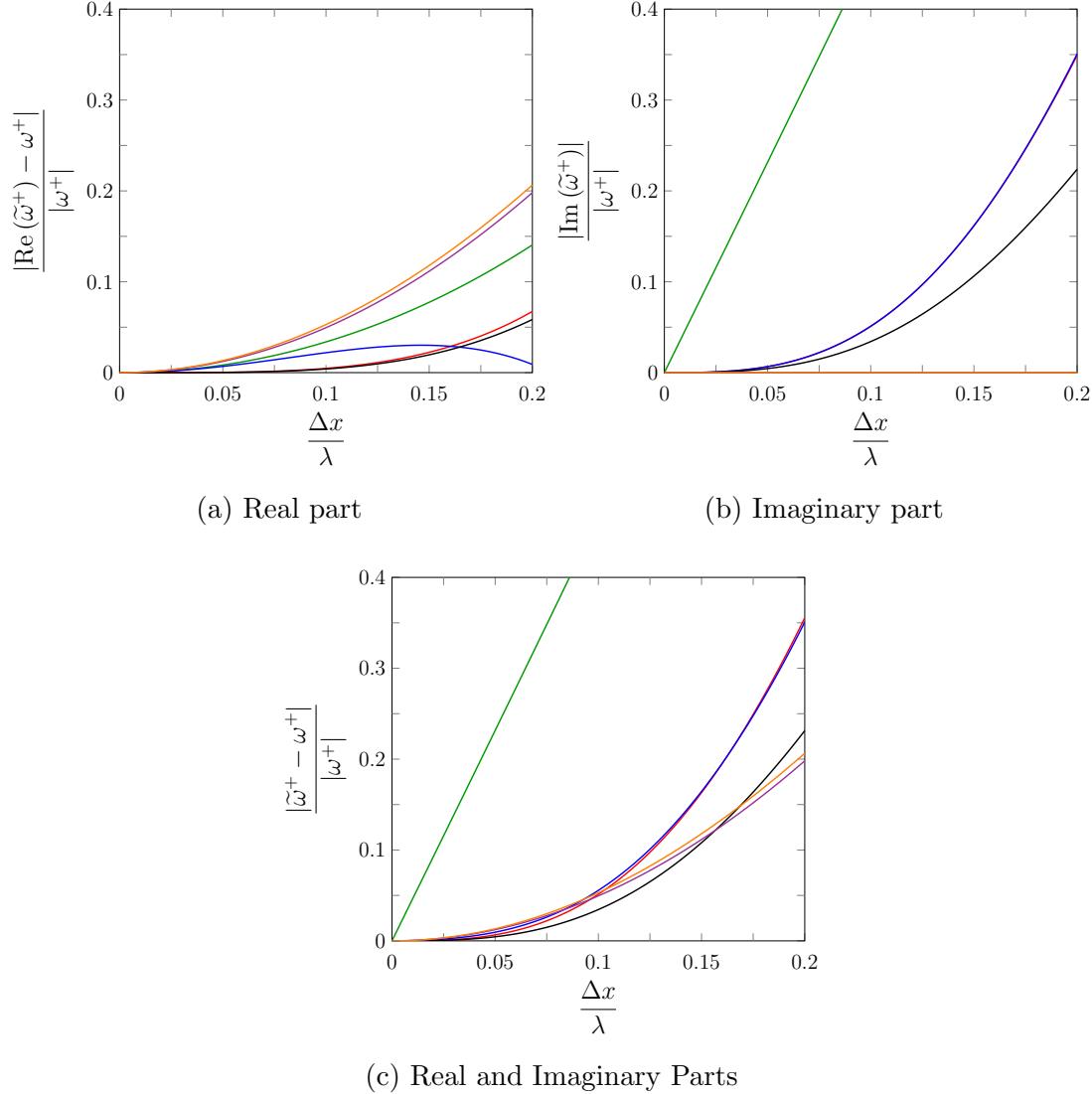


Figure 4.4: Relative dispersion error against  $\Delta x/\lambda$  when  $H = 1m$ ,  $k = 2.5m^{-1}$  and  $U = 0m/s$  for FDVM<sub>1</sub> (green), FDVM<sub>2</sub> (red), FEVM<sub>2</sub> (blue), FDVM<sub>3</sub> (black),  $\mathcal{D}$  (purple) and  $\mathcal{W}$  (orange).

$U$  can have a significant impact on the conclusions drawn from the dispersion analysis. Furthermore, taking the imaginary part of the dispersion error into account is important as  $\omega^\pm$  determines not only the speed of waves but also their amplitude. For instance the FDVM<sub>1</sub> performs very well for the real part of the dispersion error and poorly for the imaginary part, and so false conclusions about the accuracy of the method could be drawn from only considering the real part of the dispersion error.

The Taylor series expansion of  $\tilde{\omega}^\pm$  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 4.2 when  $-1 \leq Fr \leq 1$  for the FDVM and FEVM. In Table 4.2 it is clear that these schemes estimated  $\omega^+$  with the expected order of accuracy in both space and time. This was also the case for  $\omega^-$ .

We also present the lowest order terms of the Taylor series for  $\tilde{\omega}^+ - \omega^+$  for both  $Fr < -1$  and  $Fr > 1$  in Table 4.3. We only present the errors that are different from those reported in Table 4.2, this was only the case for the spatial error of the first- and third-order numerical methods. From these tables it is clear that the FDVM and the FEVM retain their order of accuracy when approximating  $\omega^+$  when the flow is supercritical, this was also the case for  $\omega^-$ .

Finally we present the lowest order terms of the Taylor series for  $\tilde{\omega}^+ - \omega^+$  for the finite difference methods in Table 4.4. These methods do not change depending on the value of the physical quantities. The two finite difference methods retain their order of accuracy in space and time when approximating  $\omega^+$ .

Because all methods were demonstrated to have the expected order of accuracy in approximating  $\omega^\pm$  for the linearised Serre equations this implies that for small  $\Delta x$  values the order of accuracy will be the primary driver of the dispersion error, as was observed.

In this chapter the convergence and dispersion properties of the numerical methods were studied using a linear analysis. The results of this analysis demonstrated the superiority of the high-order accurate FDVM and FEVM over the finite difference methods. The numerical methods will now be validated against analytic and forced solutions.

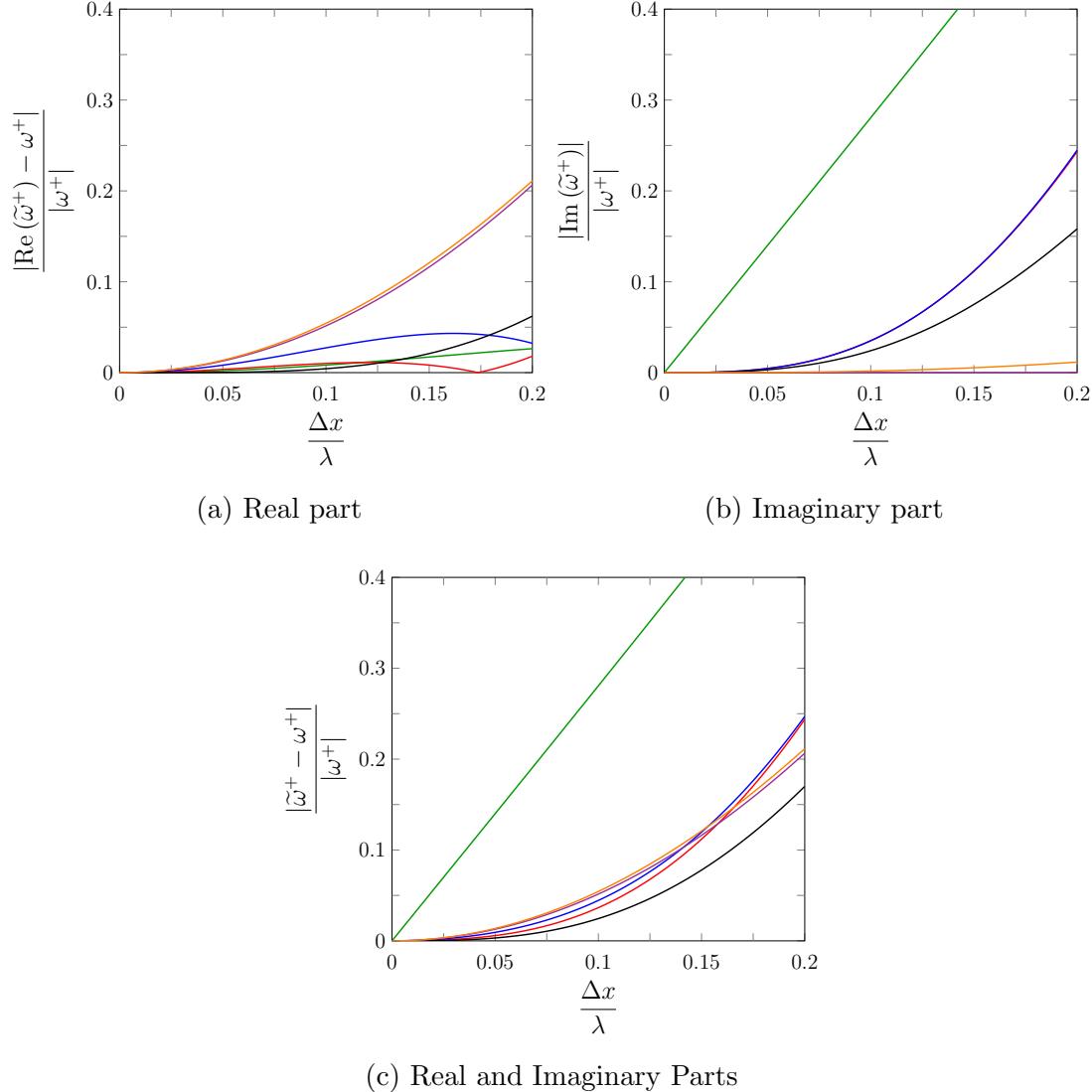


Figure 4.5: Relative dispersion error against  $\Delta x/\lambda$  when  $H = 1m$ ,  $k = 2.5m^{-1}$  and  $U = 1m/s$  for FDVM<sub>1</sub> (green), FDVM<sub>2</sub> (red), FEVM<sub>2</sub> (blue), FEVM<sub>3</sub> (black),  $\mathcal{D}$  (purple) and  $\mathcal{W}$  (orange).

Scheme	Lowest Order Terms of $\tilde{\omega}^+ - \omega^+$	
	$\Delta x$	$\Delta t$
FDVM <sub>1</sub>	$-\left(2\sqrt{gH} - \sqrt{\frac{3U}{\beta}}\right) \frac{ik^2}{4} \Delta x$	$\frac{i(\omega^+)^2}{2} \Delta t$
FDVM <sub>2</sub>	$\frac{2\beta U - 3\sqrt{3gH\beta}}{\beta^2} \frac{k^3}{24} \Delta x^2$	$-\frac{(\omega^+)^3}{6} \Delta t^2$
FEVM <sub>2</sub>	$\left(U + \frac{(42 + 15k^2H^2)\sqrt{3gH\beta}}{20\beta^2}\right) \frac{k^3}{12} \Delta x^2$	$-\frac{(\omega^+)^3}{6} \Delta t^2$
FDVM <sub>3</sub>	$-\left(2\sqrt{gH} - \sqrt{3\beta}U\right) \frac{ik^4}{24} \Delta x^3$	$-\frac{i(\omega^+)^4}{24} \Delta t^3$

Table 4.2: Lowest order terms for Taylor series of  $\tilde{\omega}^+ - \omega^+$  for all FDVM and the FEVM. With  $-1 \leq Fr \leq 1$  and  $\beta = 3 + H^2k^2$ .

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

Table 4.3: Lowest order  $\Delta x$  term for Taylor series of  $\tilde{\omega}^+ - \omega^+$  for all FDVM for supercritical Froude numbers where different from Table 4.2. With  $\beta = 3 + H^2k^2$ .

Scheme	Lowest Order Terms of $\tilde{\omega}^+ - \omega^+$	
	$\Delta x$	$\Delta t$
$\mathcal{D}$	$-\chi\Delta x^2$	$-\frac{(\omega^+)^3}{3}\Delta t^2$
$\mathcal{W}$	$\chi\Delta x^2$	$\begin{aligned} & \frac{1}{\beta^2} \left( \beta U^2 [9\sqrt{3gH\beta} + 4\beta U] \right. \\ & \left. + 3gH^2 [\sqrt{3gH\beta} + 6\beta U] \right) \frac{k^3}{18} \Delta t^2 \end{aligned}$

Table 4.4: Lowest order terms of the Taylor series of  $\tilde{\omega}^+ - \omega^+$  for  $\mathcal{D}$  and  $\mathcal{W}$ .  
With  $\beta = 3 + H^2 k^2$  and  $\chi = \left( U + \frac{(4 + H^2 k^2) \sqrt{3gH\beta}}{4\beta^2} \right)$ .

# Chapter 5

## Numerical Validation

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

- The second-order Finite Element Volume Method (FEVM) termed FEVM<sub>2</sub> which is described in Chapter 3.
- The first-, second- and third-order Finite Difference Volume Methods (FDVM) termed FDVM<sub>1</sub>, FDVM<sub>2</sub> and FDVM<sub>3</sub> respectively. Descriptions of these methods were published by Zoppou et al. [14]. Additionally, FDVM<sub>2</sub> was extended to handle dry beds using the technique outlined in Chapter 3.
- The second-order finite difference methods  $\mathcal{D}$  and  $\mathcal{W}$  whose descriptions were published by Pitt et al. [17].

To verify that the numerical methods have the expected convergence and conservation properties we make use of the analytic and forced solutions described in Chapter 2. To assess these properties we first introduce the measures of convergence and conservation for a numerical solution. These measures are then used to compare all the numerical methods using the solitary travelling wave solution. The convergence and conservation properties of FDVM<sub>2</sub> and FEVM<sub>2</sub> are compared using the lake at rest solution. Currently, the FDVM<sub>2</sub> and FEVM<sub>2</sub> are the only methods in this thesis that incorporate varying bathymetry.

Finally we validate FDVM<sub>2</sub> and FEVM<sub>2</sub> using forced solutions which test the accuracy of their approximations to all terms in the Serre equations. The forced solutions are obtained by adding terms to the Serre equations (2.13) to force any desired solution. This allows the method to be validated against more flow scenarios than possible given the limited number of currently known analytic

solutions. These forced solutions can be any functions and so the forced Serre equations are no longer strictly conservative. Therefore, the forced solutions are only used to assess the convergence properties of these numerical methods.

## 5.1 Measuring Convergence and Conservation

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

### 5.1.1 Measure of Convergence

By measuring the relative difference between the numerical and analytic solutions as  $\Delta x$  varies, the convergence of the numerical methods can be investigated. To measure the relative difference we use the  $L_2$  vector norm to compare the numerical and analytic solutions at the cell midpoints  $x_j$  at the end of the simulations. For a quantity  $q$ , the vector of its values  $\mathbf{q}$  at the cell midpoints  $x_j$  and the corresponding numerical solution at those locations  $\mathbf{q}^*$ ; the  $L_2$  norm is

$$L_2(\mathbf{q}, \mathbf{q}^*) = \begin{cases} \frac{\|\mathbf{q}^* - \mathbf{q}\|_2}{\|\mathbf{q}\|_2} & \|\mathbf{q}\|_2 > 0 \\ \|\mathbf{q}^*\|_2 & \|\mathbf{q}\|_2 = 0. \end{cases}$$

### 5.1.2 Measures of Conservation

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

$$C(q, \mathbf{q}^*) = \begin{cases} \frac{|\mathcal{C}^*(\mathbf{q}^*) - \mathcal{C}(q(x, 0))|}{|\mathcal{C}(q(x, 0))|} & |\mathcal{C}(q(x, 0))| > 0 \\ |\mathcal{C}^*(\mathbf{q}^*)| & |\mathcal{C}(q(x, 0))| = 0 \end{cases} \quad (5.1)$$

is used. The total amount in a numerical solution  $\mathcal{C}^*(\mathbf{q}^*)$  was calculated using 3 point Gaussian quadrature over the  $j^{th}$  cell and summing these cell integrals for all  $j$ . The value of  $q$  at 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  $q_{j+2}$  at the surrounding cell midpoints. Gaussian quadrature using three points is 5<sup>th</sup> order accurate and interpolation by quartics is 5<sup>th</sup> order accurate for the quantity  $q$  and 4<sup>th</sup> 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  $h$ ,  $uh$ ,  $G$  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 approximate  $\mathcal{C}(q(x, 0))$  with  $\mathcal{C}^*(\mathbf{q}^0)$  in (5.1); where  $\mathbf{q}^0$  is the vector of the quantity at the cell midpoints used as the initial conditions of our numerical method. We denote the numerical approximation to the conservation error (5.1) by  $C^*$ . Errors generated by the discretisation will be similar for  $\mathcal{C}^*(\mathbf{q}^0)$  and  $\mathcal{C}^*(\mathbf{q}^*)$  and will therefore cancel when studying  $C^*$ . Hence, the completely numerically calculated conservation error will be useful in cases where the errors introduced by the discretisation dominate.

## 5.2 Solitary Travelling Wave Solution

To assess the ability of our numerical methods to solve the Serre equations with a horizontal bed we use the solitary travelling wave solution (2.11) described in Chapter 2. This is a particular member of the family of periodic travelling wave solutions [31]. Every member of this family of solutions except the trivial stationary one have the same non-zero terms and thus provide similar tests for the numerical methods. Hence, it is sufficient to only study the solitary travelling wave solution.

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

For our numerical tests we used the solitary travelling wave solution (2.11) with  $a_0 = 1m$ ,  $a_1 = 0.7m$  and  $g = 9.81m/s^2$  at  $t = 0s$  as the initial conditions. The spatial domain was  $[-250m, 250m]$  and the problem was solved until  $t = 50s$ .

This was done for a range of  $\Delta x$  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  $\Delta t = Cr\Delta x / \sqrt{g(a_0 + a_1)}$ . For FDVM<sub>2</sub> and FEVM<sub>2</sub> the limiting parameter  $\theta = 1.2$  was used in the generalised minmod limiter (3.2) employed by both methods during the reconstruction step. While FDVM<sub>3</sub> used the Koren limiter [55] in its reconstruction [14], which has no limiting parameter.

For the parameters  $a_0 = 1m$  and  $a_1 = 0.7m$  the non-linearity is  $\epsilon = a_1/a_0 = 0.7$ ; this is large but beneath most of the well known breaking thresholds for water waves  $\epsilon \leq 0.8$  [56]. Because  $\epsilon$  is large the non-linear effects are large and therefore, so are the balancing dispersive effects making this particular analytic solution a rigorous test of the numerical methods. For this spatial domain and final time  $t = 50s$  there is no interaction of the wave with the boundary, therefore the Dirichlet boundary conditions were appropriate.

The results of this analytic solution validation were published by Zoppou et al. [14] for FDVM<sub>1</sub>, FDVM<sub>2</sub> and FDVM<sub>3</sub>. I produced the results in that paper which have been expanded here to include an investigation of the convergence of  $G$  and the conservation of  $h$ ,  $uh$  and  $G$ .

An example numerical solution with  $\Delta x = 100/2^{11}m \approx 0.049m$  from all methods was plotted in Figure 5.1 against the analytic solution at  $t = 50s$ . We have only plotted an illustrative amount of the points in the numerical solution. From these plots it is clear that FDVM<sub>1</sub> performs significantly worse than the high-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 high-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  $L_2$  error was calculated for  $h$ ,  $u$  and  $G$  for all numerical solutions and was plotted against  $\Delta x$  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 as a function of  $\Delta x$  is determined by the order of accuracy of the numerical scheme. All methods demonstrate the expected order of accuracy given the order of accuracy of the approximations used in the method. This was also true for the results of the linear analysis in Chapter 4 and Appendix C.

All methods more accurately reproduced the analytic solution for  $h$  than either

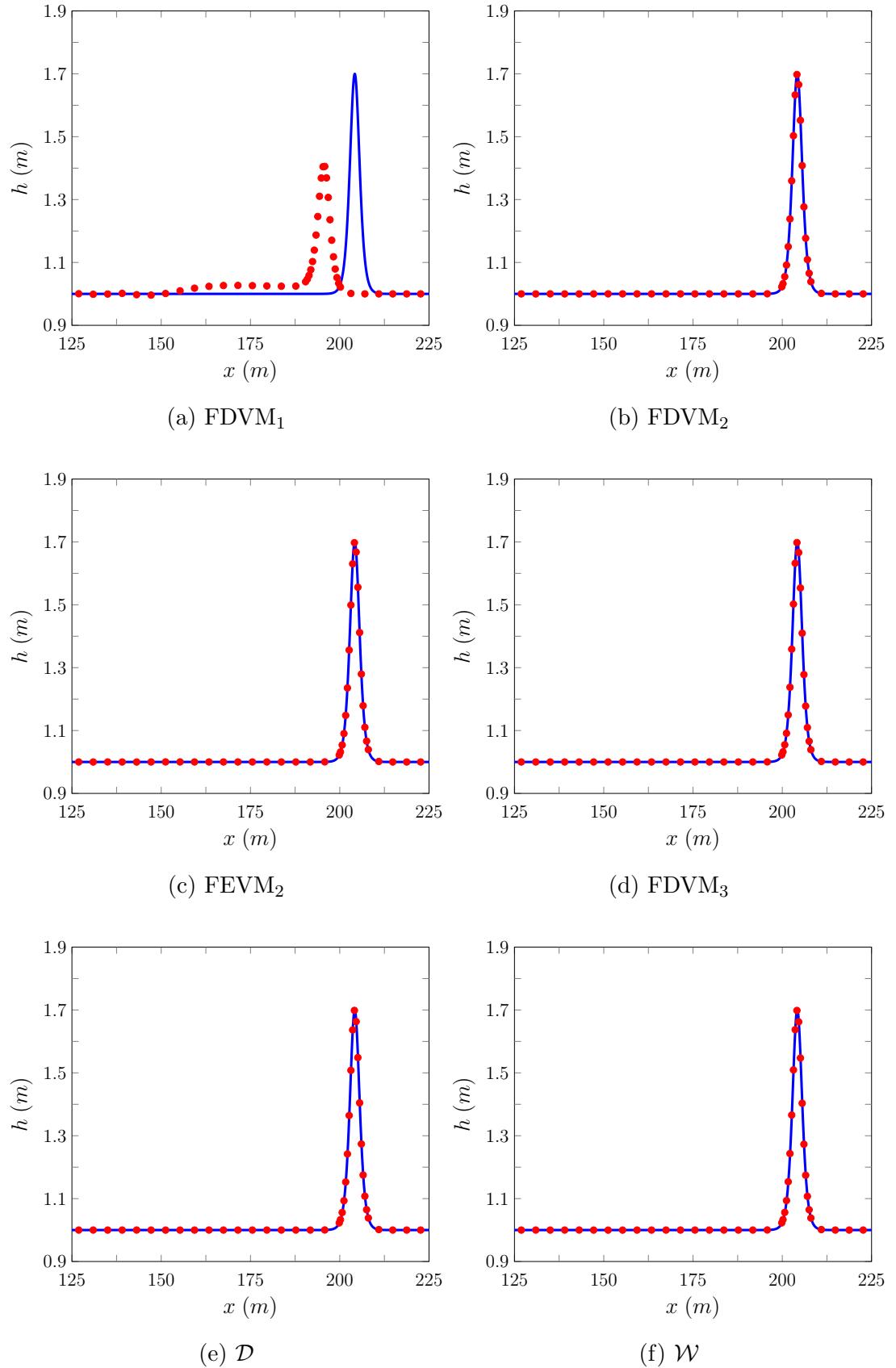


Figure 5.1: Comparison of the analytic solution (—) and numerical solution with  $\Delta x = 100/2^{11}m$  (•) for the solitary travelling wave solution at  $t = 50s$  for all methods.

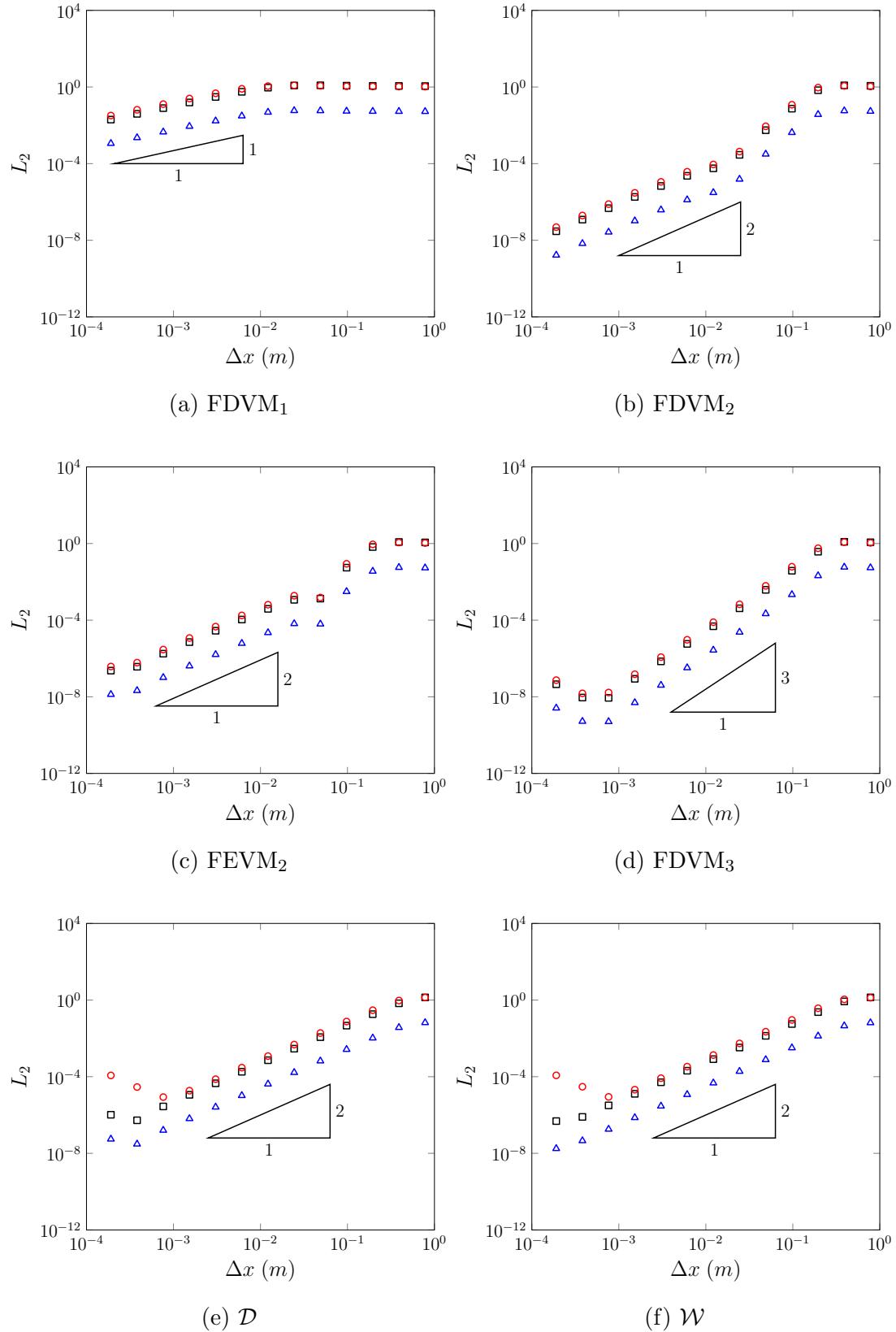


Figure 5.2: Convergence as measured by the  $L_2$  norm against  $\Delta x$  for  $h$  ( $\Delta$ ),  $u$  ( $\square$ ) and  $G$  ( $\circ$ ) for the soliton problem for all methods.

$G$  or  $u$  across all  $\Delta x$  values. This is due to the simplicity of the evolution equation for  $h$  (2.6a) compared to the evolution equations of  $G$  (2.6b) and  $uh$  (2.4b). With the error in  $G$  transmitted to  $u$  when solving (2.7) in the finite volume based methods.

Increasing the order of accuracy of our numerical methods leads to smaller errors when comparing two numerical solutions for the same  $\Delta x$  value, as Figure 5.2 clearly demonstrates. This is consistent with the example numerical solution in Figure 5.1, where the lowest order accuracy scheme, FDVM<sub>1</sub> had the poorest reproduction of the analytic solution. However, there is only a slight benefit from moving from the second-order FEVM<sub>2</sub> and FDVM<sub>2</sub> to the third-order FDVM<sub>3</sub>.

For the second-order methods we find that FDVM<sub>2</sub> consistently produces the smallest  $L_2$  error followed by FEVM<sub>2</sub>,  $\mathcal{W}$  and  $\mathcal{D}$ . The difference between the FDVM<sub>2</sub> and FEVM<sub>2</sub> is significant with the errors of FEVM<sub>2</sub> being 2 to 4 times larger than those of FDVM<sub>2</sub>. Therefore, FDVM<sub>2</sub> is reproducing the solitary wave solution more accurately than FEVM<sub>2</sub>.

The finite difference methods produce very similar errors which are twice as large as the errors from FEVM<sub>2</sub>. Additionally, the round-off effects dominate the  $L_2$  error of the finite difference methods at larger  $\Delta x$  values than for the finite volume based methods.

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

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

For small  $\Delta x$  values the round-off errors dominate the conservation error, particularly for the finite difference methods. Interestingly, FDVM<sub>3</sub> has an accumulation of round-off error increasing the conservation error for  $h$  and  $G$  as  $\Delta x$  decreases. This was found to be caused by the Runge-Kutta coefficients of the third-order time stepping method [14] not being exactly represented as floating point numbers. For the third-order SSP Runge-Kutta time stepping method the

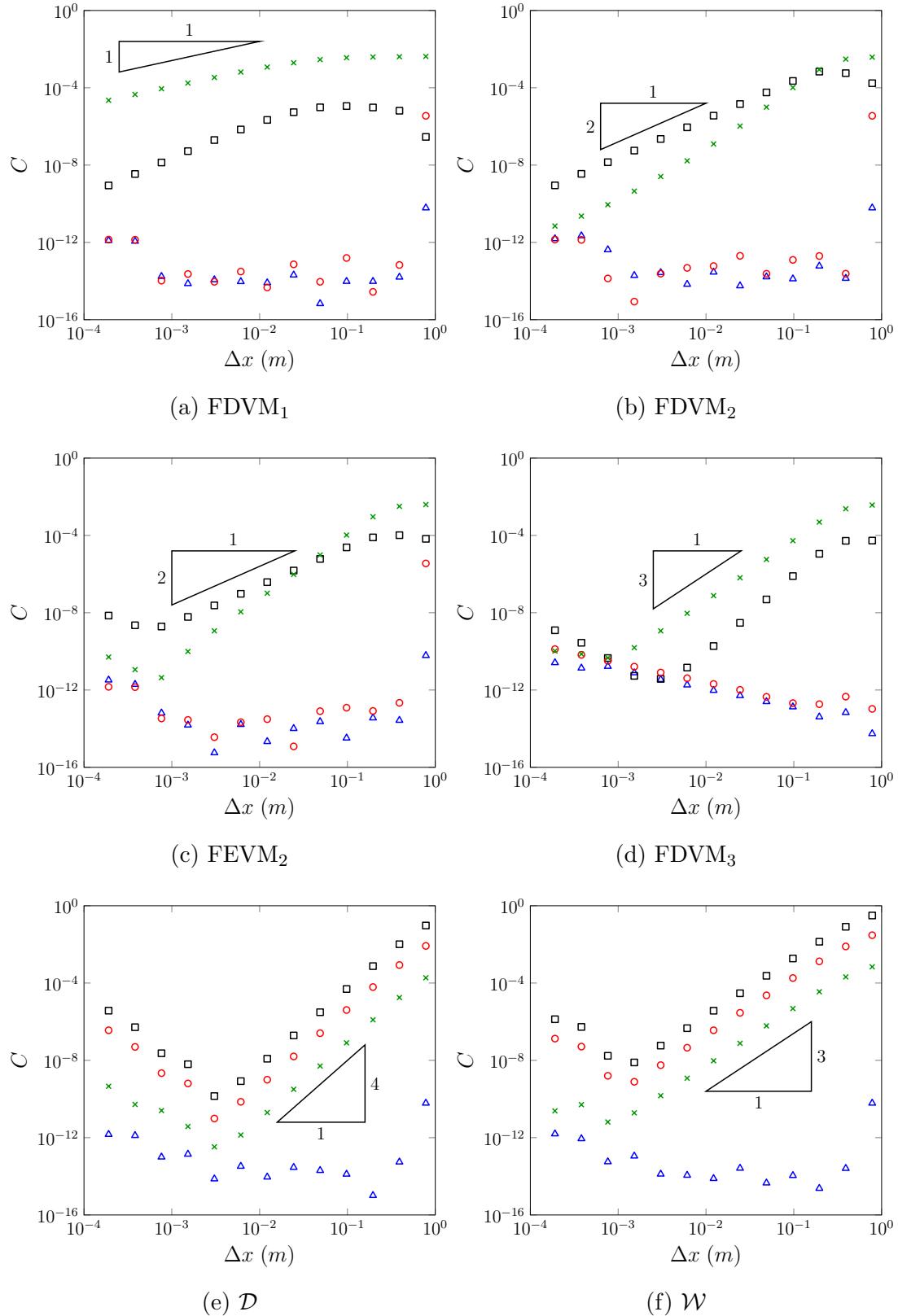


Figure 5.3: Conservation error  $C$  against  $\Delta x$  for  $h$  ( $\Delta$ ),  $uh$  ( $\square$ ),  $G$  ( $\circ$ ) and  $\mathcal{H}$  ( $\times$ ) for the solitary travelling wave solution for all methods.

coefficients in the last step are  $1/3$  and  $2/3$ . Since these numbers are not exactly represented in floating point they are approximated with a small error that when summed does not maintain the conservation properties of  $h$  and  $G$ . Thus, every time step accumulates a small conservation error of machine precision size leading to the observed increase as  $\Delta x$  becomes small and the number of time steps increases. Remedies for this were attempted such as using the coefficients  $1/3$  and  $(1 - 1/3)$  and bringing the common divisor out but this problem persisted. Some other numerical techniques are required to resolve this issue such as those of Higham [57]. Ultimately, since the convergence of FDVM<sub>3</sub> was only slightly better than FEVM<sub>2</sub> and FDVM<sub>2</sub> [14] it was not developed further and this issue was not resolved.

These results demonstrate the need for high-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. This was also the conclusion of the analytic validation by Zoppou et al. [14]. Finally, they demonstrate the ability of FEVM and FDVM to conserve  $h$  and  $G$  up to machine precision, as desired. Given these results, only FEVM<sub>2</sub> and FDVM<sub>2</sub> 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 consider FEVM<sub>2</sub> and FDVM<sub>2</sub>.

### 5.3 Lake at Rest Solution

To verify the validity of our numerical methods for the Serre equations with variable bathymetry and assess the well-balancing method we compare various numerical solutions to the lake at rest solution (2.12).

The particular lake at rest solution (2.12) associated with the bed profile

$$b(x) = a_1 \sin(a_2 x) \quad (5.2)$$

was chosen for this validation to ensure that all terms with derivatives of the bed were tested. To demonstrate the capability of the methods in the presence of dry and wet beds the parameter values  $a_0 = 0m$ ,  $a_1 = 1m$  and  $a_2 = 2\pi/50m^{-1}$  were chosen. These parameter values result in wet regions with a horizontal free surface where the stage  $w(x, t) = h(x, t) + b(x) = a_0 = 0$  (2.12). Therefore, we have a periodic bed where water submerges the troughs of the bed while the peaks of the bed are dry.

For the numerical solutions the spatial domain was  $x \in [-112.5m, 87.5m]$  and the final time was  $t = 10s$ , with the standard gravitational acceleration  $g = 9.81m/s^2$ . The spatial resolution of the method was varied so that  $\Delta x = 100/2^k m$  with  $k \in [8, \dots, 17]$  and the CFL condition (3.23) was satisfied by having  $\Delta t = Cr\Delta x/\sqrt{g}$  with condition number  $Cr = 0.5$ . The standard limiting parameter  $\theta = 1.2$  was used in the generalised minmod limiter, (3.2) for both FEVM<sub>2</sub> and FDVM<sub>2</sub>. Dirichlet boundary conditions were used at both ends as the analytic solution is stationary.

The numerical methods are assessed by using the specified lake at rest solution as initial conditions and comparing the numerical solutions of FEVM<sub>2</sub> and FDVM<sub>2</sub> at  $t = 10s$  to the analytic solution, which are the initial conditions. To demonstrate the utility of the well-balancing method the results from two versions of FEVM<sub>2</sub> and FDVM<sub>2</sub> are presented, where the well-balancing method described in Chapter 3 is and is not employed.

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

Examination of the  $L_2$  errors depicted in Figure 5.5 reveals that only the well-balanced methods have accurately recovered the analytic solution. With both well-balanced versions of the methods reproducing  $h$ ,  $G$  and  $u$  precisely, accounting for round-off errors. With the increasing error in  $h$ ,  $G$  and  $u$  as  $\Delta x \rightarrow 0$  caused by the accumulation of round-off errors.

For methods without well-balancing; the errors are significantly larger, yet they are converging to the analytic solution. However, the order of accuracy of the convergence in  $u$  and  $G$  has degraded and is not the expected second-order accuracy observed for  $h$ . The poor convergence of  $u$  and  $G$  is a result of the errors in  $u$  and  $G$  not being damped by the method. Thus errors generated by the imbalance between the flux and source terms increase over time degrading the order of accuracy. The second-order accuracy in  $h$  is retained for the presented  $\Delta x$  values as these errors introduced in  $u$  and  $G$  are small for a single cell, although for smaller  $\Delta x$  values these errors in  $u$  and  $G$  will begin to dominate the errors in  $h$ .

Using the expressions in Appendix A for the total amounts of the conserved quantities the conservation error  $C$  was calculated for FEVM<sub>2</sub> and FDVM<sub>2</sub> with the results plotted in Figure 5.6. The error in conservation of these methods affirms the superiority of the well-balanced version of the methods. In particular,

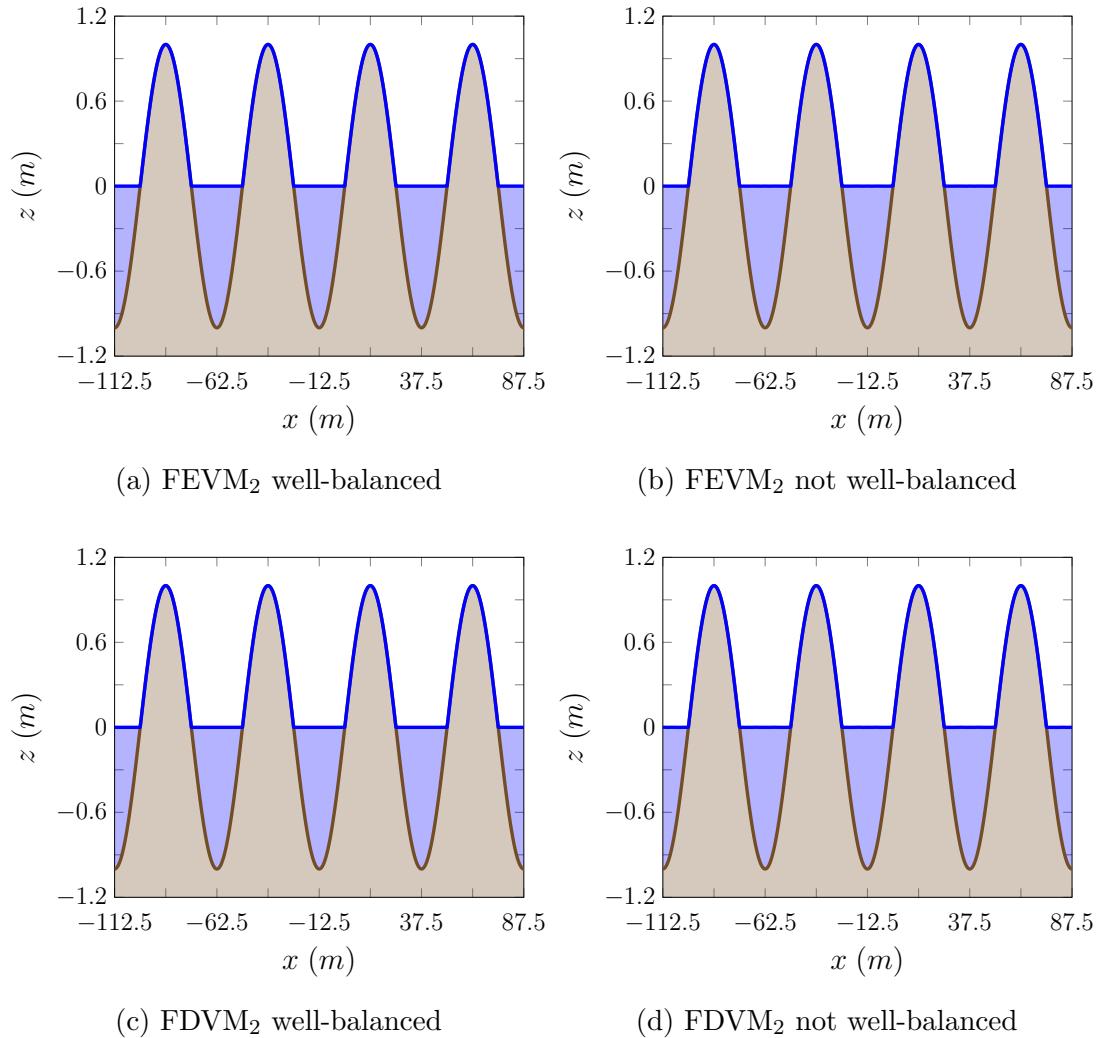


Figure 5.4: Numerical solutions for  $w$  (■) and  $b$  (□) with  $\Delta x = 100/2^{10}m$  for the lake at rest problem at  $t = 10s$  for FEVM<sub>2</sub> and FDVM<sub>2</sub>.

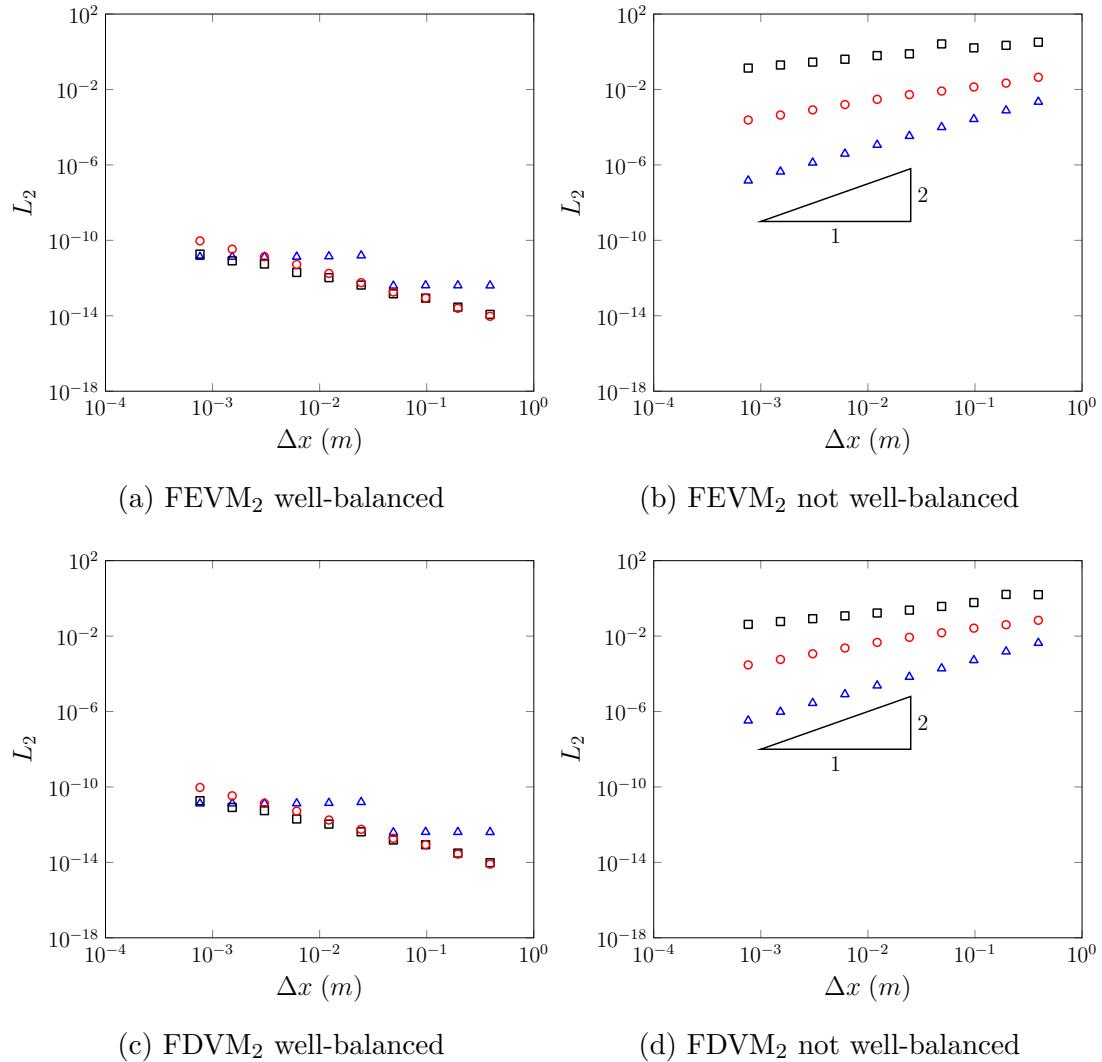


Figure 5.5: Convergence as measured by the  $L_2$  norm against  $\Delta x$  for  $h$  ( $\Delta$ ),  $u$  ( $\square$ ) and  $G$  ( $\circ$ ) for the lake at rest problem at  $t = 10s$  for FEVM<sub>2</sub> and FDVM<sub>2</sub>.

we see that the total amounts of  $uh$  and  $G$  are only conserved within machine precision when well-balancing is employed. Since  $uh$  and  $G$  are uniformly zero in the initial conditions the well-balanced methods have only introduced round-off errors into these quantities, whilst without well-balancing large errors in these quantities are introduced in the methods.

The errors in conservation of  $h$  and  $\mathcal{H}$  for the well-balanced methods are large, but do converge at the order of accuracy of the scheme or better. These errors are caused by the discretisation of the initial conditions, primarily the approximation of the boundaries of the wet regions by the numerical grids. This initial discretisation error can be removed by comparing the total amounts of the conserved quantities in the numerical solution to their numerically calculated total amounts in the initial conditions with  $C^*$  as in Figure 5.7. For the completely numerically calculated conservation error  $C^*$  we observe that all the conserved quantities are conserved at machine precision for the well-balanced methods. With  $\mathcal{H}$  being conserved exactly for most numerical solutions, hence its disappearance from the log-log plot. The conservation error of  $\mathcal{H}$  is small for the lake at rest solution since  $u$  is very small. Hence,  $\mathcal{H}$  is essentially the gravitational potential energy which since mass is well conserved is also well conserved.

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

## 5.4 Forced Solutions

The previous analytic solution validations do not provide a stringent test for all terms present in the Serre equations and there are currently no known analytic solutions that do. To remedy this the forced solutions introduced in Chapter 2 were used to validate the numerical methods. Since the added terms in the modified Serre equations, (2.13) can be determined and accounted for analytically, the only source of error in the numerical solutions are the numerical methods themselves and thus the theoretical second-order accuracy of FEVM<sub>2</sub> and FDVM<sub>2</sub> should be recovered.

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

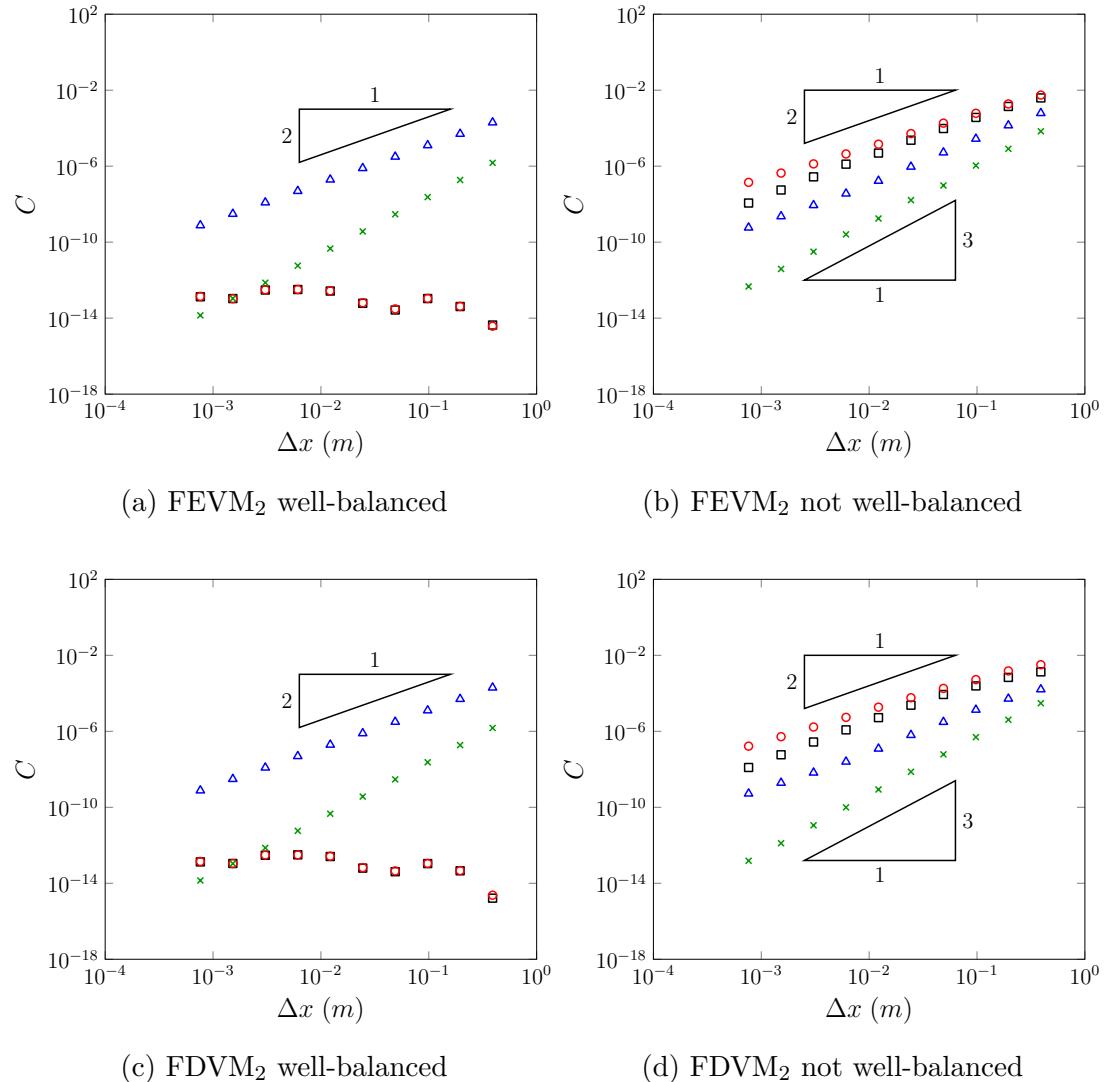


Figure 5.6: Conservation error  $C$  against  $\Delta x$  for  $h$  ( $\Delta$ ),  $uh$  ( $\square$ ),  $G$  ( $\circ$ ) and  $\mathcal{H}$  ( $\times$ ) for the lake at rest problem at  $t = 10s$  for FEVM<sub>2</sub> and FDVM<sub>2</sub>.

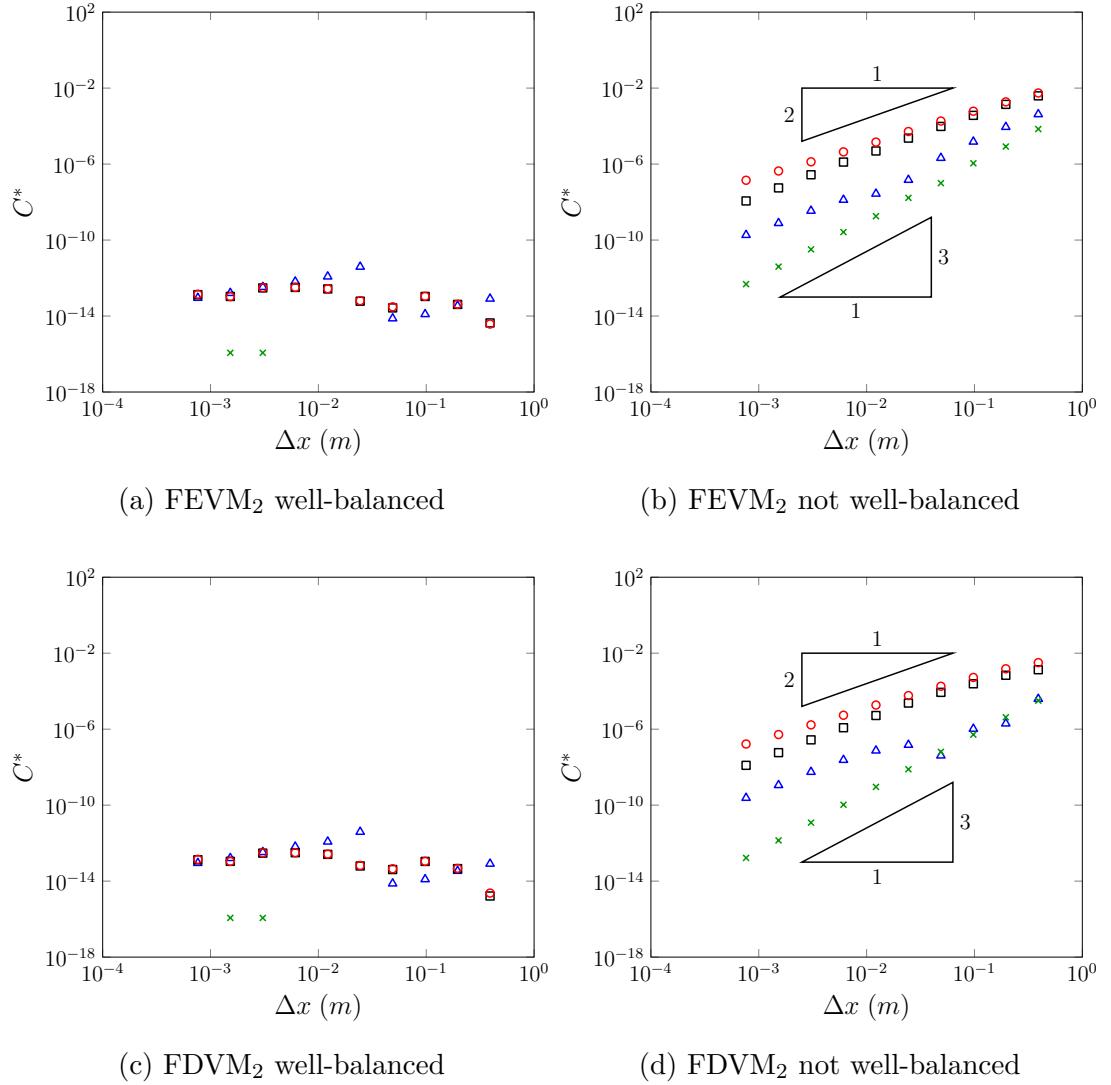


Figure 5.7: Conservation error using only numerical calculations  $C^*$  against  $\Delta x$  for  $h$  ( $\Delta$ ),  $uh$  ( $\square$ ),  $G$  ( $\circ$ ) and  $\mathcal{H}$  ( $\times$ ) for the lake at rest problem at  $t = 10s$  for FEVM<sub>2</sub> and FDVM<sub>2</sub>.

equations were accurately approximated in the numerical method the functions

$$h^*(x, t) = a_0 + a_1 \exp\left(-\frac{[(x - a_2 t) - a_3]^2}{2a_4}\right), \quad (5.3a)$$

$$u^*(x, t) = a_5 \exp\left(-\frac{[(x - a_2 t) - a_3]^2}{2a_4}\right), \quad (5.3b)$$

$$b^*(x) = a_6 \sin(a_7 x) \quad (5.3c)$$

for the primitive variables were chosen. These functions produce a Gaussian bump with an amplitude of  $a_1$  for  $h$  and  $u$  that travels at a fixed speed  $a_2$  over a periodic bed. Thus,  $h$  and  $u$  will have constant shape and travel to the right over time. However, this is not the case for  $G$  as  $u$  and  $h$  have constant shape but the bed is periodic. With the bed terms in  $G$  (2.7) changing the shape of  $G$  as the Gaussian bump in  $h$  and  $u$  encounters different bed slopes.

For non-trivial 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 the values  $a_1 = 0.5m$ ,  $a_2 = 2\pi/(10a_7)m/s$ ,  $a_3 = -3\pi/(2a_7)m$ ,  $a_4 = \pi/(16a_7)m^2$ ,  $a_5 = 0.5m/s$ ,  $a_6 = 1.0m$  and  $a_7 = \pi/25m^{-1}$  with  $a_0 = 1m$  for the finite water depth forced solution and  $a_0 = 0m$  for the dry bed forced solution. These parameter values result in a Gaussian bump in  $h$  and  $u$  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  $k \in [8, \dots, 17]$ . To satisfy the CFL condition, (3.23) the temporal resolution  $\Delta t = Cr\Delta x / (a_2 + a_5 + \sqrt{g(a_0 + a_1)})$  was chosen with condition number  $Cr = 0.5$ . The value  $\theta = 1.2$  was used in the generalised minmod limiter (3.2) for both FEVM<sub>2</sub> and FDVM<sub>2</sub> and Dirichlet boundary conditions were applied at the boundaries of the domain.

### 5.4.1 Results for a Wet Bed

For the non-zero water depth case where  $a_0 = 1m$  an example of the numerical solutions of FEVM<sub>2</sub> and FDVM<sub>2</sub> are given in Figures 5.8 and 5.9 respectively for  $\Delta x = 100/2^{10}m \approx 0.0977m$  at various times. The numerical solutions accurately

reproduced the forced solution as it travels over the bed. Thus  $h$  and  $u$  maintain their constant shape while  $G$  does not due to the influence of the periodic bed.

The  $L_2$  error of  $h$ ,  $u$  and  $G$  for the FEVM<sub>2</sub> and FDVM<sub>2</sub> are given in Figure 5.10. Both methods recover the expected second-order accuracy for  $h$ ,  $u$  and  $G$ . The additional term of the forced Serre equations is added analytically and all terms of the Serre equations must be accurately approximated by the method to reproduce this forced solution. Therefore, these results demonstrate that FEVM<sub>2</sub> and FDVM<sub>2</sub> are second-order accurate for all terms when the bed is wet everywhere, as desired.

#### 5.4.2 Results with a Dry Bed

To demonstrate the capability of the methods to handle wetting and drying of a bed, a series of numerical simulations of the forced solutions (5.3a) where  $a_0 = 0m$  were conducted using both FEVM<sub>2</sub> and FDVM<sub>2</sub>.

Example numerical solutions demonstrating the evolution of the wave are given in Figure 5.11 for FEVM<sub>2</sub> and in Figure 5.12 for FDVM<sub>2</sub> with  $\Delta x = 100/2^{10}m \approx 0.0977m$  at various times. The methods accurately reproduce the analytic solution for the stage  $w$ ,  $h$  and  $G$ . However, both fail to accurately reproduce  $u$  when  $h$  is small, particularly behind the Gaussian bump. So that now  $h$  is the only quantity that maintains a constant shape in the numerical solutions, as  $G$  changes due to the periodic bed and  $u$  changes due to numerical errors.

These large errors in  $u$  when  $h$  is small are caused by the particular choices  $h_{base} = 10^{-8}$  and  $h_{tol} = 10^{-12}$  used in the desingularisation transformation applied to the finite element method (3.12). By choosing larger values of these quantities the errors in  $u$  can be significantly damped. However, if  $h_{base}$  and  $h_{tol}$  are larger they begin to dominate the  $L_2$  errors in  $h$ ,  $G$  and  $uh$  making the convergence less obvious. This trade-off is present in all desingularisation transforms.

For our purposes the chosen desingularisation transform (3.27) with small  $h_{base}$  and  $h_{tol}$  values was sufficient, resulting in large observed errors in  $u$  when  $h$  is small.

The  $L_2$  errors for  $h$ ,  $u$ ,  $uh$  and  $G$  for both methods are given in Figure 5.13. Both methods exhibit second-order convergence in all the quantities except  $u$ . The large errors in  $u$  only occur when  $h$  is small. This can be seen by restricting our convergence to only compare regions where  $h > 10^{-3}m$  as in Figure 5.14. It can be seen that the expected second-order accuracy in  $u$  is recovered when  $h$  is

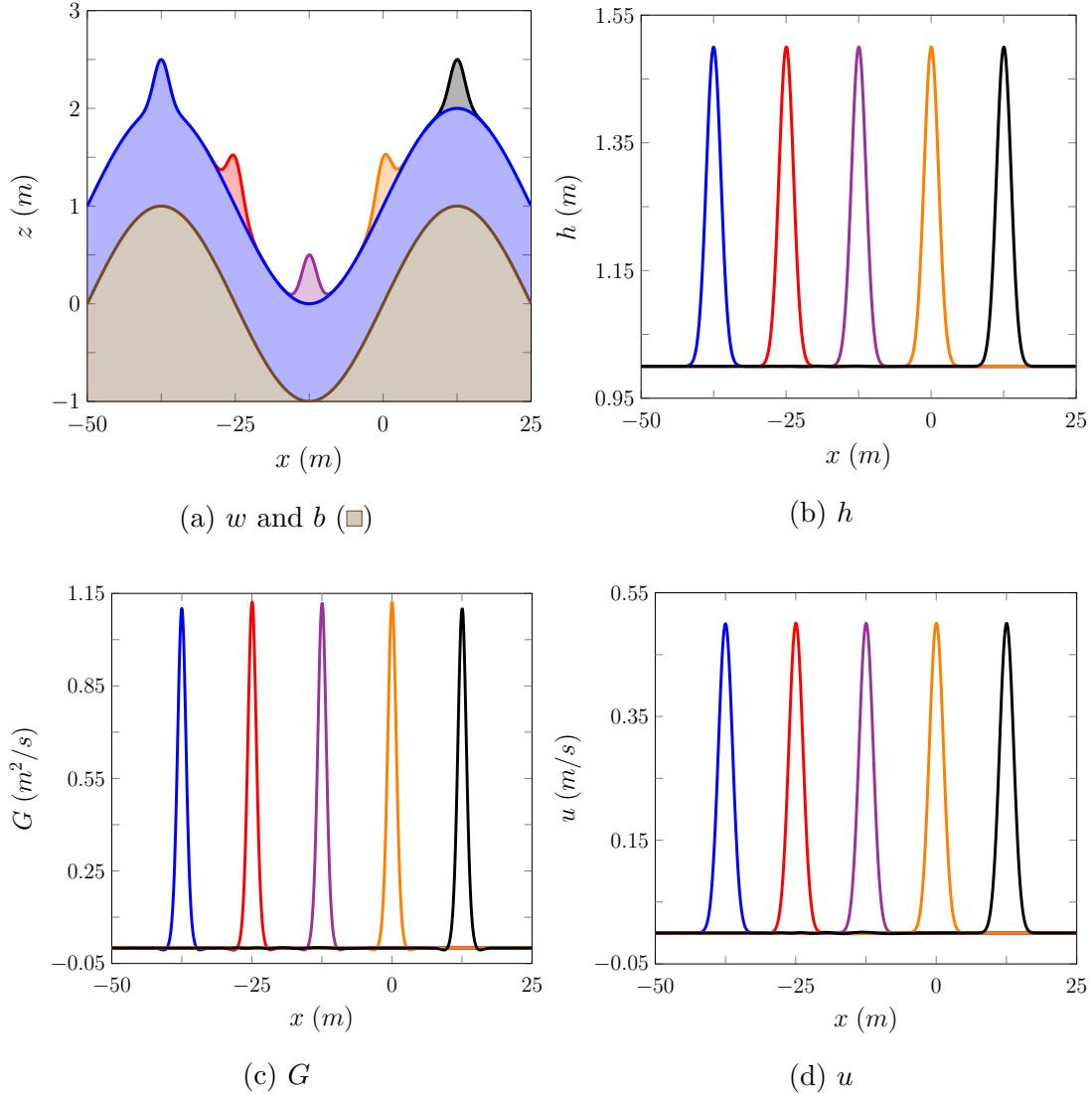


Figure 5.8: Numerical solutions for  $w$ ,  $b$ ,  $h$ ,  $G$  and  $u$  produced by FEVM<sub>2</sub> with  $\Delta x = 100/2^{10}m$  at  $t = 0s$  (— / □),  $2.5s$  (— / ■),  $5.0s$  (— / ▨),  $7.5s$  (— / □),  $10.0s$  (— / □) to the wet bed forced solution problem, where  $a_0 = 1m$ .

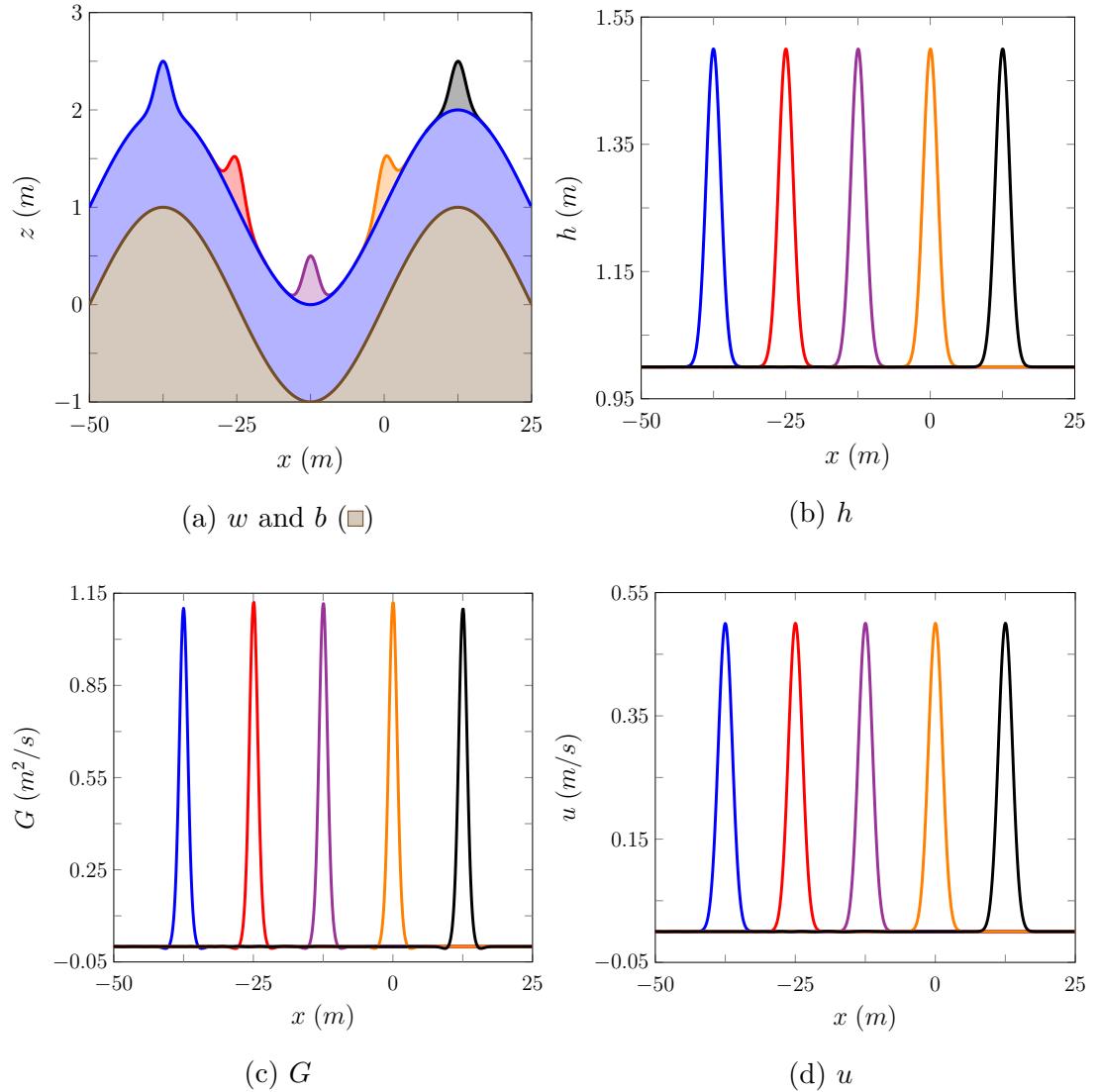


Figure 5.9: Numerical solutions for  $w$ ,  $b$ ,  $h$ ,  $G$  and  $u$  produced by FDVM<sub>2</sub> with  $\Delta x = 100/2^{10}m$  at  $t = 0s$  (— / ■),  $2.5s$  (— / □),  $5.0s$  (— / ▨),  $7.5s$  (— / ▢),  $10.0s$  (— / ▭) to the wet bed forced solution problem, where  $a_0 = 1m$ .

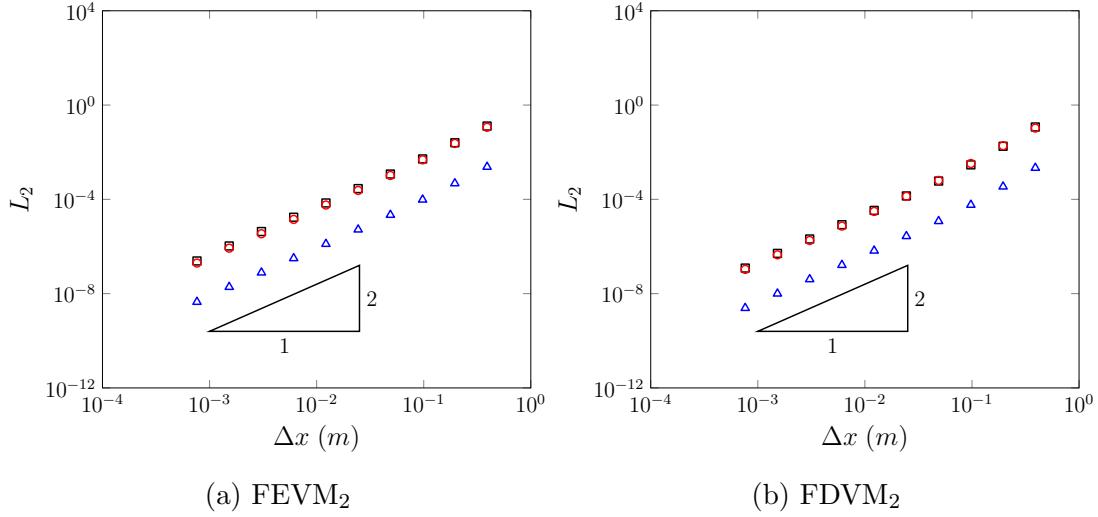


Figure 5.10: Convergence as measured by the  $L_2$  norm against  $\Delta x$  for  $h$  ( $\Delta$ ),  $u$  ( $\square$ ) and  $G$  ( $\circ$ ) for the wet bed forced solution problem for FEVM<sub>2</sub> and FDVM<sub>2</sub> at  $t = 10s$ .

not small. Since all the flux and source terms of the Serre equations (2.6) 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  $uh$ .

Therefore, these methods can accurately handle the dry bed problem, even with small  $h_{base}$  and  $h_{tol}$  values. Although, in such cases the velocity may have large errors in regions where  $h$  is small. For physical applications where large errors in  $u$  when  $h$  is small are not acceptable we recommend altering the dry bed handling of the scheme by increasing the  $h_{base}$  and  $h_{tol}$  values or altering the desingularisation transformation [52].

In this chapter the analytic and forced solutions were used to assess the numerical methods. It was found that the finite volume based methods performed better than the finite difference methods and that second-order methods were sufficient to accurately reproduce the analytic solutions. Finally the second-order accuracy of FEVM<sub>2</sub> and FDVM<sub>2</sub> was confirmed for the wetting and drying of variable beds using forced solutions to the Serre equations. The analytically validated FEVM<sub>2</sub> and FDVM<sub>2</sub> will now be validated against experimental results.

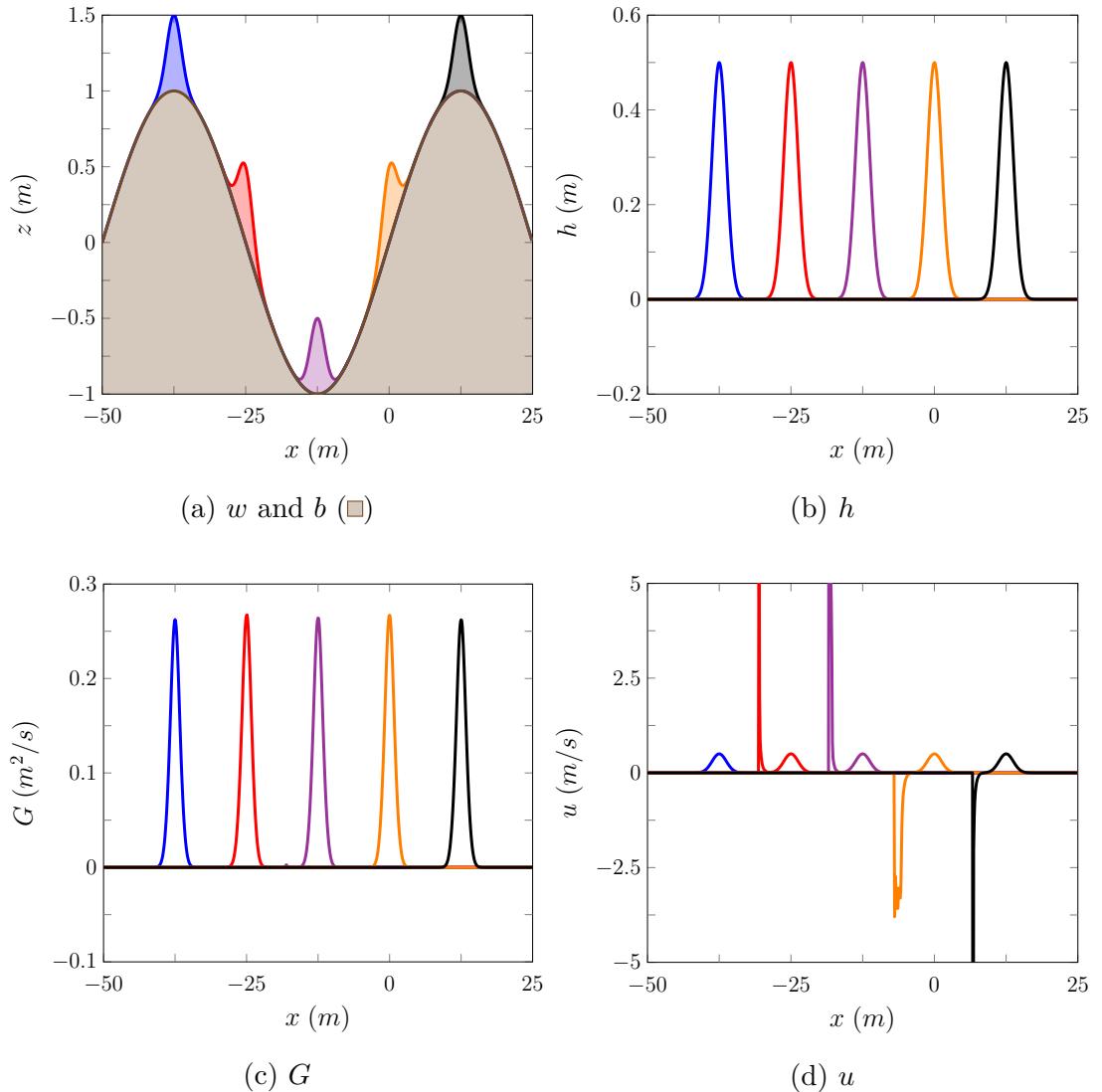


Figure 5.11: Numerical solutions for  $w$ ,  $b$ ,  $h$ ,  $G$  and  $u$  produced by FEVM<sub>2</sub> with  $\Delta x = 100/2^{10}m$  at  $t = 0s$  (— / □),  $2.5s$  (— / ▢),  $5.0s$  (— / ▣),  $7.5s$  (— / ▤),  $10.0s$  (— / ▥) to the dry bed forced solution problem, where  $a_0 = 0m$ .

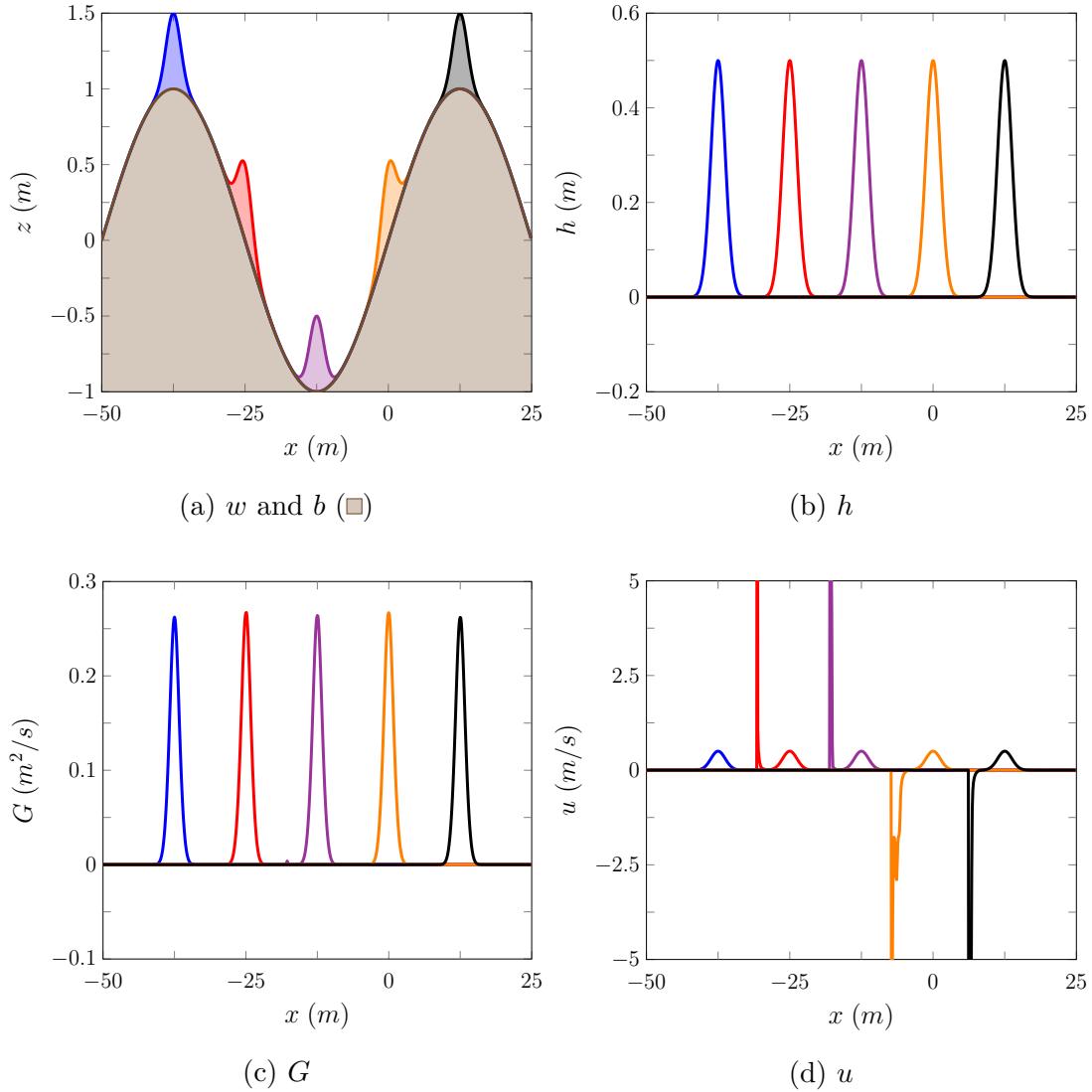


Figure 5.12: Numerical solutions for  $w$ ,  $b$ ,  $h$ ,  $G$  and  $u$  produced by FDVM<sub>2</sub> with  $\Delta x = 100/2^{10}m$  at  $t = 0s$  (— / ■),  $2.5s$  (— / □),  $5.0s$  (— / ▨),  $7.5s$  (— / ▤),  $10.0s$  (— / ▨) to the dry bed forced solution problem, where  $a_0 = 0m$ .

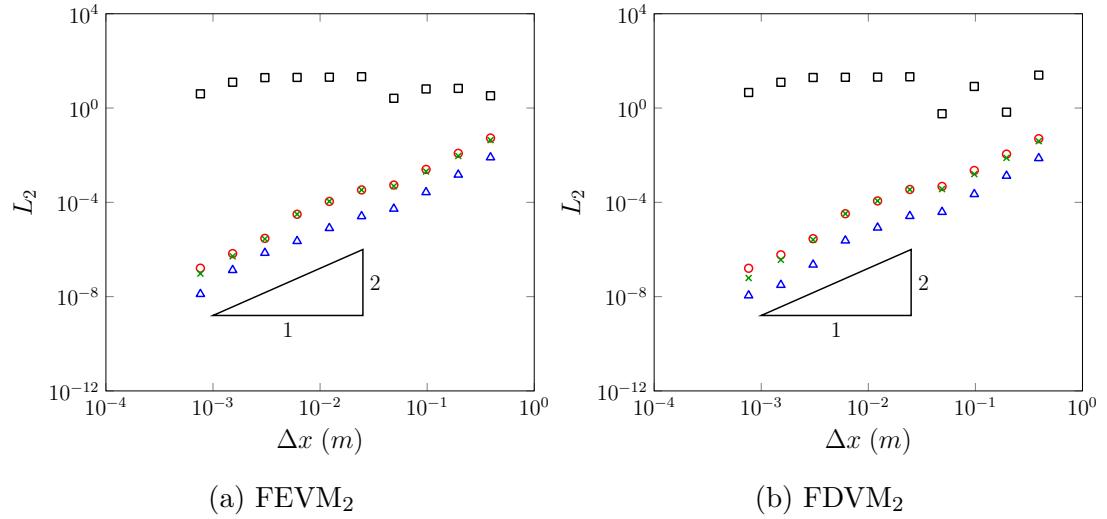


Figure 5.13: Convergence as measured by the  $L_2$  norm against  $\Delta x$  for  $h$  ( $\Delta$ ),  $u$  ( $\square$ ),  $uh$  ( $\times$ ) and  $G$  ( $\circ$ ) for the dry bed forced solution problem for FEVM<sub>2</sub> and FDVM<sub>2</sub> at  $t = 10s$ .

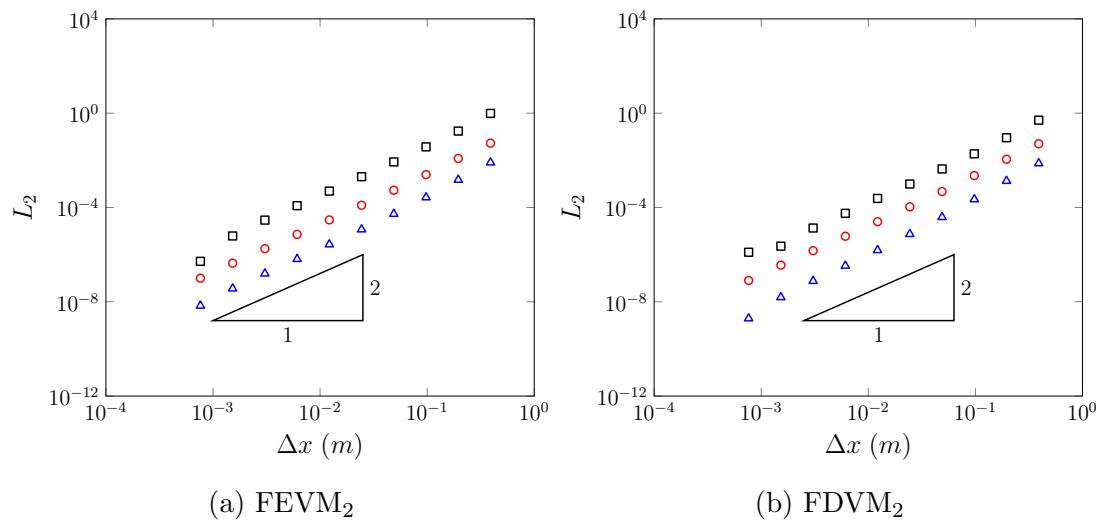


Figure 5.14: Convergence for regions where  $h > 10^{-3}m$  as measured by the  $L_2$  norm against  $\Delta x$  for  $h$  ( $\Delta$ ),  $u$  ( $\square$ ) and  $G$  ( $\circ$ ) for the dry bed forced solution problem for FEVM<sub>2</sub> and FDVM<sub>2</sub> at  $t = 10s$ .



# Chapter 6

## Experimental Validation

The second-order Finite Difference Volume Method (FDVM) denoted FDVM<sub>2</sub> and the second-order Finite Element Volume Method (FEVM) termed FEVM<sub>2</sub> are experimentally validated by comparing their numerical solutions to experimental data. A description of FEVM<sub>2</sub> was provided in Chapter 3 while FDVM<sub>2</sub> was described by Zoppou et al. [14]. Note that the dry-bed handling technique outlined in Chapter 3 for FEVM<sub>2</sub> was also applied to FDVM<sub>2</sub>.

The chosen experiments allow the methods capability to model a variety of physical situations to be tested. These situations include the presence of steep gradients in the flow, the interaction of strong dispersive waves with varying bathymetry, shoaling and wave-breaking and finally the wetting and drying of a sloping beach. Thus, the ability of these methods to robustly reproduce all the experimental results well strongly demonstrates their capability to model a variety of physical situations.

### 6.1 Evolution of a Negative Rectangular Wave

A series of experiments studying the evolution of a negative rectangular wave in the free-surface were conducted by Hammack and Segur [15]. These experiments were performed in a wave tank that was 0.394m wide, 31.6m long and 0.61m high. The rectangular negative waves were generated using a piston 0.61m long with its left edge against the wave tank wall. The 0.1m deep water is initially stationary with a horizontal free surface and the piston in the up position. The experiment begins when the piston suddenly moves down. This creates a sudden negative rectangular wave on the water surface generating a dispersive wave train that is recorded at Wave Gauges (WG) located 0m, 5m, 10m, 15m and 20m away

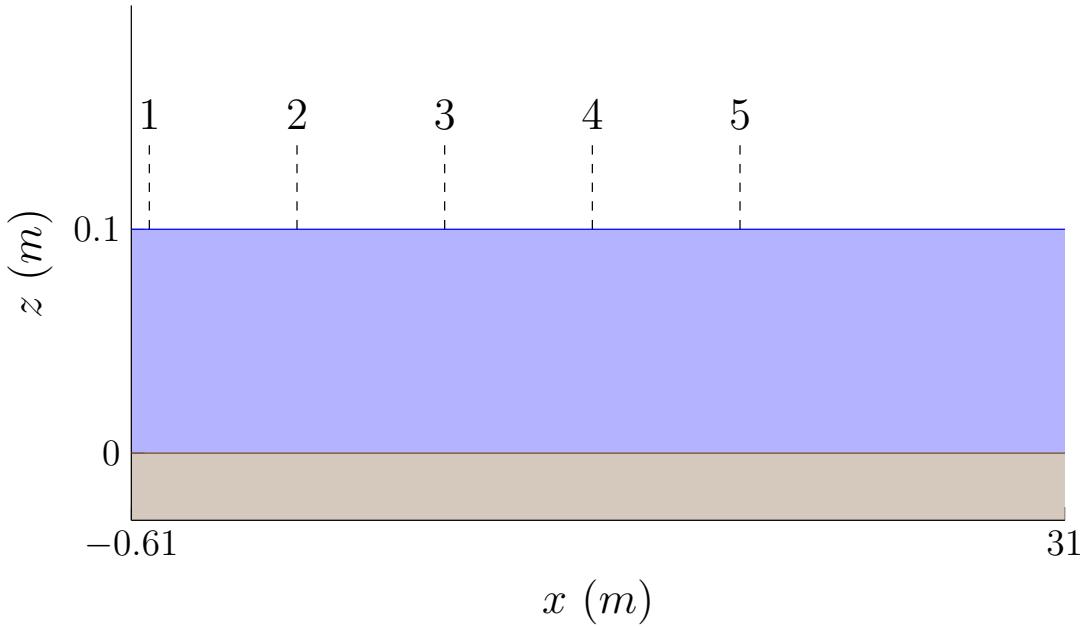


Figure 6.1: Diagram showing a longitudinal section of the wave tank for the negative rectangular wave experiments with the water (blue), the bed (brown) and the WG locations marked.

from the right edge of the piston. A diagram of the longitudinal section of the wave tank with the WG 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. These experiments are affected by bed friction, viscosity and the inability of the piston and water to move vertically instantaneously and slip free. The Serre equations do not contain viscosity and bed friction was neglected in this thesis. Additionally, discontinuous initial conditions are used to model the negative rectangular wave. Therefore, we expect that the numerical solutions of the Serre equations may produce many more oscillations in the dispersive wave train than are observed experimentally [17].

Hammack and Segur [15] report the results for two different initial negative wave amplitudes  $0.01m$  and  $0.03m$ , resulting in the non-linearity parameters  $\epsilon = 0.1$  and  $\epsilon = 0.3$  respectively. Since these non-linearity parameters are relatively small there was no breaking of waves throughout the experiment.

This experiment was modelled numerically using the reflected problem, with the left wall of the wave tank as the axis of symmetry. In the numerical experiments the domain is  $[-60m, 60m]$  and the experiment is run for  $50s$  with

$g = 9.81 \text{ m/s}^2$ . For the spatial resolution we set  $\Delta x = 0.01 \text{ m}$  while the CFL condition (3.23) was satisfied using  $\Delta t = 0.5\Delta x/\sqrt{g0.1}$ . The limiting parameter  $\theta = 1.2$  was used in the reconstruction (3.2) in FEVM<sub>2</sub> and FDVM<sub>2</sub>.

The numerical results of FDVM<sub>2</sub> for these experiments produced by me have been published [14]. They have been extended here with the inclusion of the conservation error of  $h$ ,  $G$  and  $uh$  in the numerical simulation of both experiments.

### 6.1.1 Results for the 0.01m Negative Rectangular Wave

Plots comparing the numerical and experimental WG data for the 0.01m negative rectangular wave are displayed in Figures 6.2 and 6.3 for FEVM<sub>2</sub> and FDVM<sub>2</sub> respectively. We present this data using the same dimensionless scales as reported in the original paper [15]. Tables 6.1 and 6.2 which display the completely numerically calculated error in conservation  $C^*$  are also provided.

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

The numerical solutions produce larger and consequently faster waves and generate oscillations in the dispersive wave train which are not present in the experimental data of WG1. Moreover, as expected the methods produce many more oscillations than were observed experimentally. These discrepancies can be attributed to the lack of viscosity and the omission of bed friction for the Serre equations in this thesis (2.6). Furthermore, it is highly likely the experiment produced some smooth approximation to a discontinuous jump in the water depth with the down-stroke of the piston. Such a smoothing of the initial conditions will significantly attenuate the high frequency waves in the generated dispersive wave train [17]. Given these differences the numerical methods do a very good job of replicating the experimental behaviour.

These numerical solutions compare well to those of Zoppou et al. [14] for FDVM<sub>2</sub> with  $\Delta x = 0.005 \text{ m}$ ,  $\Delta t = 0.2\Delta x/\sqrt{g0.1} \text{ s}$  and  $\theta = 1$ . Where a higher resolution and a more diffusive value of the limiting parameter were used. These differences had little impact on the WG results. However, the more diffusive value of the limiting parameter smoothed the initial discontinuity which negatively impacted the conservation of  $\mathcal{H}$  reported by Zoppou et al. [14].

Both FEVM<sub>2</sub> and FDVM<sub>2</sub> have produced indistinguishable results at this scale and have demonstrated very good conservation of all the quantities see, Tables 6.1 and 6.2. Given the extensive examination of several numerical schemes for

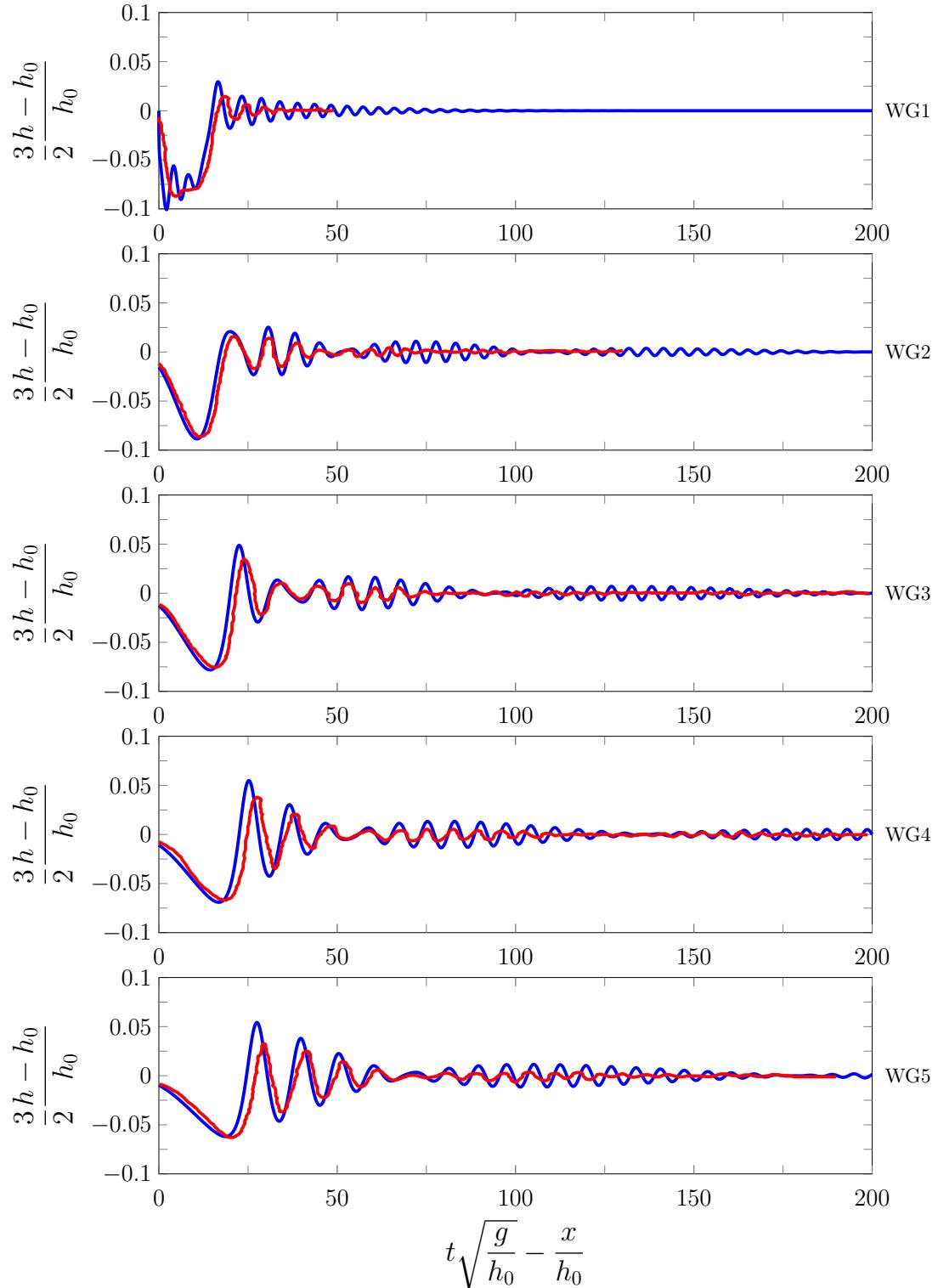


Figure 6.2: Time series of the experimental WG data (—) and numerical results (—) of FEVM<sub>2</sub> for the 0.01m negative rectangular wave.

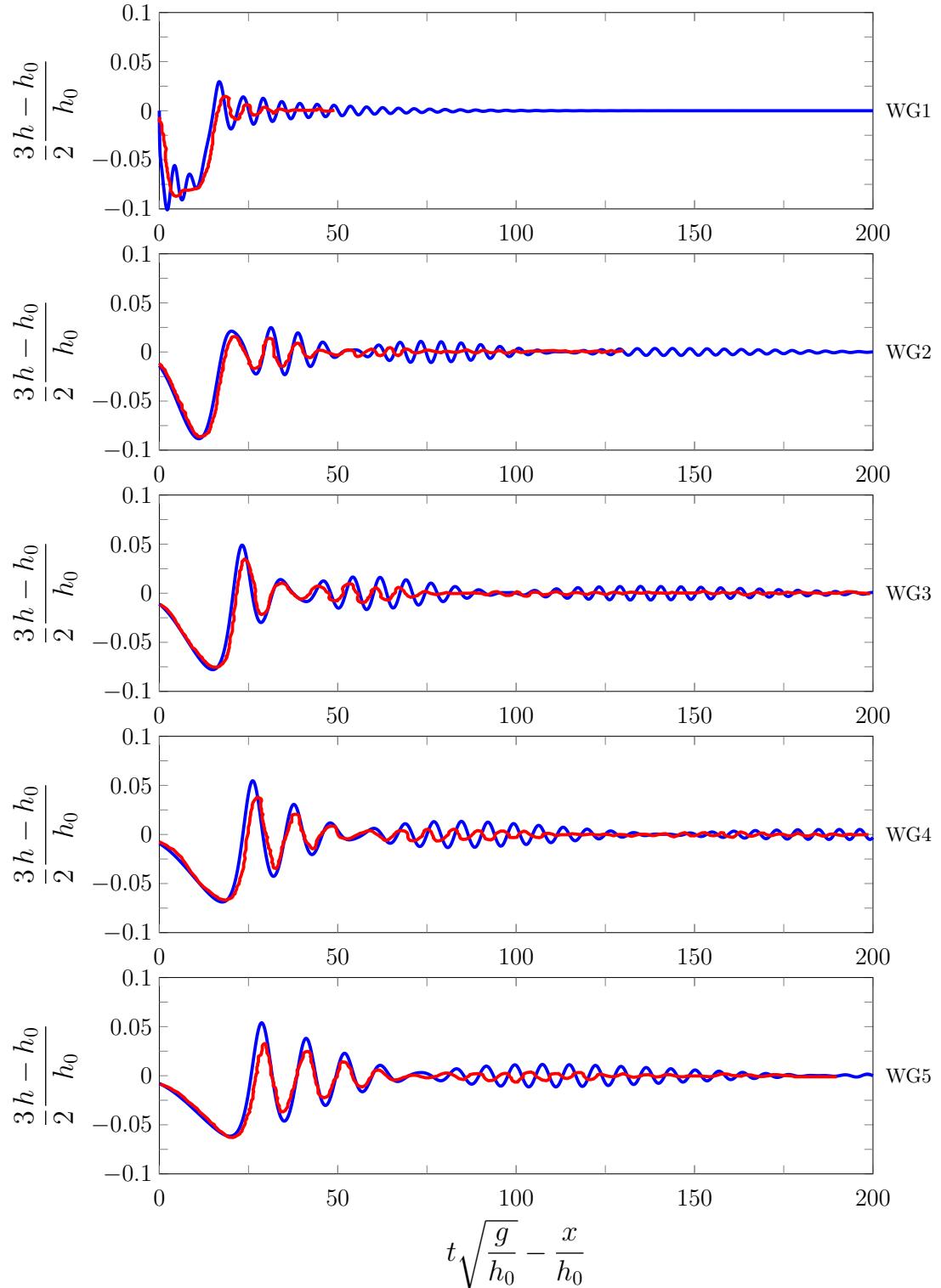


Figure 6.3: Time series of the experimental WG data (—) and numerical results (—) of FDVM<sub>2</sub> for the 0.01m negative rectangular wave.

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

Table 6.1: Initial and final total amounts and the conservation error for all conserved quantities for the numerical solution of FEVM<sub>2</sub> for the 0.01m negative rectangular wave.

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

Table 6.2: Initial and final total amounts and the conservation error for all conserved quantities for the numerical solution of FDVM<sub>2</sub> for the 0.01m negative rectangular wave.

steep gradient problems [17], this indicates that these solutions are indicative of true solutions of the Serre equations. However, these results do not demonstrate the superiority of one of these methods over the other.

### 6.1.2 Results for the 0.03m Negative Rectangular Wave

The WG data for the numerical and experimental results for the evolution of the 0.03m negative rectangular wave are displayed in Figures 6.4 and 6.5 for FEVM<sub>2</sub> and FDVM<sub>2</sub> respectively. These results are reported using the same dimensionless scales as in the original paper [15]. The completely numerically calculated conservation error  $C^*$  of all the conserved quantities are given in Tables 6.3 and 6.4 for FEVM<sub>2</sub> and FDVM<sub>2</sub> respectively.

Both methods again reproduce the overall behaviour of this experiment very

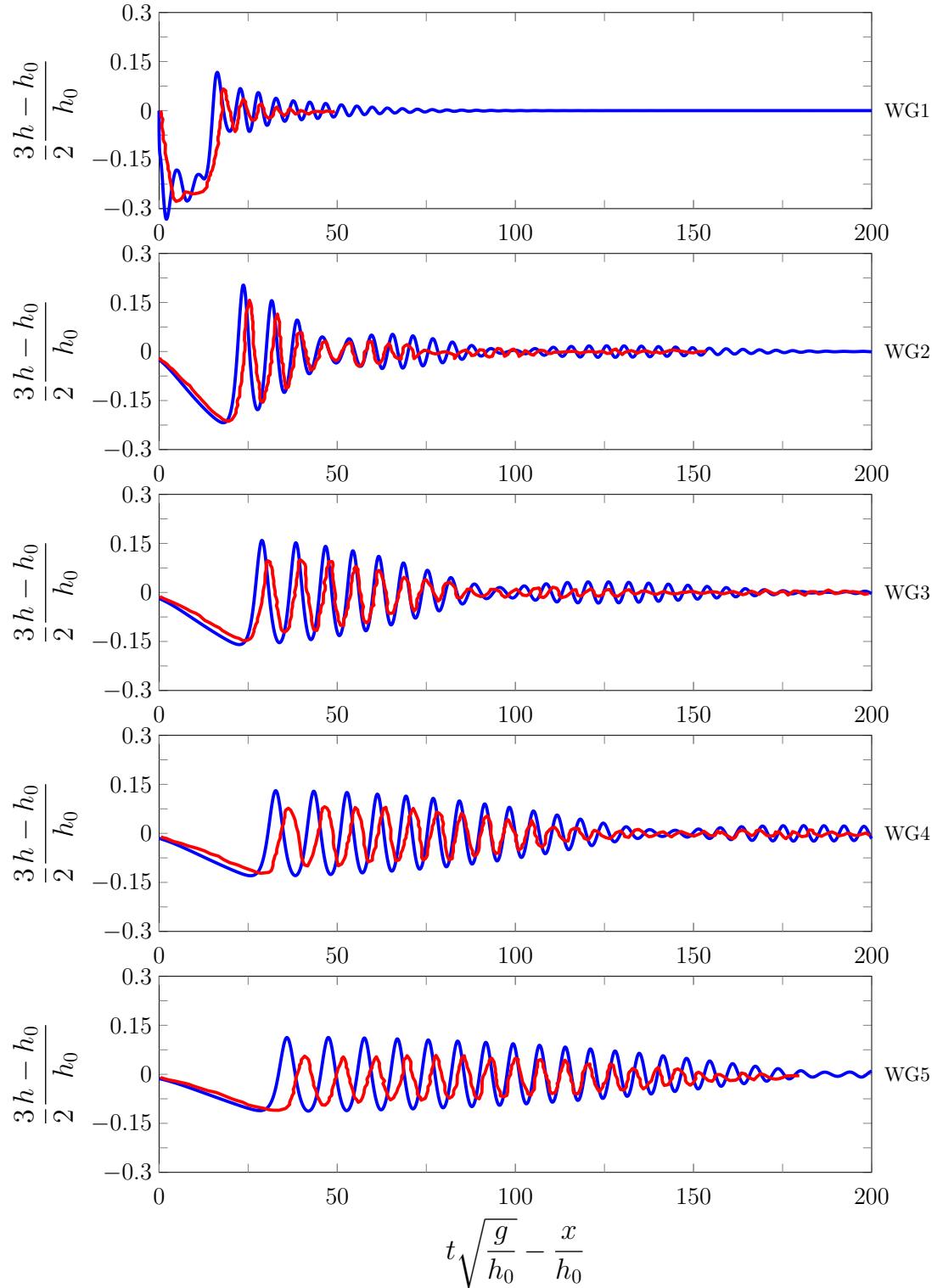


Figure 6.4: Time series of the experimental WG data (—) and numerical results (—) of FEVM<sub>2</sub> for the 0.03m negative rectangular wave.

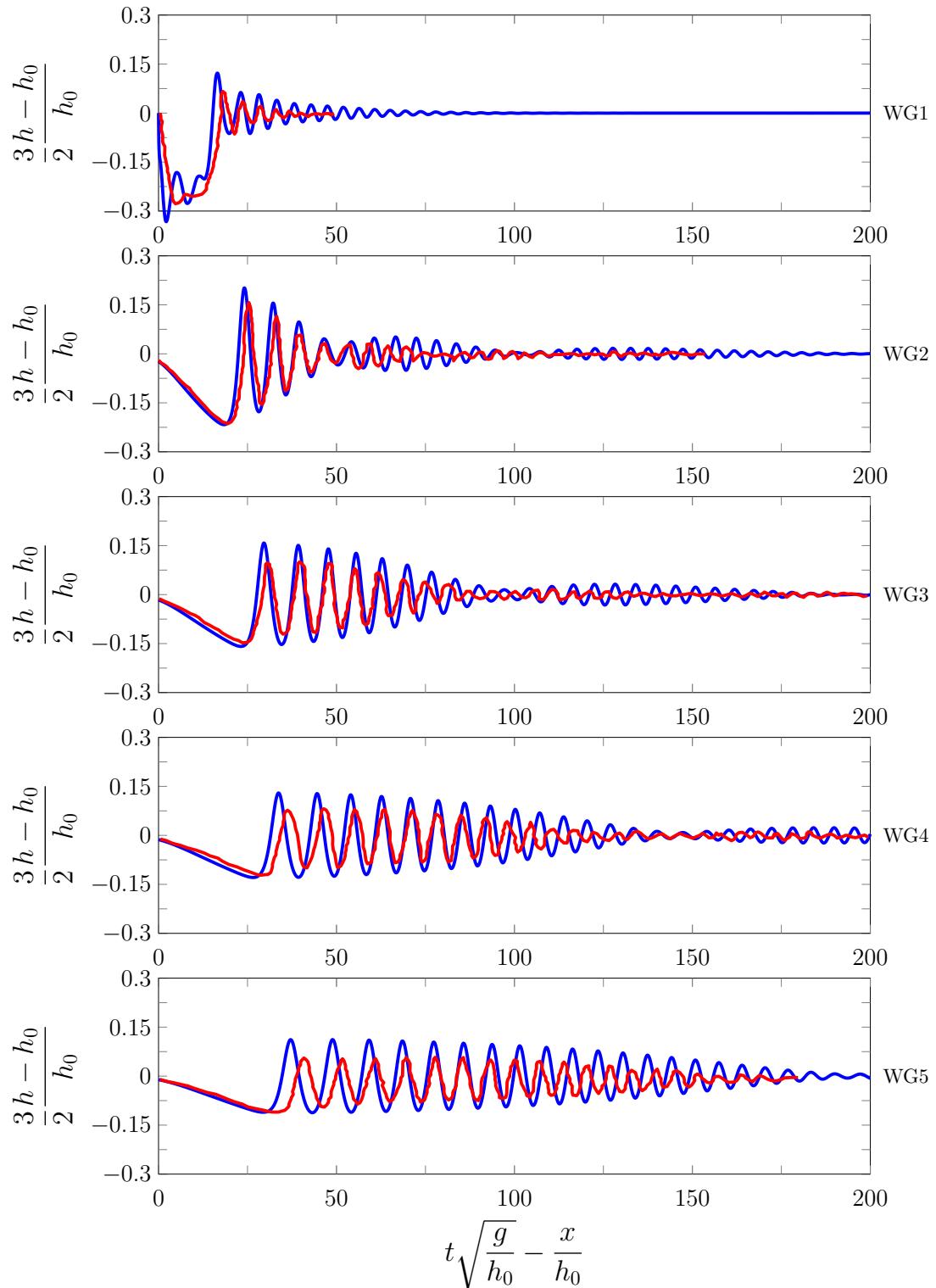


Figure 6.5: Time series of the experimental WG data (—) and numerical results (—) of FDVM<sub>2</sub> for the 0.03m negative rectangular wave.

well. Because the rectangular wave is deeper, this experiment provides a more rigorous test for the numerical methods. However, increasing the depth also strengthens the causes of the discrepancy between the experimental results and the numerical solutions of the Serre equations. Leading to the observed differences for the amplitude and speed of the generated waves.

Since the negative rectangular wave is larger than in the previous example, the numerical methods have a larger error in conservation for all the quantities as compared to the  $0.01m$  negative rectangular wave. The exception is the mass which is conserved exactly. For  $G$  and momentum these errors are around machine precision 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.

These results compare well with those by Zoppou et al. [14] for FDVM<sub>2</sub> with  $\Delta x = 0.005m$ ,  $\Delta t = 0.2\Delta t/\sqrt{g0.1}s$  and  $\theta = 1$ . The WG data of those numerical solutions and the solutions displayed here are indistinguishable. The error in conservation in  $\mathcal{H}$  is very similar as well, although in the numerical solutions displayed here the numerical grid is coarser. This is because the limiting parameter  $\theta = 1$  diffuses the discontinuous jump, leading to poorer conservation of  $\mathcal{H}$  [17] than expected given the higher resolution of the grid.

Conserving  $\mathcal{H}$  is difficult for numerical methods solving steep gradient problems [17]. More so for this simulation where two steep gradients interact with one another over short time spans. Increasing the resolution of the numerical method improves the conservation of  $\mathcal{H}$  as demonstrated for the analytic solutions. However, the results for the WG data for higher resolution solutions will be indistinguishable from the results presented here and so the provided solutions are sufficient.

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

## 6.2 Periodic Waves Over A Submerged Bar

Beji and Battjes conducted a series of experiments investigating the effect of submerged bars on the propagation of periodic waves [16, 58]. The behaviour of these experiments were mainly driven by the dispersion properties of the waves

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

Table 6.3: Initial and final total amounts and the conservation error for all conserved quantities for the numerical solution of FEVM<sub>2</sub> for the 0.03m negative rectangular wave.

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

Table 6.4: Initial and final total amounts and the conservation error for all conserved quantities for the numerical solution of FDVM<sub>2</sub> for the 0.03m negative rectangular wave.

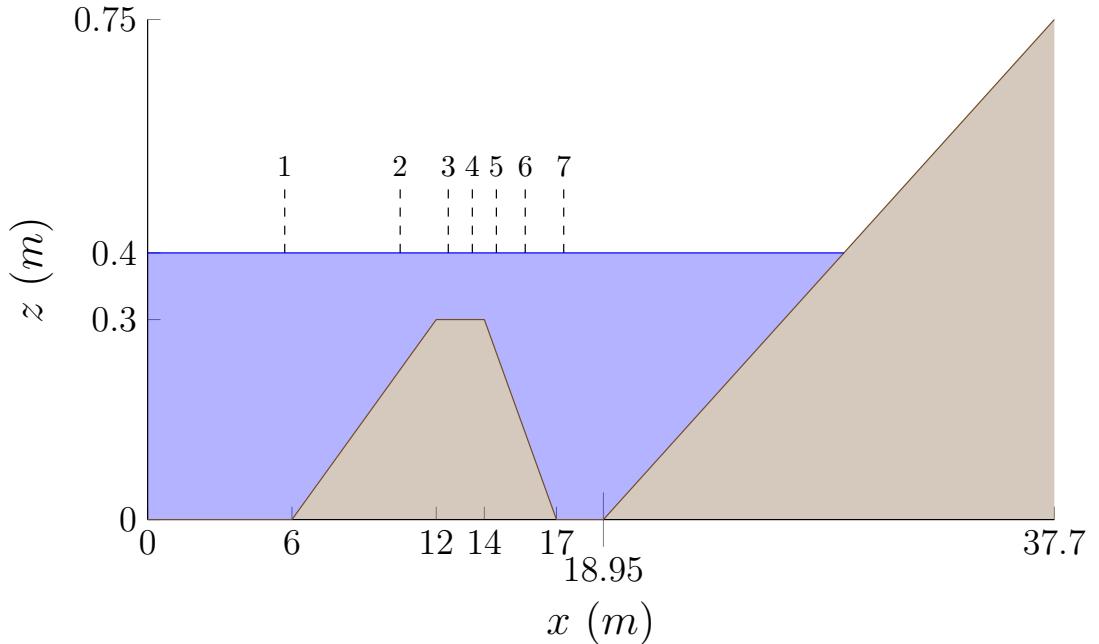


Figure 6.6: Diagram showing a longitudinal section of the wave tank for the periodic waves over a submerged bar experiments with the water (blue), the bed (brown) and the WG locations marked.

and their interaction with variations in bathymetry. Therefore, these experiments serve as a benchmark for the ability of the numerical schemes to accurately model the interaction of variable bathymetry and dispersive waves. For our purposes we will focus on the monochromatic wave experiments of Beji and Battjes [16].

The experiments of Beji and Battjes [16] 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 WG at the following locations; 5.7m, 10.5m, 12.5m, 13.5m, 14.5m, 15.7m and 17.3m. Waves are generated from a piston-type wave maker located at 0m and travel on the initially still water 0.4m deep to the right, over the submerged trapezoidal bar and are absorbed by a sloping beach.

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

For the numerical solutions the spatial domain was  $[5.7m, 150m]$  with  $\Delta x = 0.1/2^4m \approx 0.0063m$  and  $\Delta t = Sp/2^5s \approx 0.0012s$  where  $Sp = 0.039s$  is the

experimental sampling period. These  $\Delta x$  and  $\Delta t$  values satisfy the CFL condition, (3.23). In our numerical experiments only the submerged trapezoidal bar is present, and the sloping beach is replaced with a very long horizontal bed that ensures that we do not observe any effects from the Dirichlet boundary conditions at the downstream boundary.

To simulate the incoming waves at the upstream boundary we used WG1 as our left boundary condition together with linear extrapolation to calculate the other required  $h$  values in the left ghost cell. The velocity boundary conditions were calculated from the height values in the same way as Beji and Battjes [16] using

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

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

The results of FDVM<sub>2</sub> with the same resolutions and limiting parameter published by Zoppou et al. [14] were obtained by me. They have been reproduced here and serve as a comparison for the results of FEVM<sub>2</sub>.

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

### 6.2.1 Low Frequency Results

A comparison of the wave heights  $\eta$  of the experimental and numerical results are located in Figures 6.7 and 6.8 for FEVM<sub>2</sub> and Figures 6.9 and 6.10 for FDVM<sub>2</sub>. These numerical schemes produce indistinguishable results for all WG and so this benchmark does not help us discriminate between these two methods.

These results demonstrate the ability of these numerical methods to reproduce the experimental results, particularly for WG1 to WG5 where the agreement between experimental and numerical results is best. Results at these gauges validate the numerical schemes for simulating shoaling of dispersive waves as these WG are all located on the windward side of the submerged bar where shoaling occurs in the experiment.

The numerical results for WG6 and WG7 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 actual water waves [16, 59].

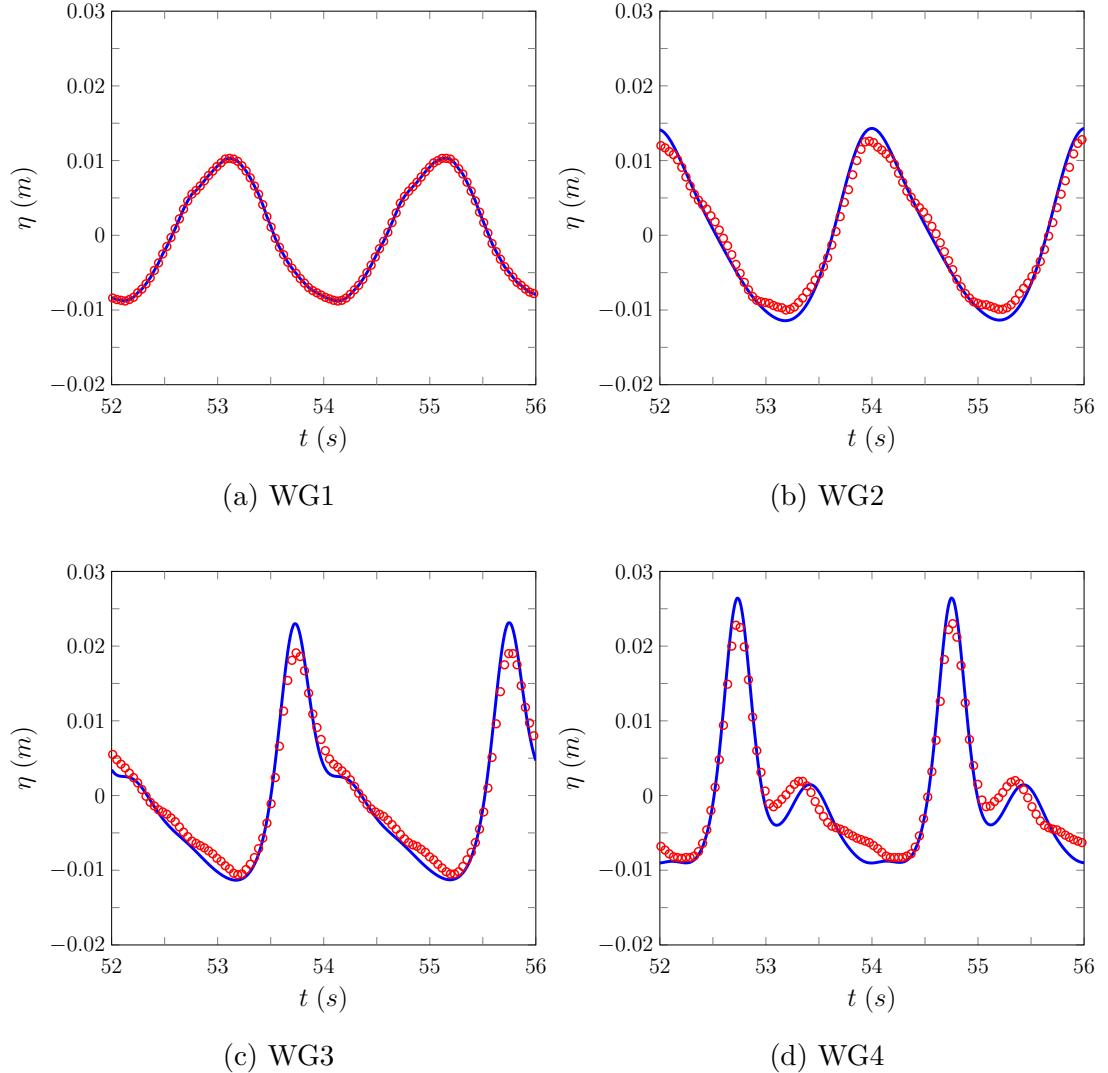


Figure 6.7: Time series of the wave heights  $\eta$  of the numerical results of FEVM<sub>2</sub> (—) and the experimental results (○) for WG1 - WG4 for the low frequency experiment.

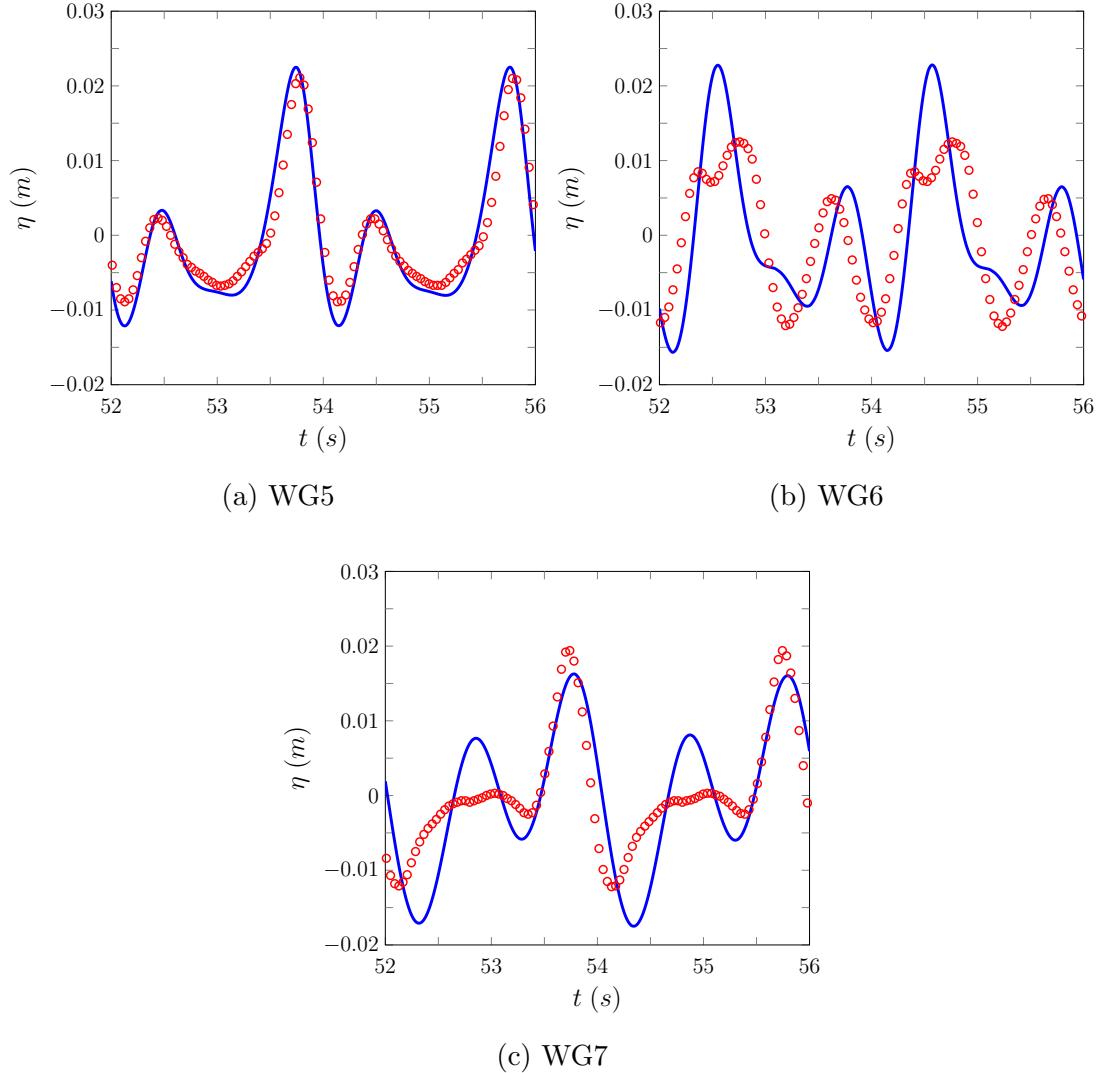


Figure 6.8: Time series of the wave heights  $\eta$  of the numerical results of FEVM<sub>2</sub> (—) and the experimental results (○) for WG5 - WG7 for the low frequency experiment.

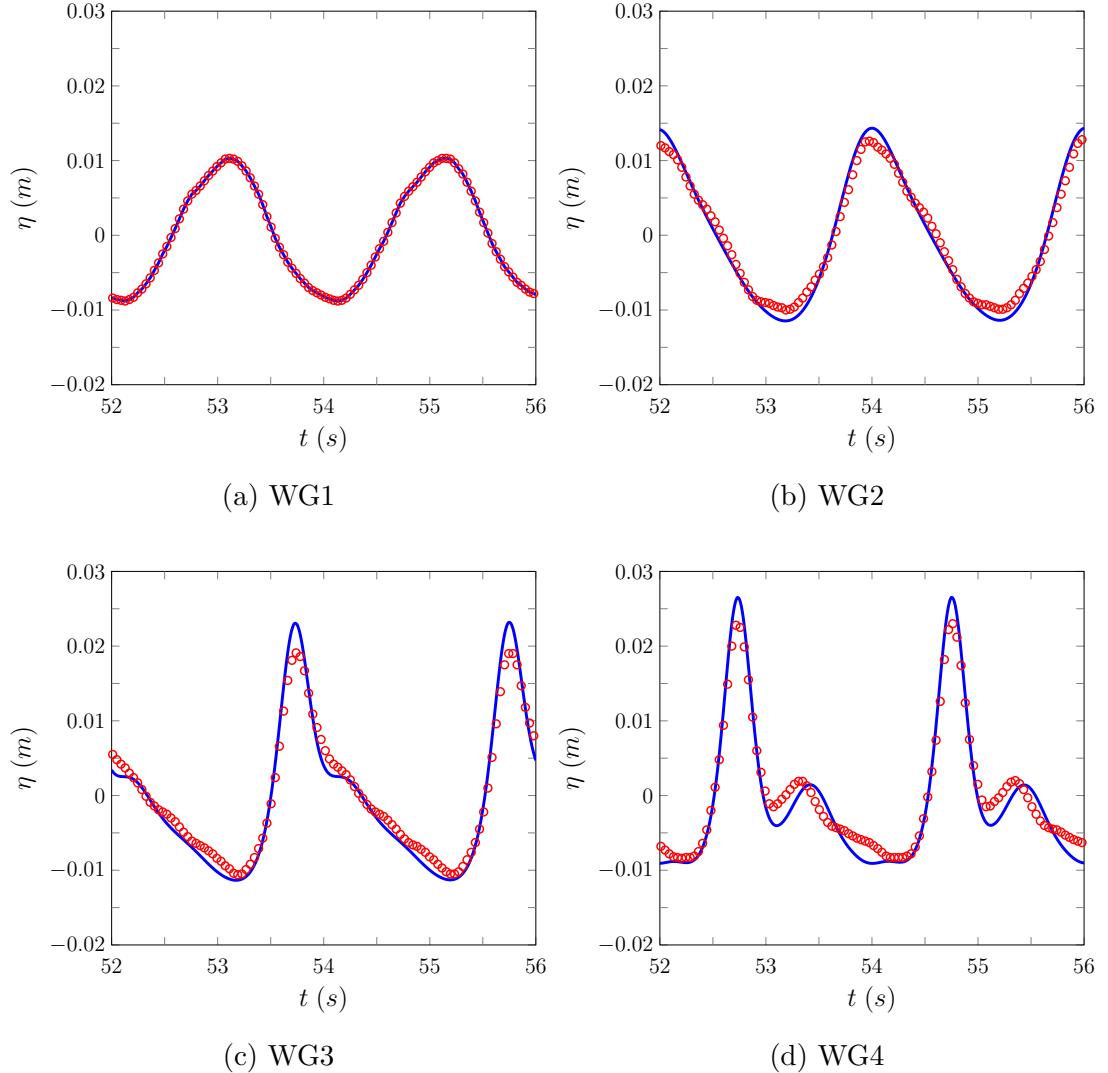


Figure 6.9: Time series of the wave heights  $\eta$  of the numerical results of FDVM<sub>2</sub> (—) and the experimental results (○) for WG1 - WG4 for the low frequency experiment.

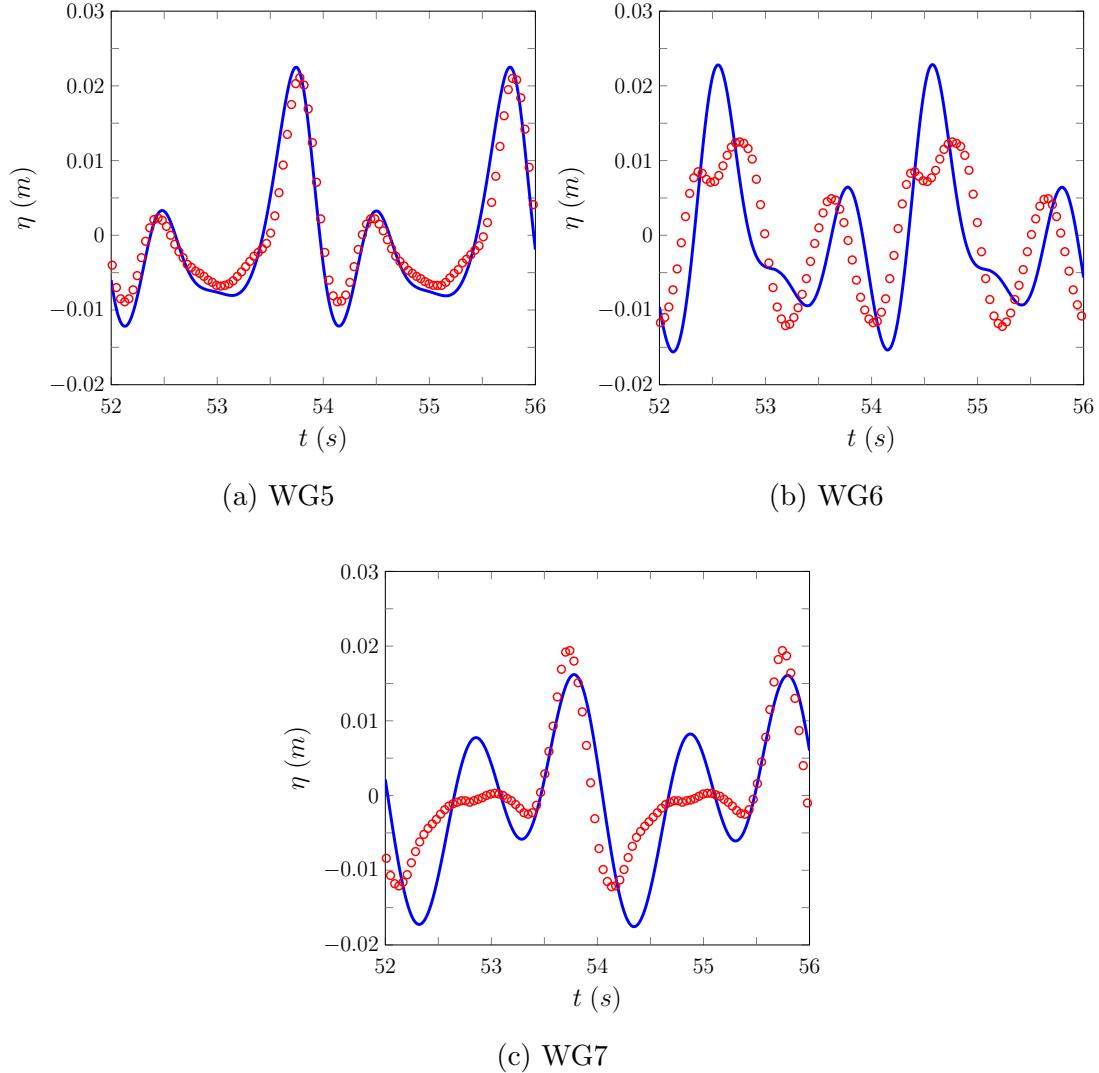


Figure 6.10: Time series of the wave heights  $\eta$  of the numerical results of FDVM<sub>2</sub> (—) and the experimental results (○) for WG5 - WG7 for the low frequency experiment.

The dispersion terms in the Serre equations are vital to recreating the experimental results for WG2 to WG5, as non-dispersive equations such as the SWWE are not capable of accurately simulating this experiment [60].

### 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<sub>2</sub>. While the results for FDVM<sub>2</sub> are given in Figures 6.13 and 6.14. As for the low frequency experiment FEVM<sub>2</sub> and FDVM<sub>2</sub> produce indistinguishable results for all WG 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 WG1 to WG4 but perform poorly for the leeward side of the slope for WG5 to WG7. With the high frequency experiment we see the divergence between the numerical and experimental results earlier than the low frequency experiment, so that now WG5 which is on the leeward side exhibits a significant difference between the numerical and experimental results. As in the low frequency example this is caused by the difference in the dispersion relations of the Serre equations and the linear theory for water waves [16, 59]. Because the difference between the dispersion relation of the Serre equations and water waves is largest for higher frequency waves [23] the earlier divergence between experimental and numerical results is not surprising.

These numerical results for the FDVM<sub>2</sub> and FEVM<sub>2</sub> agree well with other numerical results for weakly dispersive equations for the simulation of periodic waves over a submerged bar in the literature [16, 27, 59, 61]. 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 [16].

## 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 [62]. These experiments were performed in a wave tank 3.66m wide, 83.7m long and 4.57m high with a removable bed that allowed for the wide range of experiments reported by Roeber [62]. We have computationally modelled the experiment with the bathymetry displayed in Figure 6.15, where a solitary wave is generated from the wave maker at 0m and is recorded at the

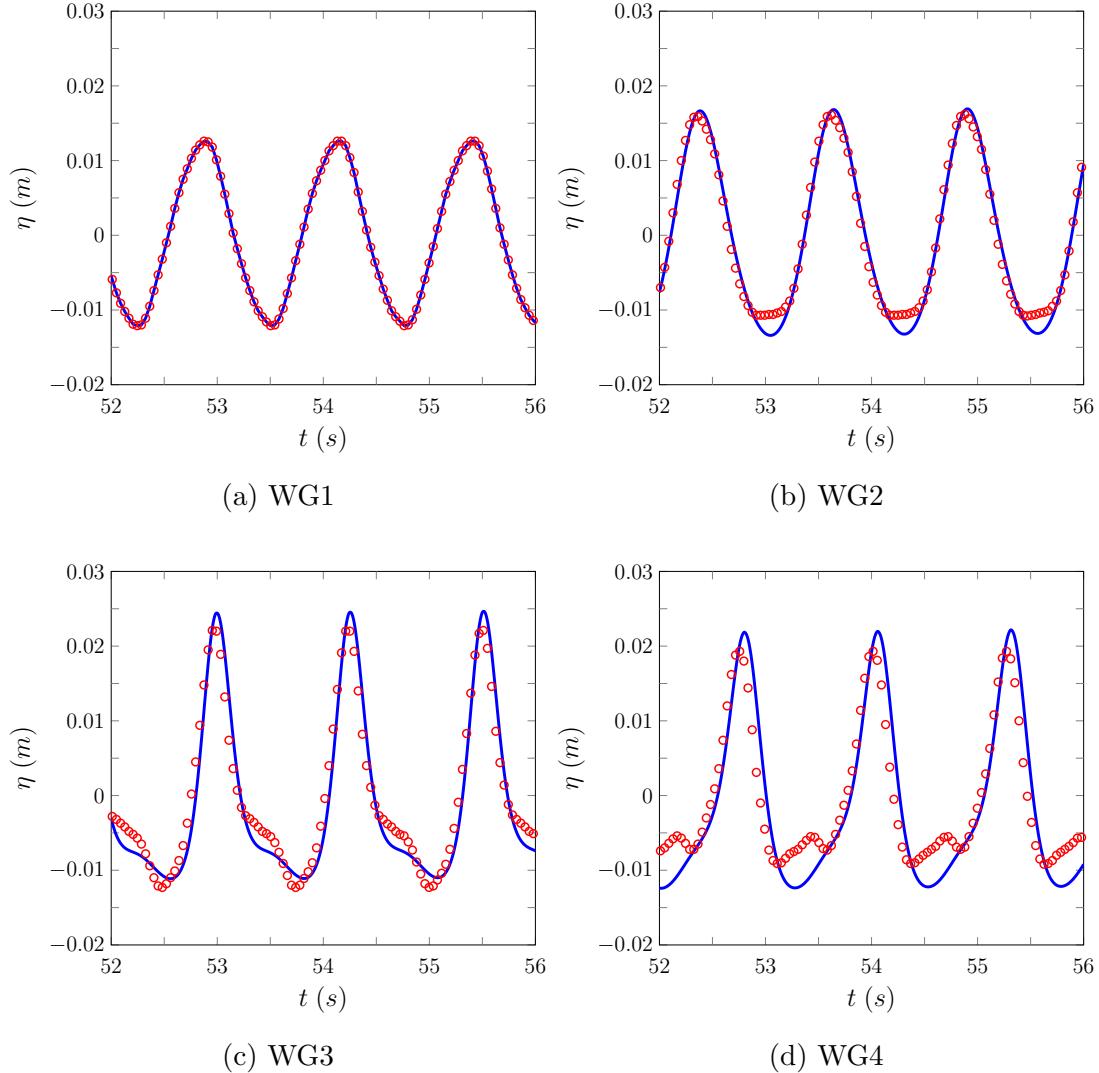


Figure 6.11: Time series of the wave heights  $\eta$  of the numerical results of FEVM<sub>2</sub> (—) and the experimental results (○) for WG1 - WG4 for the high frequency experiment.

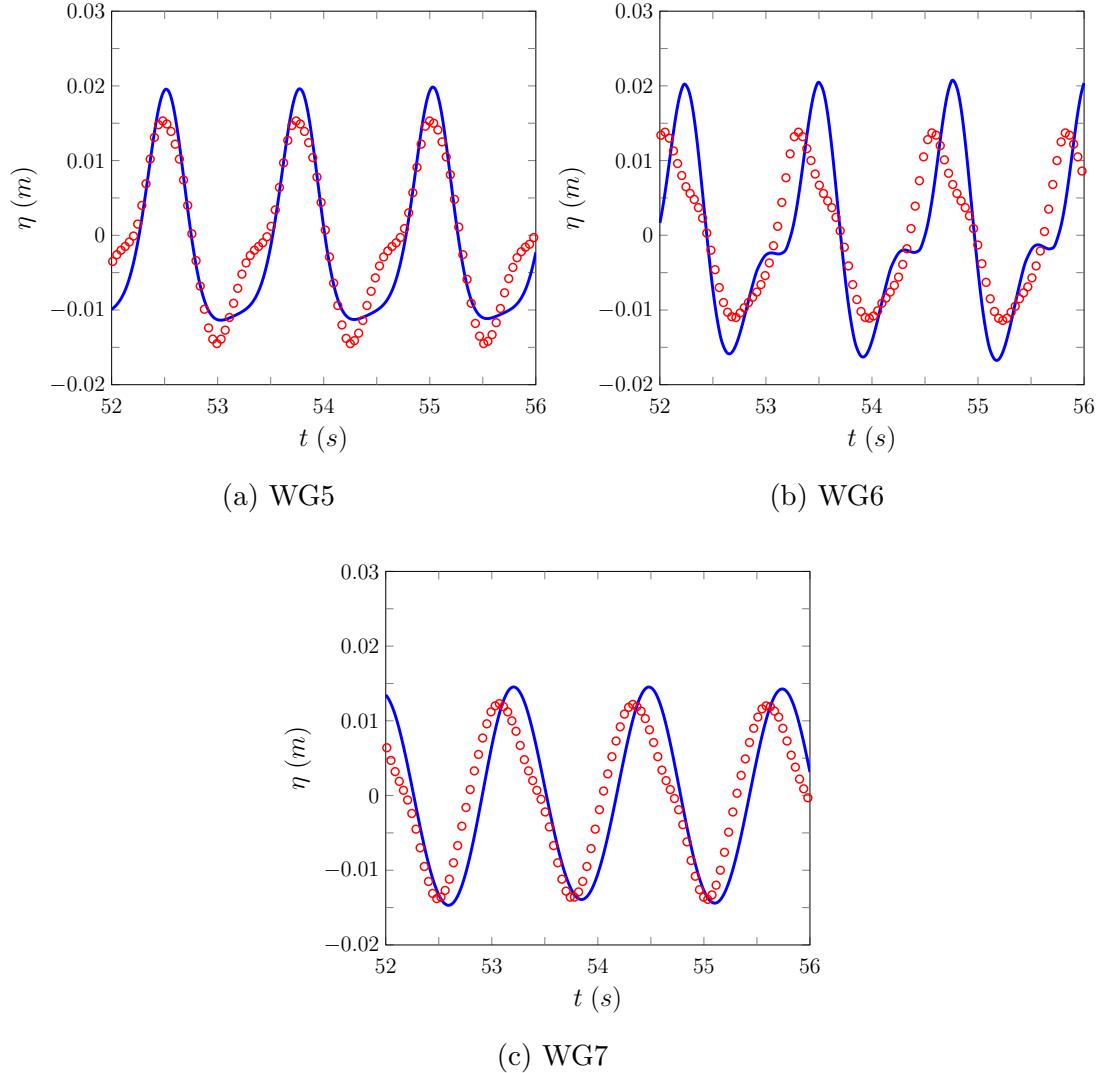


Figure 6.12: Time series of the wave heights  $\eta$  of the numerical results of FEVM<sub>2</sub> (—) and the experimental results (○) for WG5 - WG7 for the high frequency experiment.

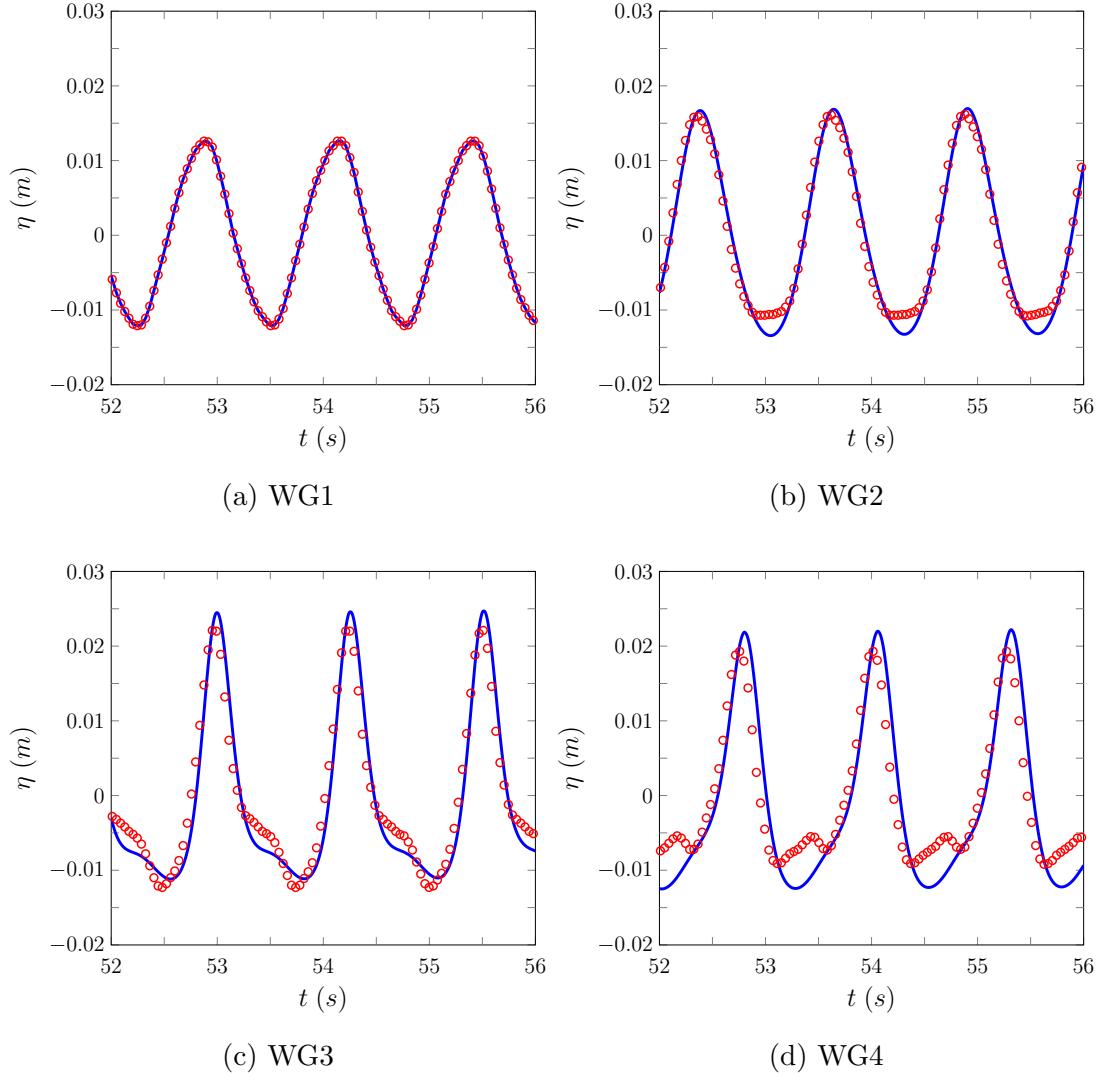


Figure 6.13: Time series of the wave heights  $\eta$  of the numerical results of FDVM<sub>2</sub> (—) and the experimental results (○) for WG1 - WG4 for the high frequency experiment.

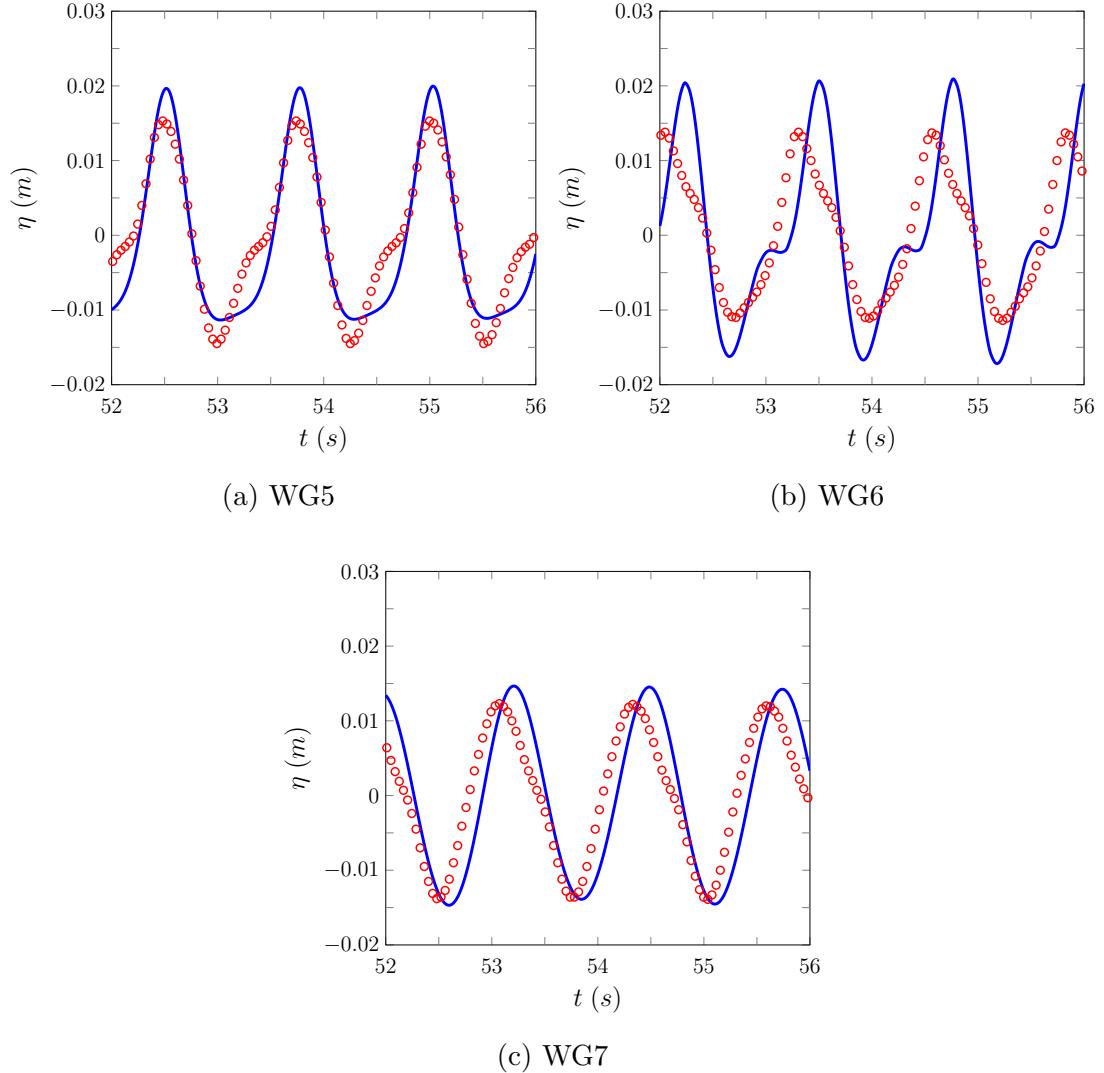


Figure 6.14: Time series of the wave heights  $\eta$  of the numerical results of FDVM<sub>2</sub> (—) and the experimental results (○) for WG5 - WG7 for the high frequency experiment.

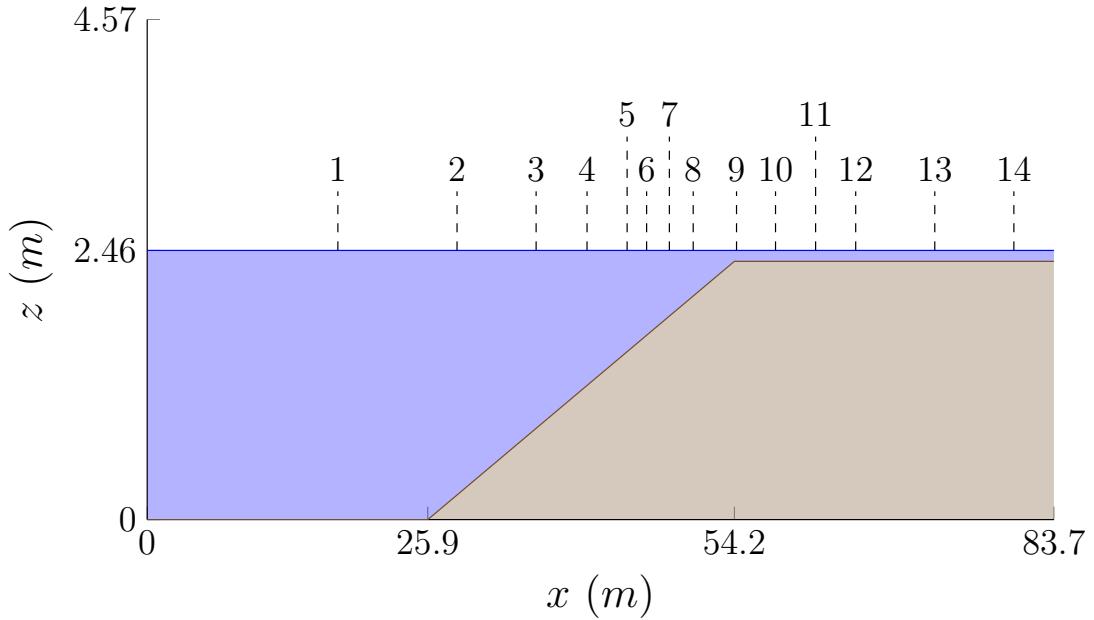


Figure 6.15: Diagram showing a longitudinal section of the wave tank for the solitary wave over a fringing reef experiment with the water (■), the bed (□) and the WG locations marked.

WG located 17.6m, 28.6m, 35.9m, 40.6m, 44.3m, 46.1m, 48.2m, 50.4m, 54.4m, 58.0m, 61.7m, 65.4m, 72.7m and 80.0m downstream of the wave maker.

This experiment investigates the behaviour of a wave with high non-linearity  $\epsilon \approx 1.23/2.46 = 0.5$  as it shoals over a linear bed into a very shallow body of water with a depth of only 0.1m. Given the high non-linearity 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  $t \approx 33s$  [62]. As with other depth averaged equations, the Serre equations are only appropriate up to wave-breaking 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 WG1 could be used as the left boundary conditions. Where the technique for the incoming boundary condition in Section 6.2 for the periodic waves over a submerged bar experiment was employed. The spatial resolution was  $\Delta x = 0.025m$  and the temporal resolution was  $\Delta t = Sp/8s = 0.0025$  where  $Sp = 0.02s$  was the sampling period of the WG. These spatial and temporal

resolutions satisfy the CFL condition (3.23) for this experiment. The right edge of the domain used Dirichlet boundary conditions, since the domain was large no effects from the downstream boundary were observed throughout the numerical simulation.

The WG results comparing the numerical and experimental data are displayed in Figures 6.16, 6.17 and 6.18 for FEVM<sub>2</sub> and 6.19 and 6.20 for FDVM<sub>2</sub>.

Both methods accurately reproduce the shoaling of the solitary wave, particularly in WG1 through WG8 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 WG and since the numerical simulation did not have reflective boundaries these waves are not replicated in their numerical solutions.

When breaking begins the numerical solutions perform much worse as expected; most notably FDVM<sub>2</sub> becomes unstable and the solution blows up. Because of this the numerical solution of FDVM<sub>2</sub> was only plotted until  $t = 34s$ . The instability is caused by the appearance of a very steep gradient with a large jump in the water depth compared to the depth of water that surrounds it as the wave breaks. The FEVM<sub>2</sub> 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 in the experiment.

Given the limitations of the underlying Serre equations the results for FEVM<sub>2</sub> are robust and accurately model the shoaling of the solitary wave. However, these results indicate the need for improvement to the handling of breaking waves to be able to more accurately model some physical situations.

## 6.4 Run-up Experiment

To study the run-up of incoming waves on linear beaches a series of experiments were conducted by Synolakis [63]. These experiments consisted of a number of run-up events for a wide array of breaking and non-breaking waves where snapshots of the entire water surface were taken at certain times. These experiments were all performed on the beach profile depicted in Figure 6.21, where all the quantities are non-dimensionalised [63]. To denote that a quantity is non-dimensionalised we use a prime; for example for a generic quantity  $q$  its non-dimensionalised version is  $q'$ . To assess the computational models we recreated

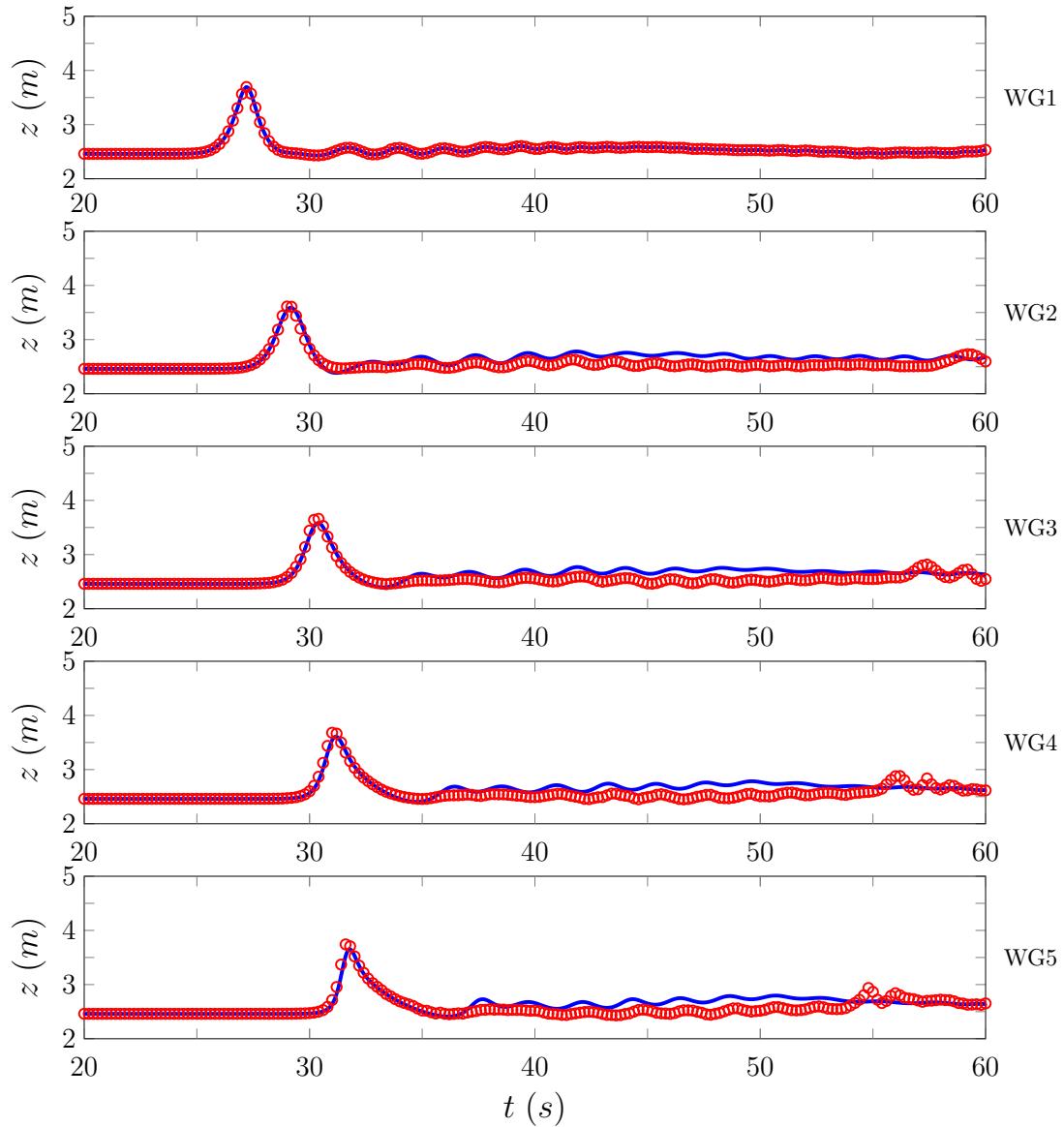


Figure 6.16: Time series of the experimental (○) and numerical (—) WG data produced by FEVM<sub>2</sub> for WG1 to WG5.

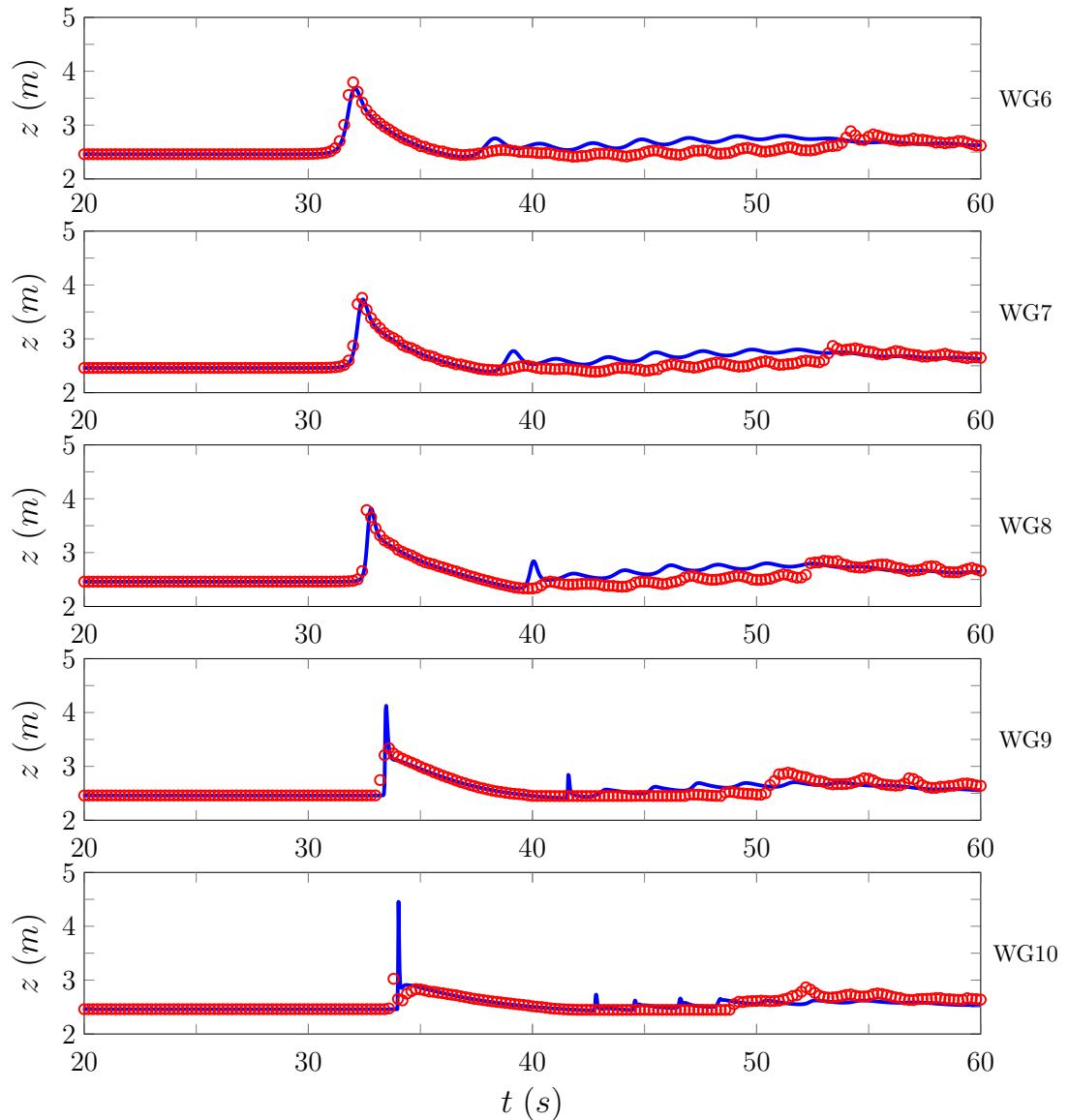


Figure 6.17: Time series of the experimental (○) and numerical (—) WG data produced by FEVM<sub>2</sub> for WG6 to WG10.

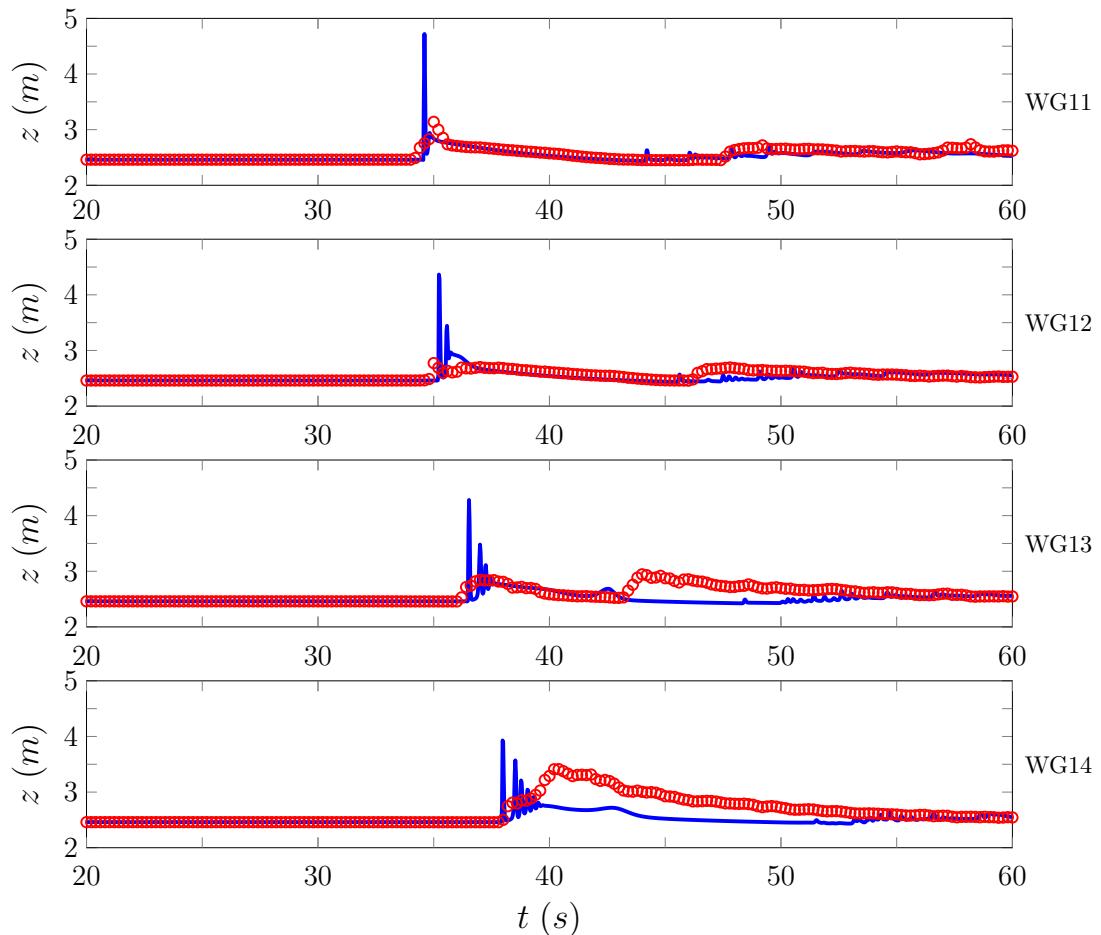


Figure 6.18: Time series of the experimental (○) and numerical (—) WG data produced by FEVM<sub>2</sub> for WG11 to WG14.

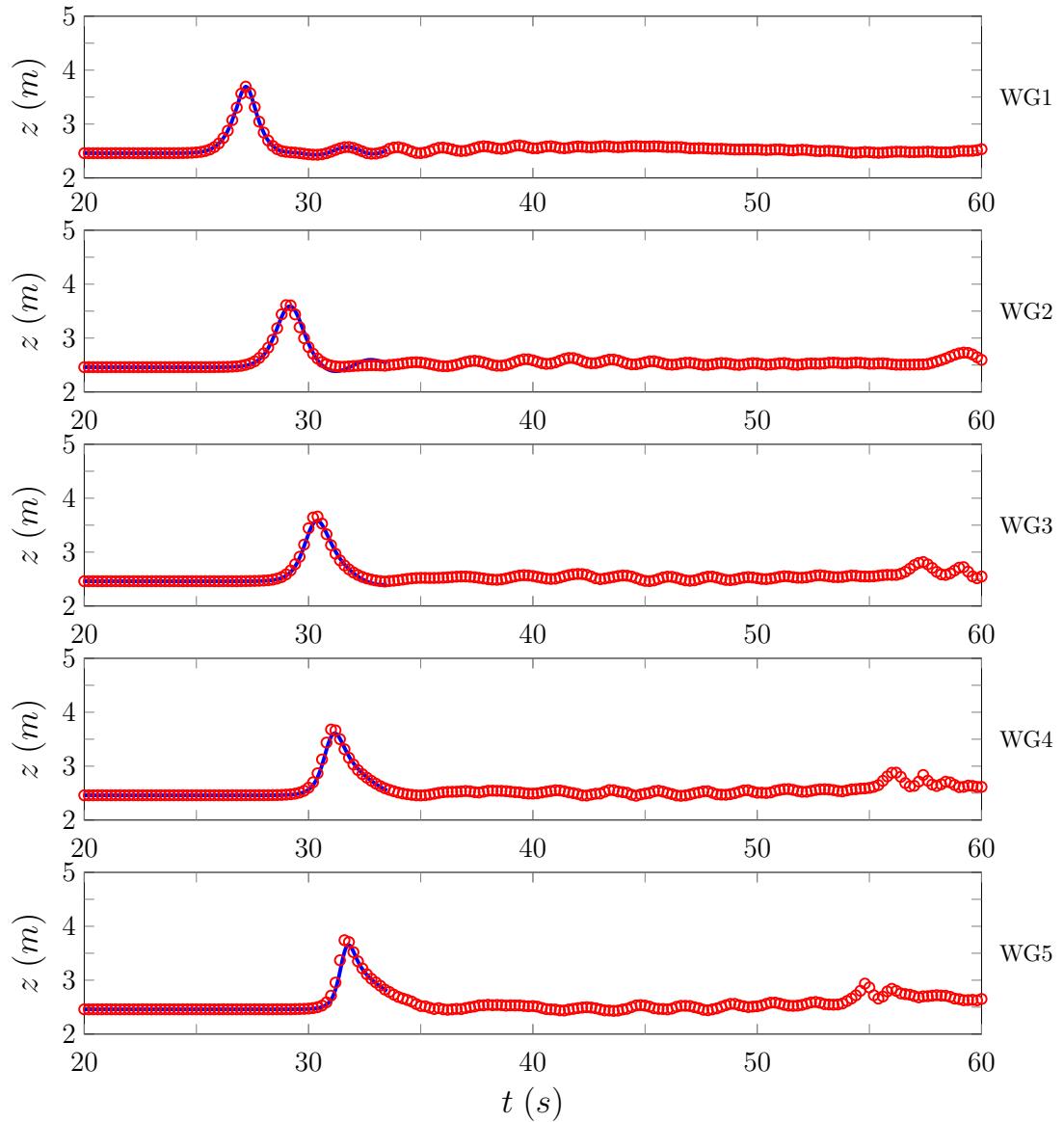


Figure 6.19: Time series of the experimental (○) and numerical (—) WG data produced by FDVM<sub>2</sub> for WG1 to WG7.

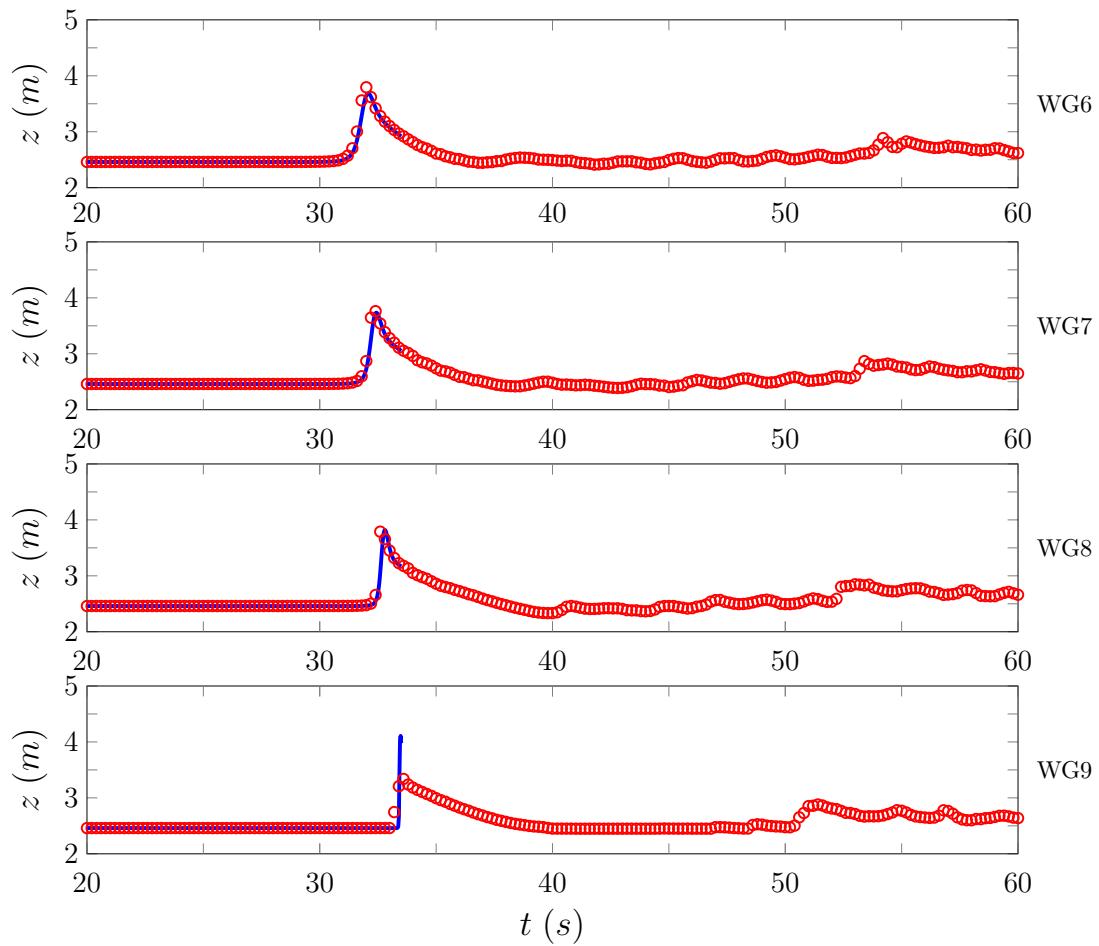


Figure 6.20: Time series of the experimental (○) and numerical (—) WG data produced by FDVM<sub>2</sub> for WG6 to WG9.

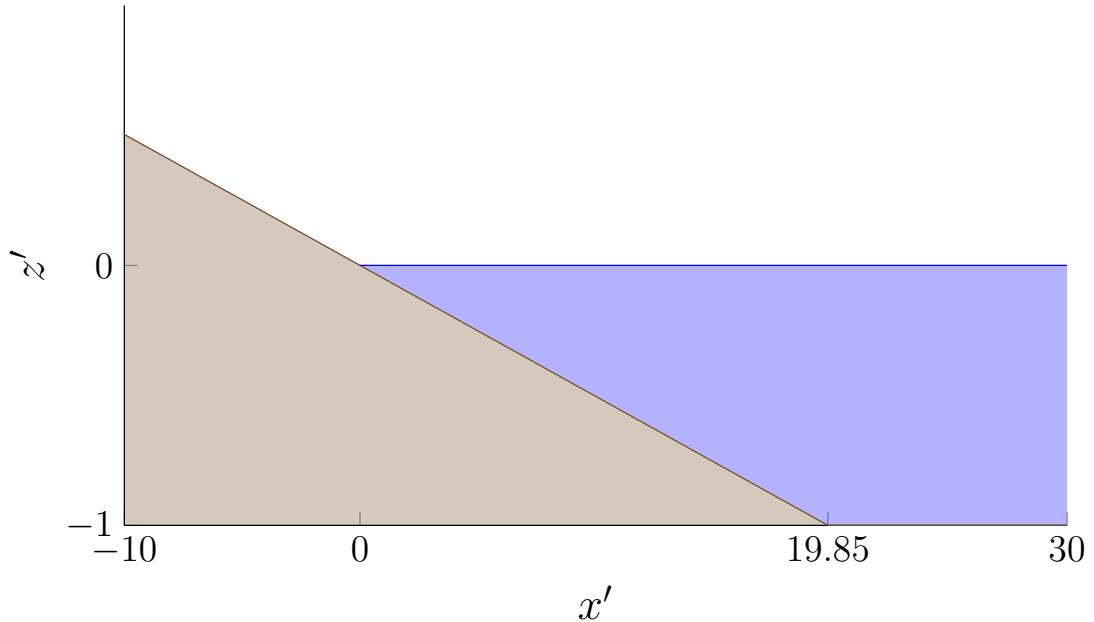


Figure 6.21: Diagram showing a longitudinal section of the wave tank for the run-up experiment with the water (blue) and the bed (brown) where the coordinates have been non-dimensionalised [63].

one of these experiments, which captured the run-up of a non-breaking solitary wave with a non-linearity 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 [64]. Therefore, the effect of the extra dispersive terms included by the Serre equations on the inundation process is not well tested by this experiment. However, this experiment does demonstrate the robustness of the numerical methods during the wetting and drying of the bed.

The numerical experiments used the non-dimensionalised quantities reported by Synolakis [63] 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 (3.23) satisfied by setting  $\Delta t = 0.1\Delta x$ . The spatial reconstruction used the input parameter  $\theta = 1.2$  and acceleration due to gravity  $g = 1$  was chosen to match the non-dimensionalisation.

The non-dimensionalised water surface data is given at the various times in Figure 6.23 for FDVM<sub>2</sub> and 6.22 for FEVM<sub>2</sub>. The error in conservation of  $h'$  and

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

Table 6.5: Initial and final total amounts and the conservation error for  $h'$  and  $\mathcal{H}'$  for the numerical solution of FEVM<sub>2</sub> for the run-up experiment.

$\mathcal{H}'$  as measured by  $C^*$  are given in Tables 6.5 and 6.6 for FEVM<sub>2</sub> and FDVM<sub>2</sub> respectively.

The results for FEVM<sub>2</sub> and FDVM<sub>2</sub> are indistinguishable, replicating the incoming wave properties and the maximum run-up 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 run-down is likely caused by the omission of bed friction for the Serre equations in this thesis.

Both  $h'$  and  $\mathcal{H}'$  are well conserved by the method throughout the run-up and run-down of the wave, particularly  $h'$ . The total energy  $\mathcal{H}'$  of the method is also well conserved, however  $\mathcal{H}'$  appears to have slightly increased in the simulation during the run-up process due to the methods 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  $u'h'$  and  $G'$  will only be conserved in this experiment after the wave has completely reflected from the beach. Full reflection of the wave has not occurred by  $t' = 70$  and so the conservation results for  $uh$  and  $G$  were omitted from Tables 6.5 and 6.6.

These numerical solutions demonstrate good agreement with experimental results and display the capability of the method to model the inundation of non-breaking waves.

In this chapter FEVM<sub>2</sub> and FDVM<sub>2</sub> were validated using experimental data. It was found that for most experiments the solutions of FEVM<sub>2</sub> and FDVM<sub>2</sub> were indistinguishable although FEVM<sub>2</sub> is the preferred method due to its greater robustness.

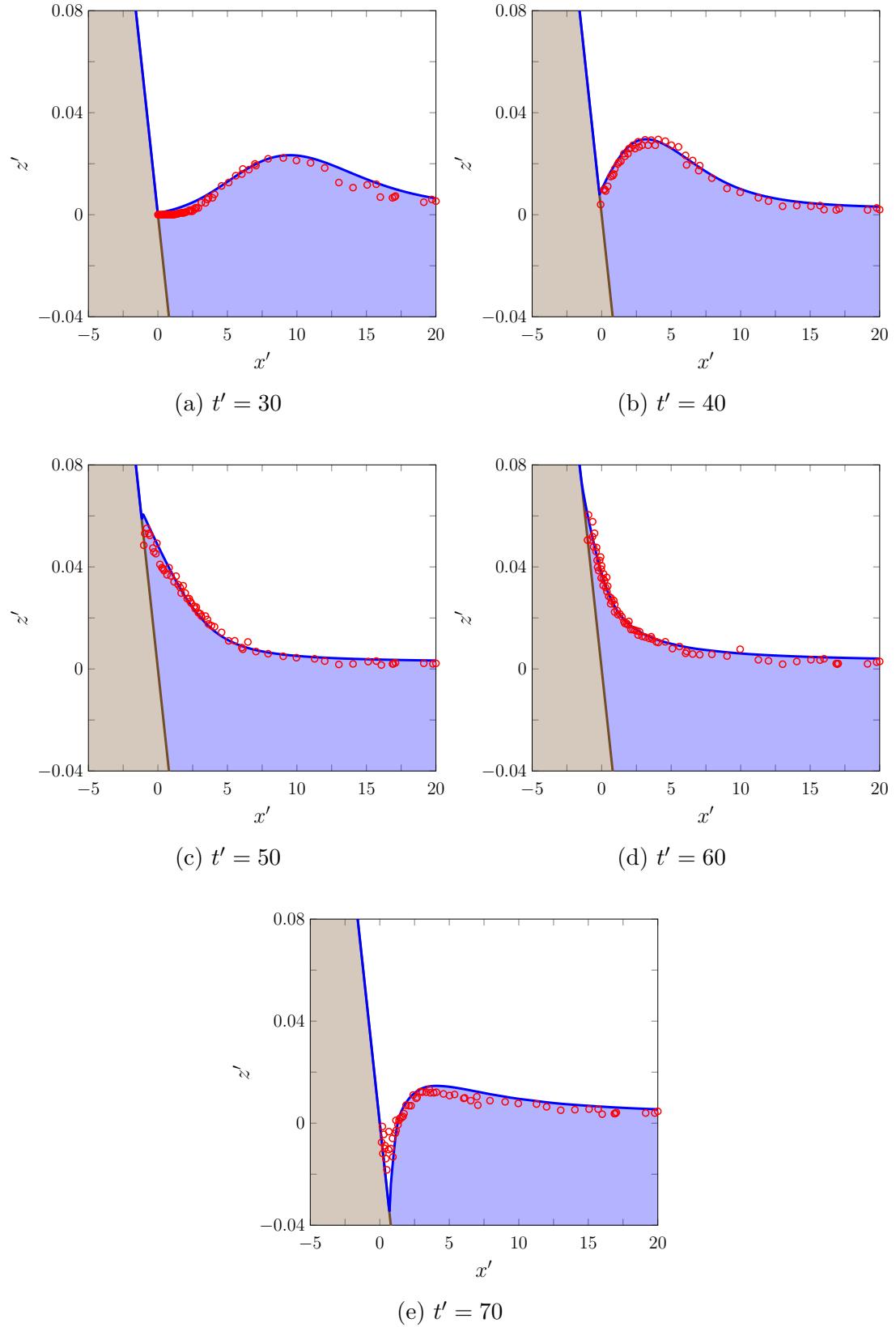


Figure 6.22: A comparison of the water surface profiles  $w'(x', t')$  for the experiment (○) and the numerical solution (■) produced by FEVM<sub>2</sub> over the bed (■) at various times.

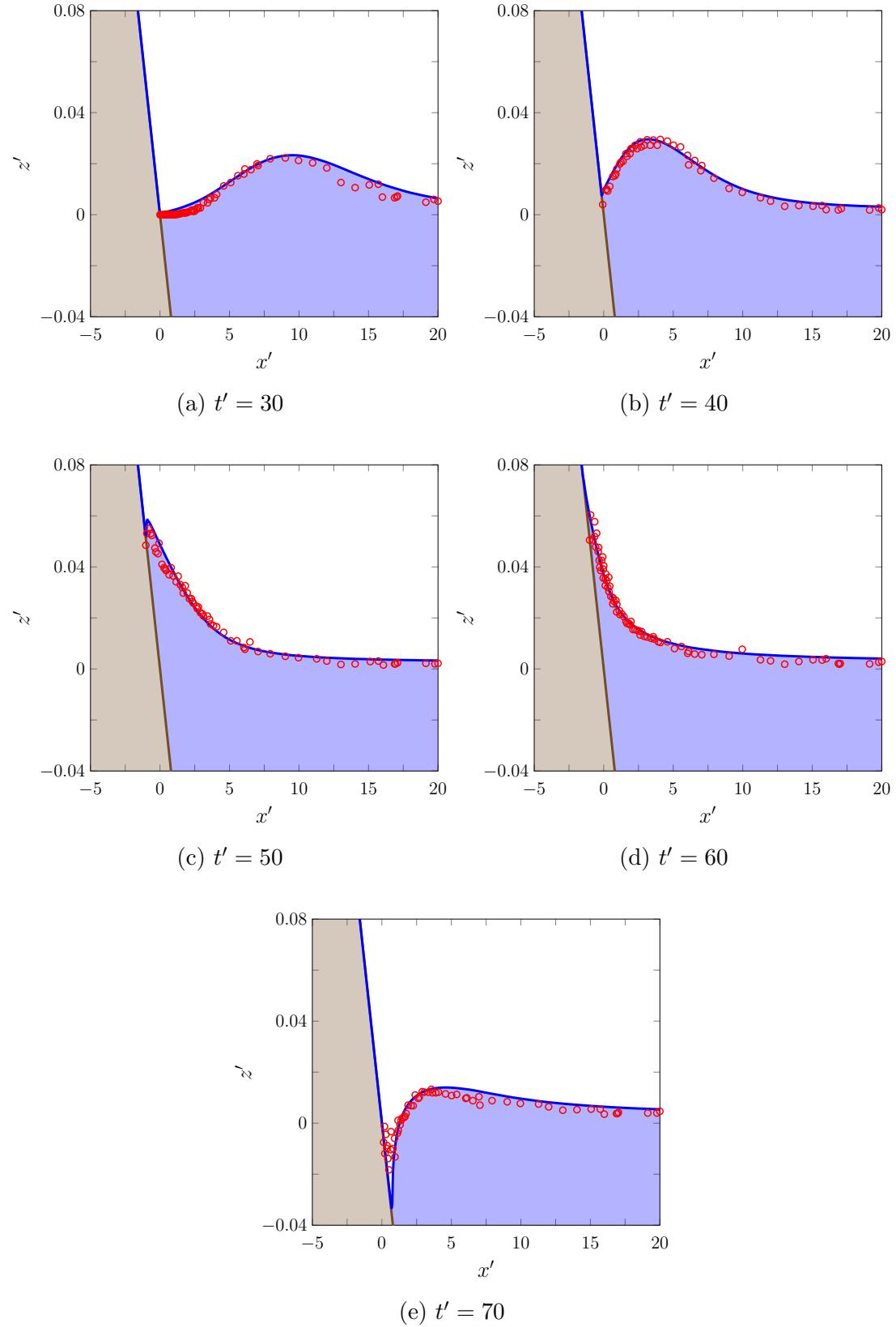


Figure 6.23: A comparison of the water surface profiles  $w'(x', t')$  for the experiment (○) and the numerical solution (■) produced by FDVM<sub>2</sub> over the bed (■) at various times.

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

Table 6.6: Initial and final total amounts and the conservation error for  $h'$  and  $\mathcal{H}'$  for the numerical solution of FDVM<sub>2</sub> for the run-up experiment.



# Chapter 7

## Conclusion

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

A well-balanced second-order Finite Element Volume Method (FEVM) termed FEVM<sub>2</sub> was described for the one-dimensional Serre equations. The method uses a finite element and a finite volume method and thus is robust to steep gradients present in the conserved variables  $h$  and  $G$ . The use of these methods also makes the FEVM well suited to solve the two-dimensional Serre equations using unstructured meshes with parallelised code.

A linear analysis of the convergence and dispersion properties of FEVM<sub>2</sub> was performed. The results of the linear analysis were also given for the Finite Difference Volume Methods of first- (FDVM<sub>1</sub>), second- (FDVM<sub>2</sub>) and third-order (FDVM<sub>3</sub>) described by Zoppou et al. [14]. Also provided were the linear results for the second-order finite difference methods  $\mathcal{D}$  and  $\mathcal{W}$  described by Pitt et al. [17]. The analysis demonstrated that FDVM<sub>1</sub>, FDVM<sub>2</sub>, FDVM<sub>3</sub>, FEVM<sub>2</sub> and  $\mathcal{D}$  are convergent methods, while  $\mathcal{W}$  is only convergent when the mean background flow velocity is zero. The dispersion analysis demonstrated that all methods approximated the dispersion relation of the Serre equations with the expected order of accuracy. This analysis extended a previous analysis of the dispersion relationships of numerical methods [39] by allowing non-zero mean flow, combining the spatial and temporal analyses and comparing the real and imaginary parts of the dispersion error.

A comparison of the various numerical methods and the analytic solitary travelling wave solution of the Serre equations was performed. The expected order of accuracy and conservation properties of all the methods was observed.

However, these results also demonstrated that the increase in accuracy achieved by a third-order method over a second-order method did not warrant the extra computational effort, justifying the further development of second-order methods over third-order methods for future work. For this reason only the second-order FDVM<sub>2</sub> and FEVM<sub>2</sub> were developed further to allow varying bathymetry and dry beds.

The second-order FDVM<sub>2</sub> and FEVM<sub>2</sub> were then validated against the lake at rest steady state and the forced solutions. These results demonstrated that these methods are well-balanced and accurately approximate all terms in the Serre equations in the presence of dry beds.

Finally the second-order FDVM<sub>2</sub> and FEVM<sub>2</sub> were compared to experimental data; demonstrating their modelling capabilities across a wide array of physical scenarios. These results established the greater robustness of FEVM<sub>2</sub>; as FDVM<sub>2</sub> was found to be unstable for the solitary wave over a fringing reef experiment due to the presence of large jumps in the water surface as the wave broke.

To summarise the major contributions of my research are:

- Observation and justification of a new structure in the solution of the Serre equations to the dam-break problem.
- Development and description of the well-balanced FEVM<sub>2</sub> that can handle dry beds and conserves  $h$  and  $G$ .
- Linear analysis of the convergence properties of FEVM<sub>2</sub>, FDVM<sub>1</sub>, FDVM<sub>2</sub>, FDVM<sub>3</sub>,  $\mathcal{D}$  and  $\mathcal{W}$ .
- Analysis of the dispersion properties of FEVM<sub>2</sub>, FDVM<sub>1</sub>, FDVM<sub>2</sub>, FDVM<sub>3</sub>,  $\mathcal{D}$  and  $\mathcal{W}$ . This analysis allowed for non-zero mean flow velocity and accounted for the total dispersion error.
- Validation of FEVM<sub>2</sub> against analytic and forced solutions and experimental results.

## 7.1 Future Work

Following the work conducted in this thesis; some possible extensions are:

- Inclusion of wave-breaking in the model.
- Implementation of different boundary conditions.

- Incorporation of discontinuous bed profiles.
- Incorporation of bed friction in FEVM<sub>2</sub>.
- A complete analysis of the convergence properties of FEVM<sub>2</sub>.
- Extension of FEVM<sub>2</sub> to the two-dimensional Serre equations on unstructured meshes.



# Appendix A

## Additional Conservation Information

To calculate the conservation errors requires an analytic expression for the total amount of  $h$ ,  $uh$ ,  $G$  and  $\mathcal{H}$  present in the initial conditions. Therefore, to facilitate the validation tests against the analytic solutions performed in Chapter 5 we present these analytic expressions for the initial conditions of the solitary travelling wave and the lake at rest solutions described in Chapter 2. To allow for the simple calculation of the integrals in a concise way for any domains we present the integrals in indefinite form.

### A.1 Solitary Travelling Wave Solution

The solitary wave solution (2.11) is

$$h(x, t) = a_0 + a_1 \operatorname{sech}^2(\kappa [x - ct]),$$

$$\begin{aligned} u(x, t) &= c \left( 1 - \frac{a_0}{h(x, t)} \right), \\ b(x) &= 0. \end{aligned}$$

When  $t = 0$  the indefinite spatial integrals of the conserved quantities are

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

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

$$\int G(x, 0) \, dx = \frac{ca_1}{3\kappa} \left( 3 + 2a_0^2\kappa^2 \operatorname{sech}^2(\kappa x) + 2a_0a_1\kappa^2 \operatorname{sech}^4(\kappa x) \right) \tanh(\kappa x) + \text{constant}, \quad (\text{A.1c})$$

$$\begin{aligned} \int \mathcal{H}(x, 0) \, dx &= \frac{1}{2} \left( \int g [h(x, 0)]^2 \, dx + \int h(x, 0) [u(x, 0)]^2 \, dx \right. \\ &\quad \left. + \int [h(x, 0)]^3 \left[ \frac{\partial u(x, 0)}{\partial x} \right]^2 \, dx \right) \end{aligned} \quad (\text{A.1d})$$

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

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

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

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

Therefore, we have the analytic values of the total amounts of our conserved quantities for the solitary travelling wave solution (2.11) when  $t = 0s$ , as desired.

## A.2 Lake At Rest Solution

The lake at rest solution (2.12) for an arbitrary bed profile  $b(x)$  is

$$\begin{aligned} h(x, t) &= \max \{a_0 - b(x), 0\}, \\ u(x, t) &= 0, \\ G(x, t) &= 0. \end{aligned}$$

The total amount of  $uh$  and  $G$  in the system are straightforward to calculate as both are zero everywhere and so we have

$$\begin{aligned} \int u(x, 0)h(x, 0) dx &= 0 + \text{constant}, \\ \int G(x, 0) dx &= 0 + \text{constant}. \end{aligned}$$

To calculate the total  $h$  and  $\mathcal{H}$  in the solution we must partition the domain into wet regions where  $b(x) < a_0$  and dry regions where  $b(x) \geq a_0$ . For the dry regions  $h$  and  $\mathcal{H}$  are zero everywhere and so we have

$$\begin{aligned} \int h(x, 0) dx &= 0 + \text{constant}, \\ \int \mathcal{H}(x, 0) dx &= 0 + \text{constant} \end{aligned}$$

whilst in wet regions we have

$$\int h(x, 0) dx = a_0 x - \int b(x) dx + \text{constant}, \quad (\text{A.2a})$$

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

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



# Appendix B

## Finite Element Method Definitions

The definitions of the basis functions of the finite element method used by FEVM<sub>2</sub> described in Chapter 3 and the function spaces mentioned in Chapter 3 are provided here. Beginning with the basis function definitions.

### B.1 Basis Function Definitions

Since all integrals of the basis functions are calculated with respect to the variable  $\xi$ , the basis functions are given in terms of  $\xi$ . The mapping from the  $x$ -space of the numerical grid to the canonical  $\xi$ -space is

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

This mapping takes the  $j^{th}$  cell  $[x_{j-1/2}, x_{j+1/2}]$  in the  $x$ -space to the interval  $[-1, 1]$  in the  $\xi$ -space.

The basis functions  $\psi$  for  $h$  and  $G$  shown in Figure B.1 are

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

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

The basis functions  $\phi$  for  $u$  and the test function  $v$  displayed in Figure B.2

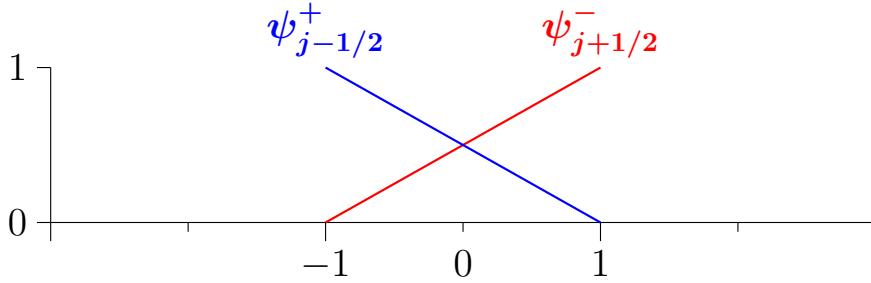


Figure B.1: Support of the discontinuous linear basis functions  $\psi$  which are non-zero over the  $j^{th}$  cell plotted in the  $\xi$ -space.

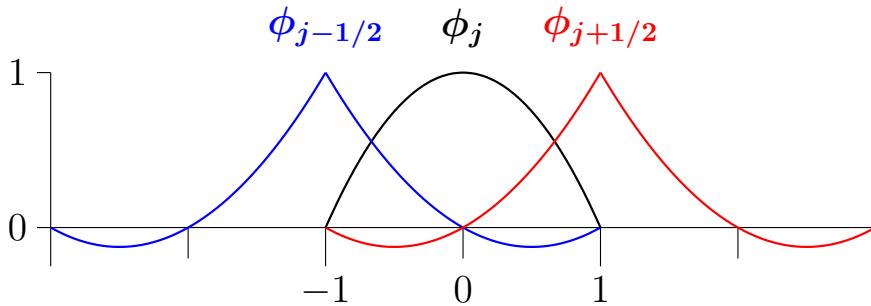


Figure B.2: Support of the continuous piecewise quadratic basis functions  $\phi$  which are non-zero over the  $j^{th}$  cell plotted in the  $\xi$ -space.

are given by

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

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

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

Finally the basis functions  $\gamma$  for the bed profile  $b$  displayed in Figure B.3 are

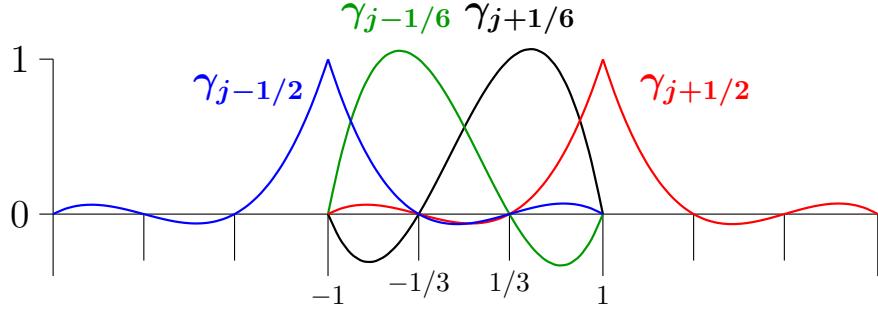


Figure B.3: Support of the continuous piecewise cubic basis functions  $\gamma$  which are non-zero over the  $j^{th}$  cell plotted in the  $\xi$ -space.

given by

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

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

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

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

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

## B.2 Function Space Definitions

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

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

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

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

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

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

# Appendix C

## Linear Analysis Results

In this appendix we provide additional details supplementing the linear analysis results reported in Chapter 4. We provide the evolution matrices  $\mathbf{E}$  for all methods in this thesis, details on how to calculate the convergence and dispersion properties from the evolution matrix where appropriate and the consistency tables not already in Chapter 4.

The linear analysis in Chapter 4 studies the convergence and dispersion properties of the numerical methods solving the linearised Serre equations with a horizontal bed (4.3). The linearisation occurs for a small perturbation  $\eta$  on the mean water depth  $H$  and a small perturbation  $\mu$  on the mean background flow velocity  $U$ . The linearised Serre equations with a horizontal bed (4.3) are

$$\frac{\partial \eta}{\partial t} + H \frac{\partial \mu}{\partial x} + U \frac{\partial \eta}{\partial x} = 0, \quad (\text{C.1a})$$

$$H \frac{\partial \mu}{\partial t} + gH \frac{\partial \eta}{\partial x} + UH \frac{\partial \mu}{\partial x} - \frac{H^3}{3} \left( U \frac{\partial^3 \mu}{\partial x^3} + \frac{\partial^3 \mu}{\partial x^2 \partial t} \right) = 0 \quad (\text{C.1b})$$

which can be written in conservation law form (4.4) for  $\eta$  and  $G$  as

$$\frac{\partial \eta}{\partial t} + \frac{\partial}{\partial x} (H\mu + U\eta) = 0, \quad (\text{C.2a})$$

$$\frac{\partial G}{\partial t} + \frac{\partial}{\partial x} (UG + UH\mu + gH\eta) = 0 \quad (\text{C.2b})$$

where

$$G = UH + U\eta + H\mu - \frac{H^3}{3} \frac{\partial^2 \mu}{\partial x^2}. \quad (\text{C.2c})$$

The linear analysis assumes that  $\eta$  and  $\mu$  are Fourier modes (4.5). A generic quantity  $q$  is a Fourier mode if

$$q(x, t) = q(0, 0) e^{i(\omega \pm t + kx)} \quad (\text{C.3})$$

where  $\omega^\pm$  is the frequency of the wave given by the dispersion relation (2.9) and  $k$  is the wavenumber. There are two possible frequencies of waves corresponding to upwind and downwind travelling waves these are denoted with the  $-$  and  $+$  superscripts respectively. Since  $\eta$  and  $\mu$  are Fourier modes (C.3) then  $G$  as well as the cell averages  $\bar{\eta}$ ,  $\bar{\mu}$  and  $\bar{G}$  are also Fourier modes.

For a fixed grid such as those used by all numerical methods in this thesis we can relate these Fourier mode quantities at different nodal values to one another. So that for a generic Fourier mode  $q$  we have

$$q_{j+l}^{n+m} = q_j^n e^{i(m\omega^\pm \Delta t + lk\Delta x)} \quad (\text{C.4})$$

where  $\Delta t$  is the length of a time step and  $\Delta x$  is the length of a cell.

Applying the numerical method to the linearised Serre equations (C.1) and using (C.4) one can derive an expression relating the quantities at the current and earlier times to the quantities at the next time via the evolution matrix  $\mathbf{E}$  of the numerical method. This evolution matrix can then be analysed to obtain the convergence and dispersion properties of the numerical method.

This appendix will provide the process to calculate  $\mathbf{E}$  for the first-, second- and third-order Finite Difference Volume Methods (FDVM) which are named FDVM<sub>1</sub>, FDVM<sub>2</sub> and FDVM<sub>3</sub> respectively. Descriptions of these methods were published by Zoppou et al. [14]. Additionally, we provide the evolution matrices for the two second-order Finite Difference Methods (FDM)  $\mathcal{D}$  and  $\mathcal{W}$ , whose descriptions were published by Pitt et al. [17].

The linear analysis for the second-order finite element volume method FEVM<sub>2</sub> was performed in Chapter 4. Given that work the evolution matrix for other finite volume based methods can be written in terms of the factors given by the reconstruction operators and the velocity solve operators of the method. We demonstrate this process and provide the expressions for these factors. This allows others to reproduce the evolution matrices for all finite volume based methods in this thesis.

For  $\mathcal{D}$  and  $\mathcal{W}$  the evolution matrix itself is provided. Since these FDM use the quantities at previous time steps, the shape of the evolution matrix is different. Hence, the process to obtain the convergence and dispersion properties from these particular evolution matrices is explained.

Finally, the consistency results for FDVM<sub>1</sub>, FDVM<sub>2</sub>, FDVM<sub>3</sub>,  $\mathcal{D}$  and  $\mathcal{W}$  will be provided. The results for FEVM<sub>2</sub> were provided in Chapter 4.

Method	<b>E</b>
FDVM <sub>1</sub>	$\mathbf{I} - \Delta t \mathbf{F}$
FDVM <sub>2</sub> and FEVM <sub>2</sub>	$\mathbf{I} - \Delta t \mathbf{F} + \frac{1}{2} \Delta t^2 \mathbf{F}^2$
FDVM <sub>3</sub>	$\mathbf{I} - \Delta t \mathbf{F} + \frac{1}{2} \Delta t^2 \mathbf{F}^2 - \frac{1}{6} \Delta t^3 \mathbf{F}^3$

Table C.1: Formula for **E** given **F** determined by the SSP Runge-Kutta time stepping method.

## C.1 Finite Difference Volume Methods

For the finite volume based methods the evolution matrix is a  $2 \times 2$  matrix that gives the following relationship (4.1)

$$\begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^{n+1} = \mathbf{E} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n.$$

This evolution matrix **E** is obtained by applying the numerical method to (C.2) and making use of (C.4) for all quantities.

In Chapter 4 the evolution matrix for FEVM<sub>2</sub> is given in terms of the flux matrix **F** obtained from the flux approximation of the method. This expression relating **E** and **F** comes from the employed SSP Runge-Kutta time stepping of the method. The expressions for all employed SSP Runge-Kutta time stepping schemes and their associated numerical methods are summarised in Table C.1. Since FDVM<sub>2</sub> and FEVM<sub>2</sub> both use second-order SSP Runge-Kutta time stepping, their expressions are the same.

The flux matrix **F** from the finite volume methods approximation to the flux terms (4.12) is

$$\mathbf{F} = -\frac{(1 - e^{-ik\Delta x})}{\Delta x} \begin{bmatrix} \mathcal{F}^{\eta,\eta} & \mathcal{F}^{\eta,G} \\ \mathcal{F}^{G,\eta} & \mathcal{F}^{G,G} \end{bmatrix}. \quad (\text{C.5})$$

Where  $\mathcal{F}^{\eta,\eta}$ ,  $\mathcal{F}^{\eta,G}$ ,  $\mathcal{F}^{G,\eta}$ ,  $\mathcal{F}^{G,G}$  can be written in terms of factors given by the reconstruction and the velocity solve operators. These expressions written in

Froude Number	$\mathcal{F}^{\eta,\eta}$
$Fr < -1$	$H\mathcal{G}^\eta + U\mathcal{R}_{j+1/2}^+$
$-1 \leq Fr \leq 1$	$H\mathcal{G}^\eta + \frac{U}{2} \left( \mathcal{R}_{j+1/2}^- + \mathcal{R}_{j+1/2}^+ \right) - \frac{\sqrt{gH}}{2} \left( \mathcal{R}_{j+1/2}^+ - \mathcal{R}_{j+1/2}^- \right)$
$1 < Fr$	$H\mathcal{G}^\eta + U\mathcal{R}_{j+1/2}^-$

Table C.2: Factor  $\mathcal{F}^{\eta,\eta}$  that multiples  $\eta$  in the flux function for  $\eta$  for all finite volume based methods.

terms of these factors are the same for all finite volume based methods in the thesis, as they all used the flux approximation of Kurganov et al. [48].

The term  $\mathcal{F}^{\eta,G}$  in the flux matrix does depend on the Froude number  $Fr = U/gH$  and is

$$\mathcal{F}^{\eta,G} = H\mathcal{G}^G.$$

The expressions for the other terms of  $\mathbf{F}$  which do depend on the Froude number are summarised in Tables C.2-C.4 for all values of  $Fr$  with  $\mathcal{F}^{\eta,\eta}$  in Table C.2,  $\mathcal{F}^{G,\eta}$  in Table C.3 and  $\mathcal{F}^{G,G}$  in Table C.4.

Using the appropriate expressions for the factors generated by the reconstruction operators  $\mathcal{R}_j$ ,  $\mathcal{R}_{j-1/2}^+$ ,  $\mathcal{R}_{j+1/2}^-$  and the velocity solve operators  $\mathcal{G}^\eta$  and  $\mathcal{G}^G$  used by the method, one can obtain all the terms of  $\mathbf{F}$  (C.5) for the finite volume based methods.

The expressions for these fundamental factors of all finite volume based methods are summarised in Table C.5 for  $\mathcal{R}_j$ , Table C.6 for  $\mathcal{R}_{j-1/2}^+$ , Table C.7 for  $\mathcal{R}_{j+1/2}^-$  and Table C.8 for  $\mathcal{G}^G$ . Since  $\mathcal{G}^\eta = -U\mathcal{G}^G$  we have only provided the table for  $\mathcal{G}^G$ . Since the perturbations are Fourier modes (4.5) the reconstruction factor  $\mathcal{R}_{j+1/2}^+ = e^{ik\Delta x} \mathcal{R}_{j-1/2}^+$ .

Tables C.5-C.8 also include the factors for FEVM<sub>2</sub> summarising the work in Chapter 4. The analytic value of the factors for an exact method are also provided. Additionally, the lowest order term of the Taylor series of the difference between the factor from a numerical method and the exact factor are also provided. The reported Taylor series results demonstrate that all methods use

Froude Number	$\mathcal{F}^{G,\eta}$
$Fr < -1$	$UH\mathcal{G}^\eta + gH\mathcal{R}_{j+1/2}^+$
$-1 \leq Fr \leq 1$	$\frac{U\sqrt{gH}}{2} (\mathcal{R}_{j+1/2}^- - \mathcal{R}_{j+1/2}^+) + UH\mathcal{G}^\eta + \frac{gH}{2} (\mathcal{R}_{j+1/2}^- + \mathcal{R}_{j+1/2}^+)$
$1 < Fr$	$UH\mathcal{G}^\eta + gH\mathcal{R}_{j+1/2}^-$

Table C.3: Factor  $\mathcal{F}^{G,\eta}$  that multiples  $\eta$  in the flux function for  $G$  for all finite volume based methods.

Froude Number	$\mathcal{F}^{G,G}$
$Fr < -1$	$U\mathcal{R}_{j+1/2}^+ + UH\mathcal{G}^G$
$-1 \leq Fr \leq 1$	$UH\mathcal{G}^G + \frac{U}{2} (\mathcal{R}_{j+1/2}^- + \mathcal{R}_{j+1/2}^+) - \frac{\sqrt{gH}}{2} (\mathcal{R}_{j+1/2}^+ - \mathcal{R}_{j+1/2}^-)$
$1 < Fr$	$U\mathcal{R}_{j+1/2}^+ + UH\mathcal{G}^G$

Table C.4: Factor  $\mathcal{F}^{G,G}$  that multiples  $G$  in the flux function for  $G$  for all finite volume based methods

Method	$\mathcal{R}_j$	Lowest Order Term of Method - Exact
Exact	$\frac{k\Delta x}{2 \sin\left(k\frac{\Delta x}{2}\right)}$	-
FDVM <sub>1</sub>	1	$-\frac{1}{24}k^2\Delta x^2$
FDVM <sub>2</sub> and FEVM <sub>2</sub>	1	$-\frac{1}{24}k^2\Delta x^2$
FDVM <sub>3</sub>	$\frac{26 - 2 \cos(k\Delta x)}{24}$	$-\frac{3}{640}k^4\Delta x^4$

Table C.5: Factor  $\mathcal{R}_j$  from reconstructing the nodal value at the midpoint and the lowest order term of the Taylor series of the factor in the method minus the exact factor for all finite volume based methods.

approximations with the appropriate order of accuracy or better.

The terms of the flux matrix  $\mathbf{F}$  (C.5) can be calculated from these expressions for the factors summarised in Tables C.5 - C.8. The evolution matrix  $\mathbf{E}$  can then be calculated from  $\mathbf{F}$  based on the SSP Runge-Kutta time stepping expressions summarised in Table C.1. Thus all the evolution matrices of the finite volume based methods can be obtained, as desired.

Method	$\mathcal{R}_{j-1/2}^+$	Lowest Order Term of Method - Exact
Exact	$\frac{k\Delta x}{2 \sin\left(\frac{k\Delta x}{2}\right)} e^{-\frac{ik\Delta x}{2}}$	-
FDVM <sub>1</sub>	1	$\frac{i}{2}k\Delta x$
FDVM <sub>2</sub> and FEVM <sub>2</sub>	$1 - \frac{i \sin(k\Delta x)}{2}$	$\frac{1}{12}k^2\Delta x^2$
FDVM <sub>3</sub>	$\frac{1}{6}(5 + 2e^{-ik\Delta x} - e^{ik\Delta x})$	$\frac{i}{12}k^3\Delta x^3$

Table C.6: Factor  $\mathcal{R}_{j-1/2}^+$  from the reconstruction of  $\eta$  and  $G$  at  $x_{j+1/2}$  from the  $(j+1)^{th}$  cell and the lowest order term of the Taylor series of the factor in the method minus the exact factor for all finite volume based methods.

Method	$\mathcal{R}_{j+1/2}^-$	Lowest Order Term of Method - Exact
Exact	$\frac{k\Delta x}{2 \sin\left(\frac{k\Delta x}{2}\right)} e^{\frac{ik\Delta x}{2}}$	-
FDVM <sub>1</sub>	1	$-\frac{i}{2}k\Delta x$
FDVM <sub>2</sub> and FEVM <sub>2</sub>	$1 + \frac{i \sin(k\Delta x)}{2}$	$\frac{1}{12}k^2\Delta x^2$
FDVM <sub>3</sub>	$\frac{1}{6}(5 - e^{-ik\Delta x} + 2e^{ik\Delta x})$	$-\frac{i}{12}k^3\Delta x^3$

Table C.7: Factor  $\mathcal{R}_{j+1/2}^-$  from the reconstruction of  $\eta$  and  $G$  at  $x_{j+1/2}$  using the  $j^{th}$  cell and the lowest order term of the Taylor series of the factor in the method minus the exact factor for all finite volume based methods.

Method	$\mathcal{G}^G$	Lowest Order Term of Method - Exact
Exact	$\frac{3}{3H + H^3 k^2} \frac{k\Delta x}{2 \sin\left(\frac{k\Delta x}{2}\right)} e^{\frac{ik\Delta x}{2}}$	-
FDVM <sub>1</sub>	$\frac{3\Delta x^2 (1 + e^{ik\Delta x})}{6\Delta x^2 H - 2H^3 (2 \cos(k\Delta x) - 2)}$	$-\frac{6 + H^2 k^2}{4H (3 + H^2 k^2)^2} k^2 \Delta x^2$
FDVM <sub>2</sub>	$\frac{3\Delta x^2 (1 + e^{ik\Delta x})}{6\Delta x^2 H - 2H^3 (2 \cos(k\Delta x) - 2)}$	$-\frac{6 + H^2 k^2}{4H (3 + H^2 k^2)^2} k^2 \Delta x^2$
FEVM <sub>2</sub>	$\begin{aligned} & \frac{\Delta x}{6} \left( 1 + \frac{i \sin(k\Delta x)}{2} + e^{ik\Delta x} \left[ 1 - \frac{i \sin(k\Delta x)}{2} \right] \right) \\ & \div \left( H \frac{\Delta x}{30} \left[ 4 \cos\left(\frac{k\Delta x}{2}\right) - 2 \cos(k\Delta x) + 8 \right] \right. \\ & \left. + \frac{H^3}{9\Delta x} \left[ -16 \cos\left(\frac{k\Delta x}{2}\right) + 2 \cos(k\Delta x) + 14 \right] \right) \end{aligned}$	$\frac{12 + 5H^2 k^2}{40H (3 + H^2 k^2)^2} k^2 \Delta x^2$
FDVM <sub>3</sub>	$\frac{9\Delta x^2 (-e^{-ik\Delta x} + 9e^{ik\Delta x} - e^{2ik\Delta x} + 9)}{144\Delta x^2 H - 4H^3 (32 \cos(k\Delta x) - 2 \cos(2k\Delta x) - 30)}$	$-\frac{243 + 49H^2 k^2}{960H (3 + H^2 k^2)^2} k^4 \Delta x^4$

Table C.8: Factor  $\mathcal{G}^G$  that multiples  $G$  given by solving (4.3c) for  $v_{j+1/2}$  and the lowest order term of the Taylor series of the factor in the method minus the exact factor for all finite volume based methods.

## C.2 Finite Difference Methods

The Finite Difference Methods (FDM) solve the linearised Serre equations with a horizontal bed in non-conservative form (C.1). The FDM rely on previous time steps as well as the current time step to update the quantities. One particular way of expressing this is with a  $4 \times 4$  evolution matrix  $\mathbf{E}$  producing the following relationship

$$\begin{bmatrix} \eta^{n+1} \\ \mu^{n+1} \\ \eta^n \\ \mu^n \end{bmatrix}_j = \mathbf{E} \begin{bmatrix} \eta^n \\ \mu^n \\ \eta^{n-1} \\ \mu^{n-1} \end{bmatrix}_j \quad (\text{C.6})$$

where the time superscript was brought inside the vector to make clear the time step at which the different elements are placed. Since the FDM are used to calculate  $\eta_j^{n+1}$  and  $\mu_j^{n+1}$  given  $\eta_j^n$ ,  $\mu_j^n$ ,  $\eta_j^{n-1}$  and  $\mu_j^{n-1}$  their evolution matrices have the following structure

$$\mathbf{E} = \begin{bmatrix} E_{0,0} & E_{0,1} & E_{0,2} & E_{0,3} \\ E_{1,0} & E_{1,1} & E_{1,2} & E_{1,3} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \quad (\text{C.7})$$

Because  $\eta_j^{n-1} = e^{-i\omega^\pm \Delta t} \eta_j^n$  and  $\mu_j^{n-1} = e^{-i\omega^\pm \Delta t} \mu_j^n$  as  $\eta$  and  $\mu$  are Fourier modes (C.3) then (C.6) can be rewritten as

$$\begin{bmatrix} \eta \\ \mu \end{bmatrix}_j^{n+1} = \mathbf{E}^{(2 \times 2)} \begin{bmatrix} \eta \\ \mu \end{bmatrix}_j^n \quad (\text{C.8})$$

where  $\mathbf{E}^{(2 \times 2)}$  is a  $2 \times 2$  matrix that depends on the elements of  $\mathbf{E}$  (C.7) in the following way

$$\mathbf{E}^{(2 \times 2)} = \begin{bmatrix} E_{0,0} + e^{-i\omega^\pm \Delta t} E_{0,2} & E_{0,1} + e^{-i\omega^\pm \Delta t} E_{0,3} \\ E_{1,0} + e^{-i\omega^\pm \Delta t} E_{1,2} & E_{1,1} + e^{-i\omega^\pm \Delta t} E_{1,3} \end{bmatrix}. \quad (\text{C.9})$$

Since the evolution matrix can be represented in two ways, we now state whether  $\mathbf{E}$  or  $\mathbf{E}^{(2 \times 2)}$  were used in the convergence and dispersion analysis of the FDM.

Using the Lax-Equivalence theorem the convergence of the methods were analysed by analysing their stability and consistency separately. The stability analysis was performed by finding the spectral radius of the naive evolution matrix  $\mathbf{E}$  of

the FDM (C.6). The consistency analysis was based on comparing the  $2 \times 2$  evolution matrix  $\mathbf{E}^{(2 \times 2)}$  (C.9) to the exact evolution matrix  $e^{i\omega^\pm \Delta t} \mathbf{I}$  for (C.8). Finally, the dispersion error was based on the eigenvalues of  $\mathbf{E}$  (C.6), this matrix has an additional two eigenvalues beyond the ones given by  $e^{i\omega^+ \Delta t}$  and  $e^{i\omega^- \Delta t}$  that were ignored. We found the methods had the same stability and dispersion properties when  $\mathbf{E}^{(2 \times 2)}$  was investigated. We will now present the  $4 \times 4$  evolution matrices for  $\mathcal{D}$  and  $\mathcal{W}$ . Given these matrices the corresponding  $2 \times 2$  evolution matrix  $\mathbf{E}^{(2 \times 2)}$  can be calculated using (C.9).

By using (C.4) all the derivative approximations in the finite difference methods  $\mathcal{D}$  and  $\mathcal{W}$  can be written as factors that are constant in  $j$  and  $n$  as was done for the finite volume based methods.

The evolution matrix for  $\mathcal{D}$  is

$$\mathbf{E} = \begin{bmatrix} E_{0,0} & E_{0,1} & 1 & 0 \\ E_{1,0} & E_{1,1} & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad (\text{C.10})$$

with

$$E_{0,0} = -\frac{2i\Delta t}{\Delta x} U \sin(k\Delta x),$$

$$E_{0,1} = -\frac{2i\Delta t}{\Delta x} H \sin(k\Delta x),$$

$$E_{1,0} = -\frac{6gi\Delta x\Delta t}{3\Delta x^2 - 2H^2(\cos(k\Delta x) - 1)} \sin(k\Delta x),$$

$$E_{1,1} = -\frac{2i\Delta t}{\Delta x} U \sin(k\Delta x).$$

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

$$\mathbf{E} = \begin{bmatrix} E_{0,0} & E_{0,1} & 0 & E_{0,3} \\ E_{1,0} & E_{1,1} & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad (\text{C.11})$$

with

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

$$- \frac{\Delta t}{\Delta x} U \left( i \sin(k\Delta x) - \frac{\Delta t}{\Delta x} U (\cos(k\Delta x) - 1) \right),$$

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

$$E_{0,3} = - \frac{\Delta t}{\Delta x} H \frac{i \sin(k\Delta x)}{2},$$

$$E_{1,0} = - \frac{6gi\Delta x\Delta t}{3\Delta x^2 - 2H^2(\cos(k\Delta x) - 1)} \sin(k\Delta x),$$

$$E_{1,1} = - \frac{2i\Delta t}{\Delta x} U \sin(k\Delta x).$$

### C.3 Consistency Results

The consistency results for FDVM<sub>1</sub>, FDVM<sub>2</sub>, FDVM<sub>3</sub>,  $\mathcal{D}$  and  $\mathcal{W}$  are provided here. The consistency results for FEVM<sub>2</sub> were provided in Chapter 4.

To demonstrate the consistency of FDVM<sub>1</sub>, FDVM<sub>2</sub> and FDVM<sub>3</sub> we need to demonstrate that all the lowest order terms of the Taylor series of  $\mathbf{E} - e^{i\omega^\pm \Delta t} \mathbf{I}$  contain a factor of  $\Delta t$ . Since the results are similar for  $\omega^-$  and  $\omega^+$  we only give the results for  $\omega^+$ . All these terms are presented in Tables C.9 and C.10 for FDVM<sub>1</sub>, Table C.11 for FDVM<sub>2</sub> and Tables C.12 and C.13 for FDVM<sub>3</sub>.

From Tables C.9-C.13 it is clear that all the lowest order terms of the Taylor series of  $\mathbf{E} - e^{i\omega^\pm \Delta t} \mathbf{I}$  for the FDVM contain a factor of  $\Delta t$ . Hence, FDVM<sub>1</sub>, FDVM<sub>2</sub> and FDVM<sub>3</sub> are consistent numerical methods.

To demonstrate that  $\mathcal{D}$  and  $\mathcal{W}$  are consistent we need to demonstrate that all the lowest order terms of the Taylor series of  $\mathbf{E}^{(2 \times 2)} - e^{i\omega^\pm \Delta t} \mathbf{I}$  contain a factor of  $\Delta t$ . Since the results are similar for  $\omega^-$  and  $\omega^+$  we only give the results for  $\omega^+$ . The results for  $\omega^+$  are presented in Table C.14 for  $\mathcal{D}$  and Table C.15 for  $\mathcal{W}$ .

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

Table C.9: Lowest order  $\Delta x$  term of the Taylor series for the elements of  $\mathbf{E} - e^{i\omega^+\Delta t}\mathbf{I}$  for FDVM<sub>1</sub>. Here  $\beta = 3 + k^2H^2$ .

Note that  $\mathcal{D}$  and  $\mathcal{W}$  do not depend on  $Fr$ .

From Tables C.14 and C.15 it can be seen that all the lowest order terms of the Taylor series of  $\mathbf{E}^{(2 \times 2)} - e^{i\omega^\pm\Delta t}\mathbf{I}$  for the FDM contain a factor of  $\Delta t$ . Therefore,  $\mathcal{D}$  and  $\mathcal{W}$  are consistent numerical methods.

---

Element	Lowest Order $\Delta t$ Term of $\mathbf{E} - e^{i\omega^+ \Delta t} \mathbf{I}$ for FDVM <sub>1</sub>
$E_{0,0} - e^{i\omega^+ \Delta t}$	$\frac{\sqrt{3gH\beta} + 3U}{\beta} ik\Delta t$
$E_{0,1}$	$-\frac{3}{\beta} ik\Delta t$
$E_{1,0}$	$\left(-gH + \frac{3U^2}{\beta}\right) ik\Delta t$
$E_{1,1} - e^{i\omega^+ \Delta t}$	$\frac{\sqrt{3gH\beta} - 3U}{\beta} ik\Delta t$

---

Table C.10: Lowest order  $\Delta t$  term of the Taylor series for the elements of  $\mathbf{E} - e^{i\omega^+ \Delta t} \mathbf{I}$  for FDVM<sub>1</sub> for all values of  $Fr$ . Here  $\beta = 3 + k^2 H^2$ .

---

Element	Lowest Order Terms of $\mathbf{E} - e^{i\omega^+ \Delta t} \mathbf{I}$ for FDVM <sub>2</sub>	
	$\Delta x$	$\Delta t$
$E_{0,0} - e^{i\omega^+ \Delta t}$	$-\frac{i(27 + 9H^2k^2 + H^4k^4)}{12\beta^2} U k^3 \Delta t \Delta x^2$	$\frac{\sqrt{3gH\beta} + 3U}{\beta} ik\Delta t$
$E_{0,1}$	$\frac{3 + \beta}{4\beta^2} ik^3 \Delta t \Delta x^2$	$-\frac{3}{\beta} ik\Delta t$
$E_{1,0}$	$-\left(gH + \frac{3U^2}{\beta} + \frac{9U^2}{\beta^2}\right) \frac{k^3}{12} \Delta t \Delta x^2$	$\left(-gH + \frac{3U^2}{\beta}\right) ik\Delta t$
$E_{1,1} - e^{i\omega^+ \Delta t}$	$\frac{-9 + H^2k^2\beta}{\beta^2} \frac{k^3}{12} iU \Delta t \Delta x^2$	$\frac{\sqrt{3gH\beta} - 3U}{\beta} ik\Delta t$

---

Table C.11: Lowest order terms of the Taylor series for the elements of  $\mathbf{E} - e^{i\omega^+ \Delta t} \mathbf{I}$  for FDVM<sub>2</sub> for all values of  $Fr$ . Here  $\beta = 3 + k^2 H^2$ .

---

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

---

Table C.12: Lowest order  $\Delta x$  term of the Taylor series for the elements of  $\mathbf{E} - e^{i\omega^+\Delta t}\mathbf{I}$  for FDVM<sub>3</sub>. Here  $\beta = 3 + k^2H^2$ .

---

Element	Lowest Order $\Delta t$ Term of $\mathbf{E} - e^{i\omega^+\Delta t}\mathbf{I}$ for FDVM <sub>3</sub>	
$E_{0,0} - e^{i\omega^+\Delta t}$	$\frac{\sqrt{3gH\beta} + 3U}{\beta}ik\Delta t$	
$E_{0,1}$		$-\frac{3}{\beta}ik\Delta t$
$E_{1,0}$		$\left(-gH + \frac{3U^2}{\beta}\right)ik\Delta t$
$E_{1,1} - e^{i\omega^+\Delta t}$		$\frac{\sqrt{3gH\beta} - 3U}{\beta}ik\Delta t$

---

Table C.13: Lowest order  $\Delta t$  term of the Taylor series for the elements of  $\mathbf{E} - e^{i\omega^+\Delta t}\mathbf{I}$  for FDVM<sub>3</sub> for all values of  $Fr$ . Here  $\beta = 3 + k^2H^2$ .

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

Table C.14: Lowest order terms of the Taylor series for the elements of  $\mathbf{E}^{(2 \times 2)} - e^{i\omega^\pm \Delta t} \mathbf{I}$  for  $\mathcal{D}$ . Here  $\beta = 3 + k^2 H^2$ .

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

Table C.15: Lowest order terms of the Taylor series for the elements of  $\mathbf{E}^{(2 \times 2)} - e^{i\omega^+ \Delta t} \mathbf{I}$  for  $\mathcal{W}$ . Here  $\beta = 3 + k^2 H^2$ .



# Appendix D

## Publications

This appendix lists the publications my research contributed to in chronological order. This list expands the earlier list in Chapter 1 by providing the abstracts for the paper in place of the summary. My contribution to the paper and the relevance of the paper to this thesis are also provided.

### A Solution of the Conservation Law Form of the Serre Equations

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

#### **Abstract:**

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

expensive than solving the shallow water wave equations.

**My Contribution:**

I produced an independent reproduction of the method of my coauthors, verifying their results. This method was used to compare the computational cost of solving the Serre equations and the Shallow Water Wave Equations (SWWE).

**Relevance to Thesis:**

The method described in this paper was the foundation of the second-order finite difference volume method whose results are reported in this thesis. The main extensions of this method during my research were the inclusion of varying bathymetry [14] and dry beds. The results of a linear analysis of the second-order finite difference volume method is provided in Chapter 4 and Appendix C. Since the linear analysis is performed for a completely submerged horizontal bed, these results apply to the method described in this paper. Finally, the second-order finite difference volume method was validated against analytic and forced solutions and experimental results in Chapter 5 and Chapter 6, respectively.

## Numerical Solution of the Fully Non-Linear Weakly Dispersive Serre Equations for Steep Gradient Flows

*Applied Mathematical Modelling (2017)*

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

### Abstract:

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

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

### My Contribution:

This paper was based on my research which implemented the methods, performed the dispersion analysis and produced the numerical solutions. These results were then written up by my coauthors.

### Relevance to Thesis:

The results of the first-, second- and third-order finite difference volume methods described in this paper are the methods whose results are reported in this thesis. The results of a linear analysis for these methods can be found in Chapter 4 and Appendix C. While a validation of some of these methods against analytic and forced solutions and experimental results can be found in Chapter 5 and Chapter 6, respectively.

The linear analysis of the dispersion properties of the methods in this paper was extended in this thesis by allowing for a non-zero background mean flow velocity. The analytic solution validation of this paper was reproduced in this

thesis and extended by studying the convergence and conservation properties of more quantities. The experimental results of the second-order finite difference volume method in this paper for the negative rectangular wave experiment [15] and periodic waves over a submerged bar experiment [16] were reproduced in this thesis.

## Importance of Dispersion for Shoaling Waves

*22nd International Congress on Modelling and Simulation (2017)*

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

### **Abstract:**

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

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

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

The results of these simulations demonstrated that the Serre and shallow water

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

**My Contribution:**

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

**Relevance to Thesis:**

The results of the second-order finite difference volume method for the experiments of Beji and Battjes [16] studying periodic waves over a submerged bar are reproduced in Chapter 6.

## Behaviour of the Serre Equations in the Presence of Steep Gradients Revisited

*Wave Motion (2018)*

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

### **Abstract:**

We use numerical methods to study the short term behaviour of the Serre equations in the presence of steep gradients because there are no known analytical solutions for these problems. In keeping with the literature we study a class of initial condition problems that are a smooth approximation to the initial conditions of the dam-break problem. This class of initial condition problems allow us to observe the behaviour of the Serre equations with varying steepness of the initial conditions. The numerical solutions of the Serre equations are justified by demonstrating that as the resolution increases they converge to a solution with little error in conservation of mass, momentum and energy independent of the numerical method. We observe and justify four different structures of the converged numerical solutions depending on the steepness of the initial conditions. Two of these structures were observed in the literature, with the other two not being commonly found in the literature. The numerical solutions are then used to assess how well the analytical solution of the shallow water wave equations captures the mean behaviour of the solution of the Serre equations for the dam-break problem in the short term. Lastly the numerical solutions are compared to asymptotic results in the literature to approximate the depth and location of the front of an undular bore.

### **My Contribution:**

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

### **Relevance to Thesis:**

The results of this paper are summarised in Chapter 2. This paper demonstrates the utility of using a finite volume based method to solve the Serre equations in the presence of steep gradients. Hence, their further development in this thesis.



# Bibliography

- [1] L. Euler. Principes généraux du mouvement des fluides. *Mémoires de l'académie des sciences de Berlin*, 11:274–315, 1757.
- [2] C.L.M.H. Navier. Mémoire sur les lois du mouvement des fluides. *Mémoires de l'Académie Royale des Sciences de l'Institut de France*, 6:389–440, 1823.
- [3] G.G. Stokes. *Transactions of the Cambridge Philosophical Society*, 8:287, 1845.
- [4] A.J. Chorin. The numerical solution of the Navier-Stokes equations for an incompressible fluid. *Bulletin of the American Mathematical Society*, 73(6):928–931, 1967.
- [5] C. Taylor and P. Hood. Numerical solution of the Navier-Stokes equations using the finite element technique. *Computers & Fluids*, 1:73–100, 1973.
- [6] F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. *Journal of Computational Physics*, 131(2):267–279, 1997.
- [7] D. Lannes and P. Bonneton. Derivation of asymptotic two-dimensional time-dependent equations for surface water wave propagation. *Physics of Fluids*, 21(1):16601, 2009.
- [8] The Clawpack Development Team. Clawpack documentation, 2018. URL <http://www.clawpack.org/>.
- [9] X. Wang. Comcot, 2009. URL <http://223.4.213.26/archive/tsunami/cornell/comcot.htm>.
- [10] S.G. Roberts. ANUGA, 2018. URL <https://anuga.anu.edu.au/>.

- [11] J. Grue, E.N. Pelinovsky, D. Fructus, T. Talipova, and C. Kharif. Formation of undular bores and solitary waves in the strait of Malacca caused by the 26 December 2004 Indian Ocean tsunami. *Journal of Geophysical Research: Oceans*, 113(C5):008, 2008.
- [12] J.T. Kirby, F. Shi, B. Tehranirad, J.C. Harris, and S.T. Grilli. Dispersive tsunami waves in the ocean: Model equations and sensitivity to dispersion and coriolis effects. *Ocean Modelling*, 62:39–55, 2013.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] S. Beji and J.A. Battjes. Numerical simulation of nonlinear wave propagation over a bar. *Coastal Engineering*, 23(1):1–16, 1994.
- [17] 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.
- [18] R.M. Sorensen. *Basic coastal engineering*. New York, NY Springer, 3rd edition, 2006.
- [19] F. Serre. Contribution à l'étude des écoulements permanents et variables dans les canaux. *La Houille Blanche*, 6:830–872, 1953.
- [20] 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.
- [21] 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.

- [22] F.J. Seabra-Santos, D.P. Renouard, and A.M. Temperville. Numerical and experimental study of the transformation of a solitary wave over a shelf or isolated obstacle. *Journal of Fluid Mechanics*, 176:117–134, 1981.
- [23] E. Barthélemy. Nonlinear shallow water theories for coastal waves. *Surveys in Geophysics*, 25(3):315–337, 2004.
- [24] P. Bonneton, F. Chazel, D. Lannes, F. Marche, and M. Tissier. A splitting approach for the fully nonlinear and weakly dispersive Green-Naghdi model. *Journal of Computational Physics*, 230(4):1479–1498, 2011.
- [25] J.A. Liggett. *Fluid Mechanics*. McGraw-Hill civil engineering series. McGraw-Hill Inc., New York, 1994.
- [26] 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.
- [27] 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.
- [28] W. Choi and R. Camassa. Fully nonlinear internal waves in a two-fluid system. *Journal of Fluid Mechanics*, 396:1–36, 1999.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] 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.

- [33] D. Mitsotakis, B. Ilan, and D. Dutykh. On the Galerkin/finite-element method for the Serre equations. *Journal of Scientific Computing*, 61(1):166–195, 2014.
- [34] D. Mitsotakis, D. Dutykh, and D. Carter. On the nonlinear dynamics of the traveling-wave solutions of the Serre system. *Wave Motion*, 70(1):166–182, 2017.
- [35] J.S.A. do Carmo, J.A. Ferreira, L. Pinto, and G. Romanazzi. An improved Serre model: Efficient simulation and comparative evaluation. *Applied Mathematical Modelling*, 56:404–423, 2018.
- [36] V.A. Dougalis, A. Duran, M.A. Lopez-Marcos, and D.E. Mitsotakis. Numerical study of the stability of solitary waves of the Bona-Smith family of Boussinesq systems. *Journal of Nonlinear Science*, 17(6):569–607, 2007.
- [37] R. Cienfuegos, E. Barthélemy, and P. Bonneton. A fourth-order compact finite volume scheme for fully nonlinear and weakly dispersive Boussinesq-type equations. Part I: Model development and analysis. *International Journal for Numerical Methods in Fluids*, 51(11):1217–1253, 2006.
- [38] S.F. Bradford and B.F. Sanders. Finite volume schemes for the Boussinesq equations. In *Ocean Wave Measurement and Analysis (2001)*, pages 953–962. American Society of Civil Engineers, 2002.
- [39] 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.
- [40] J. Pitt. *A Second Order Well Balanced Hybrid Finite Volume and Finite Difference Method for the Serre Equations*. Honour’s thesis, Australian National University, Mathematical Sciences Institute, College of Physical and Mathematical Sciences, Australian National University, Canberra, ACT 2600, Australia, 2014.
- [41] P.M. Prenter. *Splines and Variational Methods*. A Wiley-Interscience publication. Wiley, New York, 1975.
- [42] K.E. Atkinson. *An introduction to numerical analysis*. John Wiley & Sons, Canada, 2nd edition, 1989.

- [43] P.L. Roe. Characteristic-based schemes for the Euler equations. *Annual Review of Fluid Mechanics*, 18(1):337–365, 1986.
- [44] 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.
- [45] A. Harten. High resolution schemes for hyperbolic conservation laws. *Journal of Computational Physics*, 49(3):357–3935, 1983.
- [46] P.J. Davis and P. Rabinowitz. *Methods of Numerical Integration*. Computer science and applied mathematics. Academic Press, London, 2nd edition, 1984.
- [47] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C*. Cambridge University Press, London, 2nd edition, 2002.
- [48] 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.
- [49] 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.
- [50] 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.
- [51] 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.
- [52] A. Kurganov and G. Petrova. A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system. *Communications in Mathematical Sciences*, 5(1):133–160, 2007.
- [53] S.D. Conte and C. De Boor. *Elementary numerical analysis: An algorithmic approach*. International Series in Pure and Applied Mathematics. McGraw-Hill Inc., New York, 3rd edition, 1980.

- [54] 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.
- [55] B. Koren. A robust upwind discretization method for advection, diffusion and source term. In C.B. Vreugdenhil and B. Koren, editors, *Numerical Methods for Advection-Diffusion Problems*, chapter 5, pages 117–138. Vieweg, Braunschweig, 1993.
- [56] A.T. Ippen and G. Kulin. The shoaling and breaking of the solitary wave. *Coastal Engineering Proceedings*, 1(5):4, 1954.
- [57] N.J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.
- [58] S. Beji and J.A. Battjes. Experimental investigation of wave propagation over a bar. *Coastal Engineering*, 19(1):151–162, 1993.
- [59] D. Lannes. *The Water Waves Problem: Mathematical Analysis and Asymptotics*, volume 188 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, 2013.
- [60] 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.
- [61] 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.
- [62] V. Roeber. *Boussinesq-type mode for nearshore wave processes in fringing reef environment*. PhD thesis, Department of Ocean and Resource Engineering, University of Hawaii, Manoa, Honolulu, HI, U.S.A, 2010.
- [63] C.E. Synolakis. The runup of solitary waves. *Journal of Fluid Mechanics*, 185:523–545, 1987.
- [64] A. Bollermann, S. Noelle, and M. Lukáčová-Medvidová. Finite volume evolution Galerkin methods for the shallow water equations with dry beds. *Communications in Computational Physics*, 10(2):371–404, 2011.