

# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>                               | <b>1</b>  |
| 1.1      | Objectives of the Thesis . . . . .                | 1         |
| 1.2      | Original Contribution of the Thesis . . . . .     | 1         |
| 1.3      | Organisation of the Thesis . . . . .              | 1         |
| <b>2</b> | <b>The Serre Equations</b>                        | <b>3</b>  |
| 2.1      | The Equations . . . . .                           | 3         |
| 2.1.1    | Alternative form of the Serre Equations . . . . . | 5         |
| 2.2      | Properties of the Serre Equations . . . . .       | 6         |
| 2.2.1    | Conservation Properties . . . . .                 | 6         |
| 2.2.2    | Dispersion Properties . . . . .                   | 7         |
| 2.2.3    | Analytic Solutions . . . . .                      | 7         |
| 2.2.4    | Forced Solutions . . . . .                        | 11        |
| 2.2.5    | Asymptotic Results . . . . .                      | 12        |
| <b>3</b> | <b>Finite Element Volume Method</b>               | <b>13</b> |
| 3.1      | Notation for Numerical Grids . . . . .            | 13        |
| 3.2      | Structure Overview . . . . .                      | 14        |
| (i)      | Reconstruction . . . . .                          | 16        |
| (ii)     | Fluid Velocity . . . . .                          | 18        |
| (iii)    | Flux Across the Cell Interfaces . . . . .         | 23        |
| (iv)     | Source Terms . . . . .                            | 27        |
| (v)      | Update Cell Averages . . . . .                    | 28        |
| (vi)     | Second-Order SSP Runge-Kutta Method . . . . .     | 28        |
| 3.3      | CFL condition . . . . .                           | 29        |
| 3.4      | Boundary Conditions . . . . .                     | 29        |
| 3.5      | Dry Beds . . . . .                                | 30        |

|          |   |           |
|----------|---|-----------|
| <b>4</b> | <b>Linear Analysis of the Numerical Methods</b>                                       | <b>33</b> |
| 4.1      | Linearised Serre equations with horizontal bed . . . . .                              | 34        |
| 4.2      | Evolution Matrix . . . . .  | 35        |
| 4.2.1    | Overview of the Evolution Step . . . . .  | 36        |
| (i)      | Reconstruct the Quantities Inside the Cells . . . . .                                 | 37        |
| (ii)     | Calculate the Velocity Over the Domain . . . . .                                      | 38        |
| (iii)    | Calculate All the Fluxes Across the Cell Interfaces . . . . .                         | 40        |
| (iv)     | Calculate All the Source Terms for the Cells . . . . .                                | 43        |
| (v)      | Update All the Cell Averages Using a Forward Euler Ap-<br>proximation . . . . .       | 43        |
| (vi)     | Update All the Cell Averages Using a Second-Order SSP<br>Runge-Kutta Method . . . . . | 44        |
| 4.3      | Convergence Analysis . . . . .  | 45        |
| 4.3.1    | Stability . . . . .   | 45        |
| 4.3.2    | Consistency . . . . .   | 47        |
| 4.4      | Dispersion Analysis . . . . .   | 52        |
| <b>5</b> | <b>Numerical Validation</b>   | <b>61</b> |
| 5.1      | Measuring Convergence and Conservation . . . . .                                      | 61        |
| 5.1.1    | Measures of Convergence . . . . .   | 61        |
| 5.1.2    | Measures of Conservation . . . . .  | 62        |
| 5.2      | Analytic Solutions . . . . .  | 63        |
| 5.2.1    | Solitary Cnoidal Wave Solution . . . . .  | 63        |
| 5.2.2    | Lake at Rest . . . . .  | 69        |
| 5.3      | Forced Solution . . . . .   | 69        |
| 5.3.1    | Travelling Gaussian . . . . .   | 69        |
| <b>6</b> | <b>Experimental Validation</b>  | <b>73</b> |
| 6.1      | Periodic Waves Over A Submerged Bar . . . . .   | 73        |
| 6.1.1    | Low Frequency Results . . . . .   | 74        |
| 6.1.2    | High Frequency Results . . . . .  | 75        |
| 6.2      | Synolakis . . . . .   | 80        |
| 6.3      | Roeber . . . . .  | 80        |
| <b>A</b> | <b>Linear Analysis Results</b>  | <b>91</b> |
|          | <b>Bibliography</b>   | <b>96</b> |

# Chapter 1

## Introduction

1.1 Objectives of the Thesis

1.2 Original Contribution of the Thesis

1.3 Organisation of the Thesis



# Chapter 2

## The Serre Equations

### 2.1 The Equations

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

From the depth-integration approach the Serre equations describe a free surface defined by its height  $h(x, t)$  above a stationary bed profile  $b(x)$  and a depth average of its horizontal velocity  $u(x, t)$  as in Figure 2.1. This is the same as the derivation of the Shallow Water Wave Equations (SWWE) [], except for the Serre equations we allow the vertical velocity to vary linearly with depth and so we get non-hydrostatic pressure terms and therefore dispersion.

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

$$\frac{\partial h}{\partial t} + \frac{\partial(uh)}{\partial x} = 0, \quad (2.1a)$$

$$\frac{\partial(uh)}{\partial t} + \frac{\partial}{\partial x} \left( u^2 h + \frac{gh^2}{2} + \frac{h^2}{2} \Psi + \frac{h^3}{3} \Phi \right) + \frac{\partial b}{\partial x} \left( gh + h\Psi + \frac{h^2}{2} \Phi \right) = 0. \quad (2.1b)$$

**Definition 2.1.** The  $\Phi$  and  $\Psi$  terms account for the non-hydrostatic part of the

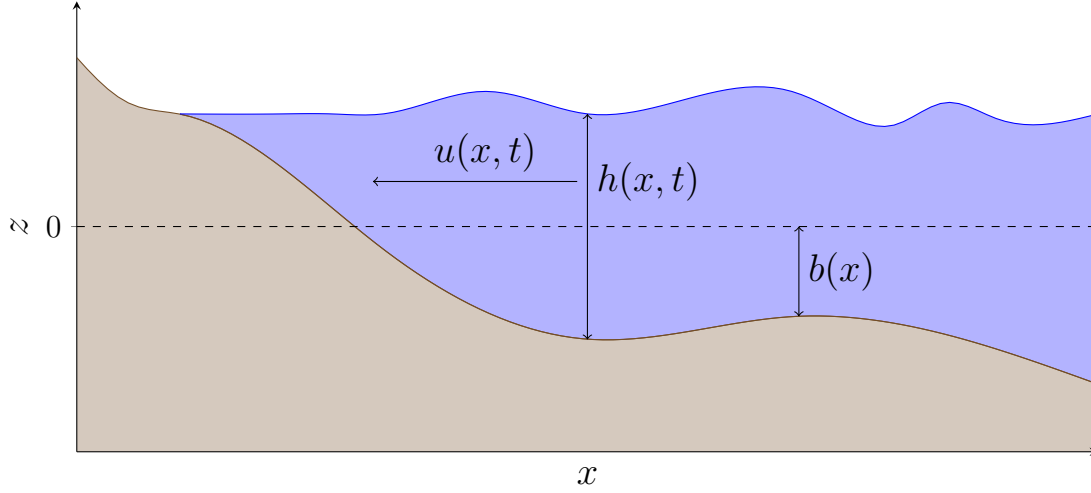


Figure 2.1: Diagram demonstrating the quantities used to describe the fluid (■) and the bed (■) for the Serre equations.

pressure and are

$$\Psi = \frac{\partial b}{\partial x} \left( \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} \right) + u^2 \frac{\partial b}{\partial x}, \quad (2.2a)$$

$$\Phi = \frac{\partial u}{\partial x} \frac{\partial u}{\partial x} - u \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial x \partial t}. \quad (2.2b)$$

When  $\Phi = \Psi = 0$  the Serre equations are equivalent to the SWWE where the vertical velocity is constant in depth, only the hydrostatic pressure is present and there is no dispersion of waves. Due to the presence of these  $\Phi$  and  $\Psi$  terms the Serre equations are much more difficult to solve analytically and numerically than the SWWE. The primary reason for this is that whilst the SWWE are hyperbolic for finite water depth, the Serre equations are neither hyperbolic nor parabolic. Furthermore the Serre equations are not in conservation law form due to the presence of temporal derivatives in  $\Phi$  and  $\Psi$ , although they are derived from conservation equations.

For a horizontal bed  $\partial b / \partial x = 0$ , thus  $\Psi = 0$  and all the source terms drop out and the Serre equations become

$$\frac{\partial h}{\partial t} + \frac{\partial(uh)}{\partial x} = 0, \quad (2.3a)$$

$$\frac{\partial(uh)}{\partial t} + \frac{\partial}{\partial x} \left( u^2 h + \frac{gh^2}{2} + \frac{h^3}{3} \Phi \right) = 0. \quad (2.3b)$$

These equations are still neither hyperbolic nor parabolic and are still not in conservation law form as  $\Phi$  contains a temporal derivative. As such even for

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

### 2.1.1 Alternative form of the Serre Equations

A major hurdle for developing numerical methods for the Serre equations is the presence of the mixed temporal and spatial derivative in  $\Phi$  (2.2b). By rewriting the Serre equations and introducing a new conserved quantity  $G$  [6, 5, 7], the mixed temporal and spatial derivative can be removed and the Serre equations can be written in conservation law form.

**Definition 2.2.** The conserved quantity  $G$  is

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

The Serre equations (2.1) are then rewritten as conservation laws for the conserved variables  $h$  and  $G$

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

$$\begin{aligned} \frac{\partial G}{\partial t} + \frac{\partial}{\partial x} \left( uG + \frac{gh^2}{2} - \frac{2}{3} h^3 \frac{\partial u^2}{\partial x} + h^2 u \frac{\partial u}{\partial x} \frac{\partial b}{\partial x} \right) \\ + \frac{1}{2} h^2 u \frac{\partial u}{\partial x} \frac{\partial^2 b}{\partial x^2} - hu^2 \frac{\partial b}{\partial x} \frac{\partial^2 b}{\partial x^2} + gh \frac{\partial b}{\partial x} = 0. \end{aligned} \quad (2.4b)$$

The conserved quantity  $G$  resembles  $h$  multiplied by the irrotationality [8, 9] and its conservation equation is equivalent to the conservation equation for momentum (2.1b).

This conservation law form makes the Serre equations well suited for a finite volume method for the conservation of mass and  $G$  equations, provided one can solve for  $u$  given  $h$  and  $G$ . Although this does solve the problem of it not being in conservation law, these equations are still considerably more difficult to solve than the SWWE.

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

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

$$\frac{\partial G}{\partial t} + \frac{\partial}{\partial x} \left( uG + \frac{gh^2}{2} - \frac{2}{3} h^3 \frac{\partial u^2}{\partial x} \right) = 0, \quad (2.5b)$$

$$G = hu - \frac{\partial}{\partial x} \left( \frac{1}{3} h^3 \frac{\partial u}{\partial x} \right). \quad (2.5c)$$

These equations are much simpler to solve than their counterparts that allow for variations in bathymetry, although they still present a number of challenges.

## 2.2 Properties of the Serre Equations

The Serre equations are significantly more complex than their dispersionless counterparts the SWWE and so fewer analytic solutions can be found in the literature. This increases the need to utilise the known properties of the Serre equations and construct forced solutions to assess our numerical methods. The relevant properties of the Serre equations are thus presented here, beginning with their conservation properties.

### 2.2.1 Conservation Properties

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

**Definition 2.3.** The total amount of a quantity  $q$  in a system occurring on the interval  $[a, b]$  at time  $t$  is

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

Conservation of a quantity  $q$  therefore implies that  $\mathcal{C}_q(0) = \mathcal{C}_q(t) \forall t$ .

Given that the Serre equations (2.1) are conservation equations for mass and momentum and that the conservation of momentum equation can be rewritten as a conservation equation for  $G$  (2.4), the Serre equations conserve these quantities. Additionally owing to the Hamiltonian structure of the Serre equations, the Serre equations possess a Hamiltonian  $\mathcal{H}$  which acts as an energy and is therefore also conserved.

**Definition 2.4.** The Hamiltonian of the Serre equations is

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

For the system to be closed for the Serre equations the flux terms of the conservation of mass and momentum equations at both boundaries must with cancel each other and the integral of the source terms over the domain must be zero.



### 2.2.2 Dispersion Properties

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

$$\omega = Uk \pm k\sqrt{gH} \sqrt{\frac{3}{(kH)^2 + 3}}. \quad (2.6)$$

Barthélemy [10] compared this dispersion relation to that of the linear theory of water waves and demonstrated its utility when  $k$  is large. However when  $k$  is small the difference between the dispersion relation of the Serre equations and that of water wave theory increases, leading to poorer reproduction of experimental observations by the Serre equations. The dispersion relation of the Serre equations can be modified by introducing terms to reduce this difference [10], but such modifications are beyond the scope of this thesis.

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

$$v_p = U \pm \sqrt{gH} \sqrt{\frac{3}{(kH)^2 + 3}}, \quad (2.7a)$$

$$v_g = U \pm \sqrt{gH} \left( \sqrt{\frac{3}{(kH)^2 + 3}} \mp (kH)^2 \sqrt{\frac{3}{((kH)^2 + 3)^3}} \right). \quad (2.7b)$$

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

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

$$U - \sqrt{gH} \leq v_p \leq U + \sqrt{gH}, \quad (2.8a)$$

$$U - \sqrt{gH} \leq v_g \leq U + \sqrt{gH}. \quad (2.8b)$$

### 2.2.3 Analytic Solutions

Due to the complexity of the Serre equations, few analytic solutions have been discovered. In particular there is a travelling wave solution for horizontal beds

and a lake at rest solution for any bathymetry. Both analytic solutions will be used to assess the capabilities of the numerical methods and so we present them here.

### Solitary Travelling Wave Solution

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

$$h(x, t) = a_0 + a_1 \operatorname{sech}(\kappa(x - ct)), \quad (2.9a)$$

$$u(x, t) = c \left( 1 - \frac{a_0}{h(x, t)} \right), \quad (2.9b)$$

$$b(x) = 0 \quad (2.9c)$$

with

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

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

This wave has an amplitude of  $a_1$ , an infinite wavelength and propagates on water  $a_0$  deep. These solitary waves are not true solitons however, due to their inelastic collisions with one another [11]. These solitary waves can also be generalised to a family of periodic solutions [12] which are all periodic travelling wave solutions.

From (2.9) we can derive  $G$  for the solitary wave solution in terms of  $h$

$$G = c(h - a_0) - \frac{ca_0}{3} \left( h \frac{\partial^2 h}{\partial x^2} + \left[ \frac{\partial h}{\partial x} \right] \right) \quad (2.10)$$

where the derivatives of  $h$  are

$$\begin{aligned} \frac{\partial h}{\partial x} &= -2a_1\kappa \operatorname{sech}^2(\kappa(x - ct)) \tanh(\kappa(x - ct)), \\ \frac{\partial^2 h}{\partial x^2} &= 2a_1\kappa^2 (\cosh(2\kappa(x - ct)) - 2) \operatorname{sech}^4(\kappa(x - ct)). \end{aligned}$$

As we would like to assess the conservation properties of our numerical methods we will require the total mass, momentum,  $G$  and Hamiltonian for this solution. In particular we require these totals at the initial time  $t = 0$ , to allow for

various domains we present the integrals in indefinite form

$$\int h(x, 0) dx = a_0 x + \frac{a_1}{\kappa} \tanh(\kappa x) + \text{constant}, \quad (2.11a)$$

$$\int u(x, 0) h(x, 0) dx = \frac{a_1 c}{\kappa} \tanh(\kappa x) + \text{constant}, \quad (2.11b)$$

$$\begin{aligned} \int G(x, 0) dx = \frac{ca_1}{3\kappa} & \left( 3 + 2a_0^2 \kappa^2 \text{sech}^2(\kappa x) \right. \\ & \left. + 2a_0 a_1 \kappa^2 \text{sech}^4(\kappa x) \right) \tanh(\kappa x) + \text{constant}, \end{aligned} \quad (2.11c)$$

$$\begin{aligned} \int \mathcal{H}(x, 0) dx = \frac{1}{2} & \left( \int g[h(x, 0)]^2 dx + \int h(x, 0) [u(x, 0)]^2 dx \right. \\ & \left. + \int [h(x, 0)]^3 \left[ \frac{\partial u(x, 0)}{\partial x} \right]^2 dx \right) \end{aligned} \quad (2.11d)$$

where the integrals of the Hamiltonian are

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

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

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

Therefore, we have the analytic values of our physical variables  $h$ ,  $u$ ,  $G$  and  $b$  for all  $t$  as well as the total amounts of our conserved quantities for the initial conditions when  $t = 0s$  as desired.

### Lake at Rest

The lake at rest solution is a rudimentary stationary solution of the Serre equations that exists for all bathymetry  $b(x)$  and is defined by

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

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

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

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. The maximum function is included for the water depth to allow for dry regions of the bed when  $b(x) > a_0$ . We write these solutions in terms of  $b(x)$  as this solution holds for all bed profiles.

For these quantities (2.12) the Serre equations (2.4) reduce to

$$\begin{aligned} \frac{\partial h}{\partial t} &= 0, \\ \frac{\partial G}{\partial t} + \frac{\partial}{\partial x} \left( \frac{gh^2}{2} \right) + gh \frac{\partial b}{\partial x} &= 0. \end{aligned}$$

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

The total momentum and  $G$  in our system is straightforward to calculate as both are zero everywhere and so we have

$$\int u(x, 0)h(x, 0) dx = 0 + \text{constant}, \quad (2.13)$$

$$\int G(x, 0) dx = 0 + \text{constant}. \quad (2.14)$$

To calculate the total mass and Hamiltonian in our system we must break up our domain into wet regions where  $b(x) < a_0$  and dry regions where  $b(x) \geq a_0$ . For the dry regions the total mass and energy are 0 and so we have

$$\int h(x, 0) dx = 0, \quad (2.15a)$$

$$\int \mathcal{H}(x, 0) dx = 0 \quad (2.15b)$$

whilst in a wet region we have

$$\int h(x, 0) dx = a_0 x - \int b(x) dx, \quad (2.16a)$$

$$\int \mathcal{H}(x, 0) dx = \frac{g}{2} \left( a_0^2 x - 2a_0 \int b(x) dx + \int b(x)^2 dx \right). \quad (2.16b)$$

Therefore as desired we have expressions for all the quantities in terms of the bed profile  $b(x)$ .

### 2.2.4 Forced Solutions

To account for the small number of analytic solutions to the Serre equations we also constructed forced solutions to assess our numerical methods. These work by introducing a source term  $S$  for both the conservation of mass and momentum (or  $G$ ) equations that balances these equation to force a solution. So for instance we add the source terms  $S_{\text{mass}}$  and  $S_G$  into (2.4) so that we have

$$\frac{\partial h}{\partial t} + \frac{\partial(uh)}{\partial x} + S_{\text{mass}} = 0, \quad (2.17a)$$

$$\begin{aligned} \frac{\partial G}{\partial t} + \frac{\partial}{\partial x} \left( uG + \frac{gh^2}{2} - \frac{2}{3}h^3 \frac{\partial u^2}{\partial x} + h^2 u \frac{\partial u}{\partial x} \frac{\partial b}{\partial x} \right) \\ + \frac{1}{2}h^2 u \frac{\partial u}{\partial x} \frac{\partial^2 b}{\partial x^2} - hu^2 \frac{\partial b}{\partial x} \frac{\partial^2 b}{\partial x^2} + gh \frac{\partial b}{\partial x} + S_G = 0. \end{aligned} \quad (2.17b)$$

as our new forced Serre equations in conservation law form. To construct the forced solution we choose some functions for  $h$ ,  $u$  and  $b$  for all  $x$  and  $t$ , which will be our forced analytic solutions to the forced Serre equations. From these chosen functions we determine  $G$ ,  $S_{\text{mass}}$  and  $S_G$  so that these equations (2.17) possess our chosen  $h$ ,  $u$  and  $b$  as solutions.

We demonstrate this for the particular forced function we employ, a travelling Gaussian wave over a periodic bed profile which has

$$h(x, t) = a_0 + a_1 \exp \left( -\frac{((x - a_2 t) - a_3)^2}{2a_4} \right), \quad (2.18a)$$

$$u(x, t) = a_5 \exp \left( -\frac{((x - a_2 t) - a_3)^2}{2a_4} \right), \quad (2.18b)$$

$$b(x) = a_6 \sin(a_7 x). \quad (2.18c)$$

From these definitions  $G$  can be calculated using Def 2.2 while  $S_{\text{mass}}$  and  $S_G$  can be calculated using

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

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

with  $h$ ,  $u$  and  $b$  from (2.18) and the corresponding  $G$ .

With these values the forced Serre equations (2.17) admit the analytic solutions (2.18). This allows us to assess our numerical solutions for a larger range of scenarios than possible with the current analytic solutions of the Serre equations.

### 2.2.5 Asymptotic Results

Beyond analytic solutions to the Serre equations there has also been studies of the long term behaviour of the Serre equations for situations that are difficult to treat analytically. One particular scenario of interest is the evolution of a moving discontinuous jump in  $h$  known as a bore, which can be observed naturally, for example tidal bores [1]. There is an analytic solution for bores for the SWWE, but not for the Serre equations.

To gain some insight into the behaviour Whitham modulation techniques were applied to the Serre equations to study the behaviour of bores of the Serre equations as  $t \rightarrow \infty$  [12]. These techniques provided an estimate of the speed  $S^+$  and amplitude  $A^+$  of the front of a bore

$$\frac{\Delta}{(A^+ + 1)^{1/4}} - \left( \frac{3}{4 - \sqrt{A^+ + 1}} \right)^{21/10} \left( \frac{2}{1 + \sqrt{A^+ + 1}} \right)^{2/5} = 0 \quad (2.19a)$$

$$S^+ = \sqrt{g(A^+ + 1)} \quad (2.19b)$$

where  $\Delta = h_b/h_0$ ,  $h_b$  is the height of the bore and  $h_0$  is the depth of still water in front of the bore. These estimates were demonstrated to agree well with numerical simulations provided that  $\Delta < 1.43$  [12].

# Chapter 3

## Finite Element Volume Method

### 3.1 Notation for Numerical Grids

We begin by defining the numerical grid in both space and time from a starting location  $x_0$  and a beginning time  $t^0$ . For any  $j, n \in \mathbb{N}$  we have

$$x_j = x_0 + j\Delta x, \quad (3.1a)$$

$$t^n = t^0 + n\Delta t \quad (3.1b)$$

which produces a uniform grid;  $x_j$  in space and  $t^n$  in time. The Finite Element Volume Method (FEVM) can be readily adapted to nonuniform grids and we restrict our description of the method to uniform grids for simplicity. We will extend this notation to describe locations and times not on the grid by allowing (3.1) to also have subscripts and superscripts in  $\mathbb{R}$ .

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

$$q_j^n = q(x_j, t^n). \quad (3.2)$$

The description of our numerical method focuses on cells which are regions surrounding the grid points. The  $j^{th}$  cell is the region  $[x_{j-1/2}, x_{j+1/2}]$  centred around  $x_j$ . For each cell we define the average of a quantity

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

in cell  $j$ .

In the FEVM we reconstruct quantities at various points inside the cell from the cell average values. We distinguish between two reconstructions that are possible at each cell edge, which both exist due to the overlap of neighbouring

cells. To do this we use superscripts so that for the cell edge  $x_{j+1/2}$  and a general quantity  $q$ , we have  $q_{j+1/2}^-$  as the reconstructed value of  $q$  at  $x_{j+1/2}$  from the  $j^{th}$  cell and  $q_{j+1/2}^+$  as the reconstructed value of  $q$  at  $x_{j+1/2}$  from the  $(j+1)^{th}$  cell. Since the particular time level is typically obvious from context and hence omitted the use of a superscript will not clutter the notation.

## 3.2 Structure Overview

To describe the FEVM we first present an overview of the evolution step and then provide the details for each component. We begin our evolution step with all the cell averages for  $h$ ,  $w$  and  $G$  at time  $t^n$  and all the nodal values of  $b$ . We write these as vectors from the  $0^{th}$  cell to the  $m^{th}$  in the following way

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

The evolution step proceeds as follows:

- (i) Reconstruction: We reconstruct the quantities  $h$ ,  $w$ ,  $G$  and  $b$  inside every cell at various points. The values of  $h$ ,  $w$  and  $G$  in the  $j^{th}$  cell are reconstructed at  $x_{j-1/2}$ ,  $x_j$  and  $x_{j+1/2}$  using the reconstruction operators  $\mathcal{R}_{j-1/2}^+$ ,  $\mathcal{R}_j$  and  $\mathcal{R}_{j+1/2}^-$  respectively. 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  $x_{j+1/2}$  using the reconstruction operators  $\mathcal{B}_{j-1/2}$ ,  $\mathcal{B}_{j-1/6}$ ,  $\mathcal{B}_{j+1/6}$  and  $\mathcal{B}_{j+1/2}$  respectively. So that

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



This generates the vectors of these quantities reconstructed for every cell;  $\hat{\mathbf{h}}$ ,  $\hat{\mathbf{w}}$ ,  $\hat{\mathbf{G}}$  and  $\hat{\mathbf{b}}$  which are

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

- (ii) Fluid Velocity: The remaining unknown quantity, the depth averaged fluid velocity,  $u$  is calculated by solving the elliptic equation (2.2) with a FEM from which we obtain  $u_{j-1/2}$ ,  $u_j$  and  $u_{j+1/2}$  at every cell. We denote the solution of the FEM by the map  $\mathcal{G}$  given  $\hat{\mathbf{h}}$ ,  $\hat{\mathbf{G}}$  and  $\hat{\mathbf{b}}$  as inputs. So that

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

- (iii) Flux Across Cell Interfaces: We calculate the average fluxes  $F_{j-1/2}$  and  $F_{j+1/2}$  across the cell boundaries  $x_{j-1/2}$  and  $x_{j+1/2}$  over time using  $\mathcal{F}_{j-1/2}$  and  $\mathcal{F}_{j+1/2}$  so that

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

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

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

- (v) Update All the Cell Averages Using a Forward Euler Approximation: We update the cell average values from time  $t^n$  to the next time level combining a fractional step method and a forward Euler approximation for the flux and source terms.

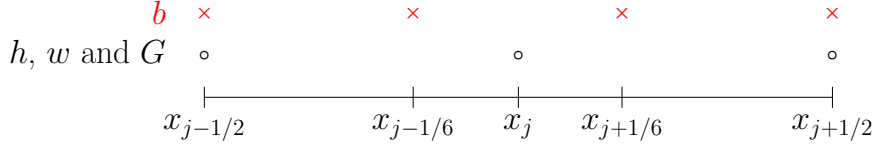


Figure 3.1: Location of the reconstructions for  $h$ ,  $w$ ,  $G$  and  $b$ .

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

### (i) Reconstruction

We now provide the details for how each of the quantities inside the  $j^{th}$  cell are calculated at the locations demonstrated in Figure 3.1. For  $h$ ,  $w$  and  $G$  the reconstructions is performed from the cell averages. While  $b$  is reconstructed from the nodal values.

#### The height, stage and $G$

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

$$q_{j-1/2}^+ = \mathcal{R}_{j-1/2}^+(\bar{q}) = \bar{q} - \frac{\Delta x}{2} d_j, \quad (3.4a)$$

$$q_j = \mathcal{R}_j(\bar{q}) = \bar{q}, \quad (3.4b)$$

$$q_{j+1/2}^- = \mathcal{R}_{j+1/2}^-(\bar{q}) = \bar{q} + \frac{\Delta x}{2} d_j \quad (3.4c)$$

where

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

with  $\theta \in [1, 2]$ . The choice of the  $\theta$  parameter changes the diffusion introduced by the reconstruction, with  $\theta = 1$  the reconstruction introduces the most diffusion and is equivalent to the basic minmod reconstruction [13] While for  $\theta = 2$  the reconstruction is equivalent to the monotized central reconstruction 14 and introduces the least diffusion.

**Definition 3.1.** The minmod function takes a list of  $x_j \in \mathbb{R}$ . If all elements have the same sign then minmod returns the element with smallest absolute value, otherwise it returns 0.

$$\text{minmod}(x_0, x_1, \dots) := \begin{cases} \min \{x_j\} & x_j > 0 \ \forall j \\ \max \{x_j\} & x_j < 0 \ \forall j \\ 0 & \text{otherwise} \end{cases}.$$

The nonlinear limiting used to calculate  $d_j$  ensures that the reconstruction of  $h$ ,  $w$  and  $G$  inside the cell does not introduce non-physical oscillations and is thus Total Variation Diminishing (TVD). It does this by constraining the slope  $d_j$  to zero near local extrema, resulting in a first-order approximation, which is always 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 method for which we get that

$$\bar{q} = \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} q \, dx = q_j + \mathcal{O}(\Delta x^2). \quad (3.5)$$

### The bed profile

To reconstruct  $b$  at  $x_{j-1/2}$ ,  $x_{j-1/6}$ ,  $x_{j+1/6}$  and  $x_{j+1/2}$  we construct the cubic polynomial  $C_j(x)$  centred around  $x_j$

$$C_j(x) = c_0 (x - x_j)^3 + c_1 (x - x_j)^2 + c_2 (x - x_j) + c_3. \quad (3.6)$$

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

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

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

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

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

We require a continuous bed profile across the cell edges and so on top of the reconstruction of the bed given by these coefficients and (3.6) we also average between the two reconstructions at the cell edge so that our reconstruction of the bed profile in the  $j^{th}$  cell is

$$b_{j-1/2} = \mathcal{B}_{j-1/2}(\mathbf{b}) = \frac{1}{2} (C_j(x_{j-1/2}) + C_{j-1}(x_{j-1/2})), \quad (3.7a)$$

$$b_{j-1/6} = \mathcal{B}_{j-1/6}(\mathbf{b}) = C_j(x_{j-1/6}), \quad (3.7b)$$

$$b_{j+1/6} = \mathcal{B}_{j+1/6}(\mathbf{b}) = C_j(x_{j+1/6}), \quad (3.7c)$$

$$b_{j+1/2} = \mathcal{B}_{j+1/2}(\mathbf{b}) = \frac{1}{2} (C_j(x_{j+1/2}) + C_{j+1}(x_{j+1/2})). \quad (3.7d)$$

## (ii) Fluid Velocity

The elliptic equation that relates the conserved variables  $h$  and  $G$  and the bed profile  $b$  to the primitive variable  $u$  was given in Def 2.2. To form the FEM we take the weak form of Def 2.2 which is

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

Integrating by parts with Dirichlet boundary conditions we get

$$\begin{aligned} \int_{\Omega} Gv \, dx = \int_{\Omega} uh \left( 1 + \frac{\partial b^2}{\partial x} \right) v \, dx + \int_{\Omega} \frac{1}{3} h^3 \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} \, dx \\ - \int_{\Omega} \frac{1}{2} h^2 \frac{\partial b}{\partial x} u \frac{\partial v}{\partial x} \, dx - \int_{\Omega} \frac{1}{2} h^2 \frac{\partial b}{\partial x} \frac{\partial u}{\partial x} v \, dx. \end{aligned} \quad (3.8)$$

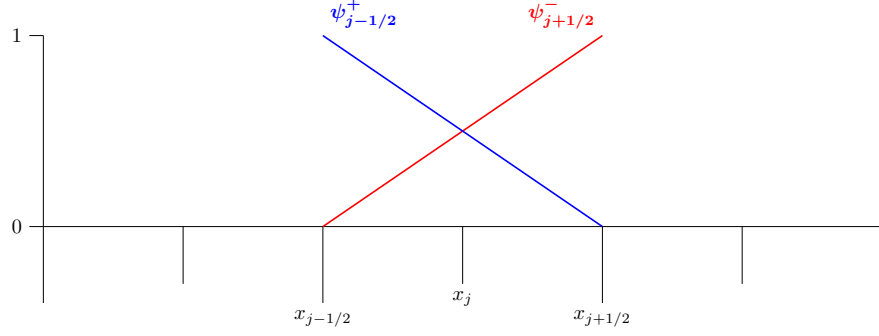


Figure 3.2: Discontinuous linear basis functions over a cell.

This weak formulation implies that if  $G$  and  $h$  are members of the function space  $\mathbb{L}^p$  and  $b$  is a member of the Sobolev space  $\mathbb{W}^{1,p}$  then  $u$  is a member of the Sobolev space  $\mathbb{W}^{1,p}$  as well [15]. Since our method requires the derivative of  $u$  to be well defined and hence that  $u \in \mathbb{W}^{1,p}$  we will restrict ourselves to only allow  $h, G \in \mathbb{L}^p$  and  $b \in \mathbb{W}^{1,p}$ .

We simplify (3.8) by performing the integration over the cells and then summing the integrals together to get the equation for the entire domain

$$\sum_j \int_{x_{j-1/2}}^{x_{j+1/2}} \left[ \left( uh \left( 1 + \frac{\partial b^2}{\partial x} \right) - \frac{1}{2} h^2 \frac{\partial b}{\partial x} \frac{\partial u}{\partial x} - G \right) v + \left( \frac{1}{3} h^3 \frac{\partial u}{\partial x} - \frac{1}{2} h^2 \frac{\partial b}{\partial x} u \right) \frac{\partial v}{\partial x} \right] dx = 0 \quad (3.9)$$

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

### Basis Function Approximations

For  $h$  and  $G$  we use the basis functions  $\psi$  which are linear inside a cell and 0 everywhere outside the cell as shown in Figure 3.2. This produces a second-order approximation to  $h$  and  $G$  inside the cell and allows for discontinuities at the cell edges. Since these basis functions are in  $\mathbb{L}^p$  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  $G$

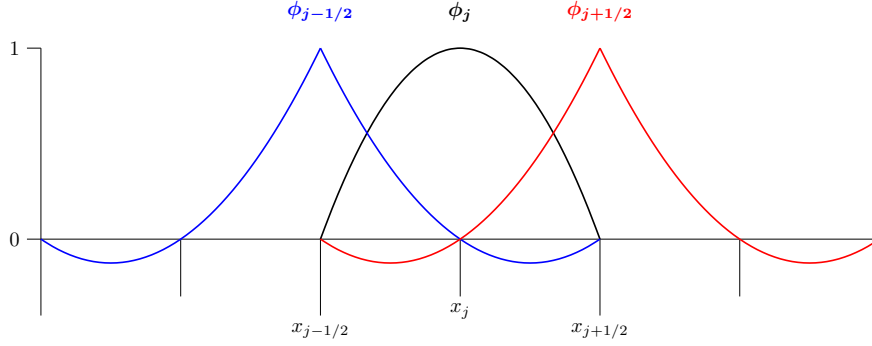


Figure 3.3: Continuous piecewise quadratic basis functions over a cell.

in our finite element method

$$h = \sum_j h_{j-1/2}^+ \psi_{j-1/2}^+ + h_{j+1/2}^- \psi_{j+1/2}^-, \quad (3.10a)$$

$$G = \sum_j G_{j-1/2}^+ \psi_{j-1/2}^+ + G_{j+1/2}^- \psi_{j+1/2}^-. \quad (3.10b)$$

For  $u$  we require a locally calculated second-order approximation to the first derivative, which forces at least a quadratic representation of  $u$  in each cell. Since we require  $u$  to be a member of  $\mathbb{W}^{1,p}$ ,  $u$  will be continuous. Therefore, we use basis functions  $\phi$  that are continuous at the cell edges  $x_{j\pm 1/2}$ , resulting in the basis functions depicted in Figure 3.3.

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

$$u = \sum_j u_{j-1/2} \phi_{j-1/2} + u_j \phi_j + u_{j+1/2} \phi_{j+1/2}. \quad (3.11)$$

For the bed term we require a local approximation to the second derivative of the bed that is second-order accurate. To allow for an appropriate second derivative of the bed profile,  $b$  must be a member of  $\mathbb{W}^{2,p}$  which is smoother than indicated by the elliptic equation (3.8). We choose the cubic basis functions  $\gamma$  which are continuous across the cell edges, as the bed profile will be continuous. These basis functions are shown in Figure 3.4 and from them we get our basis function approximation to  $b$

$$b = \sum_j b_{j-1/2} \gamma_{j-1/2} + b_{j-1/6} \gamma_{j-1/6} + b_{j+1/6} \gamma_{j+1/6} + b_{j+1/2} \gamma_{j+1/2}. \quad (3.12)$$

With  $b_{j-1/2}$  and  $b_{j+1/2}$  being well defined as our method (3.7) forces both reconstructions at the cell edges to be the identical for adjacent cells.

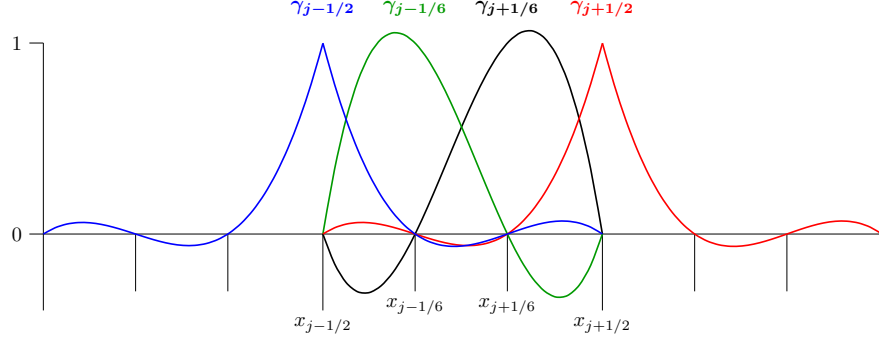


Figure 3.4: Continuous piecewise cubic basis functions over a cell.

### Calculation of Element-wise Matrices

The integral equation (3.9) holds for all  $v$ . However, since our solution space has the basis functions  $\phi$  it is sufficient to satisfy (3.9) 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 (3.9) like so

$$\int_{x_{j-1/2}}^{x_{j+1/2}} \left[ \left( uh \left( 1 + \frac{\partial b^2}{\partial x} \right) - \frac{1}{2} h^2 \frac{\partial b}{\partial x} \frac{\partial u}{\partial x} - G \right) \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} + \left( \frac{1}{3} h^3 \frac{\partial u}{\partial x} - \frac{1}{2} h^2 \frac{\partial b}{\partial x} u \right) \frac{\partial}{\partial x} \left( \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} \right) \right] dx \quad (3.13)$$

where we use our finite element approximations for  $h$  (3.10a),  $G$  (3.10b),  $u$  (3.11) and  $b$  (3.12). Because the basis functions over an element are just translations of the other basis functions, this integral can be generalised by moving to the  $\xi$  space. The mapping from the  $x$  space to the  $\xi$  space is

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

Making the change of variables from  $x$  to  $\xi$  in (3.13) we get

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

We will demonstrate the rest of the process for the  $uh$  term as an example and provide the remaining integrals in the Appendix[]. The  $uh$  term is

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

Since the integral is computed over  $[x_{j-1/2}, x_{j+1/2}]$ , there are only a few non-zero contributions from the finite element approximations to  $h$  and  $u$ , so we have

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

Calculating the integrals of all the basis function combinations we get

$$\begin{aligned} & \frac{\Delta x}{2} \int_{-1}^1 uh \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} d\xi = \\ & \frac{\Delta x}{2} \begin{bmatrix} \frac{7}{30} h_{j-1/2}^+ + \frac{1}{30} h_{j+1/2}^- & \frac{4}{30} h_{j-1/2}^+ & -\frac{1}{30} h_{j-1/2}^+ - \frac{1}{30} h_{j+1/2}^- \\ \frac{4}{30} h_{j-1/2}^+ & \frac{16}{30} h_{j-1/2}^+ + \frac{16}{30} h_{j+1/2}^- & \frac{4}{30} h_{j+1/2}^- \\ -\frac{1}{30} h_{j-1/2}^+ - \frac{1}{30} h_{j+1/2}^- & \frac{4}{30} h_{j+1/2}^- & \frac{1}{30} h_{j-1/2}^+ + \frac{7}{30} h_{j+1/2}^- \end{bmatrix} \\ & \quad \begin{bmatrix} u_{j-1/2} \\ u_j \\ u_{j+1/2} \end{bmatrix}. \quad (3.14) \end{aligned}$$



### Assembly of the Global Matrix

By combining all the matrices generated by the integral of each of the  $u$  terms we get the  $j^{th}$  cells contribution to the stiffness matrix  $\mathbf{A}_j$ . Likewise all the integrals of the remaining term  $Gv$  generate the vector  $\mathbf{g}_j$ . Therefore, (3.9) can be rewritten as

$$\sum_j \mathbf{A}_j \begin{bmatrix} u_{j-1/2} \\ u_j \\ u_{j+1/2} \end{bmatrix} = \sum_j \mathbf{g}_j. \quad (3.15)$$

This is a penta-diagonal matrix equation which can be solved by direct banded matrix solution techniques to obtain

$$\hat{\mathbf{u}} = \mathcal{G}(\hat{\mathbf{h}}, \hat{\mathbf{G}}, \hat{\mathbf{b}}) = \mathbf{A}^{-1} \mathbf{g} \quad (3.16)$$

as desired.

### (iii) Flux Across the Cell Interfaces

We use Kurganovs method [16] 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 an expression for the characteristics but do possess estimates on the maximum and minimum wave speeds (2.8a).

We will only demonstrate the calculation of the flux term for  $F_{j+1/2}$  as the process to calculate the flux term  $F_{j-1/2}$  is identical but with different cells. For a general quantity  $q$ , Kurganov's method [16] is

$$F_{j+1/2} = \frac{a_{j+1/2}^+ f(q_{j+1/2}^-) - a_{j+1/2}^- f(q_{j+1/2}^+)}{a_{j+1/2}^+ - a_{j+1/2}^-} + \frac{a_{j+1/2}^+ a_{j+1/2}^-}{a_{j+1/2}^+ - a_{j+1/2}^-} [q_{j+1/2}^+ - q_{j+1/2}^-] \quad (3.17)$$

where  $a_{j+1/2}^+$  and  $a_{j+1/2}^-$  are given by the wave speed bounds. For the Serre equations we have the wave speed bounds (2.8a) and so we obtain

$$a_{j+1/2}^- = \min \left\{ 0, u_{j+1/2}^- - \sqrt{gh_{j+1/2}^-}, u_{j+1/2}^+ - \sqrt{gh_{j+1/2}^+} \right\}, \quad (3.18)$$

$$a_{j+1/2}^+ = \max \left\{ 0, u_{j+1/2}^- + \sqrt{gh_{j+1/2}^-}, u_{j+1/2}^+ + \sqrt{gh_{j+1/2}^+} \right\}. \quad (3.19)$$

The flux functions  $f(q_{j+\frac{1}{2}}^-)$  and  $f(q_{j+\frac{1}{2}}^+)$  are evaluated using the reconstructed values from the  $j^{th}$  and  $(j+1)^{th}$  cell respectively. From the continuity equation (2.4a) we have

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

For the irrotationality equation (2.4b) we have

$$f\left(G_{j+\frac{1}{2}}^-\right) = u_{j+1/2}^- G_{j+1/2}^- + \frac{g}{2} \left(h_{j+1/2}^-\right)^2 - \frac{2}{3} \left(h_{j+1/2}^-\right)^3 \left[\left(\frac{\partial u}{\partial x}\right)_{j+1/2}^-\right]^2 \quad (3.20a)$$

$$+ \left(h_{j+1/2}^-\right)^2 u_{j+1/2}^- \left(\frac{\partial u}{\partial x}\right)_{j+1/2}^- \left(\frac{\partial b}{\partial x}\right)_{j+1/2}^-, \quad (3.20b)$$

$$f\left(G_{j+\frac{1}{2}}^+\right) = u_{j+1/2}^+ G_{j+1/2}^+ + \frac{g}{2} \left(h_{j+1/2}^+\right)^2 - \frac{2}{3} \left(h_{j+1/2}^+\right)^3 \left[\left(\frac{\partial u}{\partial x}\right)_{j+1/2}^+\right]^2 \quad (3.20c)$$

$$+ \left(h_{j+1/2}^+\right)^2 u_{j+1/2}^+ \left(\frac{\partial u}{\partial x}\right)_{j+1/2}^+ \left(\frac{\partial b}{\partial x}\right)_{j+1/2}^+. \quad (3.20d)$$

During the reconstruction process we calculated  $h_{j-1/2}^+$ ,  $h_{j+1/2}^-$ ,  $G_{j-1/2}^+$ ,  $G_{j+1/2}^-$  (3.4) while during the calculation of velocity we obtained  $u_{j+1/2}^\pm = u_{j+1/2}$ ; because  $u$  is continuous across the cell boundaries. Therefore, we further require an approximation to  $\left(\frac{\partial b}{\partial x}\right)_{j+1/2}^\pm$  and  $\left(\frac{\partial u}{\partial x}\right)_{j+1/2}^\pm$  to calculate the flux (3.20).

### Calculation of Derivatives

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

$$P_j^u(x) = p_0^u (x - x_j)^2 + p_1^u (x - x_j) + p_2^u, \quad (3.21a)$$

$$P_j^b(x) = p_0^b (x - x_j)^3 + p_1^b (x - x_j)^2 + p_2^b (x - x_j) + p_3^b, \quad (3.21b)$$

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

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

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

$$\begin{aligned} p_0^b &= \frac{-9b_{j-1/2} + 27b_{j-1/6} - 27b_{j+1/6} + 9b_{j+1/2}}{2\Delta x^3}, \\ p_1^b &= \frac{9b_{j-1/2} - 9b_{j-1/6} - 9b_{j+1/6} + 9b_{j+1/2}}{4\Delta x^2}, \\ p_2^b &= \frac{b_{j-1/2} - 27b_{j-1/6} + 27b_{j+1/6} - b_{j+1/2}}{8\Delta x}, \\ p_3^b &= \frac{-b_{j-1/2} + 9b_{j-1/6} + 9b_{j+1/6} - b_{j+1/2}}{16}. \end{aligned}$$

Taking the derivative of the polynomials (3.21) we get

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

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

$$\left( \frac{\partial u}{\partial x} \right)_{j+1/2}^- = \frac{\partial}{\partial x} P_j^u(x_{j+1/2}), \quad (3.22a)$$

$$\left( \frac{\partial u}{\partial x} \right)_{j+1/2}^+ = \frac{\partial}{\partial x} P_{j+1}^u(x_{j+1/2}), \quad (3.22b)$$

$$\left( \frac{\partial b}{\partial x} \right)_{j+1/2}^- = \frac{\partial}{\partial x} P_j^b(x_{j+1/2}), \quad (3.22c)$$

$$\left( \frac{\partial b}{\partial x} \right)_{j+1/2}^+ = \frac{\partial}{\partial x} P_{j+1}^b(x_{j+1/2}). \quad (3.22d)$$

Therefore, we possess all the terms need to calculate the approximation to the flux (3.17) for both the continuity and irrotationality equation, as desired. However, we must make a slight modification to ensure that our scheme is well balanced and recovers the lake at rest steady state solution.

### Well Balancing

To recover the lake at rest steady state solution we follow the work of Audusse et al. [17], who accomplished this for the Shallow Water Wave equations. It was demonstrated that this process could also be extended to the Serre equations [18]. To enforce well balancing  $h$  is modified at the cell edges in the following way; we first calculate

$$\dot{b}_{j+1/2}^- = w_{j+1/2}^- - h_{j+1/2}^- \quad \text{and} \quad \dot{b}_{j+1/2}^+ = w_{j+1/2}^+ - h_{j+1/2}^+. \quad (3.23)$$

Find the maximum

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

then we define

$$\dot{h}_{j+1/2}^- = \max \left\{ 0, w_{j+1/2}^- - \dot{b}_{j+1/2} \right\}, \quad (3.24a)$$

$$\dot{h}_{j+1/2}^+ = \max \left\{ 0, w_{j+1/2}^+ - \dot{b}_{j+1/2} \right\}. \quad (3.24b)$$

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

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

which we use to calculate the flux term  $F_{j+1/2}$  in (3.17) for both the continuity (2.4a) and irrotationality (2.4b) equation. Applying the same process but with different cells we obtain  $F_{j-1/2}$  and we have

$$F_{j-1/2} = \mathcal{F}_{j-1/2} \left( \dot{\mathbf{h}}, \hat{\mathbf{G}}, \hat{\mathbf{b}}, \hat{\mathbf{u}} \right), \quad (3.25a)$$

$$F_{j+1/2} = \mathcal{F}_{j+1/2} \left( \dot{\mathbf{h}}, \hat{\mathbf{G}}, \hat{\mathbf{b}}, \hat{\mathbf{u}} \right) \quad (3.25b)$$

for both the continuity and irrotationality equation as desired.

### (iv) Source Terms

The Serre equations in conservation law form (2.4) contain a flux term and a source term, to treat this numerically we use a first-order splitting method. This requires an approximation to the source term at the cell centre  $x_j$  which we denote as  $S_j$ . Since the continuity equation (2.4a) has no source term, we will just present the calculation of the source term for the irrotationality equation (2.4b).

Following the work of Audusse et al. [17], we split our approximation to  $S_j$  into the centred source term  $S_{ci}$  and the corrective interface source terms  $S_{j+\frac{1}{2}}^-$  and  $S_{j+\frac{1}{2}}^+$ . Where  $S_{ci}$  is the naive source term approximation and  $S_{j+\frac{1}{2}}^-$  and  $S_{j+\frac{1}{2}}^+$  are correction terms that ensure that the flux and source term cancel for the lake at rest steady state.

We calculate the centred source term using

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

Where we use  $h_j$  from the reconstruction process (3.4) and  $u_j$  from the solution of the elliptic equation (3.16). To calculate the derivatives we employ our polynomial representations of  $u$  and  $b$  inside a cell. However, to ensure that the terms cancel properly for a lake at rest we must modify our approximation to  $\frac{\partial b}{\partial x}$  to use  $\dot{b}_{j+1/2}^-$  and  $\dot{b}_{j+1/2}^+$  from (3.23). Therefore, the following approximations are used in the calculation of  $S_{ci}$

$$\left( \frac{\partial u}{\partial x} \right)_j = \frac{\partial}{\partial x} P_j^u(x_j), \quad (3.26a)$$

$$\left( \frac{\partial b}{\partial x} \right)_j = \frac{\dot{b}_{j+1/2}^- - \dot{b}_{j-1/2}^+}{\Delta x}, \quad (3.26b)$$

$$\left( \frac{\partial^2 b}{\partial x^2} \right)_j = \frac{\partial^2}{\partial x^2} P_j^b(x_j). \quad (3.26c)$$

The corrective interface source terms are

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

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

Which makes use of  $h_{j+\frac{1}{2}}^-$  and  $h_{j+\frac{1}{2}}^+$  obtained from the reconstruction (3.4) and the modified values  $\hat{h}_{j+\frac{1}{2}}^-$  and  $\hat{h}_{j+\frac{1}{2}}^+$  from (3.24). Combining the centred and interface source terms our approximation to the source term for the irrotationality equation is

$$S_j = \mathcal{S}_j \left( \hat{\mathbf{h}}, \hat{\mathbf{h}}, \hat{\mathbf{w}}, \hat{\mathbf{b}}, \hat{\mathbf{u}} \right) = S_{j+\frac{1}{2}}^- + \Delta x S_{ci} + S_{j-\frac{1}{2}}^+. \quad (3.27)$$

### (v) Update Cell Averages

We use the forward Euler approximation to approximate the time derivatives and obtain an update formula for the cell averages. Additionally we employ a fractional step method [] to split the flux term and source term part of the Serre equations written in conservation law form (2.4). This results in the following time-stepping method that is first-order accurate in time

$$\begin{aligned} \bar{q}'_j &= \bar{q}_j^n + \frac{\Delta t}{\Delta x} \left( F_{j+\frac{1}{2}} - F_{j-\frac{1}{2}} \right), \\ \bar{q}_j^{n+1} &= \bar{q}'_j + \frac{\Delta t}{\Delta x} S_j \end{aligned}$$

where  $F_{j+\frac{1}{2}}$ ,  $F_{j-\frac{1}{2}}$  and  $S_j$  are all calculated using the quantities at time  $t^n$ . Therefore, this method can be condensed into

$$\bar{q}_j^{n+1} = \bar{q}_j^n + \frac{\Delta t}{\Delta x} \left( F_{j+\frac{1}{2}} - F_{j-\frac{1}{2}} + S_j \right). \quad (3.28)$$

### (vi) Second-Order SSP Runge-Kutta Method

To increase the order of accuracy in time we employ the strong stability preserving Runge-Kutta method [19] which is a convex combination of multiple first-order time steps (3.28) in the following way

$$\bar{q}'_j = \bar{q}_j^n + \frac{\Delta t}{\Delta x} \left( F_{j+\frac{1}{2}} - F_{j-\frac{1}{2}} + S_j \right), \quad (3.29a)$$

$$\bar{q}''_j = \bar{q}'_j + \frac{\Delta t}{\Delta x} \left( F'_{j+\frac{1}{2}} - F'_{j-\frac{1}{2}} + S'_j \right), \quad (3.29b)$$

$$\bar{q}_j^{n+1} = \frac{1}{2} (\bar{q}_j^n + \bar{q}''_j). \quad (3.29c)$$

This results in a time stepping method that preserves the stability of the first-order method (3.28) and is second-order accurate in time. Since all the spatial approximations are second-order accurate, the steps (I-VI) result in a second-order accurate FEVM for the Serre equations, as desired.

### 3.3 CFL condition

To ensure the stability of our FEVM we use the Courant-Friedrichs-Lewy (CFL) condition [20] which is necessary for stability. The CFL condition ensures that time steps are small enough so that information is only transferred between neighbouring cells. For the Serre equations the CFL condition is

$$\Delta t \leq \frac{Cr}{\max_j \{a_{j+1/2}^\pm\}} \Delta x \quad (3.30)$$

where  $a_{j+1/2}^\pm$  are the wavespeed bounds used in Kurganovs flux approximation (3.19) and  $0 \leq Cr \leq 1$  is the Courant number. Typically, we use the conservative  $Cr = 0.5$  for our numerical experiments.

### 3.4 Boundary Conditions

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

$$\begin{aligned} \hat{\mathbf{h}}_{-1} &= \begin{bmatrix} h_{-3/2}^+ \\ h_{-1} \\ h_{-1/2}^- \end{bmatrix}, & \hat{\mathbf{h}}_{m+1} &= \begin{bmatrix} h_{m+1/2}^+ \\ h_{m+1} \\ h_{m+3/2}^- \end{bmatrix}, \\ \hat{\mathbf{w}}_{-1} &= \begin{bmatrix} w_{-3/2}^+ \\ w_{-1} \\ w_{-1/2}^- \end{bmatrix}, & \hat{\mathbf{w}}_{m+1} &= \begin{bmatrix} w_{m+1/2}^+ \\ w_{m+1} \\ w_{m+3/2}^- \end{bmatrix}, \\ \hat{\mathbf{G}}_{-1} &= \begin{bmatrix} G_{-3/2}^+ \\ G_{-1} \\ G_{-1/2}^- \end{bmatrix}, & \hat{\mathbf{G}}_{m+1} &= \begin{bmatrix} G_{m+1/2}^+ \\ G_{m+1} \\ G_{m+3/2}^- \end{bmatrix}, \\ \hat{\mathbf{u}}_{-1} &= \begin{bmatrix} u_{-3/2} \\ u_{-1} \\ u_{-1/2} \end{bmatrix}, & \hat{\mathbf{u}}_{m+1} &= \begin{bmatrix} u_{m+1/2} \\ u_{m+1} \\ u_{m+3/2} \end{bmatrix} \end{aligned}$$

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

To ensure that the solution of  $u$  by (3.16) agrees with the boundary conditions  $\hat{\mathbf{u}}_{-1}$  and  $\hat{\mathbf{u}}_m$  the element matrices  $\mathbf{A}_0$  and  $\mathbf{A}_m$  and vectors  $\mathbf{g}_0$  and  $\mathbf{g}_m$  must be modified in the following way

$$\mathbf{A}_0 = \begin{bmatrix} 1 & 0 & 0 \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad \mathbf{g}_0 = \begin{bmatrix} u_{-1/2} \\ g_1 \\ g_2 \end{bmatrix}, \quad (3.31)$$

$$\mathbf{A}_m = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{g}_m = \begin{bmatrix} g_0 \\ g_1 \\ u_{m+1/2} \end{bmatrix}. \quad (3.32)$$

These are then combined with the other element contributions to the global matrix (3.15).

### 3.5 Dry Beds

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

In previous work [21] we employed direct banded matrix solvers such as the Thomas algorithm to solve (3.16), however such methods rely on non-singular matrices and so are unsuitable for dry beds. To deal with this an LU decomposition algorithm by Press et al. [22] was used. This algorithm solves banded matrix problems using an LU decomposition with partial pivoting, which inserts small non-zero pivots when their value is below some tolerance value  $p_{tol}$ . It does this while also keeping the banded matrix structure, and so is not as memory intensive as a standard  $LU$  decomposition. This results in a FEM that could allow for  $h = 0$ , but still faced the problem of large  $u$  values when  $h$  was small.

To deal with large  $u$  values we restricted our solution of the FEM for  $u$  to cells where  $\bar{h}_j > h_{tol}$  and modified our approximation to  $h_{j-1/2}^+$  and  $h_{j+1/2}^-$  in our stiffness matrix. To restrict the domain of our solution we search through the



domain at the beginning of each first-order step and identify wet regions where  $\bar{h}_j \geq h_{tol}$  and dry regions where  $\bar{h}_j < h_{tol}$ . In the dry regions we set  $h$ ,  $G$  and  $u$  to be zero uniformly over the cell. In the wet regions we solve the FEM with the appropriate boundary conditions and the following modifications to  $h_{j-1/2}^+$  and  $h_{j+1/2}^-$

$$h_{j-1/2}^+ = h_{j-1/2}^+ \left( \frac{h_{j-1/2}^+ + h_{base}}{h_{j-1/2}^+ + h_{tol}} \right), \quad (3.33a)$$

$$h_{j+1/2}^- = h_{j+1/2}^- \left( \frac{h_{j+1/2}^- + h_{base}}{h_{j+1/2}^- + h_{tol}} \right). \quad (3.33b)$$

Where on the right hand side we mean  $h_{j-1/2}^+$  and  $h_{j+1/2}^-$  as calculated from the reconstruction (3.4). This modification ensures that as  $h \rightarrow 0$  then  $u \rightarrow 0$  by increasing the size of  $h$  in the stiffness matrix.

Typically we had  $p_{tol} = 10^{-20}$ ,  $h_{tol} = 10^{-8}$  and  $h_{base} = 10^{-4}$  which allowed for an accurate calculation of  $u$  in the presence of dry beds and very small water depths.



## Chapter 4

# Linear Analysis of the Numerical Methods

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

For linear partial differential equations the Lax-equivalence theorem states that a numerical method is convergent if and only if it is stable and consistent [23]. A numerical scheme is consistent if the error introduced by the numerical method over a time step approaches zero as the spatial and temporal resolution is increased. While a numerical method is stable if the errors from previous time steps are not amplified in subsequent time steps.

Another important property 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 []. Therefore, how well the dispersion relation of our numerical methods approximate the dispersion relation of the Serre equations is of particular interest.

We analysed the convergence and the dispersion properties of our numerical methods for the linearised Serre equations with horizontal beds. The effect of variations in the bed and nonlinear terms are important when studying the convergence properties of our methods for the full Serre equations. However, these effects greatly increase the complexity of the convergence analysis. We restrict ourselves to the study of the linearised Serre with horizontal beds to offer some

insight into the convergence properties of our numerical methods without having to deal with this greater complexity. In general we would expect that a numerical method that has poor convergence properties for the linearised Serre equations with horizontal beds will also have poor convergence properties when the bed and nonlinear terms are included. The dispersion properties are derived from the linearised Serre equations with the no effect from the bed [], therefore this analysis of dispersion properties of the numerical methods is complete.

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

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

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

We derive  $\mathbf{E}$  in (4.1) for the second-order FEVM and perform the convergence and dispersion analysis. We will then present the results of these analyses for all our numerical methods.

## 4.1 Linearised Serre equations with horizontal bed

The Serre equations with a horizontal bed (2.3) are linearised by considering waves as small perturbations  $\delta\eta$  and  $\delta v$  on a flow with a mean height  $H$  and a mean velocity  $U$  respectively. So we have

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

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

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

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

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

and for  $G$

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

These equations can be reformulated in terms of the conserved quantities  $\eta$  and  $G$

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

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

where

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

We have absorbed the  $\delta$  factor into the corresponding  $\eta$  and  $v$  terms to simplify the notation.

## 4.2 Evolution Matrix

To derive the evolution matrix we first assume that the solutions of the linearised Serre equations with horizontal beds (4.4) are periodic in space and time. In particular, we assume that  $\eta$  and  $v$  are Fourier modes, which for a general quantity  $q$  means

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

where  $k$  is the wavenumber,  $\omega$  is the frequency and  $i$  is the imaginary number. This is also the assumption made to derive the analytical dispersion relation of the linearised Serre equations []. A consequence of a quantity  $q$  being a Fourier mode represented on a uniform temporal and spatial grid is that for any real numbers  $m$  and  $l$  we have

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

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

### 4.2.1 Overview of the Evolution Step

We will now present a brief overview of an evolution step of the second-order FEVM. Given the vectors of the cell averages  $\bar{\eta}$  and  $\bar{\mathbf{G}}$  at the current time the second-order FEVM evolution step progresses in the following way:

- (i) **Reconstruction of Midpoint and Cell Interface Values:** We use the second-order accurate operator  $\mathcal{M}$  to calculate  $\eta$  and  $G$  at the cell midpoint  $x_j$  from the cell averages. We also reconstruct  $\eta$  and  $G$  at the cell interfaces  $x_{j+1/2}^-$  and  $x_{j+1/2}^+$  from the cell average values using  $\mathcal{R}^-$  and  $\mathcal{R}^+$  respectively which are both spatially second-order. So that

$$\begin{aligned} \eta_j &= \mathcal{M}(\bar{\eta}), & G_j &= \mathcal{M}(\bar{\mathbf{G}}), \\ \eta_{j+1/2}^- &= \mathcal{R}^-(\bar{\eta}), & G_{j+1/2}^- &= \mathcal{R}^-(\bar{\mathbf{G}}), \\ \eta_{j+1/2}^+ &= \mathcal{R}^+(\bar{\eta}), & G_{j+1/2}^+ &= \mathcal{R}^+(\bar{\mathbf{G}}). \end{aligned}$$

- (ii) **Calculate the Velocity at the Cell Interface:** The remaining unknown quantity,  $v_{j+1/2}$  is calculated from a spatially second-order accurate solution of the elliptic equation (4.4c). This calculation is represented by  $\mathcal{G}$  as

$$v_{j+1/2} = \mathcal{G}(H, \mathbf{G}, \eta).$$

- (iii) **Calculate the Flux Across the Cell Interface:** We calculate the average flux  $F_{j+1/2}$  across the cell boundary  $x_{j+1/2}$  over time using  $\mathcal{F}$

$$F_{j+1/2} = \mathcal{F}\left(\eta_{j+1/2}^-, G_{j+1/2}^-, \eta_{j+1/2}^+, G_{j+1/2}^+, v_{j+1/2}\right).$$

- (iv) **Time Step Using Forward Euler Approximation:** We repeat this process for each cell edge and then apply (??) to update the vectors  $\bar{\eta}$  and  $\bar{\mathbf{G}}$  from the current time level to the next time level with first-order accuracy in time.
- (v) **Complete Second-Order Time Step Using SSP Runge-Kutta Steps:** We repeat steps 1-4 and use SSP Runge-Kutta time stepping to calculate  $\bar{\eta}$  and  $\bar{\mathbf{G}}$  at the next time level with second-order accuracy in time.

We will now derive expressions for all the operators in the evolution step of the linear equations. These operators are linear because our assumption that  $\eta$  and  $v$  are Fourier modes. We will then combine these linear operators to derive  $\mathbf{E}$  for the second-order FEVM.

### (i) Reconstruct the Quantities Inside the Cells

Given  $\bar{\eta}$  and  $\bar{G}$  at  $t^n$  the first step of our numerical method is to calculate  $\eta$  and  $G$  at  $x_j$  using  $\mathcal{M}$  and at  $x_{j+1/2}^-$  and  $x_{j+1/2}^+$  using  $\mathcal{R}^-$  and  $\mathcal{R}^+$  respectively. The derivation of these operators is given in terms of a general quantity  $q$ , as they are the same for  $\eta$  and  $G$ .

#### Cell average values to nodal values: $\mathcal{M}$

For the second-order FEVM we use the fact that

$$\bar{q}_j = q_j + \mathcal{O}(\Delta x^2).$$

So to attain second-order accuracy we use

$$q_j = \bar{q}_j = \mathcal{M}\bar{q}_j \quad (4.7)$$

where the factor  $\mathcal{M} = 1$  represents the mapping between cell averages and nodal values for the numerical method.

#### Cell average values to interface values: $\mathcal{R}^-$ and $\mathcal{R}^+$

We reconstruct  $\eta$  and  $G$  at  $x_{j+1/2}^-$  and  $x_{j+1/2}^+$ . These quantities can be discontinuous across the cell interfaces in our finite volume method. However, since we are assuming that these quantities are Fourier modes and therefore smooth we do not require non-linear limiters to ensure our scheme is TVD. Therefore, the reconstruction scheme for  $\eta$  and  $G$  can be written for a general quantity  $q$  as

$$q_{j+\frac{1}{2}}^- = \bar{q}_j + \frac{-\bar{q}_{j-1} + \bar{q}_{j+1}}{4},$$

$$q_{j+\frac{1}{2}}^+ = \bar{q}_{j+1} + \frac{-\bar{q}_j + \bar{q}_{j+2}}{4}.$$

Using (4.6) and (4.7) these equations become

$$q_{j+\frac{1}{2}}^- = \bar{q}_j + \frac{-\bar{q}_j e^{-ik\Delta x} + \bar{q}_j e^{ik\Delta x}}{4} = \left(1 + \frac{i \sin(k\Delta x)}{2}\right) \bar{q}_j = \mathcal{R}^- \bar{q}_j, \quad (4.8a)$$

$$q_{j+\frac{1}{2}}^+ = \frac{\bar{q}_j e^{ik\Delta x} + \bar{q}_j + \bar{q}_j e^{2ik\Delta x}}{4} = e^{ik\Delta x} \left(1 - \frac{i \sin(k\Delta x)}{2}\right) \bar{q}_j = \mathcal{R}^+ \bar{q}_j \quad (4.8b)$$

where  $\mathcal{R}^\pm$  are the reconstruction factors for both  $\eta_{j+1/2}^\pm$  and  $G_{j+1/2}^\pm$ .

**(ii) Calculate the Velocity Over the Domain**

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

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

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

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



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

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

Using (4.6) and (4.8), we obtain

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

After simplifying

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

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

third equations. These equations are equivalent up to a translation and therefore we will only consider the third equation.

So far we have only given the contribution to the equation at  $x_{j+1/2}$  from the  $j^{th}$  cell, but there is also a contribution from the adjacent  $(j+1)^{th}$  cell because  $\phi_{j+1/2}$  is non-zero over both cells. Accounting for this we get

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

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

$$\begin{aligned} v_{j+1/2} &= \left[ \left( \frac{\Delta x}{6} (\mathcal{R}^- + \mathcal{R}^+) \right) \bar{G}_j - U \left( \frac{\Delta x}{6} (\mathcal{R}^- + \mathcal{R}^+) \right) \bar{\eta}_j - \frac{\Delta x}{3} UH \right] \\ &\quad \div \left[ H \frac{\Delta x}{30} \left( 4 \cos \left( \frac{k\Delta x}{2} \right) - 2 \cos(k\Delta x) + 8 \right) \right. \\ &\quad \left. + \frac{H^3}{9\Delta x} \left( -16 \cos \left( \frac{k\Delta x}{2} \right) + 2 \cos(k\Delta x) + 14 \right) \right] \\ &= \mathcal{G}^G \bar{G}_j + \mathcal{G}^\eta \bar{\eta}_j + \mathcal{G}^c. \end{aligned} \tag{4.9}$$

### (iii) Calculate All the Fluxes Across the Cell Interfaces

The average intercell flux  $F_{j+1/2}$  is approximated using Kurganov's method [16]. For the linearised Serre equations we have the wave speed bounds (2.8a), so that

$$a_{j+1/2}^- = \min \left\{ 0, U - \sqrt{gH} \right\} \quad \text{and} \quad a_{j+1/2}^+ = \max \left\{ 0, U + \sqrt{gH} \right\}. \tag{4.10}$$

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

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

For left supercritical flow;  $Fr < -1$  and therefore  $U + \sqrt{gH} < 0$  so we have from (4.10) that  $a_{j+1/2}^- = U - \sqrt{gH}$  and  $a_{j+1/2}^+ = 0$ . For these values the Kurganov flux approximation for a general quantity  $q$  [] reduces to

$$F_{j+\frac{1}{2}} = f\left(q_{j+\frac{1}{2}}^+\right). \quad (4.11)$$

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

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

since  $v$  is continuous and therefore  $v_{j+1/2} = v_{j+1/2}^+ = v_{j+1/2}^-$ . Using the FEM for  $v_{j+1/2}$  (4.9) and the reconstruction (4.8) we have

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

Substituting the flux function for irrotationality equation (4.4b) into the Kurganov flux approximation (4.11) we obtain

$$F_{j+\frac{1}{2}}^G = U G_{j+1/2}^+ + U H v_{j+1/2} + g H \eta_{j+1/2}^+$$

Using the FEM (4.9) to calculate  $v_{j+1/2}$  and our interface reconstruction (4.8) we have

$$\begin{aligned} F_{j+\frac{1}{2}}^G &= U G_{j+1/2}^+ + U H (\mathcal{G}^G \bar{G}_j + \mathcal{G}^\eta \bar{\eta}_j + \mathcal{G}^c) + g H \eta_{j+1/2}^+ \\ &= (U H \mathcal{G}^\eta + g H \mathcal{R}^+) \bar{\eta}_j + (U \mathcal{R}^+ + U H \mathcal{G}^G) \bar{G}_j + U H \mathcal{G}^c \\ &= \mathcal{F}_{j+\frac{1}{2}}^{G, \eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{G, G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{G, c} \end{aligned} \quad (4.13)$$

**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 Kurganov flux approximation for a general quantity  $q$  [] is

$$F_{j+\frac{1}{2}} = \frac{U}{2\sqrt{gH}} \left[ f(q_{j+\frac{1}{2}}^-) - f(q_{j+\frac{1}{2}}^+) \right] + \frac{1}{2} \left[ f(q_{j+\frac{1}{2}}^-) + f(q_{j+\frac{1}{2}}^+) \right] + \frac{U^2 - gH}{2\sqrt{gH}} \left[ q_{j+\frac{1}{2}}^+ - q_{j+\frac{1}{2}}^- \right]. \quad (4.14)$$

Substituting in the flux function from the continuity equation (4.4a) we get

$$F_{j+\frac{1}{2}}^\eta = \frac{U}{2\sqrt{gH}} \left[ H v_{j+1/2} + U \eta_{j+\frac{1}{2}}^- - H v_{j+1/2} - U \eta_{j+\frac{1}{2}}^+ \right] + \frac{1}{2} \left[ H v_{j+1/2} + U \eta_{j+\frac{1}{2}}^- + H v_{j+1/2} + U \eta_{j+\frac{1}{2}}^+ \right] + \frac{U^2 - gH}{2\sqrt{gH}} \left[ \eta_{j+\frac{1}{2}}^+ - \eta_{j+\frac{1}{2}}^- \right]. \quad (4.15)$$

Using the reconstruction factors (4.8) and the elliptic solver (4.9) we get

$$F_{j+\frac{1}{2}}^\eta = \left( H \mathcal{G}^\eta + \frac{U}{2} [\mathcal{R}^- + \mathcal{R}^+] - \frac{\sqrt{gH}}{2} [\mathcal{R}^+ - \mathcal{R}^-] \right) \bar{\eta}_j + H \mathcal{G}^G \bar{G}_j + H \mathcal{G}^c = \mathcal{F}_{j+\frac{1}{2}}^{\eta,\eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{\eta,c}. \quad (4.16)$$

For the flux function of the irrotationality equation (4.4b) the flux approximation (4.14) becomes

$$F_{j+\frac{1}{2}}^G = \frac{U}{2\sqrt{gH}} \left[ U G_{j+\frac{1}{2}}^- + U H v_{j+1/2} + g H \eta_{j+\frac{1}{2}}^- - U G_{j+\frac{1}{2}}^+ - U H v_{j+1/2} - g H \eta_{j+\frac{1}{2}}^+ \right] + \frac{1}{2} \left[ U G_{j+\frac{1}{2}}^- + U H v_{j+1/2} + g H \eta_{j+\frac{1}{2}}^- + U G_{j+\frac{1}{2}}^+ + U H v_{j+1/2} + g H \eta_{j+\frac{1}{2}}^+ \right] + \frac{U^2 - gH}{2\sqrt{gH}} \left[ G_{j+\frac{1}{2}}^+ - G_{j+\frac{1}{2}}^- \right]. \quad (4.17)$$

By using the reconstruction factors (4.8) and the elliptic solver (4.9) we get

$$F_{j+\frac{1}{2}}^G = \left( \frac{U\sqrt{gH}}{2} [\mathcal{R}^- - \mathcal{R}^+] + U H \mathcal{G}^\eta + \frac{gH}{2} [\mathcal{R}^- + \mathcal{R}^+] \right) \bar{\eta}_j + \left( U H \mathcal{G}^G + \frac{U}{2} [\mathcal{R}^- + \mathcal{R}^+] - \frac{\sqrt{gH}}{2} [\mathcal{R}^+ - \mathcal{R}^-] \right) \bar{G}_j + U H \mathcal{G}^c = \mathcal{F}_{j+\frac{1}{2}}^{G,\eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,c}. \quad (4.18)$$

**Right supercritical flow  $Fr > 1$ :**

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

$$F_{j+\frac{1}{2}} = f\left(q_{j+\frac{1}{2}}^-\right). \quad (4.19)$$

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

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

$$\begin{aligned} F_{j+\frac{1}{2}}^G &= (U H \mathcal{G}^\eta + g H \mathcal{R}^-) \bar{\eta}_j + (U \mathcal{R}^- + U H \mathcal{G}^G) \bar{G}_j + U H \mathcal{G}^c \\ &= \mathcal{F}_{j+\frac{1}{2}}^{G,\eta} \bar{\eta}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,G} \bar{G}_j + \mathcal{F}_{j+\frac{1}{2}}^{G,c}. \end{aligned} \quad (4.21)$$

**(iv) Calculate All the Source Terms for the Cells**

□

**(v) Update All the Cell Averages Using a Forward Euler Approximation**

We have obtained the operators for the flux functions for all three cases, supercritical flow in the left or right direction and subcritical flow. By substituting the appropriate flux approximation for the physical situation into our update scheme (??) our second-order FEVM can be written as

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

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

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

This can be written in matrix form as

$$\begin{aligned} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^{n+1} &= \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n - (1 - e^{-ik\Delta x}) \frac{\Delta t}{\Delta x} \begin{bmatrix} \mathcal{F}^{\eta,\eta} & \mathcal{F}^{\eta,G} \\ \mathcal{F}^{G,\eta} & \mathcal{F}^{G,G} \end{bmatrix} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n \\ &= (\mathbf{I} - \Delta t \mathbf{F}) \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n \end{aligned} \quad (4.22)$$

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

### (vi) Update All the Cell Averages Using a Second-Order SSP Runge-Kutta Method

The second-order SSP Runge Kutta time stepping uses two forward Euler steps to accomplish a temporally second-order accurate method in the following way

$$\begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^1 = (\mathbf{I} - \Delta t \mathbf{F}) \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n, \quad (4.23a)$$

$$\begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^2 = (\mathbf{I} - \Delta t \mathbf{F}) \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^1, \quad (4.23b)$$

$$\begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^{n+1} = \frac{1}{2} \left( \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n + \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^2 \right). \quad (4.23c)$$

Substituting (4.23a) and (4.23b) into (4.23c) we can write this in terms of the flux matrix  $\mathbf{F}$  and our cell averages at  $t^n$  as

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

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

$$\begin{aligned} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^{n+1} &= \left( \mathbf{I} - \Delta t \mathbf{F} + \frac{1}{2} \Delta t^2 \mathbf{F}^2 \right) \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n \\ &= \mathbf{E} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n \end{aligned} \quad (4.24)$$

which is in the desired form (4.1).

This is the evolution matrix  $\mathbf{E}$  for the second-order FEVM. The matrix  $\mathbf{E}$  is dependent on the flux matrix  $\mathbf{F}$  and therefore will depend on the Froude number. The Froude number is constant and so we can analyse these flow scenarios individually.

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

### 4.3 Convergence Analysis

The linearised Serre equations are linear partial differential equations and therefore we can apply the Lax-equivalence theorem to demonstrate the convergence of our numerical methods by establishing their consistency and stability. We use a Von Neumann stability analysis to demonstrate stability. Consistency is demonstrated for Fourier mode solutions (4.5), which are eigenfunctions of the linearised Serre equations. Together this stability and consistency condition imply convergence of the numerical method under the  $L_2$  norm.

#### 4.3.1 Stability

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

$$\rho(\mathbf{E}) \leq 1 \quad (4.25)$$

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 (4.25) 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  $U$ . We used  $g = 9.81 \text{ m/s}^2$  and chose  $\Delta t = 0.5 / (U + \sqrt{gH}) \Delta x$  to satisfy the CFL condition []. This is the common choice of  $\Delta t$  in our numerical experiments.

The values  $H = 1 \text{ m}$ ,  $k = \frac{\pi}{10} \text{ m}^{-1}$  and  $U = 0 \text{ m/s}$  and  $1 \text{ m/s}$  shown in Figure 4.1 where chosen because they represent the general behaviour of these plots. For these  $k$  and  $H$  values our shallowness parameter  $\sigma = \frac{1}{20}$  and so the Serre equations are applicable [].

In Figure 4.1 it can be seen that all methods have  $\rho(\mathbf{E}) \leq 1$  for  $U = 0 \text{ m/s}$  and are therefore stable. The two finite difference methods overlap and have

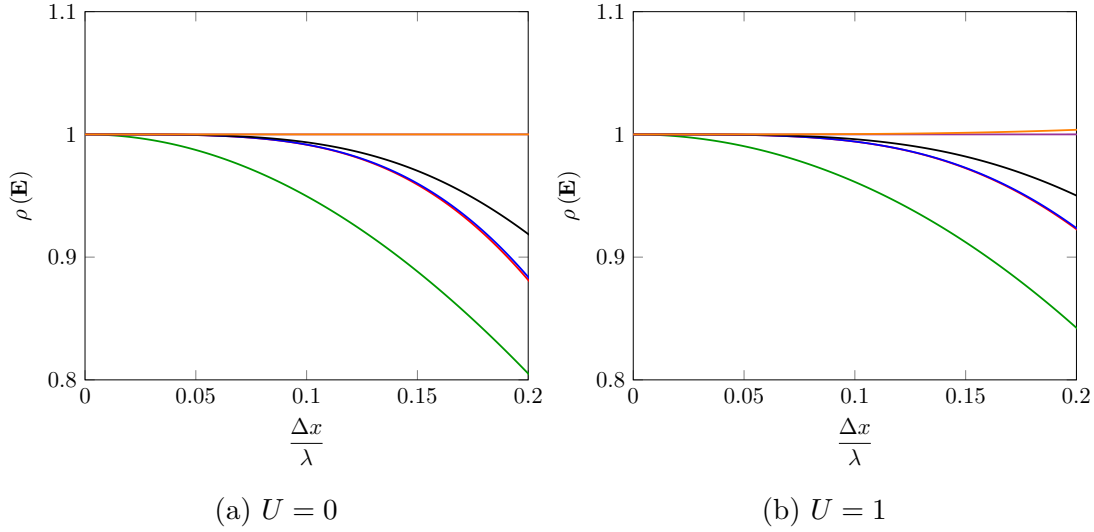


Figure 4.1: Spectral radius of  $\mathbf{E}$  for first-order FDVM (—), second-order FDVM(—), second-order FEVM (—), third-order FDVM (—),  $\mathcal{D}$  (—) and  $\mathcal{W}$  (—). With  $H = 1m$  and  $k = \frac{\pi}{10}$ .

$\rho(\mathbf{E}) = 1$  for all  $\Delta x$  values, while the second-order FDVM and the second-order FEVM also overlap. However, when  $U \neq 0m/s$  then  $\mathcal{W}$  has  $\rho(\mathbf{E}) > 1$  for all  $\Delta x$  and is therefore unstable.

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

$$\begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^{n+1} = e^{i\omega\Delta t} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n.$$

Therefore, the analytic growth factor is

$$\rho(\mathbf{E}) = |e^{i\omega\Delta t}| = \sqrt{\cos^2(\omega\Delta t) + \sin^2(\omega\Delta t)} = 1 \quad (4.26)$$

since  $\omega \in \mathbb{R}$ . Therefore numerical methods with  $\rho(\mathbf{E})$  closer to 1 are closer to the analytic value. In this sense the two finite difference methods are best, although  $\mathcal{W}$  is unstable. While for the FDVM we can see that the higher-order methods better approximate the analytic value, as expected. We can see in Figure 4.1 that  $\lim_{\Delta x \rightarrow 0} \rho(\mathbf{E}) = 1$  for all methods, as expected.

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



### 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. We will demonstrate this only for Fourier mode solutions of the linearised Serre equations. Therefore, we can demonstrate consistency by investigating the evolution matrix  $\mathbf{E}$ . The error introduced for a single time step from  $t^n$  to  $t^{n+1}$ ,  $\mathcal{T}^n$  is

$$\mathcal{T}^n = \mathbf{E} \left[ \frac{\bar{\eta}}{\bar{G}} \right]_j^n - \left[ \frac{\bar{\eta}}{\bar{G}} \right]_j^{n+1}. \quad (4.27)$$

To ensure consistency we must have that  $\lim_{\Delta x, \Delta t \rightarrow 0} \|\mathcal{T}^n\| = 0$  for all  $n$ . Taking the norm of both sides of (4.27) we get

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

Making use of (4.6) we obtain

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

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

$$\|\mathcal{T}^n\| \leq \|\mathbf{E} - e^{i\omega\Delta t}\mathbf{I}\| \left\| \left[ \frac{\bar{\eta}}{\bar{G}} \right]_j^n \right\|. \quad (4.28)$$

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

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

We observe for all of our methods 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 we have that for all methods

| Element                         | Lowest Order Term of Error                       |   |
|---------------------------------|--|---|
|                                 | $\Delta x$                                       | $\Delta t$  |
| $E_{00} - e^{i\omega+\Delta t}$ | $-\frac{1}{2}\sqrt{gH}k^2\Delta t\Delta x$       | $\frac{\sqrt{3gH}\beta + 3U}{\beta}ik\Delta t$    |
| $E_{01}$                        | $\frac{3+\beta}{4\beta^2}ik^3\Delta t\Delta x^2$ | $-\frac{3}{\beta}ik\Delta t$                      |
| $E_{10}$                        | $-\frac{1}{2}\sqrt{gH}k^2\Delta t\Delta x$       | $\left(-gH + \frac{3U^2}{\beta}\right)ik\Delta t$ |
| $E_{11} - e^{i\omega+\Delta t}$ | $-\frac{1}{2}\sqrt{gH}k^2\Delta t\Delta x$       | $\frac{\sqrt{3gH}\beta - 3U}{\beta}ik\Delta t$    |

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

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

Choosing a particular vector norm such as the  $L_1$  or  $L_\infty$  and its induced matrix norm we can see from Tables 4.1-4.8 that  $\mathbf{A}$  is independent of  $\Delta t$  and finite so that as  $\Delta t \rightarrow 0$  we have  $|\Delta t| \left( \|\mathbf{A}_0\| + \left\| \begin{bmatrix} \mathcal{O}(\Delta t) & \mathcal{O}(\Delta t) \\ \mathcal{O}(\Delta t) & \mathcal{O}(\Delta t) \end{bmatrix} \right\| \right) \rightarrow 0$  and therefore  $\|\mathbf{E} - e^{i\omega+\Delta t}\mathbf{I}\| \rightarrow 0$ . Therefore, for all our numerical methods we have  $\lim_{\Delta x, \Delta t \rightarrow 0} \|\mathcal{T}^n\| = 0$  and so all our numerical methods are consistent for Fourier mode solutions as desired.

| Scheme                           | Lowest Order $\Delta x$ Term of Error |                                    |
|----------------------------------|---------------------------------------|------------------------------------|
|                                  | $U < -\sqrt{gH}$                      | $\sqrt{gH} < U$                    |
| $E_{00} - e^{i\omega_+\Delta t}$ | $\frac{1}{2}k^2U\Delta t\Delta x$     | $-\frac{1}{2}k^2U\Delta t\Delta x$ |
| $E_{01}$                         | $\frac{1}{2}gHk^2\Delta t\Delta x$    | $\frac{1}{2}gHk^2\Delta t\Delta x$ |
| $E_{11} - e^{i\omega_+\Delta t}$ | $\frac{1}{2}k^2U\Delta t\Delta x$     | $-\frac{1}{2}k^2U\Delta t\Delta x$ |

Table 4.2: Table of the lowest order term of the Taylor series for the elements of  $\mathbf{E} - e^{i\omega_+\Delta t}\mathbf{I}$  for the first-order FDVM which are different than those in Table 4.1 with  $\beta = 3 + k^2H^2$ .

| Element                          | Lowest Order Term of Error   |   |
|----------------------------------|--|---|
|                                  | $\Delta x$   | $\Delta t$  |
| $E_{00} - e^{i\omega_+\Delta t}$ | $-\frac{i(27 + 9H^2k^2 + H^4k^4)}{12\beta^2}Uk^3\Delta x^2$                                    | $\frac{\sqrt{3gH\beta} + 3U}{\beta}ik\Delta t$    |
| $E_{01}$                         | $\frac{3 + \beta}{4\beta^2}ik^3\Delta t\Delta x^2$   | $-\frac{3}{\beta}ik\Delta t$                      |
| $E_{10}$                         | $-\left(gH + \frac{3U^2}{\beta} + \frac{9U^2}{\beta^2}\right)\frac{k^3}{12}\Delta t\Delta x^2$ | $\left(-gH + \frac{3U^2}{\beta}\right)ik\Delta t$ |
| $E_{11} - e^{i\omega_+\Delta t}$ | $\frac{-9 + H^2k^2\beta}{\beta^2}\frac{k^3}{12}iU\Delta t\Delta x^2$                           | $\frac{\sqrt{3gH\beta} - 3U}{\beta}ik\Delta t$    |

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

| Element                          | Lowest Order Term of Error                         |   |
|----------------------------------|--|---|
|                                  | $\Delta x$   | $\Delta t$  |
| $E_{00} - e^{i\omega_+\Delta t}$ | $-\frac{1}{12}\sqrt{gH}k^4\Delta t\Delta x^3$      | $\frac{\sqrt{3gH}\beta + 3U}{\beta}ik\Delta t$    |
| $E_{01}$                         | $\frac{\sqrt{gH}}{4\beta}ik^5\Delta t^2\Delta x^3$ | $-\frac{3}{\beta}ik\Delta t$                      |
| $E_{10}$                         | $-\frac{1}{12}\sqrt{gH}k^4\Delta t\Delta x^3$      | $\left(-gH + \frac{3U^2}{\beta}\right)ik\Delta t$ |
| $E_{11} - e^{i\omega_+\Delta t}$ | $-\frac{1}{12}\sqrt{gH}k^4\Delta t\Delta x^3$      | $\frac{\sqrt{3gH}\beta - 3U}{\beta}ik\Delta t$    |

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

| Scheme                           | Lowest Order $\Delta x$ Term of Error       |  |
|----------------------------------|---|--|
|                                  | $U < -\sqrt{gH}$                            | $\sqrt{gH} < U$                              |
| $E_{00} - e^{i\omega_+\Delta t}$ | $\frac{1}{12}k^4U\Delta t\Delta x^3$        | $-\frac{1}{12}k^4U\Delta t\Delta x^3$        |
| $E_{01}$                         | $\frac{1}{4\beta}iUk^5\Delta t^2\Delta x^3$ | $-\frac{1}{4\beta}iUk^5\Delta t^2\Delta x^3$ |
| $E_{10}$                         | $\frac{1}{12}gHk^4\Delta t^2\Delta x^3$     | $-\frac{1}{12}gHk^4\Delta t^2\Delta x^3$     |
| $E_{11} - e^{i\omega_+\Delta t}$ | $\frac{1}{12}k^4U\Delta t\Delta x^3$        | $-\frac{1}{12}k^4U\Delta t\Delta x^3$        |

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

| Element                          | Lowest Order Term of Error   |   |
|----------------------------------|--|---|
|                                  | $\Delta x$   | $\Delta t$  |
| $E_{00} - e^{i\omega_+\Delta t}$ | $-\frac{i(54 + 45H^2k^2 + 10H^4k^4)}{120\beta^2}Uk^3\Delta t\Delta x^2$                        | $\frac{\sqrt{3gH\beta} + 3U}{\beta}ik\Delta t$    |
| $E_{01}$                         | $\frac{\beta - 3}{\beta^2}\frac{ik^3}{40}\Delta t\Delta x^2$                                   | $-\frac{3}{\beta}ik\Delta t$                      |
| $E_{10}$                         | $-\left(gH - \frac{15U^2}{\beta} + \frac{9U^2}{\beta}\right)\frac{k^3}{120}\Delta t\Delta x^2$ | $\left(-gH + \frac{3U^2}{\beta}\right)ik\Delta t$ |
| $E_{11} - e^{i\omega_+\Delta t}$ | $\frac{126 + 75H^2k^2 + 10H^4k^4}{\beta^2}\frac{k^3}{120}iU\Delta t\Delta x^2$                 | $\frac{\sqrt{3gH\beta} - 3U}{\beta}ik\Delta t$    |

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

| Element                          | Lowest Order Term of Error                            |                                       |
|----------------------------------|---|---------------------------------------|
|                                  | $\Delta x$  | $\Delta t$                            |
| $E_{00} - e^{i\omega_+\Delta t}$ | $\frac{ik^3}{3}U\Delta t\Delta x^2$                   | $\sqrt{\frac{3gH}{\beta}}2ik\Delta t$ |
| $E_{01}$                         | $\frac{iHk^3}{3}\Delta t\Delta x^2$                   | $-2Hik\Delta t$                       |
| $E_{10}$                         | $\frac{ig(3 + \beta)}{2\beta^2}k^3\Delta t\Delta x^2$ | $-\frac{6igk}{\beta}\Delta t$         |
| $E_{11} - e^{i\omega_+\Delta t}$ | $\frac{ik^3}{3}U\Delta t\Delta x^2$                   | $\sqrt{\frac{3gH}{\beta}}2ik\Delta t$ |

Table 4.7: Table of the lowest order term of the Taylor series for the elements of  $\mathbf{E} - e^{i\omega\Delta t}\mathbf{I}$  for  $\mathcal{D}$  with  $\beta = 3 + k^2H^2$ .

| Element                          | Lowest Order Term of Error                          |                                       |
|----------------------------------|---|---------------------------------------|
|                                  | $\Delta x$  | $\Delta t$                            |
| $E_{00} - e^{i\omega_+\Delta t}$ | $\frac{ik^3}{6}U\Delta t\Delta x^2$                 | $\sqrt{\frac{3gH}{\beta}}ik\Delta t$  |
| $E_{01}$                         | $\frac{iHk^3}{6}\Delta t\Delta x^2$                 | $-Hik\Delta t$                        |
| $E_{10}$                         | $\frac{ig(3+\beta)}{2\beta^2}k^3\Delta t\Delta x^2$ | $-\frac{6igk}{\beta}\Delta t$         |
| $E_{11} - e^{i\omega_+\Delta t}$ | $\frac{ik^3}{3}U\Delta t\Delta x^2$                 | $\sqrt{\frac{3gH}{\beta}}2ik\Delta t$ |

Table 4.8: Table of the lowest order term of the Taylor series for the elements of  $\mathbf{E} - e^{i\omega_+\Delta t}\mathbf{I}$  for  $\mathcal{W}$  with  $\beta = 3 + k^2H^2$ .

## 4.4 Dispersion Analysis

To study the dispersion of our numerical methods we must calculate  $\omega$  for our numerical methods. Making use of (4.6) in (4.24) we get

$$e^{i\omega\Delta t} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n = \mathbf{E} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n. \quad (4.29)$$

Assuming that  $\mathbf{E}$  has an eigenvalue decomposition  $\mathbf{E} = \mathbf{P}^{-1}\mathbf{\Lambda}\mathbf{P}$  and substituting it into (4.29) we get

$$e^{i\omega\Delta t} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n = \mathbf{P}^{-1}\mathbf{\Lambda}\mathbf{P} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n. \quad (4.30)$$

Left multiplying (4.30) by  $\mathbf{P}$  we obtain

$$e^{i\omega\Delta t} \mathbf{P} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n = \mathbf{\Lambda}\mathbf{P} \begin{bmatrix} \bar{\eta} \\ \bar{G} \end{bmatrix}_j^n. \quad (4.31)$$

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

$$\tilde{\omega}_{\pm} = \frac{1}{i\Delta t} \log [\lambda_{\pm}]. \quad (4.32)$$

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

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

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

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

For the real part of the dispersion error the FEVM and the FDVM outperform the two finite difference methods and therefore will better approximate the speed of waves of the linearised Serre equations. However, for the dilation of waves the roles are reversed with the two finite difference methods either dilating the waves very little ( $\mathcal{W}$  for  $U > 0$ ) or not at all. When taking both effects into account with the complete error we see that the first-order FDVM has the largest dispersion error followed by  $\mathcal{W}$ ,  $\mathcal{D}$ , the second-order FEVM, the second-order FDVM and finally the third-order FDVM has the lowest dispersion error. So that the size of the total dispersion error is mainly determined by the order of accuracy of the numerical scheme. These results justify choosing these FDVM over these finite difference methods for the Serre equations.

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

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

These  $kH$  values are the same as those by Filippini et al. [24], 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. [24]. Furthermore, we extended their work by allowing for non-zero values of  $U$  and examining the imaginary and complete error in the dispersion relation.

Figure 4.5 demonstrates that the results of the real part of the dispersion error is slightly different if we allow for non-zero values of  $U$ . In particular the non-zero value of  $U$  changes the real part of the dispersion error for the first-order FDVM, most significantly when  $kH = 2.5$ . Therefore, for some methods allowing for non-zero values of  $U$  can have a significant impact on the conclusions drawn from the dispersion analysis. Furthermore taking the imaginary part of the dispersion error into account is important as  $\omega$  determines not only the speed of waves but also their amplitude. In particular it is possible that a method like the first-order FDVM performs very well for the real part of the dispersion error and poorly for the imaginary part, leading to false conclusions about the accuracy of the method.

The Taylor series expansion of  $\tilde{\omega}$  was also derived for all the numerical methods. We have compiled the lowest order terms of the Taylor series for  $\tilde{\omega}_+ - \omega_+$  in Table A.6 when  $-\sqrt{gH} \leq U \leq \sqrt{gH}$  for the FDVM and FEVM. In Table A.6 it is clear that these schemes estimated  $\omega$  with the expected order of accuracy in both space and time.

We also present the lowest order terms of the Taylor series for  $\tilde{\omega}_+ - \omega_+$  for both  $U < -\sqrt{gH}$  and  $U > \sqrt{gH}$  in Table 4.10. We only present the errors that are different from those reported in Table A.6, this was only the case for the spatial error of the odd-order numerical methods. We can see that for all the flow scenarios that our FDVM and the FEVM have the correct order of accuracy when approximating  $\omega_+$ .

Finally we present the lowest order terms of the Taylor series for  $\tilde{\omega}_+ - \omega_+$  for the finite difference methods in Table 4.11. These methods do not change depending on the value of the physical quantities. The two finite difference methods both have the correct order of accuracy in both space and time.

Because all methods were demonstrated to have the expected order of accuracy in approximating  $\omega_+$  this implies that for small  $\Delta x$  values the order of accuracy



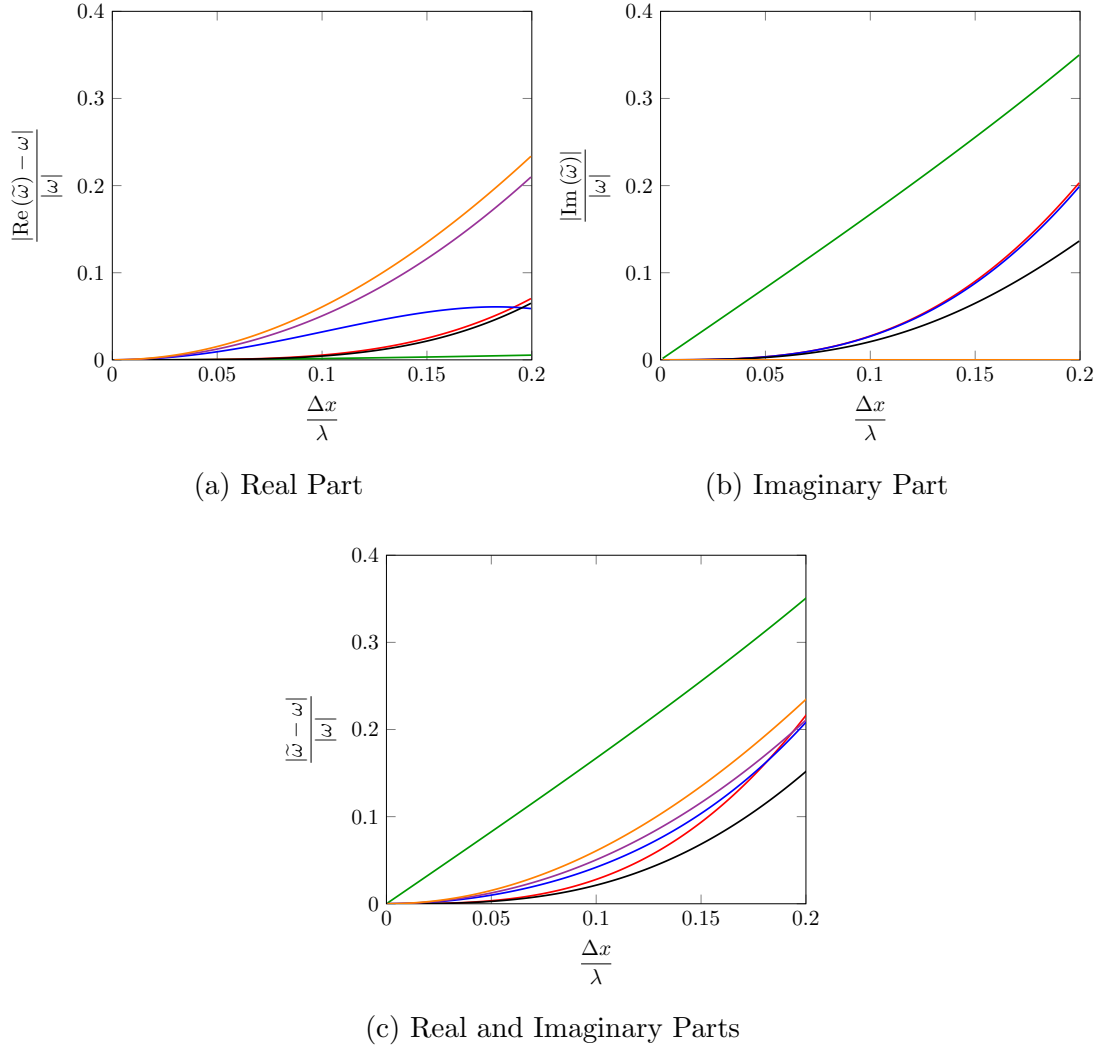


Figure 4.2: Relative dispersion error for first-order FDVM (—), second-order FDVM(—), second-order FEVM (—), third-order FDVM (—),  $\mathcal{D}$  (—) and  $\mathcal{W}$  (—). With  $H = 1m$ ,  $k = \frac{\pi}{10}$  and  $U = 0m/s$ .

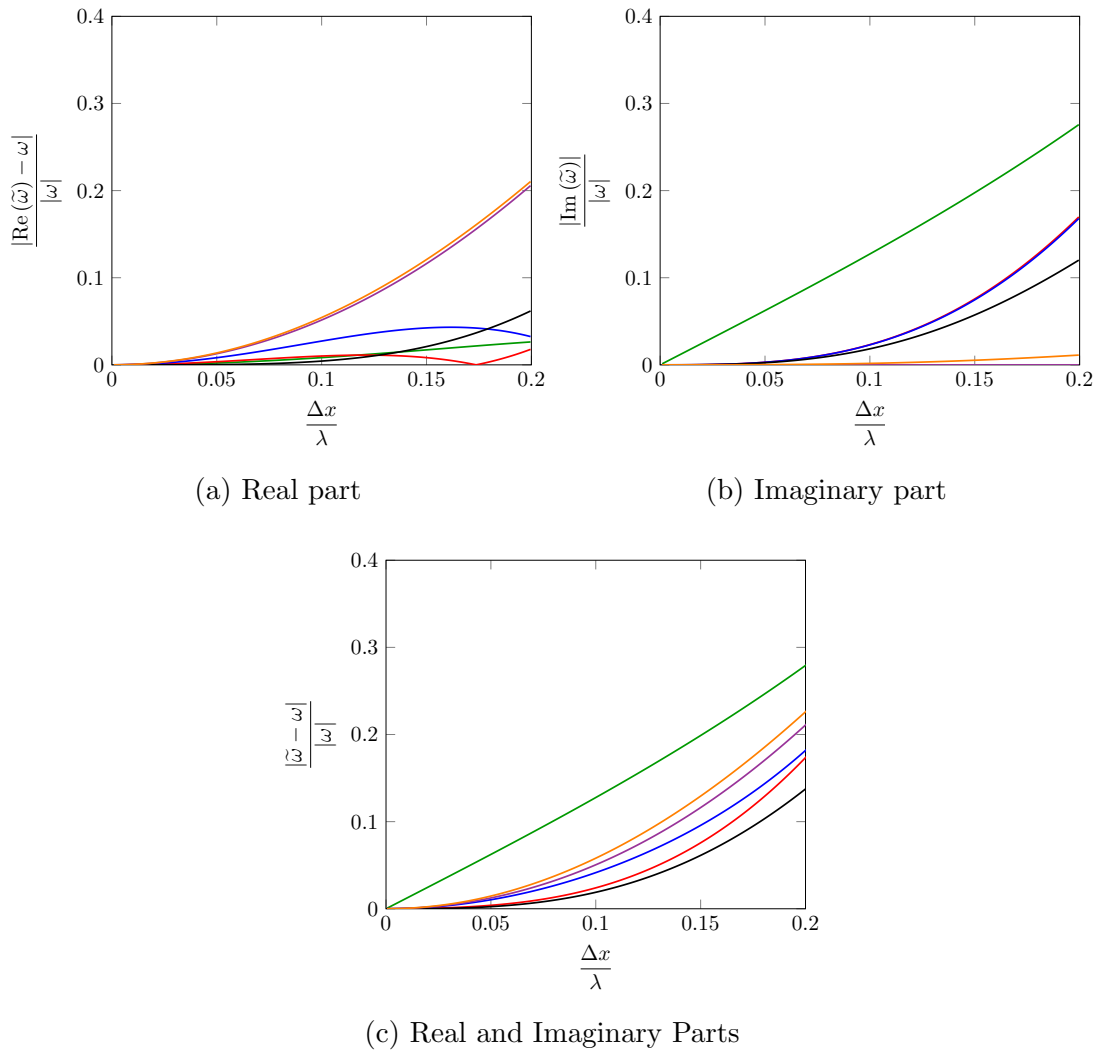


Figure 4.3: Relative dispersion error for first-order FDVM (—), second-order FDVM (—), second-order FEVM (—), third-order FDVM (—),  $\mathcal{D}$  (—) and  $\mathcal{W}$  (—). With  $H = 1m$ ,  $k = \frac{\pi}{10}$  and  $U = 1m/s$ .

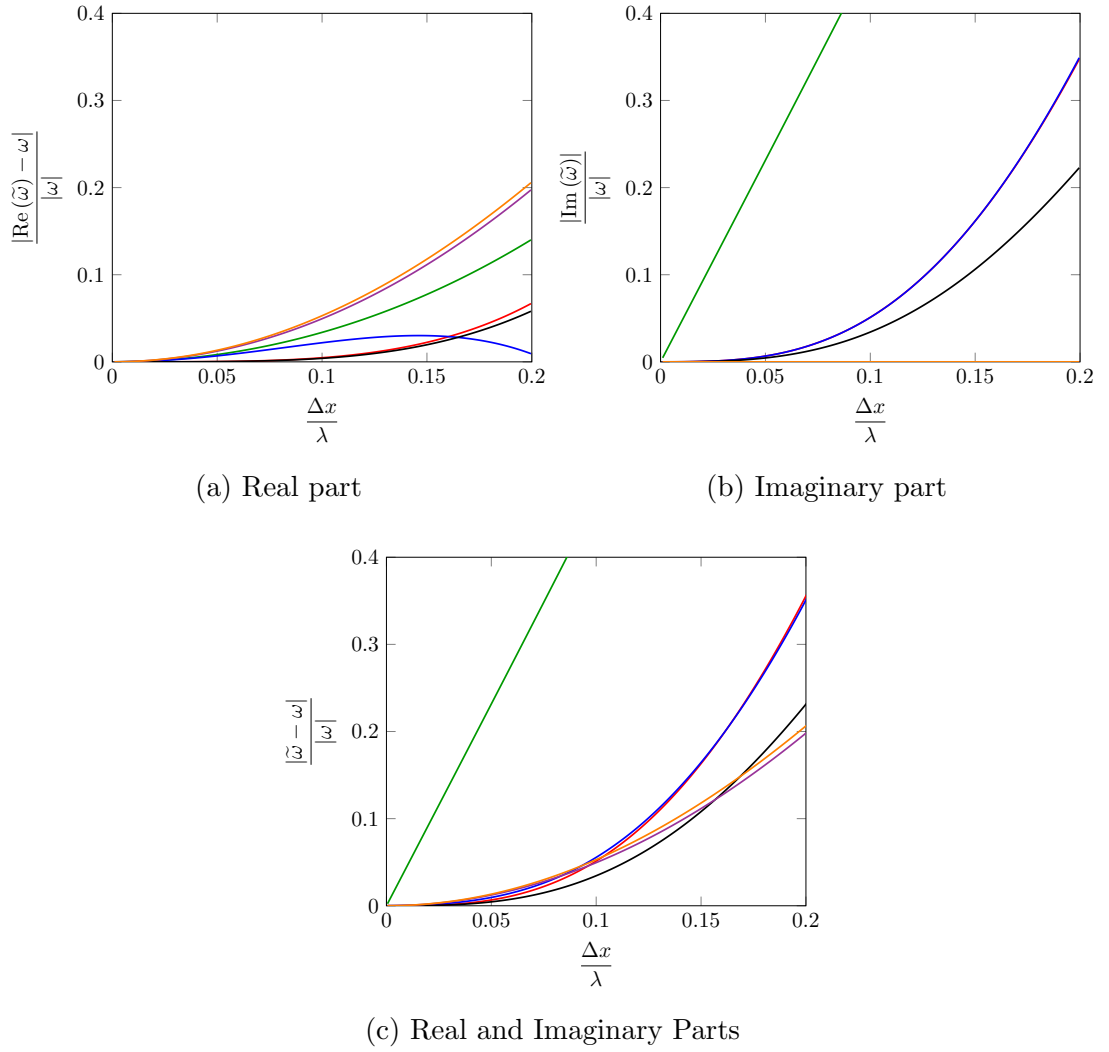


Figure 4.4: Relative dispersion error for first-order FDVM (—), second-order FDVM(—), second-order FEVM (—), third-order FDVM (—),  $\mathcal{D}$  (—) and  $\mathcal{W}$  (—). With  $H = 1m$ ,  $k = 2.5$  and  $U = 0m/s$ .

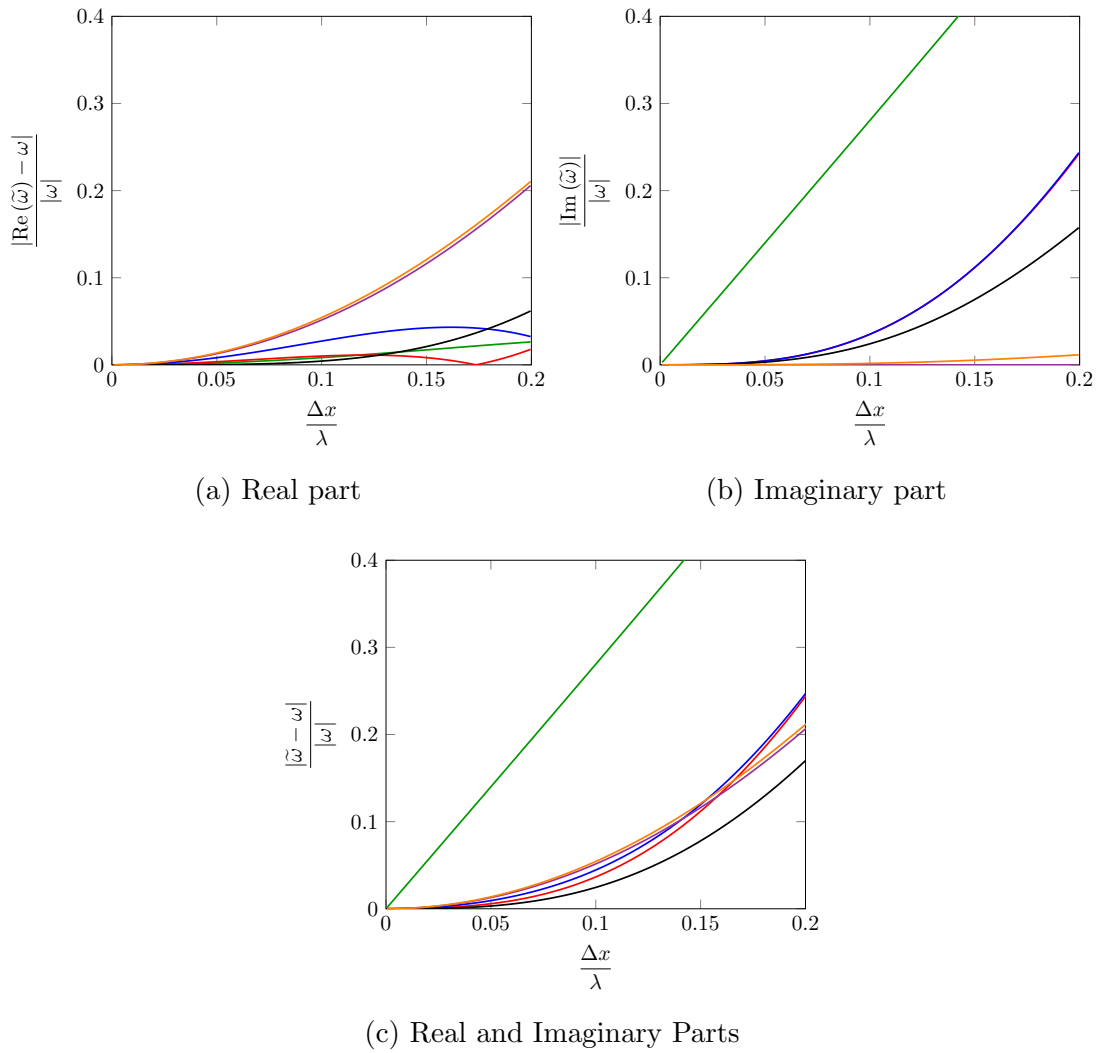


Figure 4.5: Relative dispersion error for first-order FDVM (—), second-order FDVM (—), second-order FEVM (—), third-order FDVM (—),  $\mathcal{D}$  (—) and  $\mathcal{W}$  (—). With  $H = 1m$ ,  $k = 2.5$  and  $U = 1m/s$ .

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

Table 4.9: Table showing lowest order error term for approximating  $\omega_+$  for all FDVM and the FEVM. With  $-\sqrt{gH} \leq U \leq \sqrt{gH}$  and  $\beta = 3 + H^2 k^2$ .

will be the primary driver of the dispersion error.

| Scheme            | Lowest Order $\Delta x$ Term of Error                                    |   |
|-------------------|--|---|
|                   | $U < -\sqrt{gH}$   | $\sqrt{gH} < U$   |
| 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 4.10: Table showing different lowest order spatial error term for approximating  $\omega_+$  for all FDVM and the FEVM for different values of  $U$ . With  $\beta = 3 + H^2 k^2$ .

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

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

# Chapter 5

## Numerical Validation

### 5.1 Measuring Convergence and Conservation

The numerical methods are assessed in this chapter by investigating their convergence and conservation properties. To measure the convergence of these numerical methods we study the difference between numerical and analytic solutions. While conservation is investigated by measuring the total of a conserved quantity in a numerical solution and comparing it to the total of that quantity present in the initial conditions. We introduce notation for these measures and describe their calculation here, beginning with convergence.

#### 5.1.1 Measures 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  $L_1$  vector norm; to compare the numerical and analytic solutions at the numerical grid locations  $x_j$ . For a quantity  $q$ , the vector of its values  $\mathbf{q}$  at the grid locations  $x_j$  and the corresponding numerical solution at those locations  $\mathbf{q}^*$ ; the  $L_1$  norm is

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

When no analytic solution is present, we can compare the distance between numerical solutions to gain some insight into how a sequence of numerical solutions is behaving. This allows us to demonstrate that a sequence of numerical

solutions is convergent to some solution without possessing an analytic solution. To do this the  $L_1$  vector norm is again used as in (5.1) except now both vectors are numerical solutions. Since both numerical solutions will have different grid locations, we only take the difference between the two at the common grid points. We have constructed our grids to accommodate for this, varying  $\Delta x$  by successively dividing by 2. This ensures that the grid locations generated by the larger  $\Delta x$  value are all in the grid generated by the smaller  $\Delta x$  value, and so we can compare both numerical solutions at the grid points generated by the larger  $\Delta x$  value.

### 5.1.2 Measures of Conservation

The conservation properties of the methods can be observed by calculating the total of a conserved quantity in the numerical solution  $\mathcal{C}^*(\mathbf{q}^*)$  and comparing it to the total of that quantity for the initial conditions  $\mathcal{C}(q(x, 0))$ , derived analytically. We do this again using a relative measure,

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

$\mathcal{C}^*(\mathbf{q}^*)$  was calculated using 3 point Gaussian quadrature over the  $j^{th}$  cell and summing these cell integrals for all  $j$ . The three points needed to perform the Gaussian quadrature were calculated by interpolating the  $j^{th}$  cell using a quartic that fits the nodal values  $q_{j-2}$ ,  $q_{j-1}$ ,  $q_j$ ,  $q_{j+1}$  and  $q_{j+2}$ . The Gaussian quadrature using three points is 5<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 the mass, momentum,  $G$  and  $\mathcal{H}$  will be dominated by the error introduced by the numerical solvers.

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

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



## 5.2 Analytic Solutions

The first test of a numerical method is its ability to accurately reproduce the analytic solutions of the governing partial differential equations. For the Serre equations there are only two known analytic solutions; cnoidal waves and a lake at rest. We will use both to assess our numerical methods, using the solitary cnoidal wave solution to assess all methods and a lake at rest solution to assess FEVM<sub>2</sub> and FDVM<sub>2</sub> as only these methods allow for variable bathymetry.

### 5.2.1 Solitary Cnoidal Wave Solution

To assess the numerical schemes ability to reproduce the solitary cnoidal wave solution we used (2.9) with  $a_0 = 1m$ ,  $a_0 = 0.7m$  and  $g = 9.81m/s^2$  at  $t = 0s$  as initial conditions in our numerical methods. 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  $k \in [6, 7, \dots, 19]$ . We satisfied the CFL condition with a CFL number of  $Cr = 0.5$  by setting  $\Delta t = Cr/\sqrt{g(a_0 + a_1)}$ . For FDVM<sub>2</sub> and FEVM<sub>2</sub> we used  $\theta = 1.2$  as the limiting parameter in the generalised minmod limiter.

The parameters  $a_0 = 1m$  and  $a_0 = 0.7m$  were chosen as for these values the non-linearity parameter  $\epsilon = a_1/a_0 = 0.7$  is beneath the breaking threshold for water waves []. The spatial domain was chosen so that there were no effects from the boundary throughout the numerical solution and the final time allowed for many time-steps to be performed.

### Results

An example numerical solution with  $\Delta x = 100/2^{11}m$  from all methods was plotted in Figure 5.1 against the analytic solution at  $t = 50s$ . We have only plotted an illustrative amount of the points in the numerical solution. From these plots it is clear that FDVM<sub>1</sub> performs significantly worse than the higher-order methods at reproducing the analytic solution, even for relatively fine grids where the wave is captured by more than 200 cells. This is primarily due to the numerical diffusion introduced by the method, which has caused the wave in the numerical solution to decrease in amplitude and widen significantly. The higher-order numerical methods all accurately replicate the analytic solution, with insignificant differences in these plots.

The  $L_1$  norm was calculated for  $h$ ,  $u$  and  $G$  for all numerical solutions and

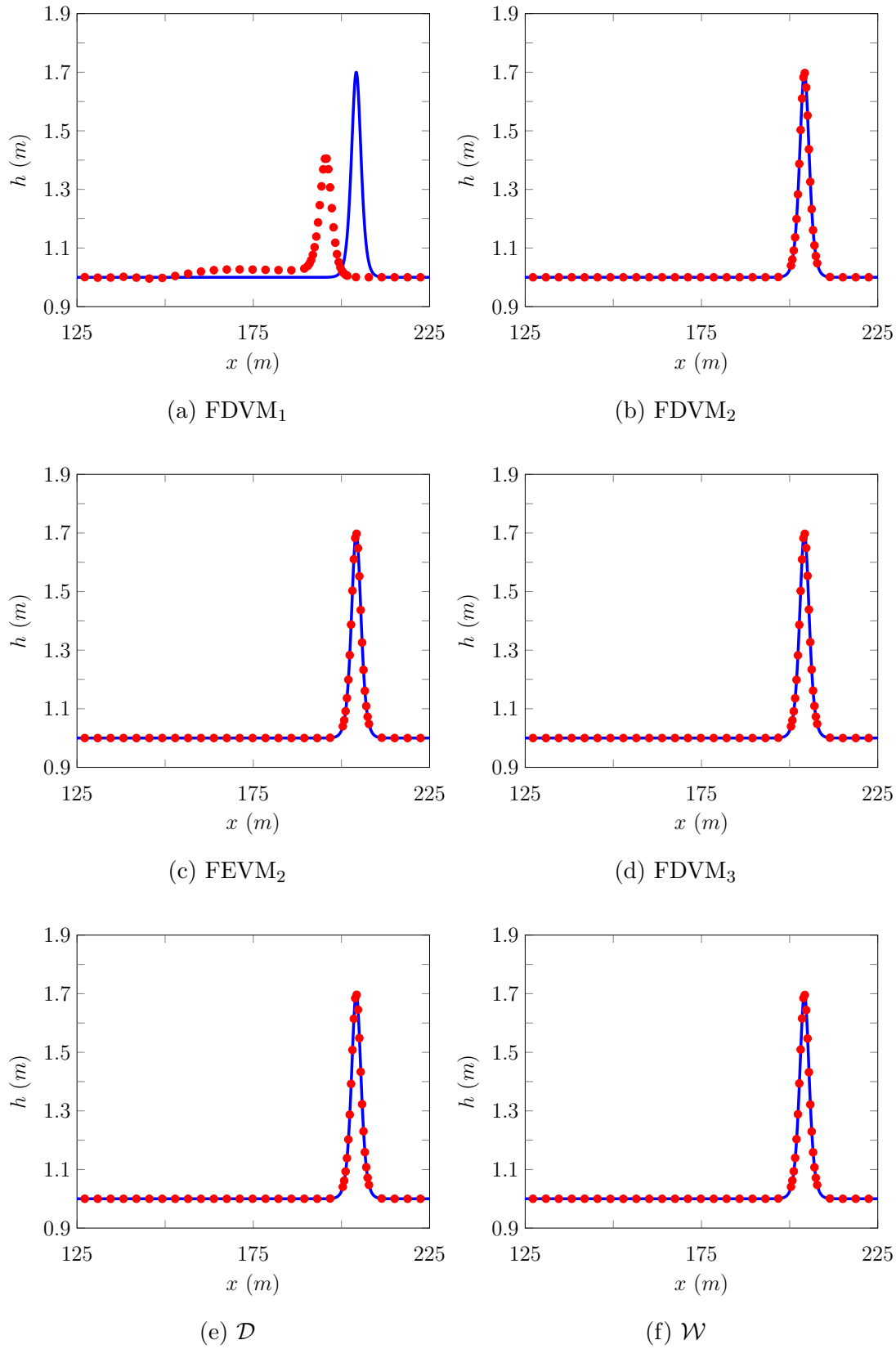


Figure 5.1: Comparison of the analytic solution ( $-$ ) and numerical solution with  $\Delta x = 100/2^{11}m$  ( $\bullet$ ) for the soliton problem at  $t = 50s$  for all methods.

was plotted against  $\Delta x$  for all numerical methods in Figure 5.2. All numerical methods demonstrate convergence for all quantities. Therefore all methods are appropriately reproducing the solitary cnoidal wave solution of the Serre equations. The rate at which the numerical solutions converge to the analytic solution over  $\Delta x$  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 and their order of accuracy from the linear analysis in Chapter 4.

All methods more accurately reproduced the analytic solution for  $h$  than either  $G$  or  $u$  across all  $\Delta x$  values. This is due to the simplicity of the continuity equation (2.4a) compared to the irrotationality equation (2.4b) and the error in  $u$  being dominated by the error in  $G$ .

Increasing the order of accuracy of our numerical methods leads to smaller errors when comparing two methods for the same  $\Delta x$  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, even though the third-order accurate FDVM<sub>3</sub> is an improvement over its second-order counterparts, this improvement is less pronounced than the improvement between first and second-order methods.

For the second-order methods we find that FDVM<sub>2</sub> consistently produces the smallest  $L_1$  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 errors of FEVM<sub>2</sub> being 2 to 4 times larger than FDVM<sub>2</sub>. Both finite difference methods produce very similar errors which are about twice as large as the errors of FEVM<sub>2</sub>.

The  $C_1$  norm was measured for mass ( $h$ ), momentum ( $uh$ ),  $G$  and  $\mathcal{H}$  using the analytic values  $\square$  and plotted in Figure 5.3. From these plots it can be seen that all methods conserve mass at round-off error as long as  $\Delta x$  is not too large, in this case requiring  $\Delta x < 0.5m$ . Since the FD methods perform just as well at conserving mass as the FDVM and FEVM, this suggests that using a finite volume method for the continuity equation (2.4a) is not necessary to conserve mass for this analytic solution. The  $C_1$  norm suggests that FDVM<sub>3</sub> is the worst method for conservation of mass at lower  $\Delta x$  values. However, this is due to the larger number of calculations required by this method leading to large accumulation of round-off errors, rather than a particular deficiency in FDVM<sub>3</sub>.

The conservation of momentum is significantly above round-off error, and for all numerical methods decreases at the rate determined by the order of accuracy or better. For the FDVM and the FEVM this is not surprising as  $u$  is calculated

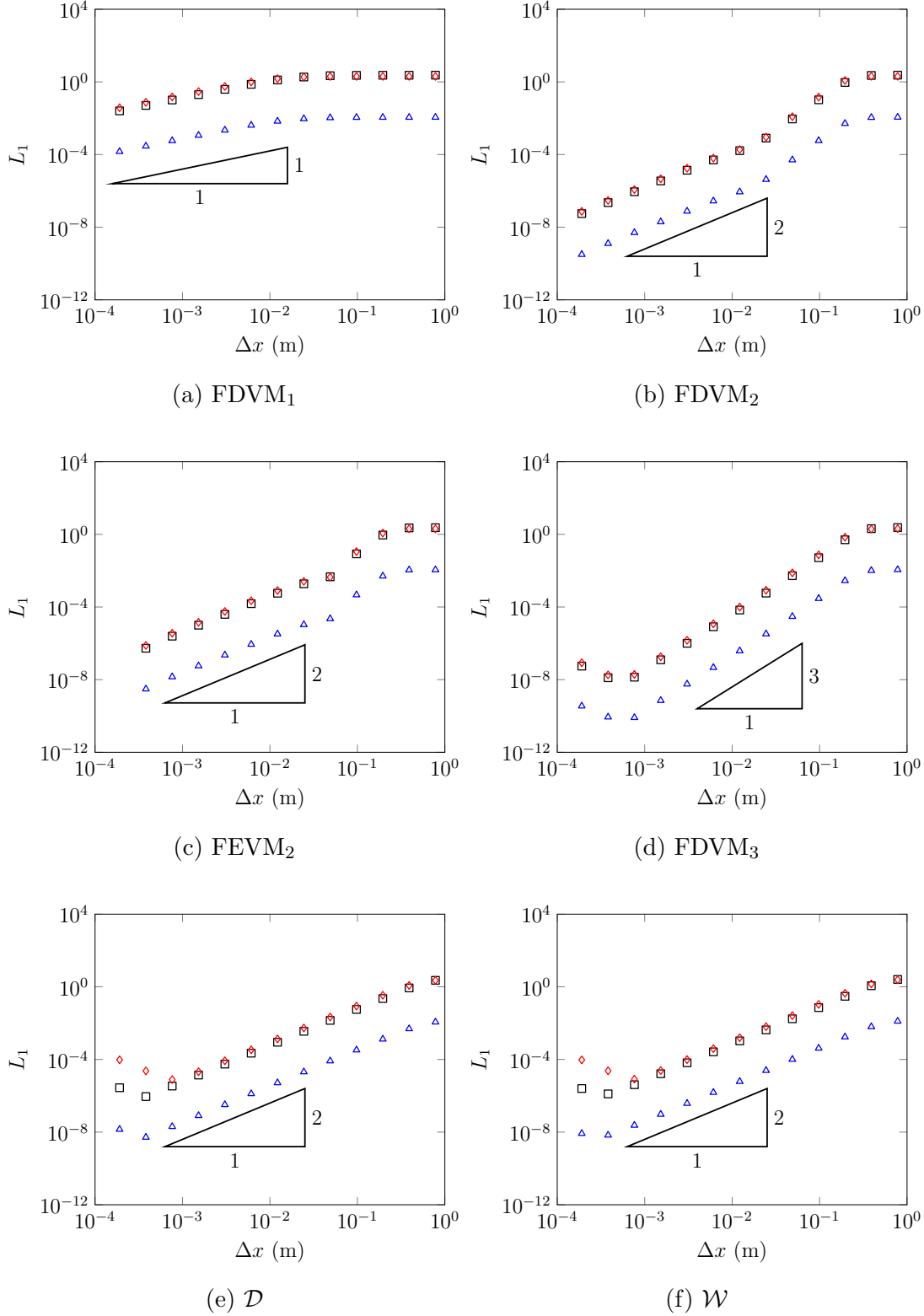


Figure 5.2: Convergence plots as measured by the  $L_1$  norm for  $h$  ( $\triangle$ ),  $u$  ( $\square$ ) and  $G$  ( $\diamond$ ) for the soliton problem for all methods.

from the elliptic equation, which will not necessarily conserve momentum in the system. While the FD methods which solve the momentum equation directly do not conserve momentum as finite difference methods are not necessarily conservative. Given the results for the  $L_1$  norm, one might expect that the  $C_1$  norm would decrease at the order of accuracy of the scheme. However, given that a quantity can be conserved and possess a large  $L_1$  error, for instance translations of the solution, the fact that  $C_1$  decreases faster than the order of accuracy is not surprising.

The second-order FEVM<sub>2</sub> performs the worst of all the methods for the conservation of momentum, even compared to the first-order FDVM<sub>1</sub>. The superiority of the FDVM over the FEVM for conservation of momentum appears to be due to the use of the finite difference method to solve the elliptic equation, as both FDVM<sub>1</sub> and FDVM<sub>2</sub> use the same elliptic solver and both conserve momentum better than FEVM<sub>2</sub>.

For the conservation of  $G$  we see that the case is very similar to the conservation of momentum, except for FEVM<sub>2</sub> which conserves  $G$  at round-off error as long as  $\Delta x$  isn't large. This implies that the FD used to solve the elliptic equation while resulting in smaller  $L_1$  errors and better conservation of momentum, performs worse for the conservation of  $G$ . Therefore, we require a FEM to solve the elliptic equation so that our scheme conserves the conservative quantities  $h$  and  $G$ . There appears to be some trade-off here between the conservation of  $G$  and the conservation of momentum, with the FEVM<sub>2</sub> improving its conservation of  $G$  by worsening its conservation of momentum.

The FD methods conserve  $G$  better than they conserve momentum, this is surprising given that their governing partial differential equations are the conservation of mass and momentum equations (2.1).

The conservation of the Hamiltonian decreases at the rate given by the order of accuracy of the method or better for all methods. Since no methods were designed using the conservation of Hamiltonian equation, its conservation is a good indication that these numerical method are appropriate for the Serre equations.

Surprisingly the two FD methods conserve the Hamiltonian best, followed by FDVM<sub>2</sub>, FEVM<sub>2</sub>, FDVM<sub>3</sub> and FDVM<sub>1</sub>. All second-order methods conserve the Hamiltonian with an order of accuracy greater than 2 while the first and third-order methods conserve the Hamiltonian at their respective order of accuracy. This suggests that even-order schemes have some advantage over odd-order schemes for the conservation of the Hamiltonian.

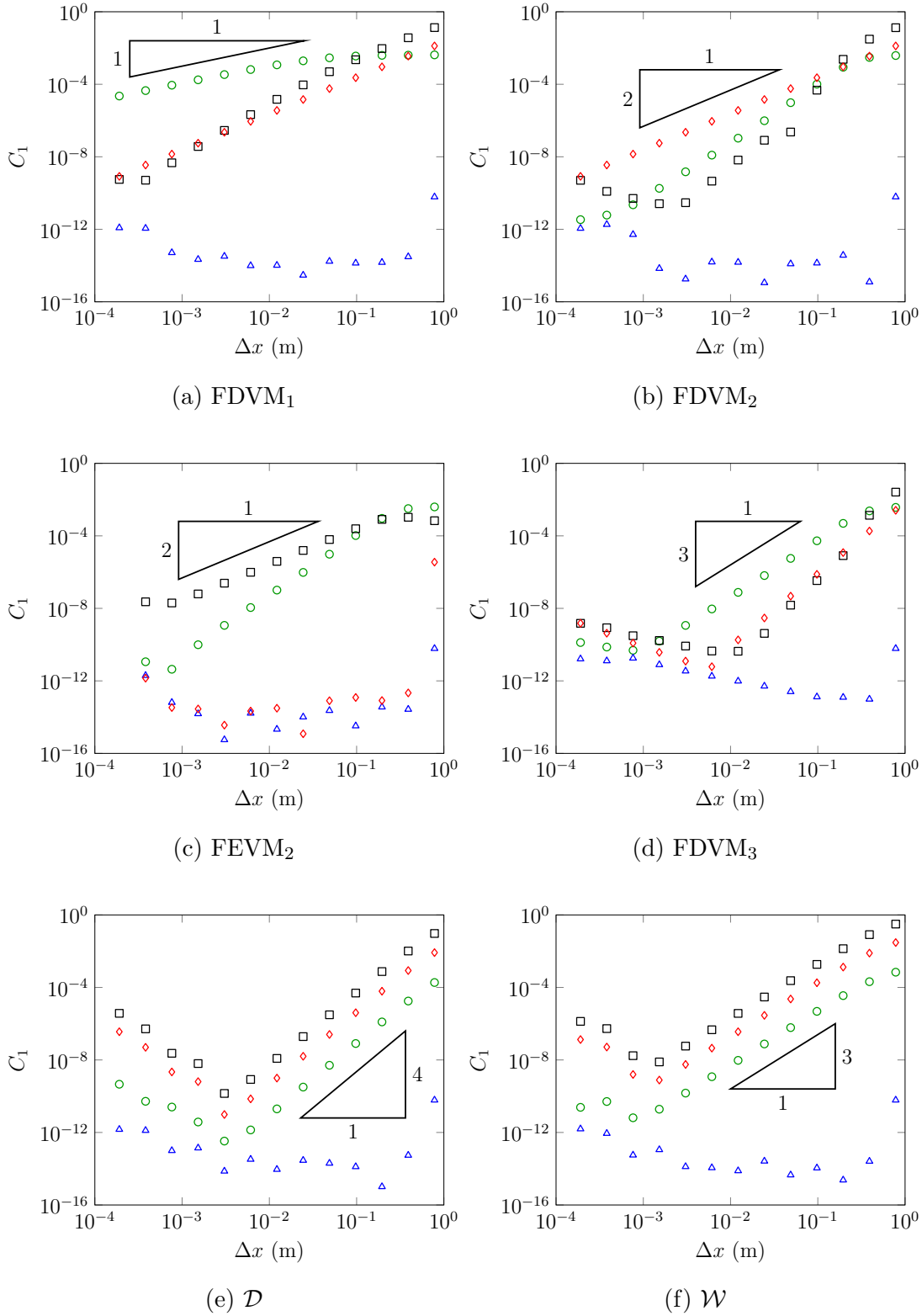


Figure 5.3: Conservation plots as measured by  $C_1$  for  $h$  ( $\triangle$ ),  $uh$  ( $\square$ ),  $G$  ( $\diamond$ ) and  $\mathcal{H}$  ( $\circ$ ) for the soliton problem for all methods.

**5.2.2 Lake at Rest**

**5.3 Forced Solution**

**5.3.1 Travelling Gaussian**

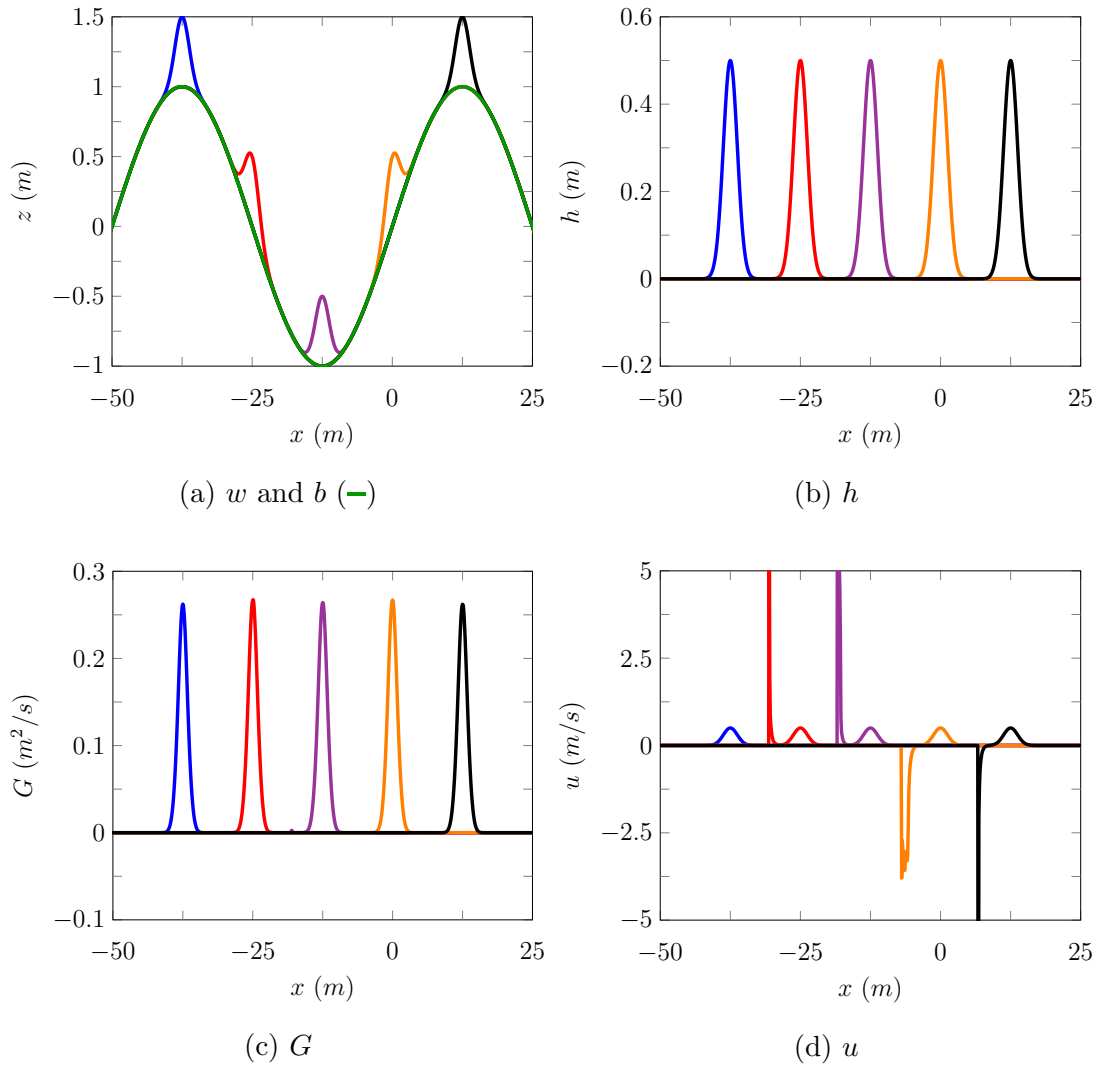


Figure 5.4: Plots of various quantities for the Forced solution at 0s (—), 2.5s (—), 5.0s (—), 7.5s (—), 10.0s (—) .



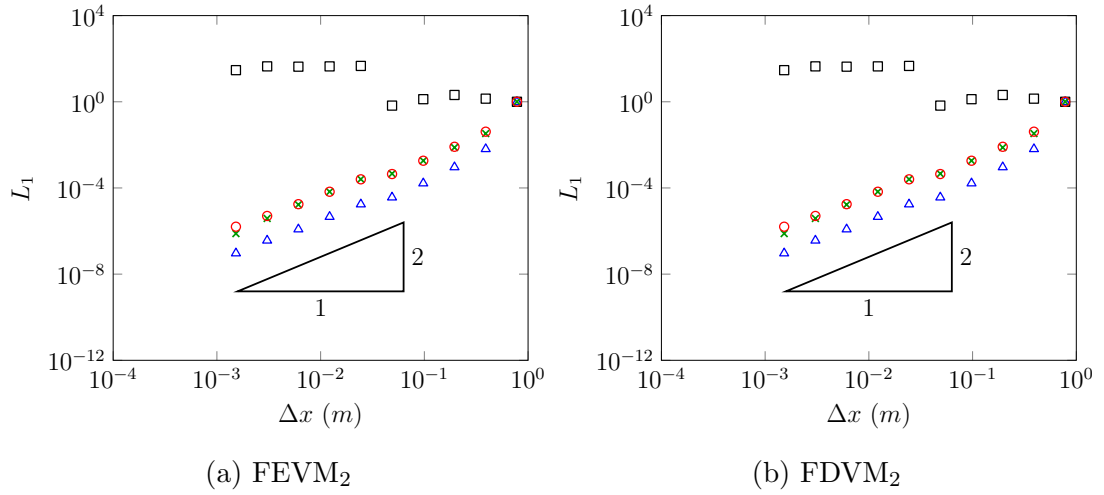


Figure 5.5: Convergence plots as measured by the  $L_1$  norm for  $h$  ( $\triangle$ ),  $u$  ( $\square$ ),  $uh$  ( $\times$ ) and  $G$  ( $\circ$ ) for the forced solution problem for FEVM and FDVM at  $t = 10s$ .

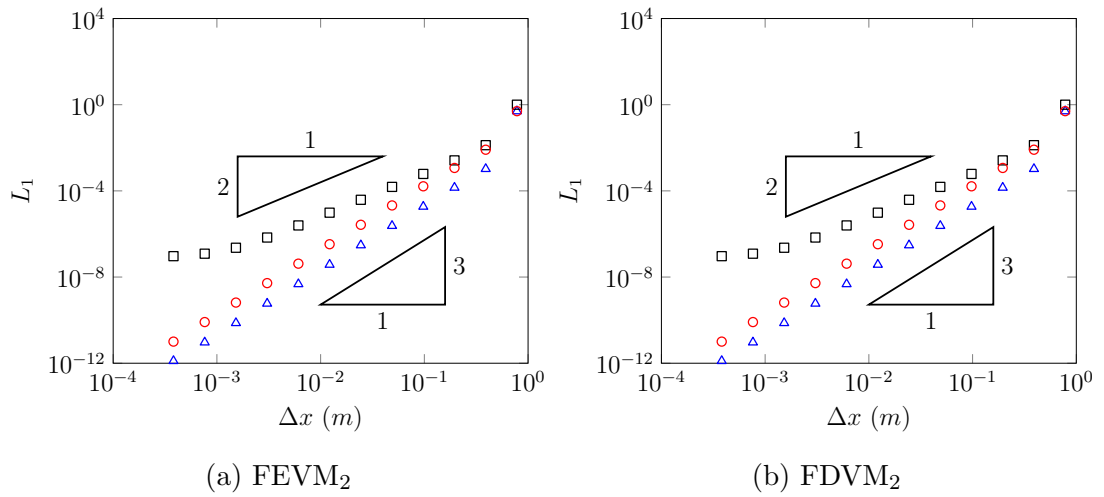


Figure 5.6: Convergence plots as measured by the  $L_1$  norm for  $h$  ( $\triangle$ ),  $u$  ( $\square$ ) and  $G$  ( $\circ$ ) for the forced solution problem for FEVM and FDVM after a single time step.

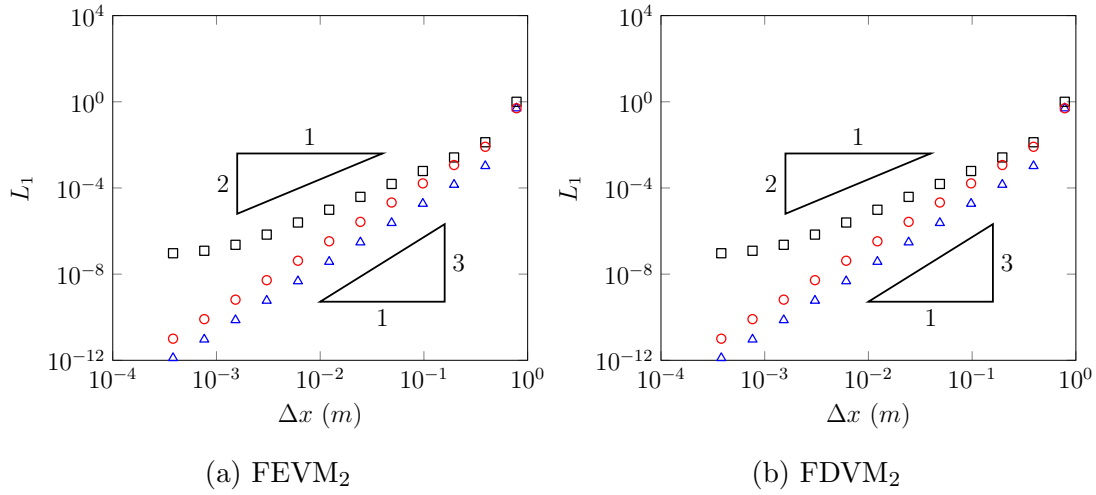


Figure 5.7: Convergence plots as measured by the  $L_1$  norm for  $h$  ( $\triangle$ ),  $u$  ( $\square$ ) and  $G$  ( $\circ$ ) for the forced solution problem for FEVM and FDVM after a single time step.

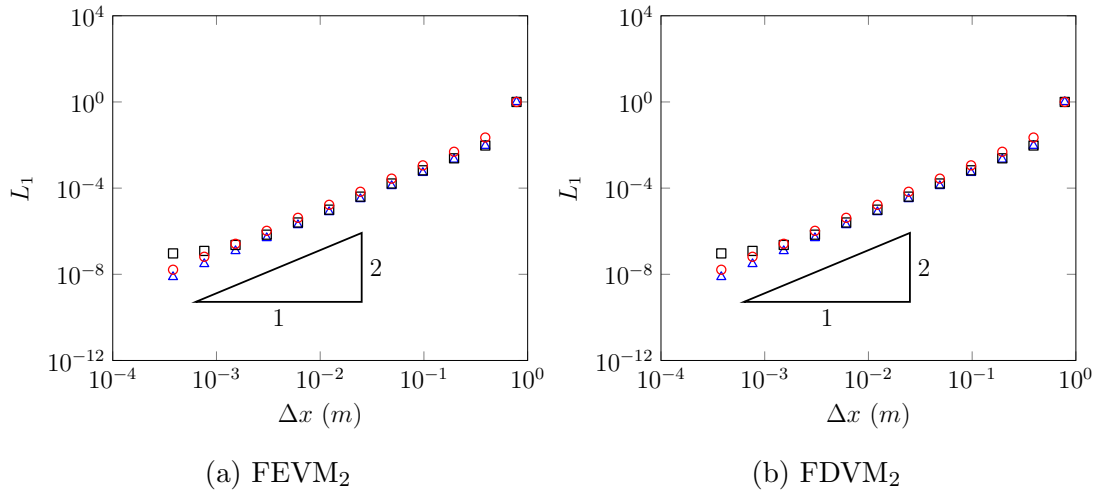


Figure 5.8: Convergence plots as measured by the  $L_1$  norm for  $h$  ( $\triangle$ ),  $u$  ( $\square$ ) and  $G$  ( $\circ$ ) for the for FEVM and FDVM after a single time step.

# Chapter 6

## Experimental Validation

### 6.1 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 [25, 26]. The behaviour of these experiments were mainly driven by the dispersion properties of the waves and their interaction with the variations in bathymetry. Therefore, these experiments serve as a good benchmark for our numerical schemes abilities to accurately model both the effect of variable bathymetry and dispersive waves. For our purposes we will focus on the monochromatic wave experiments of Beji and Battjes [26].

The experiments of Beji and Battjes [26] were conducted in a wave tank  $37.7m$  long,  $0.8m$  wide and  $0.75m$  high. A diagram of the cross section of the length of the wave tank is given in Figure 6.1. There are seven wave gauges numbered from left to right in the diagram with the following locations;  $5.7m$ ,  $10.5m$ ,  $12.5m$ ,  $13.5m$ ,  $14.5m$ ,  $15.7m$  and  $17.3m$ . Waves are generated from a piston-type wave maker located at  $0m$  and travel on the initially still water to the right, over the submerged trapezoidal bar where they are absorbed by the sloped beach.

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

We numerically simulated these experiments over the spatial domain  $[5.7m, 150m]$  with  $\Delta x = 0.1/2^4m$  and  $\Delta t = Sp/2^5$  where  $Sp = 0.039$  is the experimental sampling period. These  $\Delta x$  and  $\Delta t$  values satisfy the CFL condition (3.30) for these

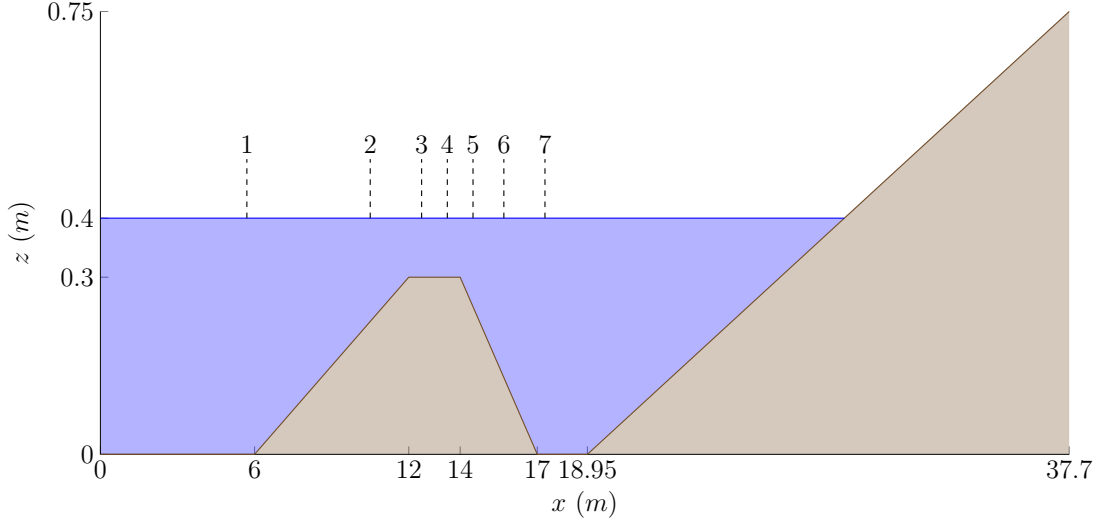


Figure 6.1: Diagram demonstrating the water (■) and the ground (■) for the Beji experiments, with the wave gauge locations marked.

experiments, ensure that the wave gauge locations are all on the spatial grid and that all sampling times occur in the temporal grid. In our numerical experiments only the submerged trapezoidal bar is present, and the sloping beach is replaced with a very long horizontal bed that ensures that we do not observe any boundary effects in our results.

To simulate the incoming waves at the upstream boundary we used the first wave gauge as our left boundary condition for  $h_{-1/2}^n$  while the other height values in the left ghost cell  $h_{-3/2}^n$ ,  $h_{-1}^n$  were calculated by linear extrapolation. The velocity boundary conditions were calculated from the height values using linear wave theory so that

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

Finally the boundary conditions for  $G$  were calculated using the boundary conditions for  $h$  and  $u$ . We shall now present our numerical results for the low and high frequency experiments.

### 6.1.1 Low Frequency Results

A comparison of wave heights of the experimental and numerical results are located in Figures 6.2 and 6.3 for FEVM<sub>2</sub> and Figures 6.4 and 6.5 for FDVM<sub>2</sub>. These numerical schemes both produce identical results for all wave gauges and so this benchmark does not help us discriminate between these two methods.

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

The numerical results for wave gauges 6 and 7 on the leeward side capture some of the wave behaviour but their agreement with the experiments results is much worse. The inadequacy of the numerical results here appears to be due to the discrepancy between the dispersion properties of the Serre equations and water waves, as the numerical solutions of improved dispersion equations [26, 27] accurately reproduce the experimental results on the leeward side.

The dispersion of the Serre equations is vital to recreating the experimental results for wave gauges 2 to 5, as non-dispersive equations such as the SWWE do a very poor job at simulating this experiment [28]. However, to properly reproduce the experimental results on the leeward side of the slope at wave gauges 6 and 7 would require improving the dispersion characteristics of the underlying Serre equations, which can be applied to these numerical methods but is beyond the scope of this thesis.

### 6.1.2 High Frequency Results

The wave heights of the experimental and numerical results are given in Figures 6.6 and 6.7 for FEVM<sub>2</sub>. While the results for FDVM<sub>2</sub> are given in Figures 6.8 and 6.9. As for the low frequency experiment, these numerical schemes FEVM<sub>2</sub> and FDVM<sub>2</sub> produce identical results for all wave gauges at this scale and so this benchmark does not discriminate between these two methods.

As in the low frequency experiment we observe that the numerical results perform well on the windward side of the slope for wave gauges 1 to 4 but perform poorly for the leeward side of the slope for wave gauges 5 to 7. With the high frequency experiment we see the divergence between the numerical and experimental results earlier than the low frequency experiment, so that now wave gauge 5 which is on the leeward side exhibits a significant difference between the numerical and experimental results. As in the low frequency example improving the dispersion properties of the governing partial differential equations lead to a much better agreement between the numerical and experimental results [26, 27]. Because the difference between the dispersion relation of the Serre equations and

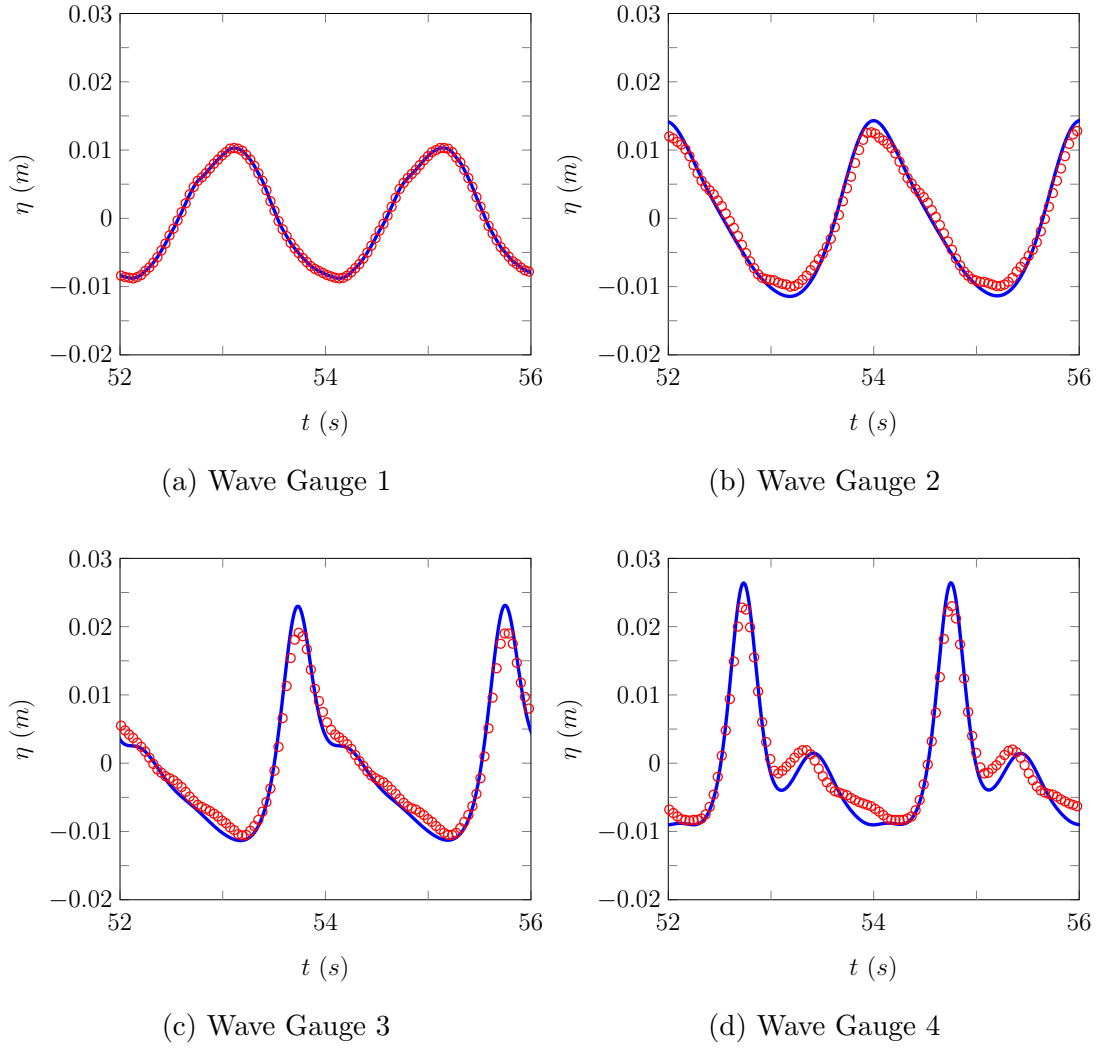


Figure 6.2: Comparison of the wave heights  $\eta$  of the numerical results for the FEVM<sub>2</sub> (—) and the experimental results (○) for wave gauges 1 - 4 for the low frequency experiment.

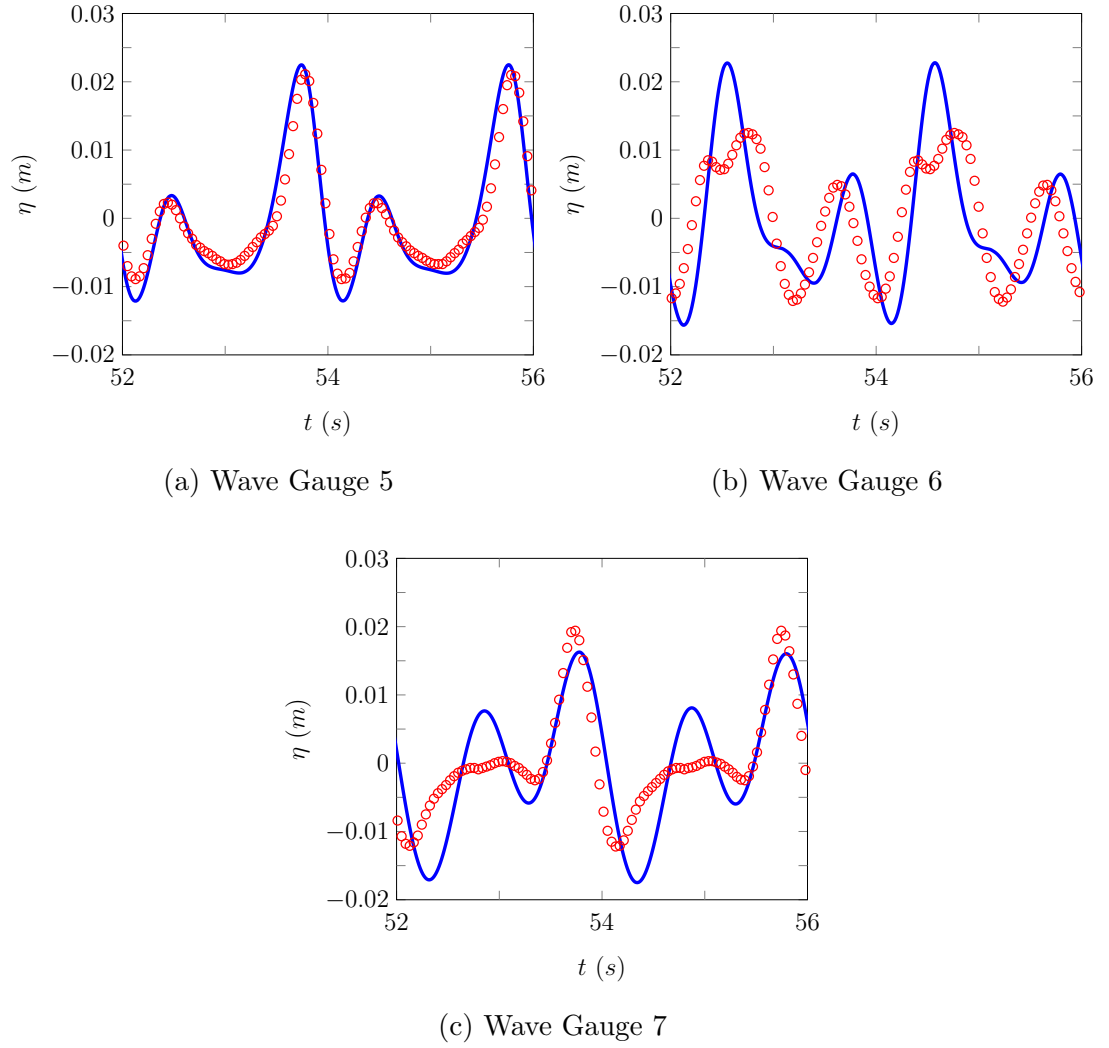


Figure 6.3: Comparison of the wave heights  $\eta$  of the numerical results for the FEVM<sub>2</sub> (—) and the experimental results (○) for wave gauges 5 - 7 for the low frequency experiment.

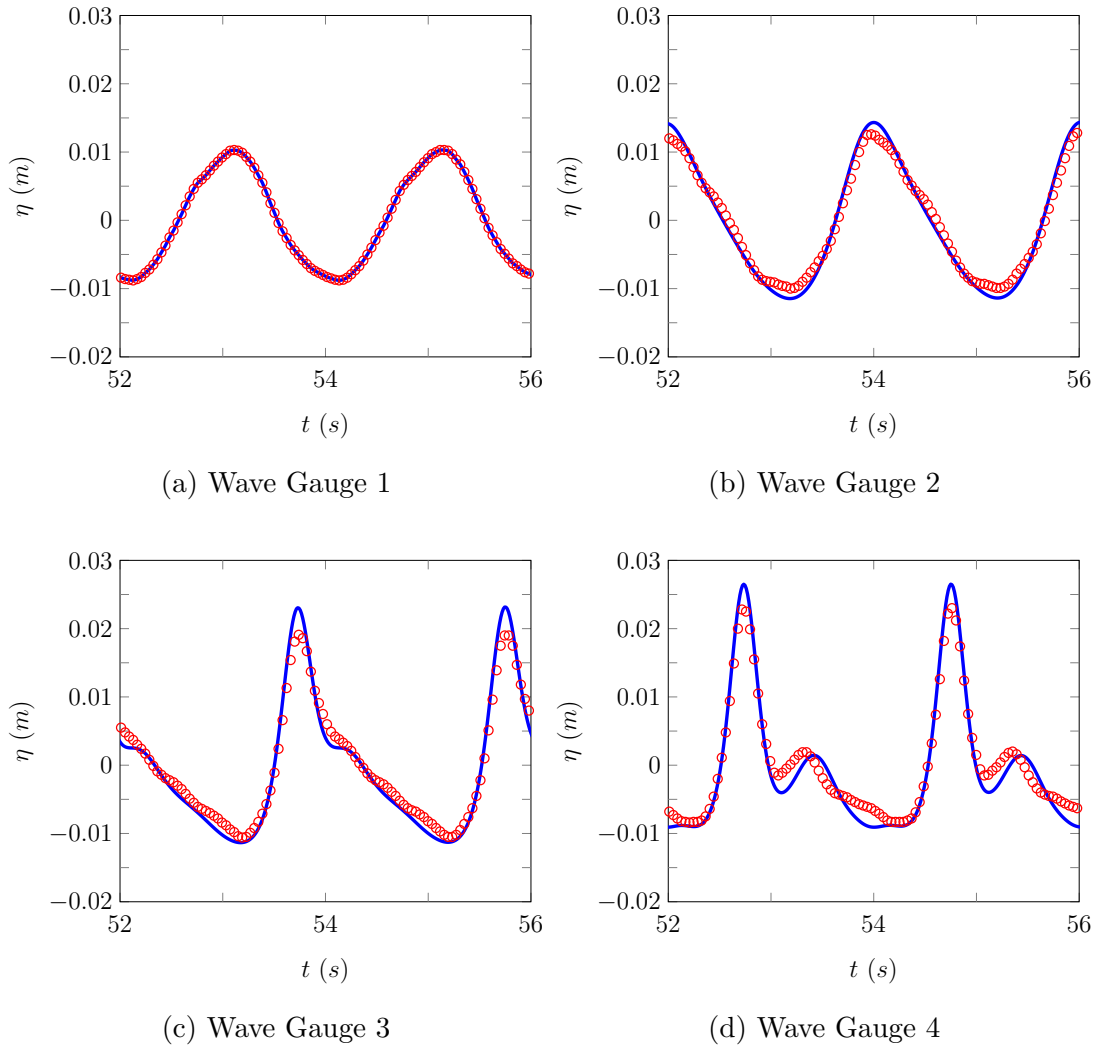


Figure 6.4: Comparison of the wave heights  $\eta$  of the numerical results for the  $\text{FDVM}_2$  (—) and the experimental results ( $\circ$ ) for wave gauges 1 - 4 for the low frequency experiment.



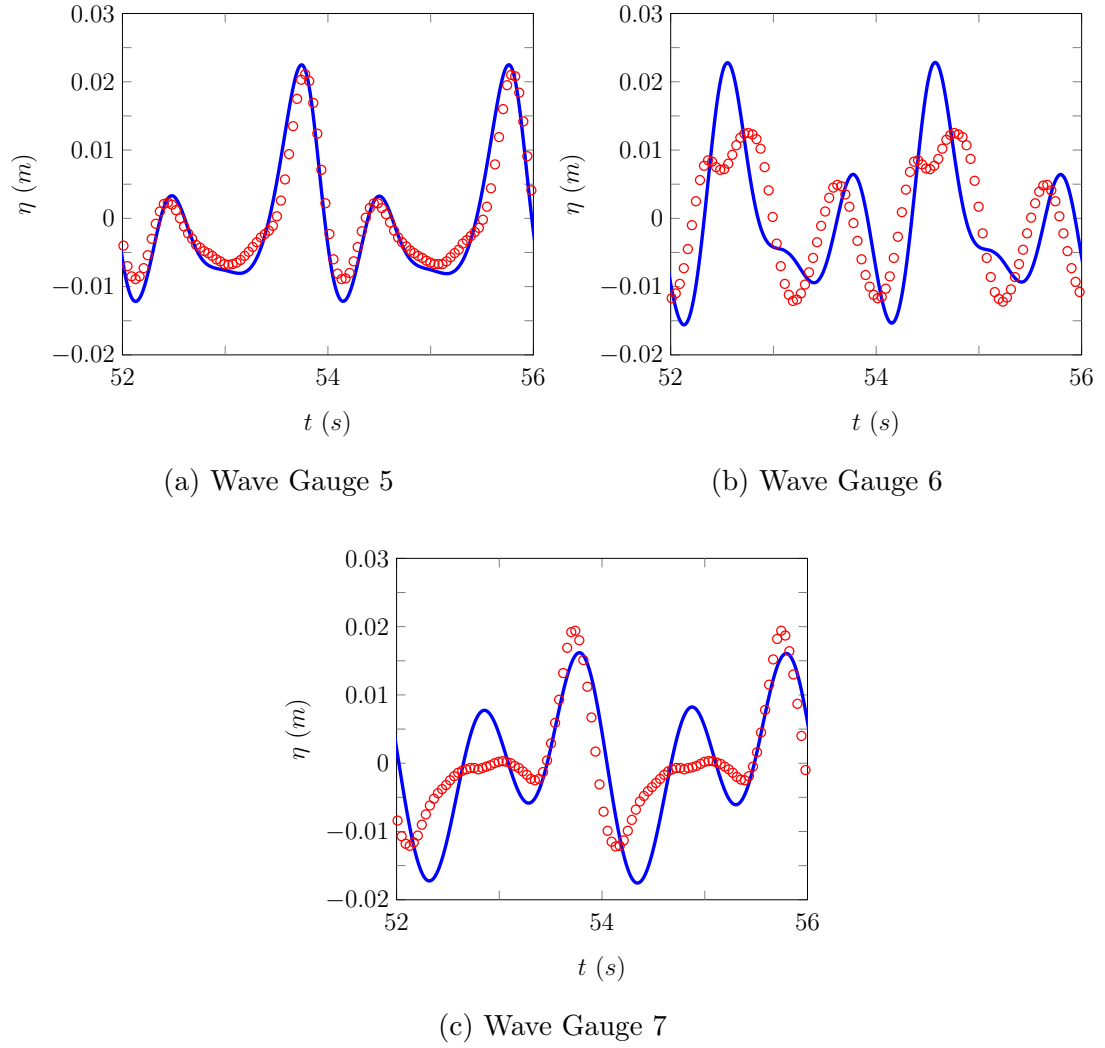


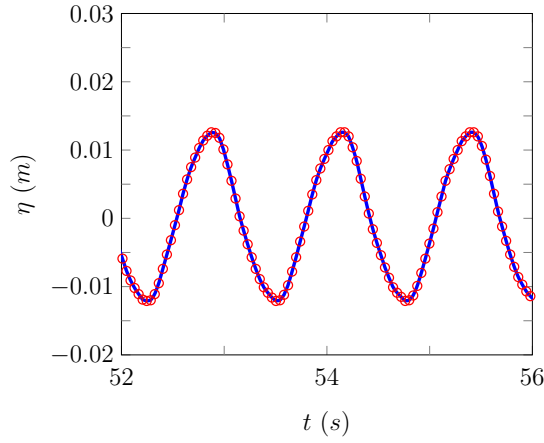
Figure 6.5: Comparison of the wave heights  $\eta$  of the numerical results for the FDVM<sub>2</sub> (—) and the experimental results (○) for wave gauges 5 - 7 for the low frequency experiment.

water waves is largest for higher frequency and therefore shorter waves the earlier divergence between experimental and numerical results is expected.

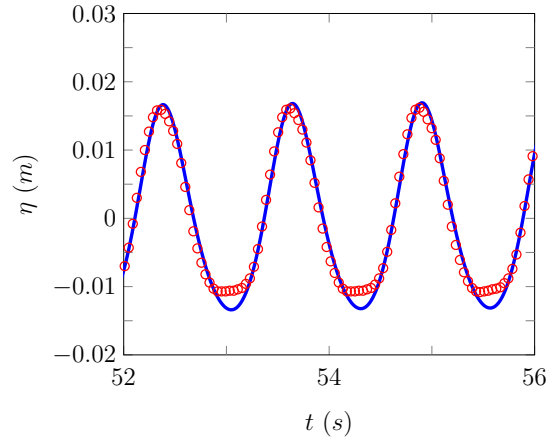
These numerical results for the FDVM<sub>2</sub> and FEVM<sub>2</sub> agree well with other numerical results for weakly dispersive equations without improved dispersion properties for the simulation of periodic waves over a submerged bar in the literature [26, 27, 7, 29]. Therefore without changing the underlying partial differential equations, our numerical methods do very well at recreating the experimental results of Beji and Battjes [26].

## 6.2 Synolakis

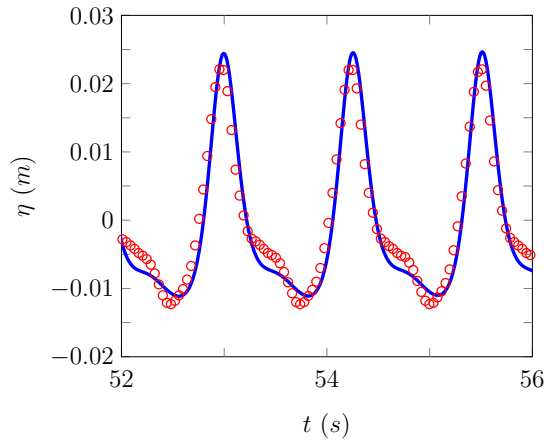
## 6.3 Roeber



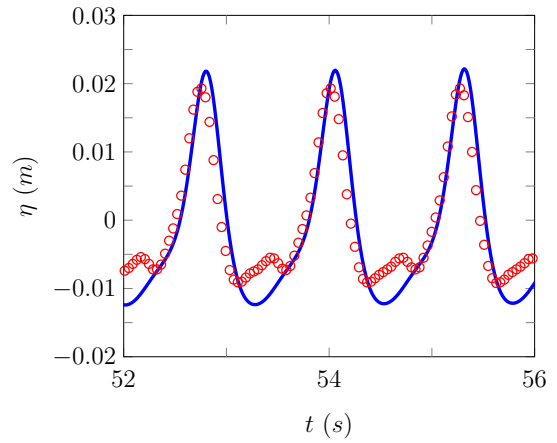
(a) Wave Gauge 1



(b) Wave Gauge 2

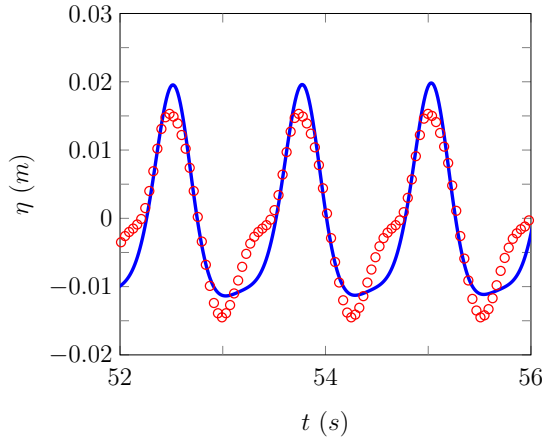


(c) Wave Gauge 3

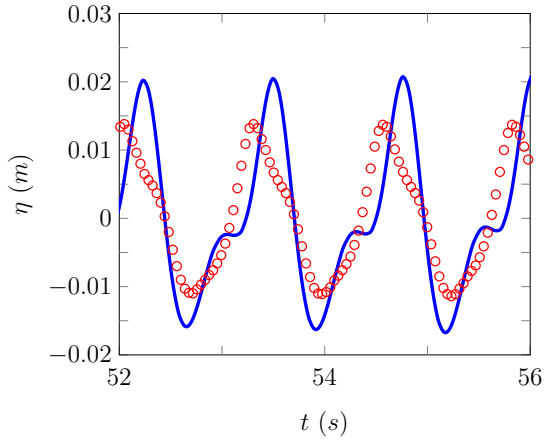


(d) Wave Gauge 4

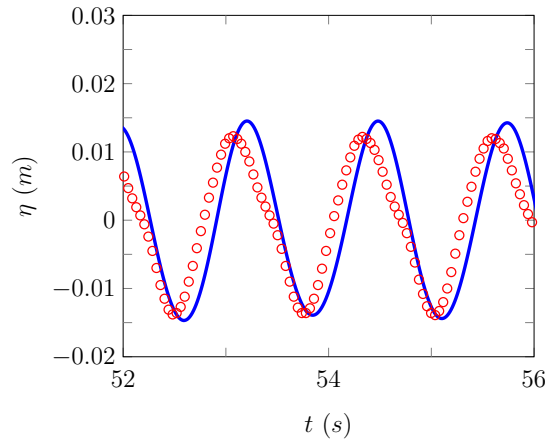
Figure 6.6: Comparison of the wave heights  $\eta$  of the numerical results for the FEVM<sub>2</sub> (—) and the experimental results (○) for wave gauges 1 - 4 for the low frequency experiment.



(a) Wave Gauge 5

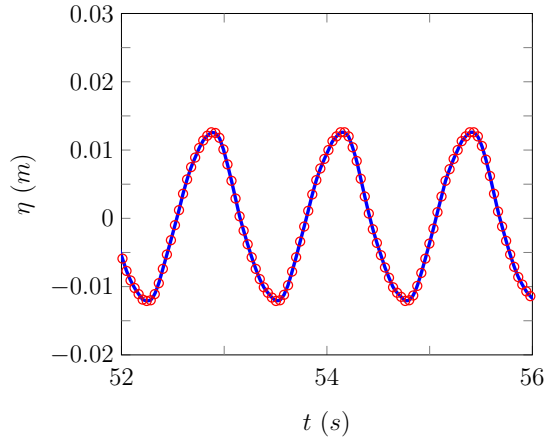


(b) Wave Gauge 6

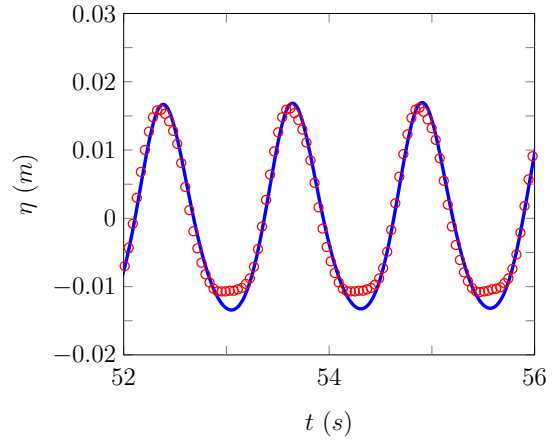


(c) Wave Gauge 7

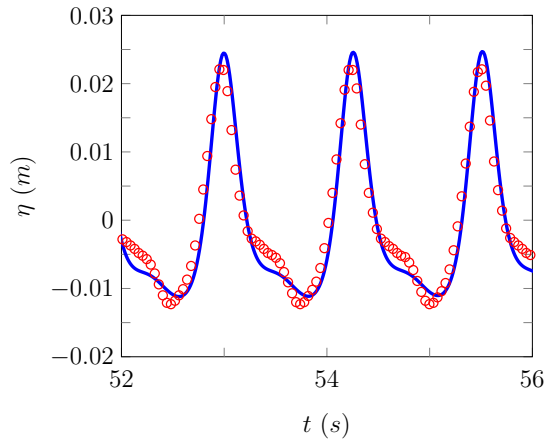
Figure 6.7: Comparison of the wave heights  $\eta$  of the numerical results for the FEVM<sub>2</sub> (—) and the experimental results (○) for wave gauges 5 - 7 for the high frequency experiment.



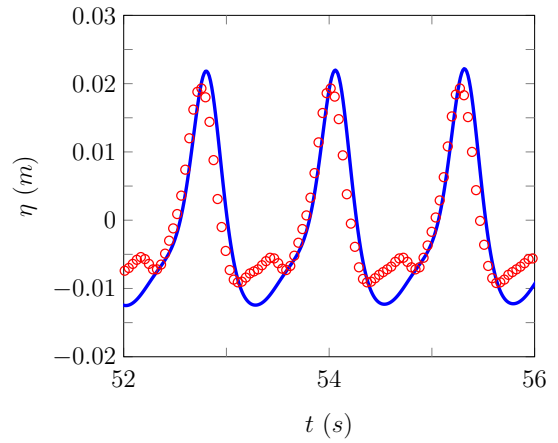
(a) Wave Gauge 1



(b) Wave Gauge 2



(c) Wave Gauge 3



(d) Wave Gauge 4

Figure 6.8: Comparison of the wave heights  $\eta$  of the numerical results for the FDVM<sub>2</sub> (—) and the experimental results (○) for wave gauges 1 - 4 for the high frequency experiment.

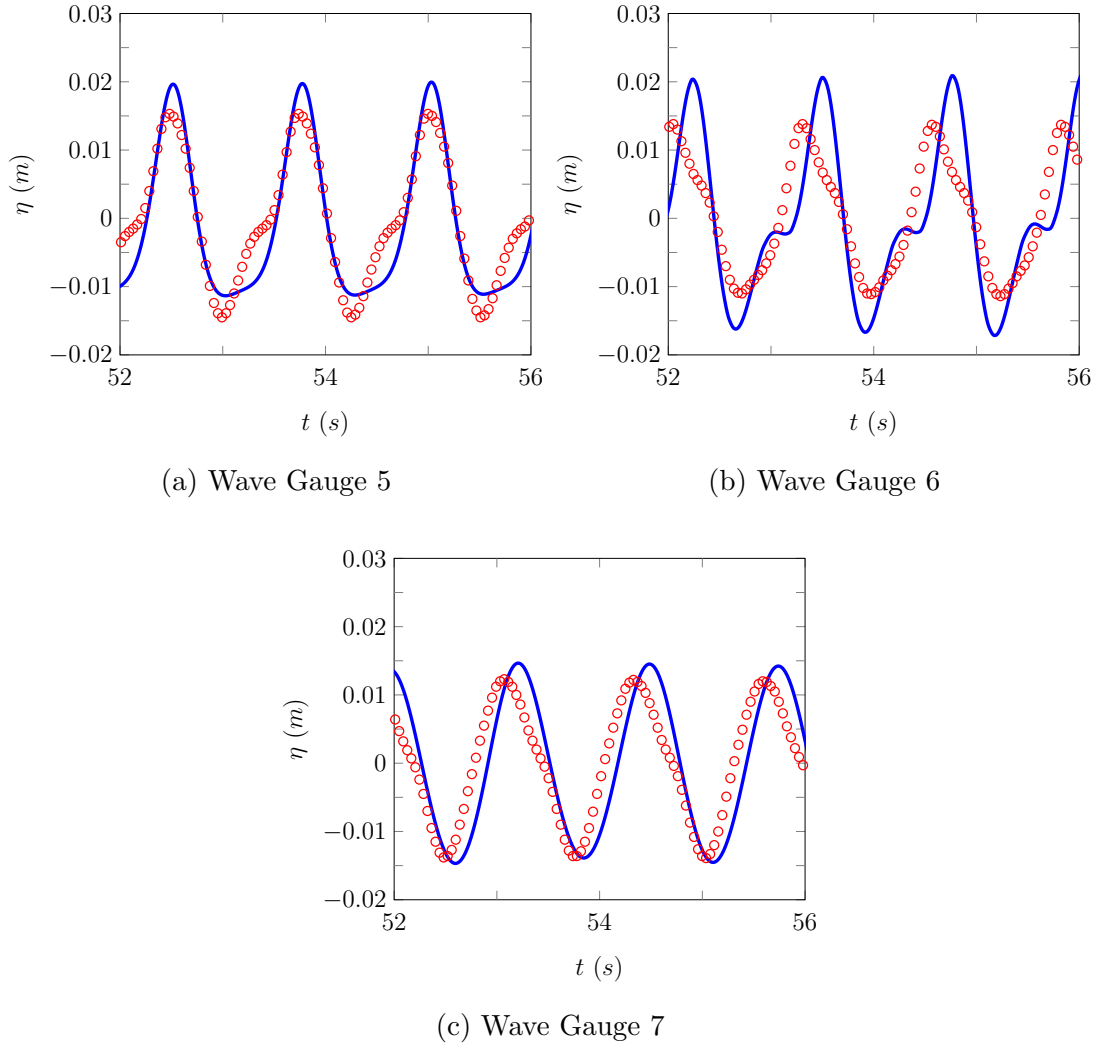


Figure 6.9: Comparison of the wave heights  $\eta$  of the numerical results for the FDVM<sub>2</sub> (—) and the experimental results (○) for wave gauges 5 - 7 for the high frequency experiment.

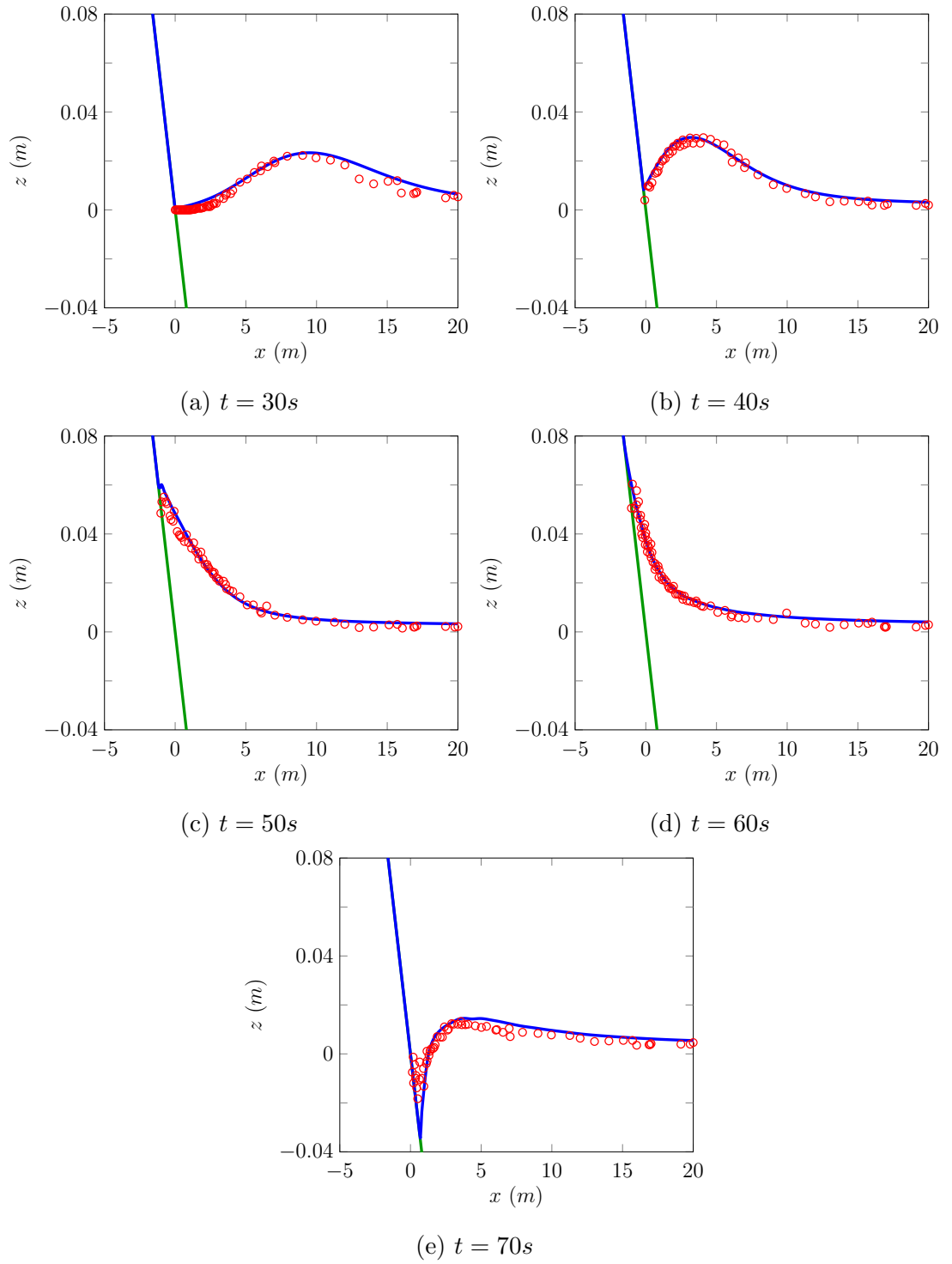


Figure 6.10: FEVM nonbreak times

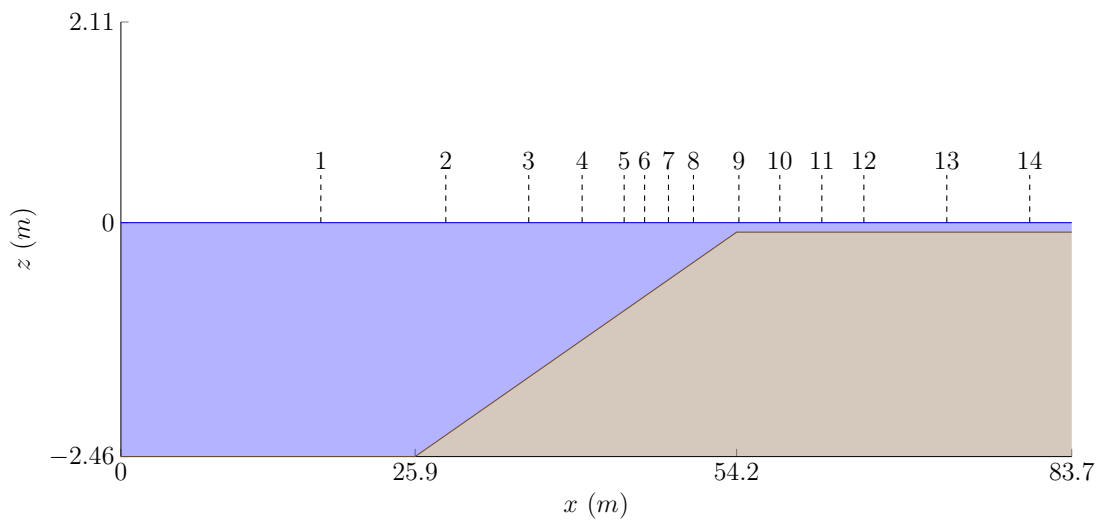
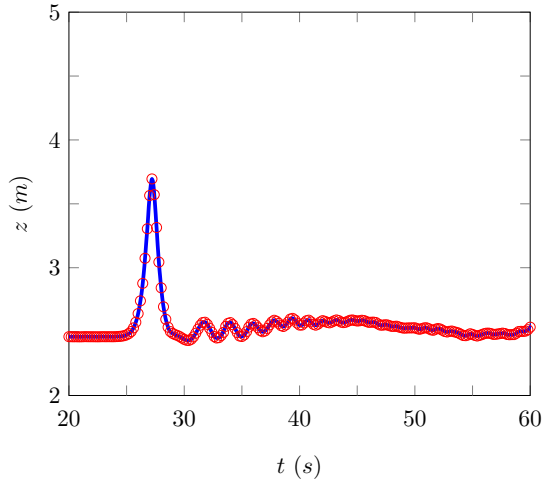
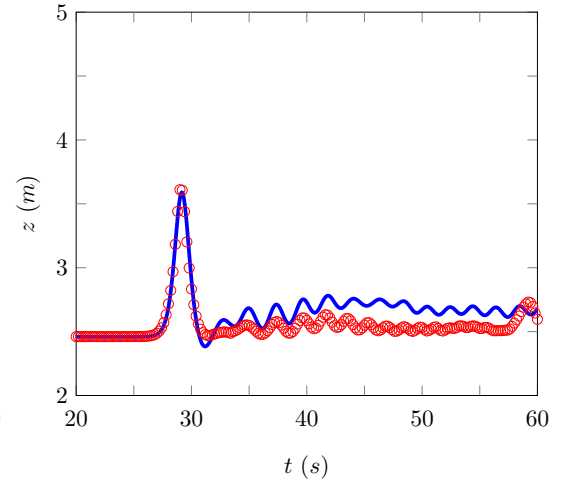


Figure 6.11: Diagram demonstrating the water (■) and the ground (■) for the Beji experiments, with the wave gauge locations marked.

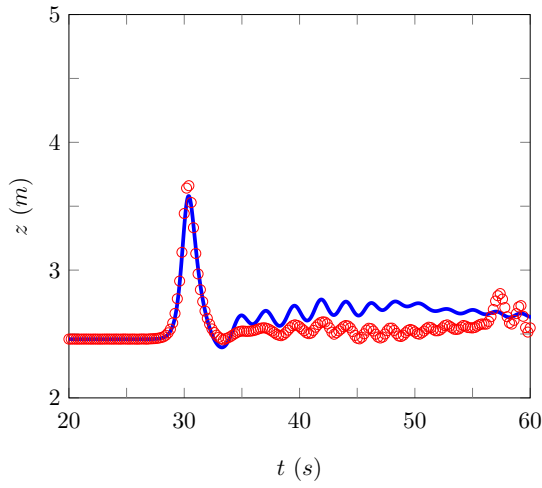




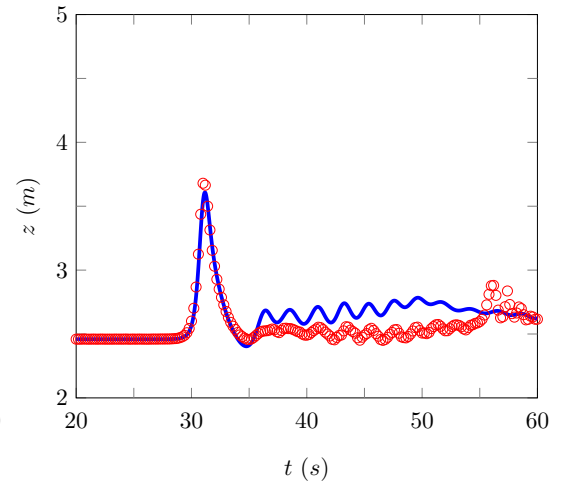
(a) Wave Gauge 1



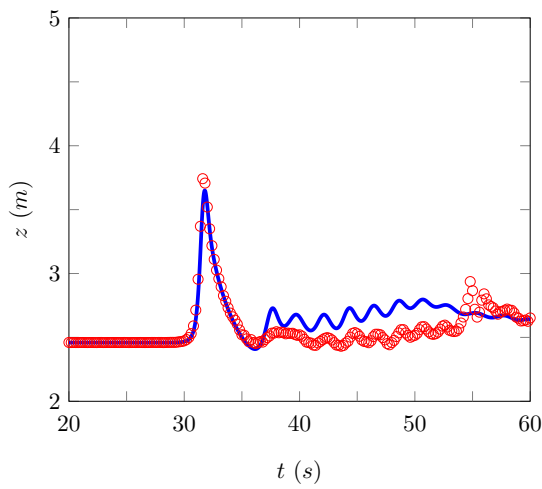
(b) Wave Gauge 2



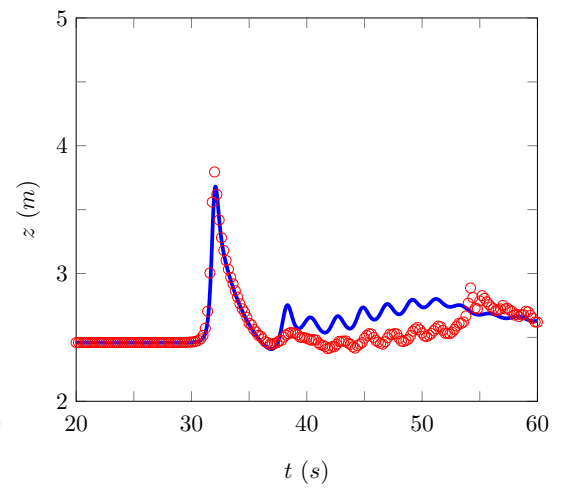
(c) Wave Gauge 3



(d) Wave Gauge 4

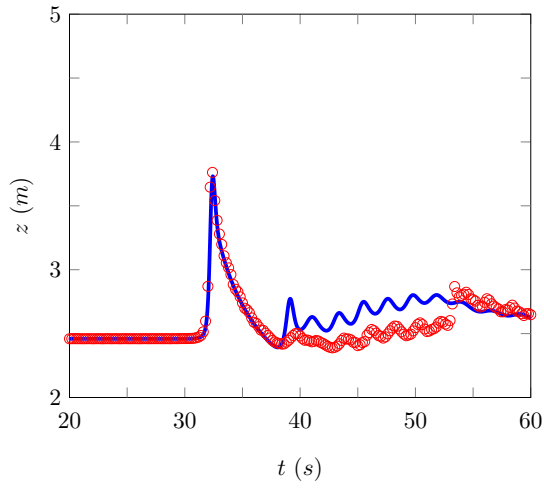


(e) Wave Gauge 5

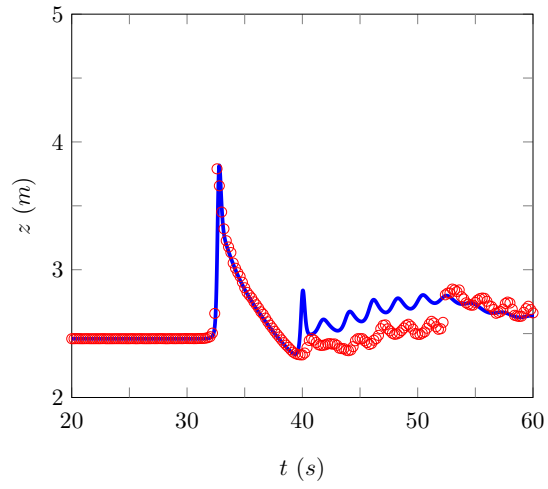


(f) Wave Gauge 6

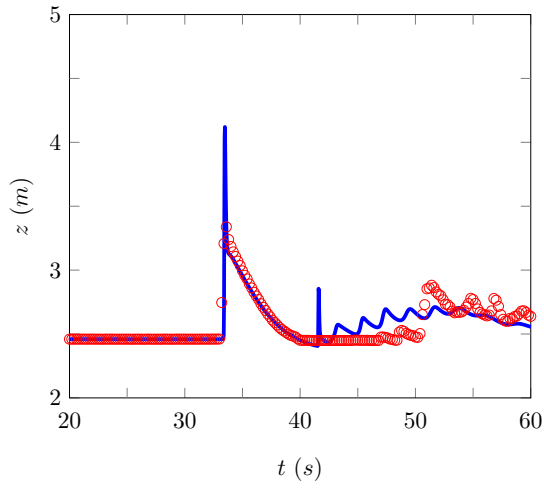
Figure 6.12: .



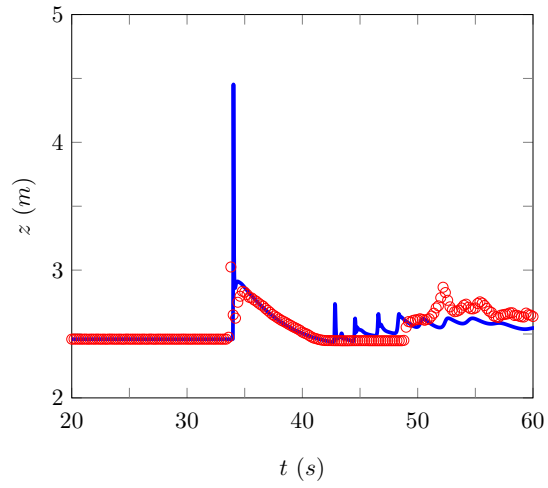
(a) Wave Gauge 7



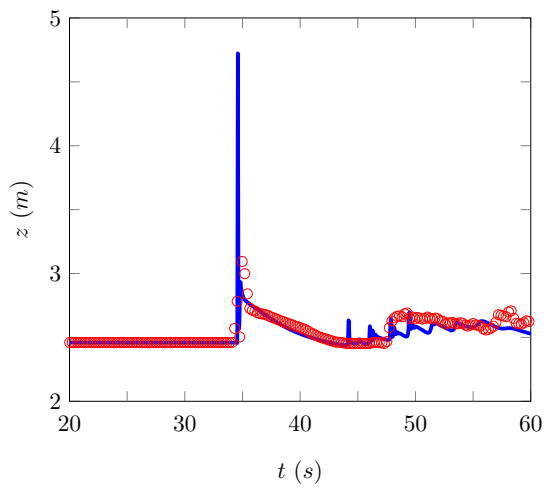
(b) Wave Gauge 8



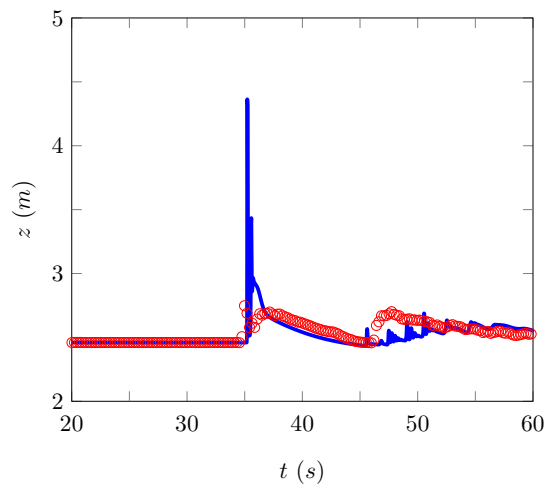
(c) Wave Gauge 9



(d) Wave Gauge 10



(e) Wave Gauge 11



(f) Wave Gauge 12

Figure 6.13: .

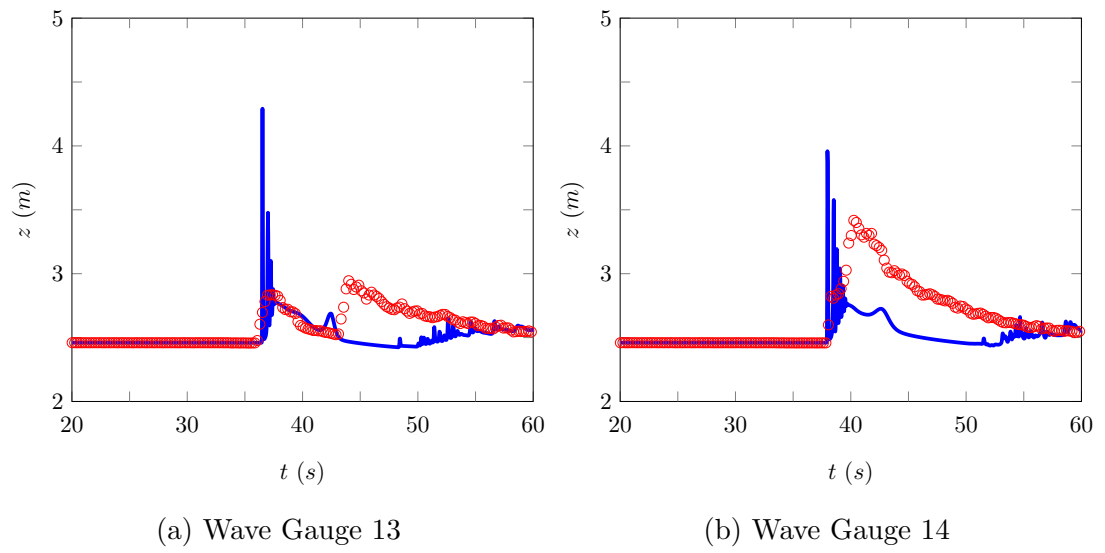


Figure 6.14: .



# Appendix A

## Linear Analysis Results

| Scheme                                  | Expression                         | Lowest Order Term of Error    |
|---|------------------------------------|-------------------------------|
| 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 A.1: Factor  $\mathcal{M}$  from transformation between nodal and cell average values. Where the analytic value is  $\mathcal{M} = \frac{k\Delta x}{2\sin\left(k\frac{\Delta x}{2}\right)}$ .

| Scheme                                  | Formula  | Lowest Order Term of Error  |
|---|--|-----------------------------|
| FDVM <sub>1</sub>                       | $e^{ik\Delta x}$   | $\frac{i}{2}k\Delta x$      |
| FDVM <sub>2</sub> and FEVM <sub>2</sub> | $e^{ik\Delta x} \left(1 - \frac{i\sin(k\Delta x)}{2}\right)$       | $\frac{1}{12}k^2\Delta x^2$ |
| FDVM <sub>3</sub>                       | $\frac{e^{ik\Delta x}}{6} (5 + 2e^{-ik\Delta x} - e^{ik\Delta x})$ | $\frac{i}{12}k^3\Delta x^3$ |

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

| Scheme                                  | Expression   | Lowest Order Term of Error   |
|---|--|------------------------------|
| 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 A.3: Factor  $\mathcal{R}^-$  from reconstruction of  $\eta$  and  $G$  at  $x_{j+1/2}^-$ . Where the analytic value is  $\mathcal{R}^- = e^{ik\Delta x/2} \frac{k\Delta x}{2 \sin(\frac{k\Delta x}{2})}$ .

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

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

| Scheme                                     | Variable   |   |   |  |
|--|--|---|---|--|
|  | $\frac{(1-e^{-ik\Delta x})}{\Delta x} \mathcal{F}^{\eta,\eta}$ | $\frac{(1-e^{-ik\Delta x})}{\Delta x} \mathcal{F}^{\eta,G}$ | $\frac{(1-e^{-ik\Delta x})}{\Delta x} \mathcal{F}^{G,\eta}$ | $\frac{(1-e^{-ik\Delta x})}{\Delta x} \mathcal{F}^{G,G}$ |
| Exact                                      | 0  | $\frac{3ik}{3+H^2k^2}$                                      | $igkH$  | 0  |
| Lowest Order Error Term from Taylor Series |  |   |   |  |
| FDVM <sub>1</sub>                          | $\frac{\sqrt{gH}}{2} k^2 \Delta x$                             | $-\frac{i(6+H^2k^2)}{4(3+H^2k^2)^2} k^3 \Delta x^2$         | $-\frac{igH}{6} k^3 \Delta x^2$                             | $\frac{\sqrt{gH}}{2} k^2 \Delta x$                       |
| FDVM <sub>2</sub>                          | $\frac{\sqrt{gH}}{8} k^4 \Delta x^3$                           | $-\frac{i(6+H^2k^2)}{4H(3+H^2k^2)^2} k^2 \Delta x^2$        | $\frac{igH}{12} k^3 \Delta x^2$                             | $\frac{\sqrt{gH}}{8} k^4 \Delta x^3$                     |
| FEVM <sub>2</sub>                          | $\frac{\sqrt{gH}}{8} k^4 \Delta x^3$                           | $\frac{i(12+5H^2k^2)}{40(3+H^2k^2)^2} k^3 \Delta x^2$       | $\frac{igH}{12} k^3 \Delta x^2$                             | $\frac{\sqrt{gH}}{8} k^4 \Delta x^3$                     |
| FDVM <sub>3</sub>                          | $\frac{\sqrt{gH}}{12} k^4 \Delta x^3$                          | $-\frac{i(243+49H^2k^2)}{960(3+H^2k^2)^2} k^5 \Delta x^4$   | $-\frac{igH}{30} k^5 \Delta x^4$                            | $\frac{\sqrt{gH}}{12} k^4 \Delta x^3$                    |

Table A.5: Spatial error for elements of  $\mathbf{F}$  for all FDVM and the FEVM.



| Scheme            | Lowest Order Term of Error                                       |   |
|-------------------|--|---|
|                   | $\Delta x$   | $\Delta t$  |
| FDVM <sub>1</sub> | $\frac{i\sqrt{gH}}{2}k^2\Delta x$                                | $-\frac{3igH}{6+2H^2k^2}k^2\Delta t$                                    |
| FDVM <sub>2</sub> | $-\frac{\sqrt{3gH}}{8(3+H^2k^2)^{3/2}}k^3\Delta x^2$             | $\frac{\sqrt{3}}{2}\left(\frac{gH}{3+H^2k^2}\right)^{3/2}k^3\Delta t^2$ |
| FEVM <sub>2</sub> | $\frac{\sqrt{3gH}(14+5H^2k^2)}{80(3+H^2k^2)^{3/2}}k^3\Delta x^2$ | $\frac{\sqrt{3}}{2}\left(\frac{gH}{3+H^2k^2}\right)^{3/2}k^3\Delta t^2$ |
| FDVM <sub>3</sub> | $\frac{i\sqrt{gH}}{12}k^4\Delta x^3$                             | $\frac{3ig^2H^2}{8(3+H^2k^2)^2}k^4\Delta t^3$                           |

Table A.6: Table showing lowest order error term for approximating  $\omega_+$  for all FDVM and the FEVM. Where the analytic value is  $\omega_+ = k\sqrt{gH}\sqrt{\frac{3}{3+k^2H^2}}$ . All cross terms,  $\Delta x^a\Delta t^b$  had  $a+b$  larger than the reported terms here.



# Bibliography

- [1] F. Serre. Contribution à l'étude des écoulements permanents et variables dans les canaux. *La Houille Blanche*, 6:830–872, 1953.
- [2] D. Lannes and P. Bonneton. Derivation of asymptotic two-dimensional time-dependent equations for surface water wave propagation. *Physics of Fluids*, 21(1):16601–16610, 2009.
- [3] A. E. Green and P. M. Naghdi. A derivation of equations for wave propagation in water of variable depth. *Journal of Fluid Mechanics*, 78(2):237–246, 1976.
- [4] C. H. Su and C. S. Gardner. Korteweg-de Vries equation and generalisations. III. Derivation of the Korteweg-de Vries equation and Burgers equation. *Journal of Mathematical Physics*, 10(3):536–539, 1969.
- [5] C. Zoppou. *Numerical Solution of the One-dimensional and Cylindrical Serre Equations for Rapidly Varying Free Surface Flows*. PhD thesis, Australian National University, Mathematical Sciences Institute, College of Physical and Mathematical Sciences, Australian National University, Canberra, ACT 2600, Australia, 2014.
- [6] O. Le Métayer, S. Gavriluk, and S. Hank. A numerical scheme for the Green-Naghdi model. *Journal of Computational Physics*, 229(6):2034–2045, 2010.
- [7] M. Li, P. Guyenne, F. Li, and L. Xu. High order well-balanced CDG-FE methods for shallow water waves by a GreenNaghdi model. *Journal of Computational Physics*, 257(1):169–192, 2014.
- [8] W. Choi and R. Camassa. Fully nonlinear internal waves in a two-fluid system. *Journal of Fluid Mechanics*, 396:1–36, 1999.

- [9] J.D Carter and R. Cienfuegos. The kinematics and stability of solitary and cnoidal wave solutions of the serre equations. *European Journal of Mechanics-B/Fluids*, 30(3):259–268, 2011.
- [10] E. Barthélemy. Nonlinear shallow water theories for coastal waves. *Surveys in Geophysics*, 25(3):315–337, 2004.
- [11] 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.
- [12] 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.
- [13] P.L Roe. Characteristic-based schemes for the euler equations. *Annual Review of Fluid Mechanics*, 18(1):337–365, 1986.
- [14] 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.
- [15] L.C. Evans and American Mathematical Society. *Partial Differential Equations*. Graduate Studies in Mathematics. American Mathematical Society, 1998.
- [16] 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.
- [17] 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.
- [18] J. Pitt. *A Second Order Well Balanced Hybrid Finite Volume and Finite Difference Method for the Serre Equations*. PhD thesis, Australian National University, Mathematical Sciences Institute, College of Physical and Mathematical Sciences, Australian National University, Canberra, ACT 2600, Australia, 2014.

- [19] S. Gottlieb, C. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *Society for Industrial and Applied Mathematics Review*, 43(1):89–112, 2001.
- [20] 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.
- [21] C. Zoppou, J. Pitt, and S. Roberts. Numerical solution of the fully non-linear weakly dispersive serre equations for steep gradient flows. *Applied Mathematical Modelling*, 48:70–95, 2017.
- [22] W.H Press, S.A Teukolsky, W.T Vetterling, and B.P Flannery. *Numerical Recipes in C*, volume 2. Cambridge University Press, 1996.
- [23] 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.
- [24] 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.
- [25] S Beji and J.A. Battjes. Experimental investigation of wave propagation over a bar. *Coastal Engineering*, 19(1):151–162, 1993.
- [26] S Beji and J.A. Battjes. Numerical simulation of nonlinear wave propagation over a bar. *Coastal Engineering*, 23(1):1–16, 1994.
- [27] D. Lannes. *The Water Waves Problem: Mathematical Analysis and Asymptotic*, volume 1 of *American Mathematical Society*. Mathematical Surveys and Monographs, 2013.
- [28] 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.
- [29] 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.