Contents

Chapter 1

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 [3]. 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

$$\begin{bmatrix} h \\ G \end{bmatrix}_j^{n+1} = \mathbf{E} \begin{bmatrix} h \\ G \end{bmatrix}_j^{n} \tag{1.1}$$

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 (1.1) 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.

1.1 Linearised Serre equations with horizontal bed

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

$$h(x,t) = H + \delta\eta(x,t) + \mathcal{O}\left(\delta^2\right), \tag{1.2a}$$
$$u(x,t) = U + \delta\upsilon(x,t) + \mathcal{O}\left(\delta^2\right), \tag{1.2b}$$

where $\delta \ll 1$. These waves are relatively small so terms of order δ^2 are negligible. We substitute (1.2) into the Serre equations and neglect terms of order δ^2 to obtain

$$\frac{\partial(\delta\eta)}{\partial t} + H\frac{\partial(\delta\upsilon)}{\partial x} + U\frac{\partial(\delta\eta)}{\partial x} = 0, \tag{1.3a}$$

$$H\frac{\partial(\delta\upsilon)}{\partial t} + gH\frac{\partial(\delta\eta)}{\partial x} + UH\frac{\partial(\delta\upsilon)}{\partial x} - \frac{H^3}{3}\left(U\frac{\partial^3(\delta\upsilon)}{\partial x^3} + \frac{\partial^3(\delta\upsilon)}{\partial x^2\partial t}\right) = 0 \tag{1.3b}$$

and for G

$$G = UH + U\delta\eta + H\delta\upsilon - \frac{H^3}{3}\frac{\partial^2(\delta\upsilon)}{\partial x^2}. \tag{1.3c}$$

These equations can be reformulated in terms of the conserved quantities η and G

$$\frac{\partial\eta}{\partial t} + \frac{\partial}{\partial x}(H\upsilon + U\eta) = 0, \tag{1.4a}$$

$$\frac{\partial G}{\partial t} + \frac{\partial}{\partial x}(UG + UH\upsilon + gH\eta) = 0. \tag{1.4b}$$

where

$$G = UH + U\eta + H\upsilon - \frac{H^3}{3}\frac{\partial^2(\upsilon)}{\partial x^2}. \tag{1.4c}$$

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

1.2 Evolution Matrix

To derive the evolution matrix we first assume that the solutions of the linearised Serre equations with horizontal beds (1.4) are periodic in space and time. In particular, we assume that η and υ are Fourier modes, which for a general quantity q means

$$q(x,t) = q(0,0)e^{i(\omega t + kx)} \tag{1.5}$$

where k is the wavenumber , ω 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

$$q_{j+l}^{n+m} = q_j^n e^{i(m\omega\Delta t + lk\Delta x)}. \tag{1.6}$$

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

1.2.1 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 $\overline{\boldsymbol{\eta}}$ and $\overline{\boldsymbol{G}}$ 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 η and G at the cell midpoint x_j from the cell averages. We also reconstruct η 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

$$\begin{aligned} \eta_j &= \mathcal{M}\left(\overline{\boldsymbol{\eta}}\right), & G_j &= \mathcal{M}\left(\overline{\boldsymbol{G}}\right), \\ \eta^-_{j+1/2} &= \mathcal{R}^-\left(\overline{\boldsymbol{\eta}}\right), & G^-_{j+1/2} &= \mathcal{R}^-\left(\overline{\boldsymbol{G}}\right), \\ \eta^+_{j+1/2} &= \mathcal{R}^+\left(\overline{\boldsymbol{\eta}}\right), & G^+_{j+1/2} &= \mathcal{R}^+\left(\overline{\boldsymbol{G}}\right). \end{aligned}$$

(ii) Calculate the Velocity at the Cell Interface: The remaining unknown quantity, $\upsilon_{j+1/2}$ is calculated from a spatially second-order accurate solution of the elliptic equation (1.4c). This calculation is represented by $\mathcal{G}$ as

$$\upsilon_{j+1/2} = \mathcal{G}\left(H, \boldsymbol{G}, \boldsymbol{\eta}\right).$$

(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 $\mathcal{F}$

$$F_{j+1/2} = \mathcal{F}\left(\eta^-_{j+1/2}, G^-_{j+1/2}, \eta^+_{j+1/2}, G^+_{j+1/2}, \upsilon_{j+1/2}\right).$$

(iv) Time Step Using Forward Euler Approximation: We repeat this process for each cell edge and then apply (**??**) to update the vectors $\overline{\boldsymbol{\eta}}$ and $\overline{\boldsymbol{G}}$ 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 $\overline{\boldsymbol{\eta}}$ and $\overline{\boldsymbol{G}}$ 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 η and υ are Fourier modes. We will then combine these linear operators to derive $\mathbf{E}$ for the second-order FEVM.

(i) Reconstruction of Midpoint and Cell Interface Values

Given $\overline{\boldsymbol{\eta}}$ and $\overline{\boldsymbol{G}}$ at t^n the first step of our numerical method is to calculate η 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 η and G.

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

For the second-order FEVM we use the fact that

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

So to attain second-order accuracy we use

$$q_j = \overline{q}_j = \mathcal{M}\overline{q}_j \tag{1.7}$$

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 η 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 η and G can be written for a general quantity q as

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

Using (1.6) and (1.7) these equations become

$$q^-_{j+\frac{1}{2}} = \overline{q}_j + \frac{-\overline{q}_j e^{-ik\Delta x} + \overline{q}_j e^{ik\Delta x}}{4} = \left(1 + \frac{i\sin\left(k\Delta x\right)}{2}\right)\overline{q}_j = \mathcal{R}^-\overline{q}_j, \tag{1.8a}$$
$$q^+_{j+\frac{1}{2}} = \frac{\overline{q}_j e^{ik\Delta x} + \overline{q}_j + \overline{q}_j e^{2ik\Delta x}}{4} = e^{ik\Delta x}\left(1 - \frac{i\sin\left(k\Delta x\right)}{2}\right)\overline{q}_j = \mathcal{R}^+\overline{q}_j \tag{1.8b}$$

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

(ii) Calculate the Velocity at the Cell Interface

To calculate $\upsilon_{j+1/2}$ we use a second-order FEM. We begin our FEM for (1.4c) with its weak formulation, obtained by multiplying (1.4c) by a test function τ and integrating over the domain Ω

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

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 τ and υ we use the basis functions $\phi_{j-1/2}$, ϕ_j and $\phi_{j+1/2}$ defined in Chapter [], so that τ and our approximation to υ 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

$$\sum_j \int_{x_{j-1/2}}^{x_{j+1/2}} \left(G^+_{j-1/2}\psi^+_{j-1/2} + G^-_{j+1/2}\psi^-_{j+1/2} \right) \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} \, dx =$$

$$\sum_j UH \int_{x_{j-1/2}}^{x_{j+1/2}} \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} \, dx + \sum_j U \int_{x_{j-1/2}}^{x_{j+1/2}} \left(\eta^+_{j-1/2}\psi^+_{j-1/2} + \eta^-_{j+1/2}\psi^-_{j+1/2} \right) \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} \, dx$$

$$+ \sum_j H \int_{x_{j-1/2}}^{x_{j+1/2}} \left(\upsilon_{j-1/2}\phi_{j-1/2} + \upsilon_j\phi_j + \upsilon_{j+1/2}\phi_{j+1/2} \right) \begin{bmatrix} \phi_{j-1/2} \\ \phi_j \\ \phi_{j+1/2} \end{bmatrix} \, dx$$

$$+ \sum_j \frac{H^3}{3} \int_{x_{j-1/2}}^{x_{j+1/2}} \left(\upsilon_{j-1/2} \frac{\partial \phi_{j-1/2}}{\partial x} + \upsilon_j \frac{\partial \phi_j}{\partial x} + \upsilon_{j+1/2} \frac{\partial \phi_{j+1/2}}{\partial x} \right) \begin{bmatrix} \frac{\partial \phi_{j-1/2}}{\partial x} \\ \frac{\partial \phi_j}{\partial x} \\ \frac{\partial \phi_{j+1/2}}{\partial x} \end{bmatrix} \, dx.$$

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

$$\sum_j \frac{\Delta x}{6}\begin{bmatrix} G^+_{j-1/2} \\ 2G^+_{j-1/2}+2G^-_{j+1/2} \\ G^-_{j+1/2}\end{bmatrix} = \sum_j UH\frac{\Delta x}{6}\begin{bmatrix}1\\4\\1\end{bmatrix} + \sum_j \frac{\Delta x}{6}U\begin{bmatrix} \eta^+_{j-1/2} \\ 2\eta^+_{j-1/2}+2\eta^-_{j+1/2} \\ \eta^-_{j+1/2}\end{bmatrix}$$
$$+\sum_j\left(H\frac{\Delta x}{30}\begin{bmatrix}4&2&-1\\2&16&2\\-1&2&4\end{bmatrix}+\frac{H^3}{9\Delta x}\begin{bmatrix}7&-8&1\\-8&16&-8\\1&-8&7\end{bmatrix}\right)\begin{bmatrix}\upsilon_{j-1/2}\\ \upsilon_j\\ \upsilon_{j+1/2}\end{bmatrix}.$$

Using (1.6) and (1.8), we obtain

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

After simplifying

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

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

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

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

From (1.6) $\upsilon_{j+1/2} = e^{ik\frac{\Delta x}{2}}\upsilon_j$ then we have

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

$$= \mathcal{G}^G\overline{G}_j + \mathcal{G}^\eta\overline{\eta}_j + \mathcal{G}^c. \tag{1.9}$$

(iii) Calculate the Flux Across the Cell Interface

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

$$a^-_{j+1/2} = \min\left\{0, U - \sqrt{gH}\right\} \quad \text{and} \quad a^+_{j+1/2} = \max\left\{0, U + \sqrt{gH}\right\}. \tag{1.10}$$

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

Left Supercritical Flow $Fr < -1$:

For left supercritical flow; $Fr < -1$ and therefore $U + \sqrt{gH} < 0$ so we have from (1.10) 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

$$F_{j+\frac{1}{2}} = f\left(q^+_{j+\frac{1}{2}}\right). \tag{1.11}$$

Substituting the flux function from the continuity equation (1.4a) into the Kurganov flux approximation we obtain

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

since υ is continuous and therefore $\upsilon_{j+1/2} = \upsilon^+_{j+1/2} = \upsilon^-_{j+1/2}$. Using the FEM for $\upsilon_{j+1/2}$ (1.9) and the reconstruction (1.8) we have

$$\begin{aligned} F^\eta_{j+\frac{1}{2}} &= H\left(\mathcal{G}^G\overline{G}_j + \mathcal{G}^\eta\overline{\eta}_j + \mathcal{G}^c\right) + U\eta^+_{j+1/2} \\ &= \left(H\mathcal{G}^\eta + U\mathcal{R}^+\right)\overline{\eta}_j + H\mathcal{G}^G\overline{G}_j + H\mathcal{G}^c \\ &= \mathcal{F}^{\eta,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{\eta,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{\eta,c}_{j+\frac{1}{2}} \end{aligned} \tag{1.12}$$

Substituting the flux function for irrotationality equation (1.4b) into the Kurganov flux approximation (1.11) we obtain

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

Using the FEM (1.9) to calculate $\upsilon_{j+1/2}$ and our interface reconstruction (1.8) we have

$$\begin{aligned} F^G_{j+\frac{1}{2}} &= UG^+_{j+1/2} + UH\left(\mathcal{G}^G\overline{G}_j + \mathcal{G}^\eta\overline{\eta}_j + \mathcal{G}^c\right) + gH\eta^+_{j+1/2} \\ &= \left(UH\mathcal{G}^\eta + gH\mathcal{R}^+\right)\overline{\eta}_j + \left(U\mathcal{R}^+ + UH\mathcal{G}^G\right)\overline{G}_j + UH\mathcal{G}^c \\ &= \mathcal{F}^{G,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{G,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{G,c}_{j+\frac{1}{2}} \end{aligned} \tag{1.13}$$

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

$$F_{j+\frac{1}{2}} = \frac{U}{2\sqrt{gH}}\left[f\left(q^-_{j+\frac{1}{2}}\right) - f\left(q^+_{j+\frac{1}{2}}\right)\right] + \frac{1}{2}\left[f\left(q^-_{j+\frac{1}{2}}\right) + f\left(q^+_{j+\frac{1}{2}}\right)\right] + \frac{U^2 - gH}{2\sqrt{gH}}\left[q^+_{j+\frac{1}{2}} - q^-_{j+\frac{1}{2}}\right]. \tag{1.14}$$

Substituting in the flux function from the continuity equation (1.4a) we get

$$F^{\eta}_{j+\frac{1}{2}} = \frac{U}{2\sqrt{gH}}\left[H\upsilon_{j+1/2} + U\eta^-_{j+\frac{1}{2}} - H\upsilon_{j+1/2} - U\eta^+_{j+\frac{1}{2}}\right] + \frac{1}{2}\left[H\upsilon_{j+1/2} + U\eta^-_{j+\frac{1}{2}} + H\upsilon_{j+1/2} + U\eta^+_{j+\frac{1}{2}}\right] + \frac{U^2 - gH}{2\sqrt{gH}}\left[\eta^+_{j+\frac{1}{2}} - \eta^-_{j+\frac{1}{2}}\right]. \tag{1.15}$$

Using the reconstruction factors (1.8) and the elliptic solver (1.9) we get

$$\begin{aligned} F^{\eta}_{j+\frac{1}{2}} &= \left(H\mathcal{G}^{\eta} + \frac{U}{2}\left[\mathcal{R}^- + \mathcal{R}^+\right] - \frac{\sqrt{gH}}{2}\left[\mathcal{R}^+ - \mathcal{R}^-\right]\right)\overline{\eta}_j + H\mathcal{G}^G\overline{G}_j + H\mathcal{G}^c \\ &= \mathcal{F}^{\eta,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{\eta,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{\eta,c}_{j+\frac{1}{2}}. \end{aligned} \tag{1.16}$$

For the flux function of the irrotationality equation (1.4b) the flux approximation (1.14) becomes

$$F^{G}_{j+\frac{1}{2}} = \frac{U}{2\sqrt{gH}}\left[UG^-_{j+\frac{1}{2}} + UH\upsilon_{j+1/2} + gH\eta^-_{j+\frac{1}{2}} - UG^+_{j+\frac{1}{2}} - UH\upsilon_{j+1/2} - gH\eta^+_{j+\frac{1}{2}}\right] + \frac{1}{2}\left[UG^-_{j+\frac{1}{2}} + UH\upsilon_{j+1/2} + gH\eta^-_{j+\frac{1}{2}} + UG^+_{j+\frac{1}{2}} + UH\upsilon_{j+1/2} + gH\eta^+_{j+\frac{1}{2}}\right] + \frac{U^2 - gH}{2\sqrt{gH}}\left[G^+_{j+\frac{1}{2}} - G^-_{j+\frac{1}{2}}\right]. \tag{1.17}$$

By using the reconstruction factors (1.8) and the elliptic solver (1.9) we get

$$\begin{aligned} F^{G}_{j+\frac{1}{2}} &= \left(\frac{U\sqrt{gH}}{2}\left[\mathcal{R}^- - \mathcal{R}^+\right] + UH\mathcal{G}^{\eta} + \frac{gH}{2}\left[\mathcal{R}^- + \mathcal{R}^+\right]\right)\overline{\eta}_j \\ &\quad + \left(UH\mathcal{G}^G + + \frac{U}{2}\left[\mathcal{R}^- + \mathcal{R}^+\right] - \frac{\sqrt{gH}}{2}\left[\mathcal{R}^+ - \mathcal{R}^-\right]\right)\overline{G}_j + UH\mathcal{G}^c \\ &= \mathcal{F}^{G,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{G,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{G,c}_{j+\frac{1}{2}}. \end{aligned} \tag{1.18}$$

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

$$F_{j+\frac{1}{2}} = f\left(q^-_{j+\frac{1}{2}}\right). \tag{1.19}$$

Substituting in the flux function for the continuity equation (1.4a) and the irrotationality equation (1.4b) we obtain

$$\begin{aligned} F^{\eta}_{j+\frac{1}{2}} &= \left(H\mathcal{G}^{\eta} + U\mathcal{R}^-\right)\overline{\eta}_j + H\mathcal{G}^G\overline{G}_j + H\mathcal{G}^c \\ &= \mathcal{F}^{\eta,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{\eta,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{\eta,c}_{j+\frac{1}{2}}, \end{aligned} \tag{1.20}$$

$$\begin{aligned} F^{G}_{j+\frac{1}{2}} &= \left(UH\mathcal{G}^{\eta} + gH\mathcal{R}^-\right)\overline{\eta}_j + \left(U\mathcal{R}^- + UH\mathcal{G}^G\right)\overline{G}_j + UH\mathcal{G}^c \\ &= \mathcal{F}^{G,\eta}_{j+\frac{1}{2}}\overline{\eta}_j + \mathcal{F}^{G,G}_{j+\frac{1}{2}}\overline{G}_j + \mathcal{F}^{G,c}_{j+\frac{1}{2}}. \end{aligned} \tag{1.21}$$

(iv) Time Step Using Forward Euler Approximation

We have obtained the operators for the flux functions for all three cases, supercrticial 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

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

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

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

This can be written in matrix form as

$$\begin{aligned}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n+1} &= \begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n} - \left(1-e^{-ik\Delta x}\right)\frac{\Delta t}{\Delta x}\begin{bmatrix}\mathcal{F}^{\eta,\eta} & \mathcal{F}^{\eta,G}\\ \mathcal{F}^{G,\eta} & \mathcal{F}^{G,G}\end{bmatrix}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n}\\ &= \left(\mathbf{I}-\Delta t\mathbf{F}\right)\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n}\end{aligned} \tag{1.22}$$

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.

(v) Complete Second-Order Time Step Using SSP Runge-Kutta Steps

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

$$\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{1} = \left(\mathbf{I}-\Delta t\mathbf{F}\right)\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n}, \tag{1.23a}$$

$$\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{2} = \left(\mathbf{I}-\Delta t\mathbf{F}\right)\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{1}, \tag{1.23b}$$

$$\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n+1} = \frac{1}{2}\left(\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n} + \begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{2}\right). \tag{1.23c}$$

Substituting (1.23a) and (1.23b) into (1.23c) we can write this in terms of the flux matrix $\mathbf{F}$ and our cell averages at t^n as

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

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

$$\begin{aligned}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n+1} &= \left(\mathbf{I}-\Delta t\mathbf{F}+\frac{1}{2}\Delta t^2\mathbf{F}^2\right)\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n}\\ &= \mathbf{E}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n}\end{aligned} \tag{1.24}$$

which is in the desired form (1.1).

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.

1.3 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 (1.5), 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.

1.3.1 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

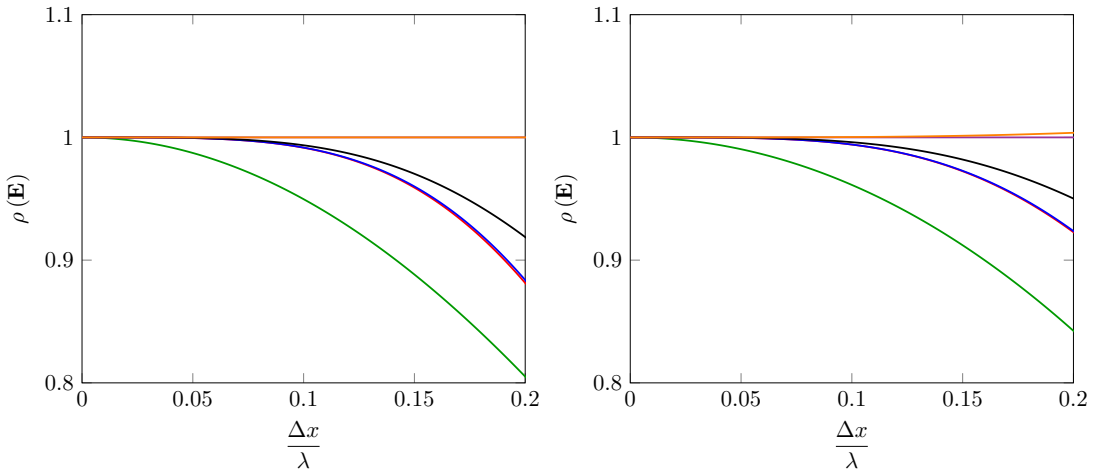
$$\rho\left(\mathbf{E}\right) \leq 1 \tag{1.25}$$

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

We calculated $\rho\left(\mathbf{E}\right)$ numerically for various values of Δx, Δt, k, H and U to check if (1.25) holds. We summarised our results in Figure 1.1 which is a plot of $\rho\left(\mathbf{E}\right)$ against $\Delta x/\lambda$ for representative values of k, H and U. We we used g $= 9.81 m/s^2$ and chose $\Delta t = 0.5/\left(U + \sqrt{gH}\right)\Delta x$ to satisfy the CFL condition []. This is the common choice of Δt in our numerical experiments.

The values $H = 1m$, $k = \frac{\pi}{10} m^{-1}$ and $U = 0m/s$ and $1m/s$ shown in Figure 1.1 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 1.1 it can be seen that all methods have $\rho\left(\mathbf{E}\right) \leq 1$ for $U = 0m/s$ and are therefore stable. The two finite difference methods overlap and have



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

Figure 1.1: 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 Δx 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 Δx and is therefore unstable.

The analytic value of $\rho(\mathbf{E})$ is given by using (1.6) to write

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

Therefore, the analytic growth factor is

$$\rho(\mathbf{E}) = \left|e^{i\omega\Delta t}\right| = \sqrt{\cos^2(\omega\Delta t) + \sin^2(\omega\Delta t)} = 1 \tag{1.26}$$

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 1.1 that $\lim_{\Delta x \to 0} \rho(\mathbf{E}) = 1$ 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 $U = 0m/s$.

1.3.2 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

$$\mathcal{T}^n = \mathbf{E}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^n - \begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n+1}. \tag{1.27}$$

To ensure consistency we must have that $\lim_{\Delta x,\Delta t\to 0}\|\mathcal{T}^n\| = 0$ for all n. Taking the norm of both sides of (1.27) we get

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

Making use of (1.6) we obtain

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

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

$$\|\mathcal{T}^n\| \leq \left\|\mathbf{E} - e^{i\omega\Delta t}\mathbf{I}\right\|\left\|\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^n\right\|. \tag{1.28}$$

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

We calculated the Taylor series of $\mathbf{E} - e^{i\omega_+\Delta t}\mathbf{I}$ for all the numerical methods for all flow scenarios. We have reported the lowest order terms of the Taylor series in Tables 1.1 and 1.2 for the first-order FDVM, Table 1.3 for the second-order FDVM, Tables 1.4 and 1.5 for the third-order FDVM, Table 1.6 for the second-order FEVM, Table 1.7 for $\mathcal{D}$ and Table 1.8 for $\mathcal{W}$. To be concise, we only reported the temporal and spatial errors for the supercritical flow scenarios that where 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 Δt. So we have that for all methods

Element	Lowest Order Term of Error	
	Δx	Δt
$E_{00} - e^{i\omega_+ \Delta t}$	$-\frac{1}{2}\sqrt{gH}k^2\Delta t\Delta x$	$\frac{\sqrt{3gH\beta}+3U}{\beta}ik\Delta t$
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 1.1: 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 $\beta = 3+k^2H^2$.

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

Choosing a particular vector norm such as the L_1 or L_∞ and its induced matrix norm we can see from Tables 1.1-1.8 that $\mathbf{A}$ is independent of Δt and finite so that as $\Delta t \to 0$ we have $|\Delta t|\left(\|\mathbf{A}_0\| + \left\|\begin{bmatrix}\mathcal{O}(\Delta t) & \mathcal{O}(\Delta t)\\ \mathcal{O}(\Delta t) & \mathcal{O}(\Delta t)\end{bmatrix}\right\|\right) \to 0$ and therefore $\left\|\mathbf{E} - e^{i\omega_+\Delta t}\mathbf{I}\right\| \to 0$. Therefore, for all our numerical methods we have $\lim_{\Delta x,\Delta t\to 0}\|\mathcal{T}^n\| = 0$ and so all our numerical methods are consistent for Fourier mode solutions as desired.

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

Table 1.2: 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 1.1 with $\beta = 3 + k^2 H^2$.

Element	Lowest Order Term of Error	
	Δx	Δt
$E_{00} - e^{i\omega_+ \Delta t}$	$-\frac{i\left(27 + 9H^2k^2 + H^4k^4\right)}{12\beta^2} U k^3 \Delta x^2$	$\frac{\sqrt{3gH\beta} + 3U}{\beta} ik\Delta t$
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 1.3: 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 $\beta = 3 + k^2 H^2$.

Element	Lowest Order Term of Error	
	Δx	Δt
$E_{00} - e^{i\omega_+ \Delta t}$	$-\frac{1}{12}\sqrt{gH}k^4\Delta t\Delta x^3$	$\frac{\sqrt{3gH\beta}+3U}{\beta}ik\Delta t$
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 1.4: 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 Δx 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 1.5: 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 1.1 with $\beta = 3+k^2H^2$.

Element	Lowest Order Term of Error	
	Δx	Δt
$E_{00} - e^{i\omega_+ \Delta t}$	$-\frac{i\left(54 + 45H^2k^2 + 10H^4k^4\right)}{120\beta^2} Uk^3 \Delta t \Delta x^2$	$\frac{\sqrt{3gH\beta} + 3U}{\beta} ik\Delta t$
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 1.6: 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	
	Δx	Δt
$E_{00} - e^{i\omega_+ \Delta t}$	$\frac{ik^3}{3} U\Delta t \Delta x^2$	$\sqrt{\frac{3gH}{\beta}} 2ik\Delta t$
E_{01}	$\frac{iHk^3}{3} \Delta t \Delta x^2$	$-2Hik\Delta t$
E_{10}	$\frac{ig\left(3 + \beta\right)}{2\beta^2} k^3 \Delta t \Delta x^2$	$-\frac{6igk}{\beta} \Delta t$
$E_{11} - e^{i\omega_+ \Delta t}$	$\frac{ik^3}{3} U\Delta t \Delta x^2$	$\sqrt{\frac{3gH}{\beta}} 2ik\Delta t$

Table 1.7: 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	
	Δx	Δt
$E_{00} - e^{i\omega_+ \Delta t}$	$\frac{ik^3}{6} U \Delta t \Delta x^2$	$\sqrt{\frac{3gH}{\beta}} ik \Delta t$
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 1.8: Table of the lowest order term of the Taylor series for the elements of $\mathbf{E} - e^{i\omega_+ \Delta t}\mathbf{I}$ for $\mathcal{W}$ with $\beta = 3 + k^2 H^2$.

1.4 Dispersion Analysis

To study the dispersion of our numerical methods we must calculate ω for our numerical methods. Making use of (1.6) in (1.24) we get

$$e^{i\omega\Delta t} \begin{bmatrix} \overline{\eta} \\ \overline{G} \end{bmatrix}_j^n = \mathbf{E} \begin{bmatrix} \overline{\eta} \\ \overline{G} \end{bmatrix}_j^n . \tag{1.29}$$

Assuming that $\mathbf{E}$ has an eigenvalue decomposition $\mathbf{E} = \mathbf{P}^{-1}\mathbf{\Lambda}\mathbf{P}$ and substituting it into (1.29) we get

$$e^{i\omega\Delta t} \begin{bmatrix} \overline{\eta} \\ \overline{G} \end{bmatrix}_j^n = \mathbf{P}^{-1}\mathbf{\Lambda}\mathbf{P} \begin{bmatrix} \overline{\eta} \\ \overline{G} \end{bmatrix}_j^n . \tag{1.30}$$

Left multiplying (1.30) by $\mathbf{P}$ we obtain

$$e^{i\omega\Delta t}\mathbf{P} \begin{bmatrix} \overline{\eta} \\ \overline{G} \end{bmatrix}_j^n = \mathbf{\Lambda}\mathbf{P} \begin{bmatrix} \overline{\eta} \\ \overline{G} \end{bmatrix}_j^n . \tag{1.31}$$

Since Λ 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

$$\widetilde{\omega}_\pm = \frac{1}{i\Delta t} \log\left[\lambda_\pm\right]. \tag{1.32}$$

By comparing $\widetilde{\omega}_{\pm}$ 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 $\widetilde{\omega}_{\pm}$ 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 $g = 9.81m/s^2$ and chose $\Delta t = 0.5/\left(U + \sqrt{gH}\right)\Delta x$ to satisfy the CFL condition [].

In Figures 1.2 and 1.3 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 1.2 and 1.3 we can see that all methods approximate the dispersion relation of the Serre equations well with the approximation improving as $\Delta x \to 0$, as expected.

For the real part of the dispersion error the FEVM and the FDVM outperform the two finite difference methods and therefore will better approximate the speed of waves of the linearised Serre equations. However, for the 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 1.2 and 1.3 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 1.4 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. [1], 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. [1]. Furthermore, we extended their work by allowing for non-zero values of U and examining the imaginary and complete error in the dispersion relation.

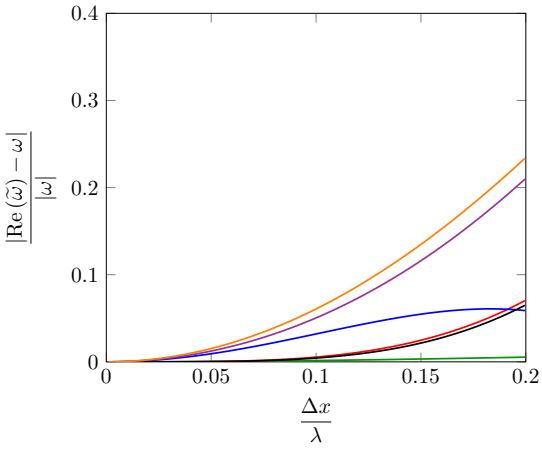
Figure 1.5 demonstrates that the results of the real part of the dispersion error is slightly different if we allow for non-zero values of U. In particular the non-zero value of U changes the real part of the dispersion error for the first-order FDVM, most significantly when $kH = 2.5$. Therefore, for some methods allowing for non-zero values of U can have a significant impact on the conclusions drawn from the dispersion analysis. Furthermore taking the imaginary part of the dispersion error into account is important as ω determines not only the speed of waves but also their amplitude. In particular it is possible that a method like the first-order FDVM performs very well for the real part of the dispersion error and poorly for the imaginary part, leading to false conclusions about the accuracy of the method.

The Taylor series expansion of $\widetilde{\omega}$ was also derived for all the numerical methods. We have compiled the lowest order terms of the Taylor series for $\widetilde{\omega}_+ - \omega_+$ in Table 1.9 when $-\sqrt{gH} \leq U \leq \sqrt{gH}$ for the FDVM and FEVM. In Table 1.9 it is clear that these schemes estimated ω with the expected order of accuracy in both space and time.

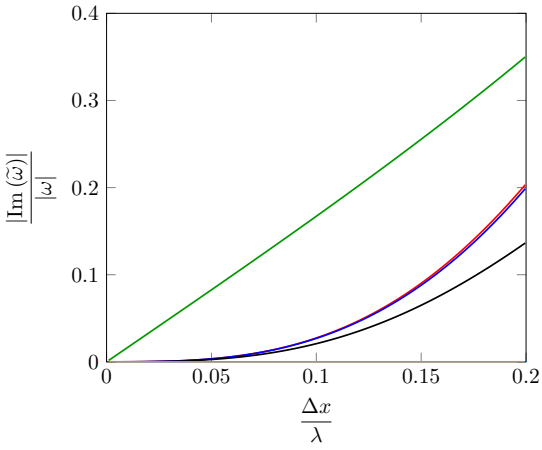
We also present the lowest order terms of the Taylor series for $\widetilde{\omega}_+ - \omega_+$ for both $U < -\sqrt{gH}$ and $U > \sqrt{gH}$ in Table 1.10. We only present the errors that are different from those reported in Table 1.9, this was only the case for the spatial error of the odd-order numerical methods. We can see that for all the flow scenarios that our FDVM and the FEVM have the correct order of accuracy when approximating ω_+.

Finally we present the lowest order terms of the Taylor series for $\widetilde{\omega}_+ - \omega_+$ for the finite difference methods in Table 1.11. 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.

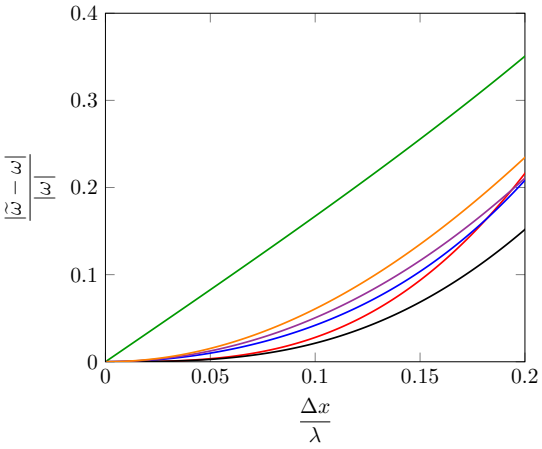
Because all methods were demonstrated to have the expected order of accuracy in approximating ω_+ this implies that for small Δx values the order of accuracy



(a) Real Part

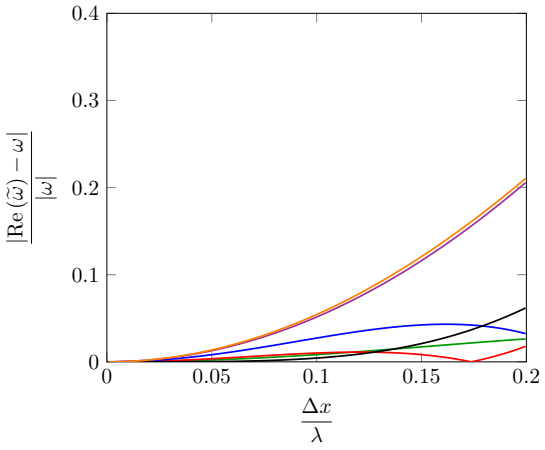


(b) Imaginary Part

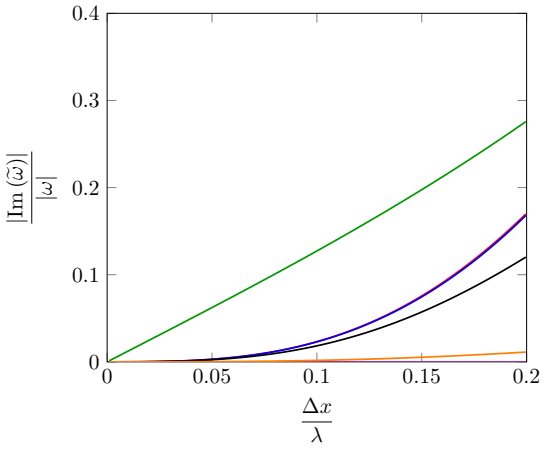


(c) Real and Imaginary Parts

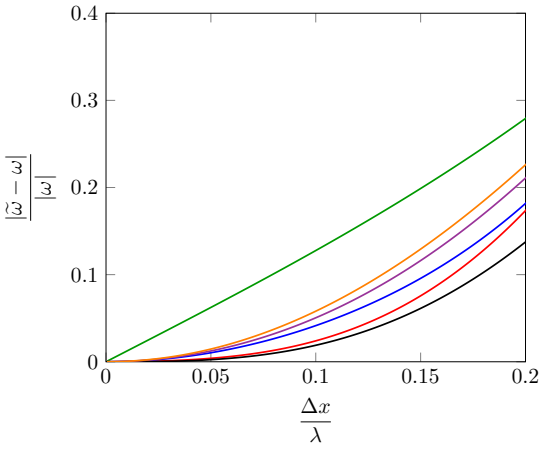
Figure 1.2: Relative dispersion error for first-order FDVM (–), second-order FDVM(–), second-order FEVM (–), third-order FDVM (–), $\mathcal{D}$ (–) and $\mathcal{W}$ (–). With $H = 1m$, $k = \frac{\pi}{10}$ and $U = 0m/s$.



(a) Real part

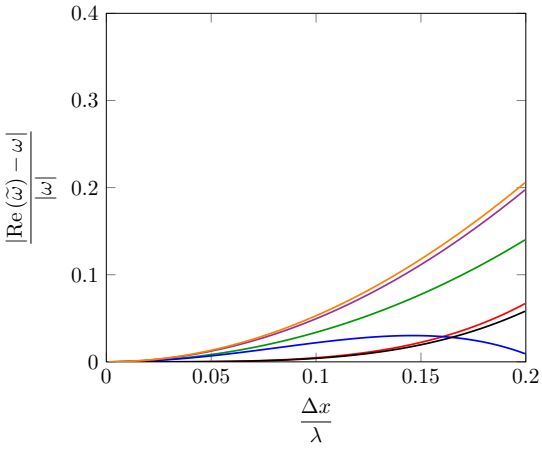


(b) Imaginary part

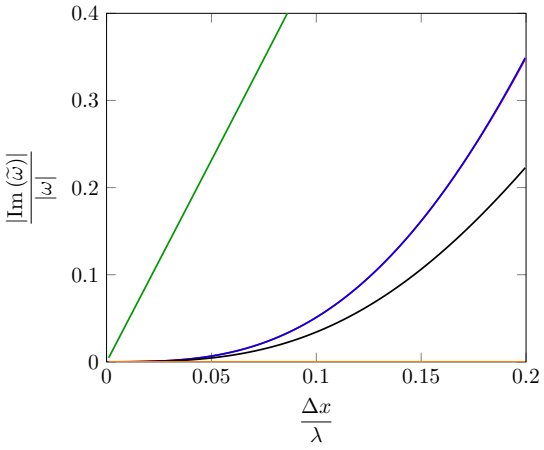


(c) Real and Imaginary Parts

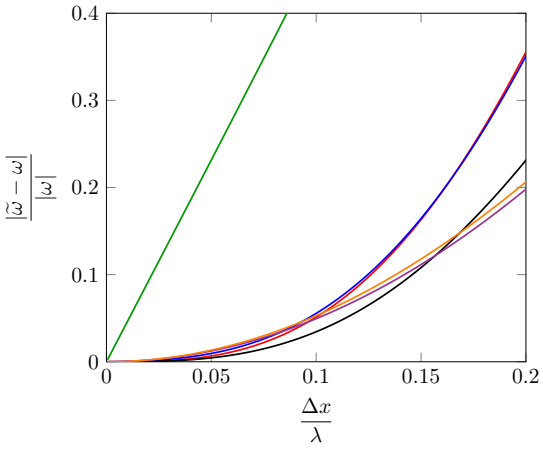
Figure 1.3: Relative dispersion error for first-order FDVM (–), second-order FDVM(–), second-order FEVM (–), third-order FDVM (–), $\mathcal{D}$ (–) and $\mathcal{W}$ (–). With $H = 1m$, $k = \frac{\pi}{10}$ and $U = 1m/s$.



(a) Real part

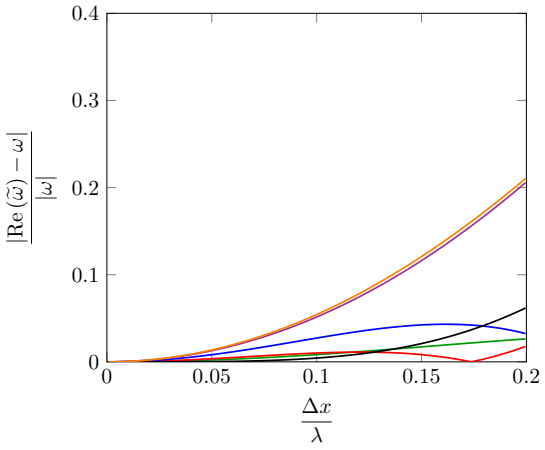


(b) Imaginary part

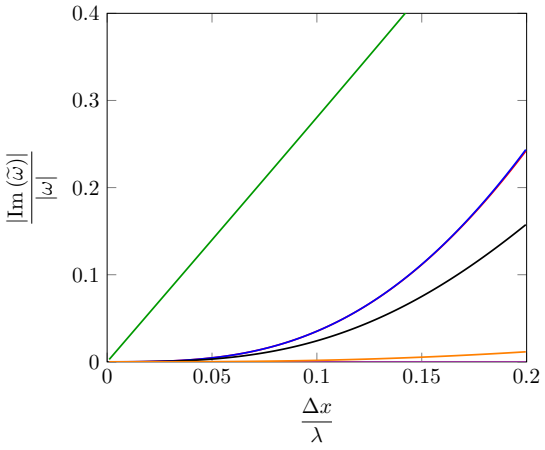


(c) Real and Imaginary Parts

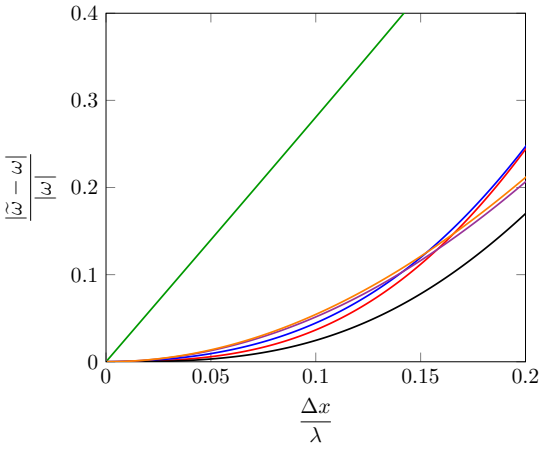
Figure 1.4: Relative dispersion error for first-order FDVM (–), second-order FDVM(–), second-order FEVM (–), third-order FDVM (–), $\mathcal{D}$ (–) and $\mathcal{W}$ (–). With $H = 1m$, $k = 2.5$ and $U = 0m/s$.



(a) Real part



(b) Imaginary part



(c) Real and Imaginary Parts

Figure 1.5: Relative dispersion error for first-order FDVM (–), second-order FDVM(–), second-order FEVM (–), third-order FDVM (–), $\mathcal{D}$ (–) and $\mathcal{W}$ (–). With $H = 1m$, $k = 2.5$ and $U = 1m/s$.

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

Table 1.9: Table showing lowest order error term for approximating ω_+ for all FDVM and the FEVM. With $-\sqrt{gH} \leq U \leq \sqrt{gH}$ and $\beta = 3 + H^2k^2$.

will be the primary driver of the dispersion error.

Scheme	Lowest Order Δx Term of Error	
	$U < -\sqrt{gH}$	$\sqrt{gH} < U$
FDVM_1	$-\left(2U + \sqrt{\frac{3gH}{\beta}}\right)\frac{ik^2}{4}\Delta x$	$\left(2U + \sqrt{\frac{3gH}{\beta}}\right)\frac{ik^2}{4}\Delta x$
FDVM_3	$-\left(2U + \sqrt{\frac{3gH}{\beta}}\right)\frac{ik^4}{24}\Delta x^3$	$\left(2U + \sqrt{\frac{3gH}{\beta}}\right)\frac{ik^4}{24}\Delta x^3$

Table 1.10: Table showing different lowest order spatial error term for approximating ω_+ for all FDVM and the FEVM for different values of U. With $\beta = 3 + H^2k^2$.

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

Table 1.11: Table showing lowest order error term for approximating ω_+ for $\mathcal{D}$ and $\mathcal{W}$.

Bibliography

[1] 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.

[2] 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.

[3] Peter D Lax and Robert D Richtmyer. Survey of the stability of linear finite difference equations. *Communications on pure and applied mathematics*, 9(2): 267–293, 1956.