# A study of two new splitting methods for the gravity part of the shallow water equations

Helge Avlesen

22 January 1998

**Abstract**

In this report two new splitting methods for the gravity part of the shallow water equations are studied. The aim of the methods is to better cope with irregular boundaries than traditional dimensional splitting methods. The proposed methods are implemented and tested in a model problem with a closed channel geometry first described in a work by T.J. Weare [9].

# 1 Introduction

The shallow water equations is a simple but important set of equations that can give surprisingly "correct" results for large scale oceanographic quantities like transports and surface elevation. The main use of these equations is as a part of complex hydrostatic 3d ocean models, where the aim is to predict the dynamics of the surface gravity waves. Furthermore it is also an attractive set of equations for local area, high resolution computations of tides and fluxes. We will focus on solving these equations fast, and with high accuracy.

During the last decades numerical methods for solving these equations have been extensively studied; finite difference methods, finite element methods, spectral methods etc. The challenges we are encountering are extremely complex geometries, fast propagating waves, open boundary conditions and so on. Even with the computing power of today, obtaining good high resolution results in reasonable time forces us to make many compromises, like coarse resolution, the size of the model problem etc.

One issue is the choice of numerical method: The finite element methods permit easy inclusion of local grid refinement and the handling of complex geometries. These methods are somewhat more expensive computationally, and much harder to set up correctly, than the much used finite difference techniques. The latter methods on the other hand, have difficulties handling irregular boundaries, and need local grid refinement. The total cost is of the same order as the finite elements, but the code is still easy to read and use for people without a strong background in numerics. Finally we have spectral methods, but these are not able to compete as of today, because of their problems handling complex geometries.

Another question is the choice of time stepping. There are several options here too. Explicit methods are very easy to program, and also computationally cheap. They also have good properties for computation of discontinuous solutions, much due to the very simple way local corrections to the solution can be computed. A drawback is the constraint on the time step, also called the CFL criterion. This is not a serious limitation for problems where the dynamics you need to resolve is faster than the CFL criterion anyway, but in our case we probably do not need time resolution better than approximately ten times the CFL criterion, in order to be able to describe the most important physical processes. Therefore an implicit method is a natural choice, also because of the minor importance of the convection terms in the shallow water equations.

The model equations are linear, and a good choice for time stepping is the Crank-Nicolson method in time. This choice forces us to solve a linear system of equations, involving all unknowns simultaneously at each time step.

An alternative class of methods known under many names: fractional step methods, locally one-dimensional methods, dimensional splitting or projection methods offer a possible strategy, in combination with implicit numerical methods, for the solution of such systems. See e.g. works by Strang [8], LeVeque & Oliger [5] and Ianenko [4].

The idea is to split the solution process into sub-steps, that finally yield the solution for the next time-step. Each of the substep operators is approximated by using efficient direct solvers to these equations to get a solution in $\mathcal{O}(\mathrm{N})$ operations, where $N$ is the number of unknowns. The idea of dimensional splitting is theoretically very tempting with respect to efficiency, but in the last decade the concern about the accuracy of the solution, especially when using large time-steps, have advocated the fully implicit solvers.

# 2 Dimensional splitting of the equations

## 2.1 Background

The standard linear shallow water equations can be written

$$\begin{aligned} \frac{\partial u}{\partial t} &= -gH(x,y)\frac{\partial \eta}{\partial x}, & (1) \\ \frac{\partial v}{\partial t} &= -gH(x,y)\frac{\partial \eta}{\partial y}, & (2) \\ \frac{\partial \eta}{\partial t} &= -\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}. & (3) \end{aligned}$$

$u$ and $v$ are depth integrated velocities, and $\eta$ is the elevation with respect to the zero level. We have neglected Coriolis forcing, friction and convection. Using the well known Arakawa C-grid [6] the system (1)-(3) is approximated in space with second order central differences.

In Figure 1 we can see one of the major weaknesses with the finite difference technique and rectangular grids, dealing with coast-lines. What happens to the accuracy when both the shape and the smoothness of the boundary are altered? We know that we make an error, but is it within acceptable limits? Following Berntsen and Espelid [1] we can describe a dimensional splitting of the equations like this: Let the vector $\vec{w}$ contain all unknowns

$$\vec{w}(t) = \begin{bmatrix} U_\Delta(t) \\ V_\Delta(t) \\ \eta_\Delta(t) \end{bmatrix}, \quad U_\Delta(t) = \begin{bmatrix} U_1(t) \\ \cdot \\ \cdot \\ U_n(t) \end{bmatrix}. \tag{4}$$

$V_\Delta, \eta_\Delta$ is on the same form as $U_\Delta$. By using "the method of lines", we can approximate (1)-(3) as:

$$\frac{d\vec{w}}{dt} = A\vec{w}(t) + \vec{f}(t). \tag{5}$$

$\vec{f}(t)$ contains contributions from the boundary. For a closed domain we have $\vec{f}(t) = 0 \ \forall \ t$. The matrix $A$, and the boundary vector $\vec{f}$ may be split up like this

$$\frac{d\vec{w}}{dt} = (A_x + A_y)\vec{w}(t) + \vec{f}_x(t) + \vec{f}_y(t). \tag{6}$$

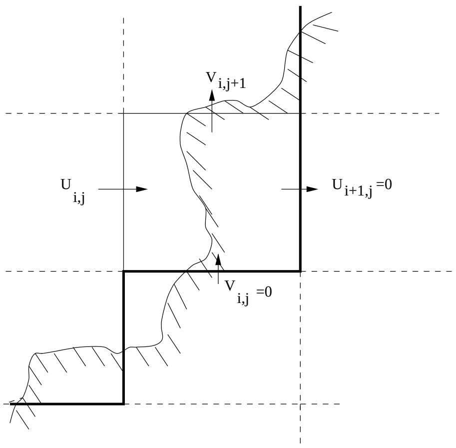


Figure 1: Choosing a boundary

Subscripts $x$ and $y$ indicate contributions from $x$ and $y$ directions respectively. We may solve this system in two sub-steps

$$\frac{d\vec{w}_x}{dt} = A_x \vec{w}_x(t) + \vec{f}_x(t), \tag{7}$$

$$\frac{d\vec{w}_y}{dt} = A_y \vec{w}_y(t) + \vec{f}_y(t). \tag{8}$$

On a closed domain we do the following approximation to our system from time $t$ to $t + \Delta t$

$$\begin{aligned} \vec{w}(t+\Delta t) = \exp((A_x + A_y)\Delta t)\vec{w}(t) \approx \\ \exp(A_x \Delta t)\exp(A_y \Delta t)\vec{w}(t). \end{aligned} \tag{9}$$

Written in operator form

$$\begin{aligned} \vec{w}_y = G_y \vec{w}(t),\ \vec{w}_x = G_x \vec{w}_y, \\ \vec{w}(t+\Delta t) \approx G_x G_y \vec{w}(t). \end{aligned} \tag{10}$$

Where $G_x$, $G_y$ are approximations to $\exp(A_x\Delta t)$ and $\exp(A_y\Delta t)$. This splitting implies second order local and first order global accuracy in time if $G_x$ and $G_y$ are at least second order approximations. Using Strang splitting we obtain second order global accuracy:

$$\vec{w}(t + 2\Delta t) \approx G_x G_y G_y G_x \vec{w}(t). \tag{11}$$

## 2.2 Obtaining tridiagonal systems

Discretisation of (7) and (8) in time with the Trapezoidal rule gives the following two systems corresponding to the operators in the x and y directions, when we

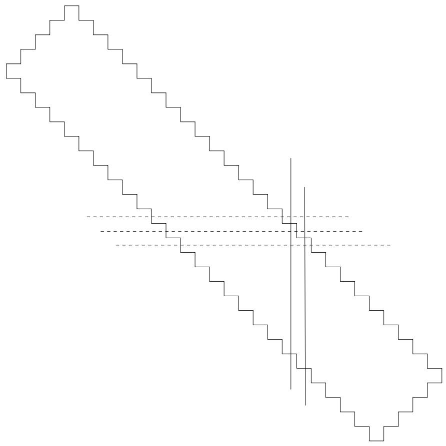

Figure 2: Sweep directions

have a closed domain:
$G_x$:

$$u^{n+1} - u^n = -\alpha_x(\delta_x \eta^{n+1} + \delta_x \eta^n), \tag{12}$$
$$\eta^{n+1} - \eta^n = -\beta_x(\delta_x u^{n+1} + \delta_x u^n). \tag{13}$$

$G_y$:

$$v^{n+1} - v^n = -\alpha_y(\delta_y \eta^{n+1} + \delta_y \eta^n), \tag{14}$$
$$\eta^{n+1} - \eta^n = -\beta_y(\delta_y v^{n+1} + \delta_y v^n). \tag{15}$$

Where $\alpha_x = \frac{gH(x,y)\Delta t}{2\Delta x}$, $\alpha_y = \frac{gH(x,y)\Delta t}{2\Delta y}$, $\beta_x = \frac{\Delta t}{2\Delta x}$, $\beta_y = \frac{\Delta t}{2\Delta y}$, and in addition $\delta_x$ and $\delta_y$ are central difference operators. In the $u$ equation the operator is evaluated in a $u$-point, similarly for the $v$- and $\eta$- equations. $\alpha_x$ and $\alpha_y$ are computed in $u$ and $v$ points respectively. Insertion of $u^{n+1}$ from (12) into (13), and $v^{n+1}$ from (14) into (15), give us a set of uncoupled tridiagonal systems for $\eta$ in the $x$ and $y$ directions. The sweeps are then performed in a sequence given by equation (11). Doing this we get second order accuracy in time, and a very cheap algorithm computationally. In order for our method to beat the explicit techniques with respect to computing time, we have to be marching with time-steps of the order ten times the CFL-criterion.

Imagine a wave propagating in the x-direction hitting the wall in Figure 2. We know that our physical boundary would give a reflection in the y-direction. Due to the one dimensional nature of our operators and the irregular grid, waves are reflected back to where they came from. Using a short time step(close to the CFL criterion) the reflected wave does not leave the wall cell before we apply the successively following operator, so we get the correction we need, and continuity is almost satisfied. At long time-steps however, the wave is handled by the following splitting operator in “the wrong cell”, implying poor accuracy.

# 3 Two new ideas for a dimensional splitting of the gravity part of the shallow water equations

## 3.1 Some motivation

Weare [9], showed how an additional error due to irregular boundaries appears when using the ADI type technique on the shallow water equations. We have repeated Weares experiment with the splitting described in (11), and we have observed the same effects.

Note that when using fully implicit methods on exactly the same grid, and with identical boundary conditions, the effect described by Weare is not seen at all. This has been the motivation for studying alternative splittings of the equations, in order to reduce or possibly eliminate this type of error.

In his paper Weare suggests a coupling of the $u$ and $v$ velocities in the boundary cells. By doing this, our previously uncoupled tridiagonal systems get coupled, and the advantage of the splitting method is reduced. We want a de-coupling of the systems into easy to solve linear systems.

## 3.2 The Weare channel

Given a domain with a coastline aligned 45 degrees to a cartesian grid. We choose a closed channel with a length of approximately 3km, and a width of 0.5km. The depth is constant $H = 10m$. Initially we have a linearly decreasing elevation in the length direction of the channel, and when releasing this we get a wave-like motion. It can be shown that when using a fractional step method

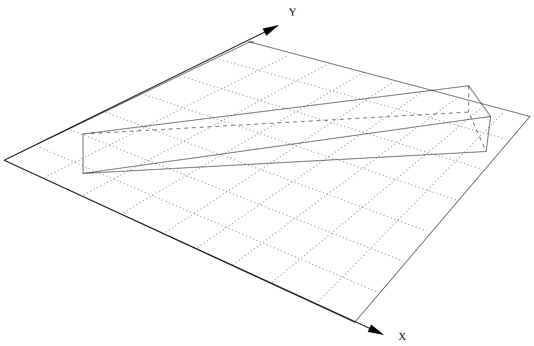


Figure 3: Geometry and elevation of the Weare channel at $t = 0s$

of the ADI type, and setting the boundary conditions as in Figure 1, this is equivalent to a no-slip boundary condition on a smooth coastline.

### 3.3 The reflection splitting

As before we set up the full discretisation of the system of equations in space using central differences. By using "the method of lines" we write this system on the form

$$\frac{d\vec{w}}{dt} = A\vec{w}(t) + \vec{f}(t). \tag{16}$$

A natural way to overcome the reflection problem must be to allow the waves to reflect on the boundary. We therefore need to solve our tridiagonal system along the assumed path of a wave like Figure 4 below shows. To describe this

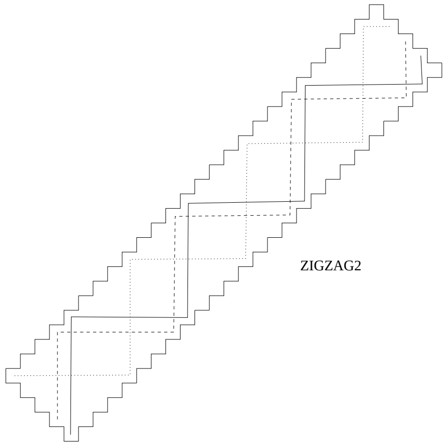


Figure 4: Reflecting systems at the boundary

formally we split the matrix $A$ into $k$ parts corresponding to the $k$ lines that together cover our domain with a rectangular grid:

$$A = A_1 + A_2 + ... + A_k. \tag{17}$$

Solving the equations of this system

$$\begin{aligned} \tfrac{d\vec{w}}{dt} &= A_1\vec{w}(t) \\ &\vdots \\ \tfrac{d\vec{w}}{dt} &= A_k\vec{w}(t) \end{aligned}$$

in sequence using the obtained solution from one equation as input for the next is equivalent to

$$\vec{w}(t+\Delta t) \approx \exp(A_1\Delta t)..\exp(A_k\Delta t)\vec{w}(t). \tag{18}$$

The exponential operators are approximated by solution operators

$$\vec{w}(t+\Delta t) \approx G_1..G_k\vec{w}(t). \tag{19}$$

Making the order of these operators symmetric, we obtain a Strang-type splitting, which formally is of second order accuracy in time:

$$\vec{w}(t+2\Delta t) \approx G_1..G_k G_k..G_1 \vec{w}(t). \tag{20}$$

Unfortunately this method seems to loose all precision for $CFL > 1$. A possible explanation for this is the way we update $\vec{w}$. By calculating the effect of one of our operators we update the values of $\vec{w}$ along a line as shown in Figure 4. Unfortunately our lines cross each other, so we have to calculate central differences from a partially updated field, intuitively this seems to be the main source of the bad behavior of this method.

A way to avoid this is not to update $\vec{w}$ until the effect of all operators are calculated. $G_1..G_k$ can then be computed in parallel. Observe that when doing this we get two solutions for each node, so we must make a choice. Choosing the last obtained solution for each point seems to be the way to get results for this test-case, but this remedy is hard to defend mathematically.

Also the properties of the method (20) with respect to parallelism are unattractive, since each of the operators have to be calculated sequentially.

### 3.4 The Zig-zag splitting

Observing that we have a wave-like motion, back and forth through the channel, and that the dynamics we need to capture is in the length direction of the channel we suggest a different approach. We construct the implicit systems in the direction of the dynamics we want to resolve. We will still make a splitting error, but this error will now be the error associated with the flow normal to the main direction, so hopefully the total error will be reduced. As before we have $\frac{d\vec{w}}{dt} = A\vec{w}(t)$. Let $A_1$ denote contributions from solid lines, $A_2$ contributions from dashed lines. Split $A$:

$$A = A_1 + A_2. \tag{21}$$

It is now possible to construct tridiagonal systems for $\eta$ along each zigzag line. Again we solve the equations in two sub-steps

$$\frac{d\vec{w}_1}{dt} = A_1\vec{w}_1(t), \tag{22}$$

$$\frac{d\vec{w}_2}{dt} = A_2\vec{w}_2(t). \tag{23}$$

Approximated as

$$\vec{w}(t+\Delta t) \approx \exp(A_1\Delta t)\exp(A_2\Delta t)\vec{w}(t). \tag{24}$$

Strang splitting gives:

$$\vec{w}(t+2\Delta t) \approx \exp(A_1\Delta t)\exp(A_2 2\Delta t)\exp(A_1\Delta t)\vec{w}(t) \tag{25}$$

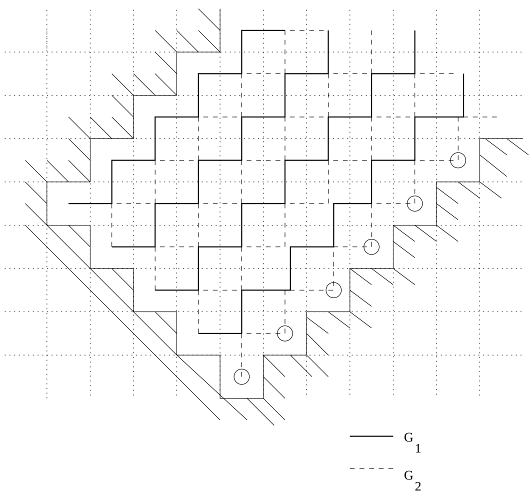


Figure 5: Layout of tridiagonal systems

or

$$\vec{w}(t + 2\Delta t) \approx G_1 G_2 G_2 G_1 \vec{w}(t). \tag{26}$$

Splitting the matrix $A$ this way using fast solvers is similar to non overlapping Domain Decomposition methods. Similar ideas can be found in e.g [7]

## 3.5 Boundary treatment in the zig-zag method

With a straightforward discretisation of our channel problem, some $\eta$-points along the boundary is left unchanged each time one of the operators $G_1$ or $G_2$ are calculated. For the $G_1$ operator these points are marked with a circle in Figure 5. Due to the large time-step the $\eta$ field loose smoothness, and this gives large errors in the solution.

In order to overcome this, two different boundary treatment ideas will be discussed in the next two subsections.

### 3.5.1 External point mirroring

Constructing a zigzag-line through the boundary in Figure 6, gives us artificial points like the one marked with a black dot. Normally the depth $H$ in a $u$ or a $v$ point on the boundary is zero. Now we give the boundary points the depth of the cell center, so that flow through the boundary is possible.

Following Forrer [3] we demand symmetry over the reflecting boundary, the value in point $(i, j)$ is equal to the average of points $(i-1, j), (i, j+1)$. Doing this, a tridiagonal system for the points marked by a circle in Figure 6 can be constructed. Unfortunately this does not work. The values at the boundary is

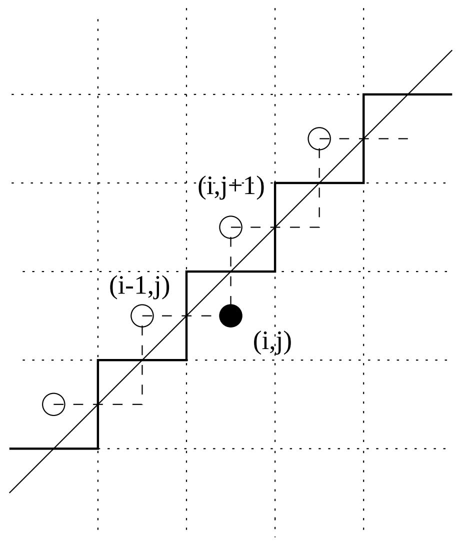


Figure 6: Mirroring of artificial points

updated, but is lagging behind in time, corrupting the solution.

### 3.5.2 Extra effort along the boundary

The next idea is crude, but it seems to work. We observe that the $G_2$ operator "fits" the boundary at the right hand side of the channel. So we use the results from one zig-zag line of this operator to approximate the missing points for the $G_1$ operator. Missing points are marked by a circle in Figure 7. A full time step with $G_1G_2$ can then be described like this:

- Calculate boundary nodes not covered by $G_1$ with $G_2$
- Calculate $G_1$ nodes
- Calculate boundary nodes not covered by $G_2$ with $G_1$
- Calculate $G_2$ nodes

Reverse the order every second time-step, to get second order accuracy.

# 4 Some Results

To compare the methods we made a reference solution by iterating the full Crank-Nicolson discretisation to single precision by the preconditioned conjugate gradients method as described in Templates [2], and the timestep $\Delta t = 1s$. For this experiment $\Delta x = 100m$, $\Delta y = 100m$, $H = 10m$. A comparison of the

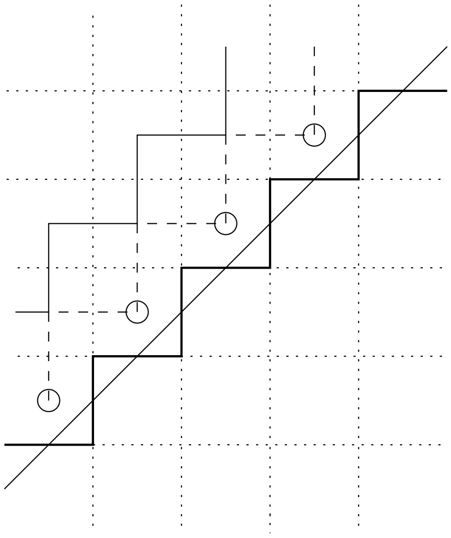

Figure 7: Approximating the solution along the boundary

time series of $\eta$ at the end of the channel for the different methods at $dt = 60s$ is shown in figures 10-13. Only the first hour is shown. It is clear from these figures that both the Berntsen & Espelid [1] splitting and the "reflection splitting" fail if we look at phase error. The splitting proposed by Berntsen & Espelid show serious phase error due to the friction effect of the irregular boundary, described by Weare [9]. Amplitudes are reasonably well computed. We must also take into consideration that this is only the first hour of the solution. Figure 8 gives an idea of what the surface looks like at this time-step, and this tell us that conventional splitting is near useless for these CFL numbers. The other methods at least yield qualitatively smooth surfaces as seen in Figure 9. The method labeled Crank in figures 10-13, is the solution obtained by 5 iterations of the preconditioned conjugate gradients method. We observe that the zig-zag splitting gives a very good solution in this context.

The figures 14-17 shows our second comparison, the time series of the depth integrated velocity $u$ in the wall cell mid-lengthwise in the channel. Also here we observe the phase error and that the Berntsen & Espelid splitting give smaller amplitude for $u$ than the other methods due to the splitting error. Our proposed splittings are not perfect in this respect either, but far better than the conventional splitting.

# 5 Conclusions and further work

- We have shown that the splitting error, due to irregular boundaries, can be reduced by constructing implicit systems in the main direction of the

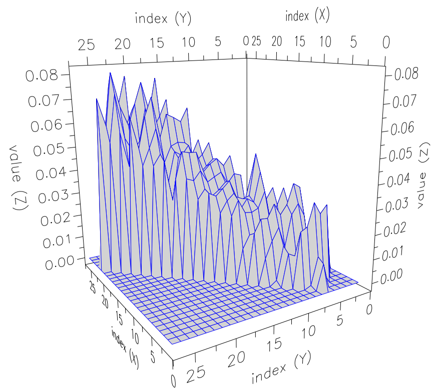


Figure 8: Snapshot of elevation $\eta$ after $t \approx 1h$ using the Espelid & Berntsen splitting

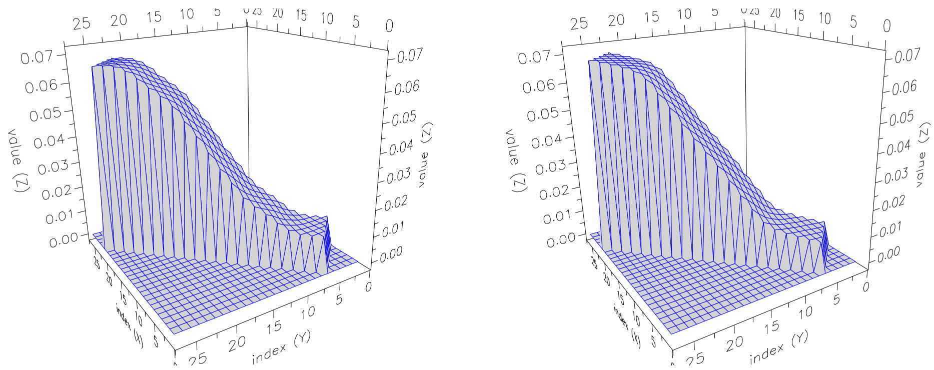


Figure 9: Snapshot of elevation $\eta$ after $t \approx 1h$ using the reflection splitting (left), and the zig-zag splitting (right)

flow.

- The described methods are too difficult to implement generally to be useful in practice, but the idea of implicit systems in the main direction of the flow seem to hold.
- The proposed splittings have similarities with domain decomposition ideas, and may prove useful as a preconditioner in combination with e.g. the conjugate gradients methods.

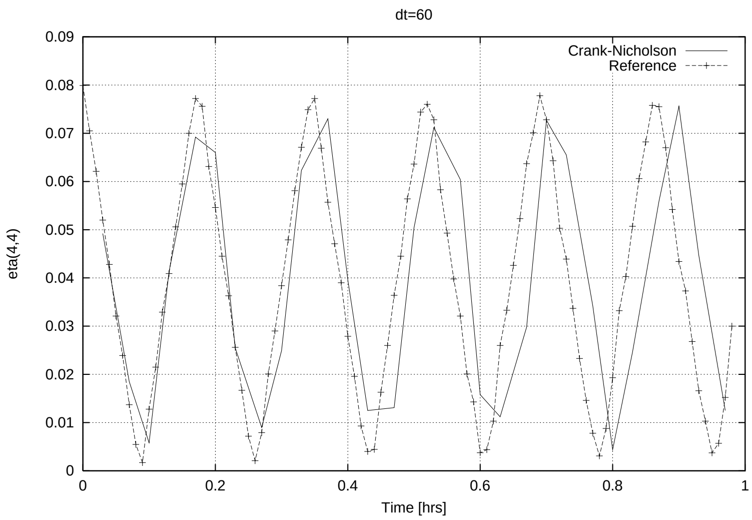


Figure 10: Time series for $\eta$ at the end of the channel for the Crank-Nicolson method, using $dt = 60s$

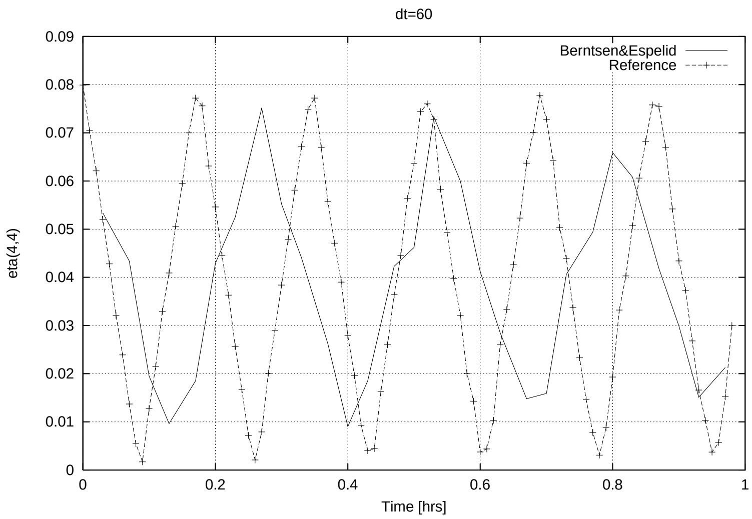


Figure 11: Time series for $\eta$ at the end of the channel for the Berntsen&Espelid splitting, using $dt = 60s$

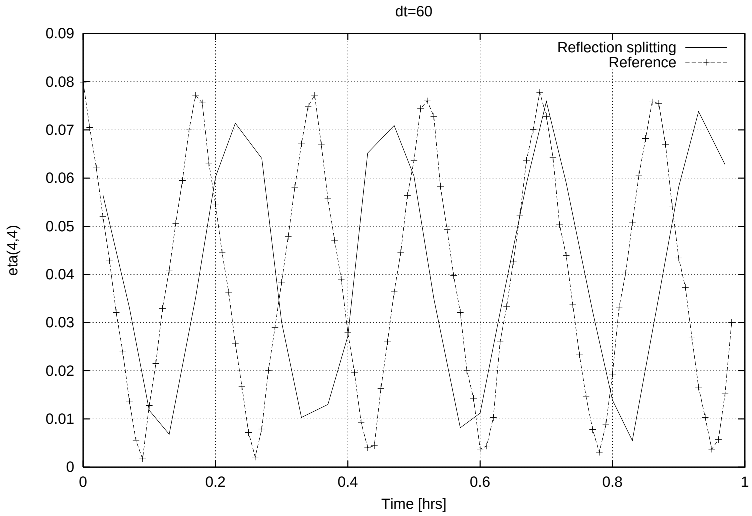


Figure 12: Time-series for $\eta$ at the end of the channel for the Reflection splitting, using $dt = 60s$

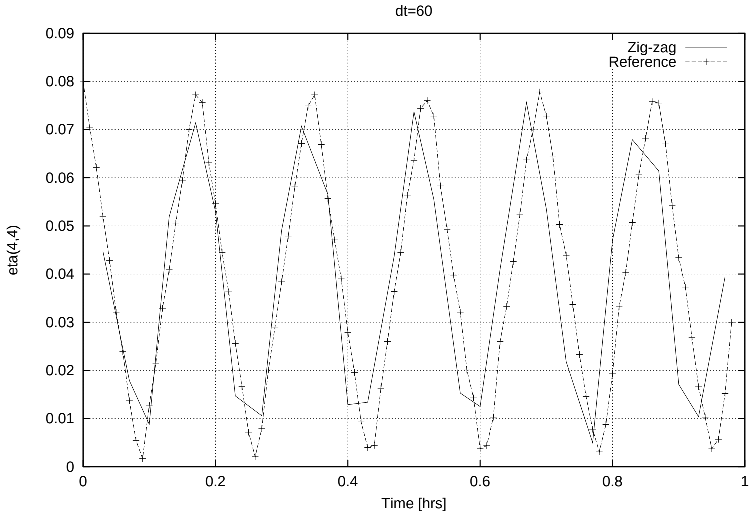


Figure 13: Time-series for $\eta$ at the end of the channel for the Zig-zag splitting, using $dt = 60s$

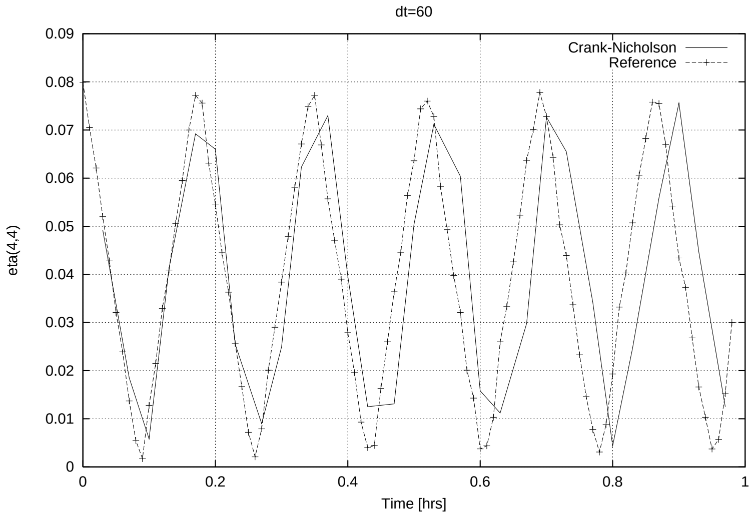


Figure 14: Timeseries for $u$ in the wall cell, mid-lengthwise of the channel, for the Crank-Nicolson method, using $dt = 60s$

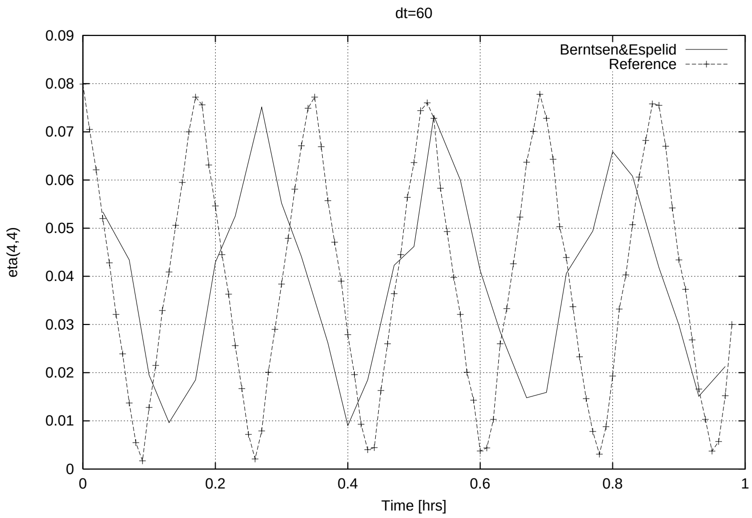


Figure 15: Timeseries for $u$ in the wall cell, mid-lengthwise of the channel, for the Berntsen&Espelid splitting, using $dt = 60s$

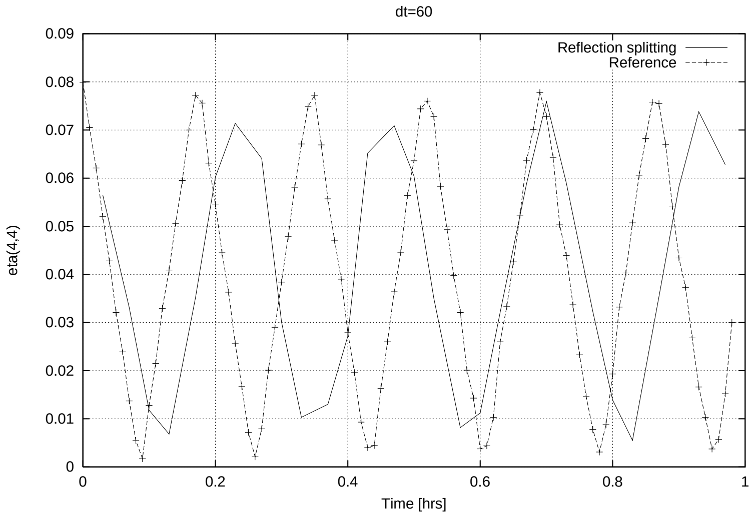


Figure 16: Time-series for $u$ in the wall cell mid-lengthwise of the channel for the Reflection splitting, using $dt = 60s$

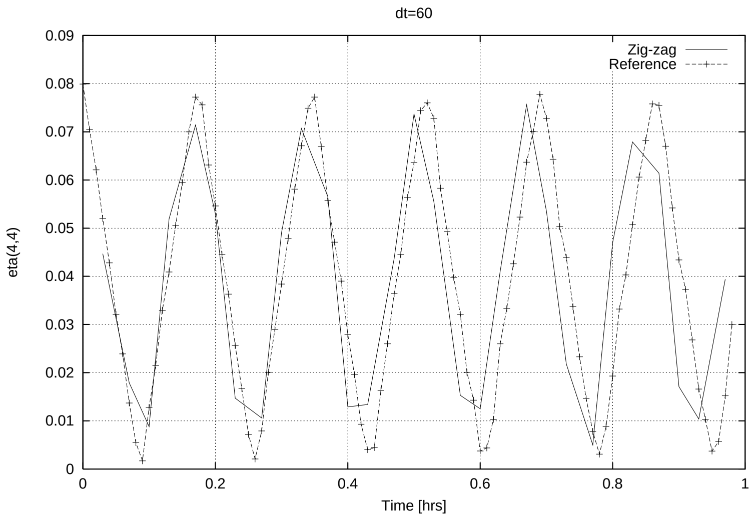


Figure 17: Time-series for $u$ in the wall cell mid-lengthwise of the channel for the Zig-zag splitting, using $dt = 60s$

# References

[1] J. Berntsen and T.O. Espelid. Propagation of shallow-water inertia-gravity waves with fractional step methods. Technical report, Dept. of Applied Math. University of Bergen, 1995.

[2] R. Barret et.al. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM. Download from Netlib.

[3] H. Forrer. *Boundary Treatments for Cartesian-Grid Methods*. PhD thesis, Swiss Federal Institute of Technology, 1997.

[4] N. N. Ianenko. *The method of fractional steps*. Springer-Verlag, Berlin, 1971.

[5] R. J. LeVeque and J. Oliger. Numerical methods based on additive splittings for hyperbolic partial differential equations. *Math. Comp.*, 40(162):469–497, 1983.

[6] F. Mesinger and A. Arakawa. *Numerical methods used in atmospheric models*. Number 17 in GARP publication series. Global Atmospheric Research Programme (GARP), 1976.

[7] B. Smith, P. Bjørstad, and W. Gropp. *Domain Decomposition: Parallel multilevel methods for elliptic partial differential equations*. Cambridge University Press, 1996.

[8] G. Strang. On the construction and comparison of different schemes. *SIAM J. Numer. Anal.*, 5(3):506–517, 1968.

[9] T. J. Weare. Errors arising from irregular boundaries in ADI solutions of the shallow-water equations. *Int. J. Numer. methods Eng.*, 14:921–931, 1979.