Numerical accuracy in the solution of the shallow-water equations

P. Broomans & C. Vuik
Delft University of Technology, Delft, Netherlands

A.E. Mynett & J. Mooiman
WL | Delft Hydraulics, Delft, Netherlands

ABSTRACT: For historical reasons computer simulations (e.g. Delft3D) are still computed in single precision, while at the same time the size and the complexity of the problems have increased. (Differences appear between results calculated on different computer platforms.) The aim of this research is to determine the influence of finite machine precision and problem size on the accuracy of the results. Therefore, proposals for possible improvements, like increasing the machine precision and row scaling, are tested. Furthermore, unstable behaviour appeared in some cases. Investigation led to the conclusion that the 3D model is conditionally stable in contrary to the 2D model, which is unconditionally stable.

1 INTRODUCTION

1.1 Modelling Cycle

The modelling cycle is the complete process of modelling a natural phenomenon. We will describe the modelling cycle for shallow flows. The modelling cycle can be divided into four steps, see Fig. 1.

In the first step the natural water flow for shallow waters (Natural System) is derived from the Navier-Stokes equations and the continuity equation through application of various assumptions and approximations. This leads to a Conceptual Model, in our case the 3D shallow-water equations (SWE).

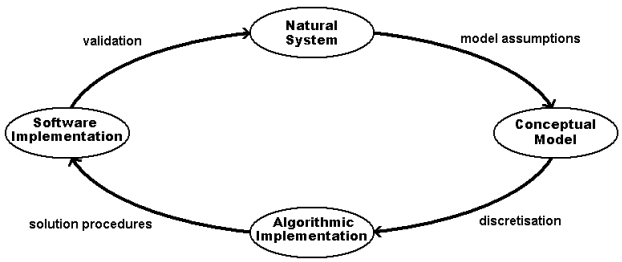

Figure 1: Modelling Cycle

Discretisations of these equations, during which truncation errors are made, leads to an Algorithmic Implementation. The solution is calculated with a certain method and a certain machine precision (Software Implementation). The final step in the modelling cycle is the validation of the solution in an application by comparison with measured data.

1.2 Discretisation of the shallow-water equations

The 3D SWE comprise two momentum equations in the horizontal directions and the continuity equation. In the vertical direction the σ-transformation has been applied (Phillips 1957).

The (barotropic) pressure terms, the eddy viscosity terms and the vertical advective terms are discretised with second order accurate central difference schemes. The horizontal advective terms are approximated with second order accurate upwind schemes. For the time integration an ADI scheme is chosen (Mitchell and Griffiths 1980).

1.3 Solution Procedure

During every iteration three linear systems of the form $A\underline{x} = \underline{b}$, have to be solved; two of them with Gaussian elimination (continuity and momentum equation) and one with a Gauss-Jacobi iteration scheme (momentum equation). During the computation of the solution vector $\underline{\hat{x}}$ rounding errors are made.

1.4 Stability

It has been proved that application of the ADI scheme results in an unconditionally stable numerical scheme (Algorithmic Implementation) for the 2D SWE, i.e. with one σ-layer (Stelling 1983). However, from our research it appears that extension to a 3D model leads to a conditionally stable scheme, even though the vertical direction is implicitly embedded in the 2D model. Some test cases in this paper show unstable

behaviour.

1.5 Problems

Results from different computer platforms show slightly different results (e.g. in the order of millimetres for computed water levels). This is one of the motivations to do this study.

Furthermore, truncation errors made during the discretisation must be smaller than the errors due to model assumptions and the cumulative rounding errors must again be smaller than the truncation errors. If, for instance, the latter is not true, refinement of the grid and the subsequent reduction of the truncation error does not lead to more accurate results, because the rounding errors are still larger.

Finally, it would be very attractive to have some quantitative information on the influence of the rounding errors on the computed results in order to obtain reliable predictions of water levels.

In this article we focus on the (the consequences of) these rounding errors, which are introduced in the solution procedures.

2 ROUNDING ERRORS

Finite machine precision μ causes rounding errors when performing floating operations. Cumulating rounding errors can cause loss of numerical accuracy.

2.1 Analysed quantities

The cumulative rounding error can be analysed by calculation of the absolute error $\|\underline{x} - \underline{\hat{x}}\|$ and the relative error $\|\underline{x} - \underline{\hat{x}}\|/\|\underline{x}\|$, where $\underline{\hat{x}}$ is the computed solution vector. Furthermore, we will also use the relative residue $\|A\underline{\hat{x}} - \underline{b}\|/\|\underline{b}\|$ as a measure for the accuracy.

2.2 Error dependence

The magnitude of the errors mainly depends on the condition number of matrix A, computational machine precision and the size of the model. In lesser degree the computer architecture (rounding procedures) and compilers (not analysed) play a role.

2.3 Upper bounds on the relative error

The exact solution $\underline{x}$ is of course not available. However, upper bounds on the relative error are. One of the most general upper bounds states:

$$\frac{\|\underline{x} - \underline{\hat{x}}\|}{\|\underline{\hat{x}}\|} \leq \kappa(A) \frac{\|A\underline{\hat{x}} - \underline{b}\|}{\|\underline{b}\|}, \tag{1}$$

where $\kappa(A)$ is the condition number of matrix A. Two of the equations are solved with Gaussian elimination. Furthermore, the matrices of those equations are diagonally dominant and tridiagonal. For these matrices another upper bound can now be constructed.

$$\frac{\|\underline{x} - \underline{\hat{x}}\|}{\|\underline{\hat{x}}\|} \leq \frac{2\|E\|\kappa(A)}{\|A\| - \|E\|\kappa(A)}, \tag{2}$$

where $\|E\| \leq 2(n-1)\mu\|A\| + (2n+2)\mu\|\hat{U}\| + \mathcal{O}(\mu^2)$. The matrix $\hat{U}$ is the computed upper matrix of the LU factorisation of matrix A.

2.4 Condition number

The condition number $\kappa(A)$ is defined as $\|A\| \cdot \|A^{-1}\|$. The condition number can be interpreted as a measure of the loss of accuracy during the computation of a matrix equation due to large differences of the matrix entries. Thus small condition numbers are preferred, because the guarantee small relative errors, see Eq. (1).

For regular problems the condition number of the discretised continuity equation is of order 10^4 to 10^5 and of the discretised momentum equations of order 10. This is fairly large.

2.5 Row-scaling

A well-known method to decrease the condition number is preconditioning (Golub and Van Loan 1996). This method prescribes that the matrix equation $A\underline{x} = \underline{b}$ is multiplied with a matrix P, resulting in $P^{-1}A\underline{x} = P{-}1\underline{b}$. The objective is to choose/compute P such that the entries in matrix A are of the same order and the condition number is as small as possible.

One of the most time-efficient preconditioning methods is row-scaling. With row-scaling every element on the same row, including the right-hand side, is divided by a certain value. The largest element on the row would be a suitable choice, because then the largest element on every row would be one. Practically, the matrices are always diagonally dominant. Thus the main diagonal element is the largest element on the row. The precondition matrix is then defined by $P = diag(A)$.

Table 1: Order of the condition number for the continuity equation.

	test case		
Row-scaling	2	3	4
no	10^4	10^5	10^5
yes	10^1	10^1	10^2

3 STABILITY

Early tests have shown unstable behaviour for some test cases in the sense that results might increase in one time step and diminish in the next (possibly due to variable/dependent matrices). Also, some test

cases react disproportionally to small changes in initial and/or boundary conditions.

In this section we will discuss the stability analysis for the 2D SWE, which we use as a basis for the stability analysis for the 3D SWE.

3.1 Two-dimensional shallow-water equations

A Fourier analysis (or Von Neumann stability analysis) has been performed for the 2D SWE (Stelling 1983). In our research we have repeated this analysis for the discretisations as they are presently implemented in Delft3D-FLOW. The analysis can also be found in other literature (Wesseling 2001).

A Fourier analysis (verwijzing naar een boek oid) shows whether small perturbations in initial or boundary conditions grow or diminish within a linear system. The analysis focuses on one of the inner points with the general coordinates (m, n). Hence, the boundary conditions themselves are not taken into account in the stability analysis.

As the SWE are non-linear (advective terms in the momentum equations and the terms in the continuity equation), approximations are needed in order to perform the stability analysis. For the advective terms this is done by assuming a constant advective current. For the terms in the continuity similar assumptions are made.

This leads to the desired linear system of the form $M\underline{w}_{mn} = \underline{b}$ at the horizontal coordinates (m, n), where M is now a constant 3×3-matrix. For the Fourier analysis we will look at the solutions of the form

$$\underline{\tilde{w}}^p_{mn} = \underline{\hat{w}}^p \, e^{i(\alpha_1 m \Delta x + \alpha_2 n \Delta y)}, \tag{3}$$

where Δx and Δy are the grid sizes in respectively the x- and y-direction; p denotes the time iterate; $\alpha_1, \alpha_2 \in \mathbb{R}$.

Substituting the range of solutions 3 into the linear system accurately, distinguishing implicit and explicit terms and taking into account the ADI scheme, gives

$$\begin{aligned} A\underline{\hat{w}}^{p+\frac{1}{2}} &= B\underline{\hat{w}}^p, \\ C\underline{\hat{w}}^{p+1} &= D\underline{\hat{w}}^{p+\frac{1}{2}}. \end{aligned} \tag{4}$$

The solution $\underline{\hat{w}}^{p+1}$ can be written as $\underline{\hat{w}}^{p+1} = G\underline{\hat{w}}^p$, with $G = C^{-1}DA^{-1}B$, the amplification matrix. It can be shown that the largest eigenvalue of G is not larger than 1, proving unconditional stability of the discretised system.

3.2 Three-dimensional shallow-water equations

In the three-dimensional case it appears that small perturbations may lead to differences between the perturbed and the unperturbed system, which are several order larger than the perturbations (Fig. 2).

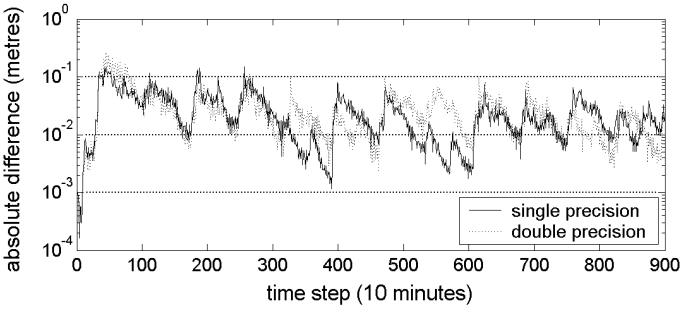

Figure 2: Absolute difference between an unperturbed and a perturbed system. Perturbation of 1 mm on the initial water level.

In contrary to the stability analysis for the 2D SWE, all points at the horizontal coordinates (m, n) are taken into account. As done before, now the 3D SWE are linearised. The advective currents are again taken constant, but assuming different values for the various σ-layers (vertical coordinates).

The boundary conditions at the bottom and the free surface are also not taken into account. However, the adapted discretisations near these boundaries are processed.

The analysis in the previous section holds also for the 3D SWE up to the equation 4. Due to the addition of the vertical direction the matrices A, B, C and D loose their advantageous form and size, making it impossible to proceed analytically.

Hence, we decided to let Matlab calculate largest eigenvalues for various test cases. Therefore, the various constants in the matrices are approximated with computed results from Delft3D-FLOW. The parameters α_1 and α_2 can assume every real value. However, they only appear in trigonometric functions, so variations within intervals (depending on Δx and Δy) will suffice. Figure 3 shows an example of the output.

It must be stated that boundary conditions, such as tidal conditions, enforce a solution on the discretised field, preventing exponential growth from occurring.

4 RESULTS

The application of double precision and row-scaling has been tested for Delft3D-FLOW on several test cases.

4.1 Test Cases

We have used three sets of test cases.

Test case 1 is a 5 m deep closed reservoir (8 by 8 km) with a wind-driven flow. The horizontal grid size is either 1000 m or 100 m. Five equidistant σ-layers are defined.

Test case 2 is a 10 m deep reservoir (8 by 8 km) with only one open boundary on which a tidal condition with an amplitude of 1 m is imposed. In the middle of the reservoir a square island (2 by 2 km)

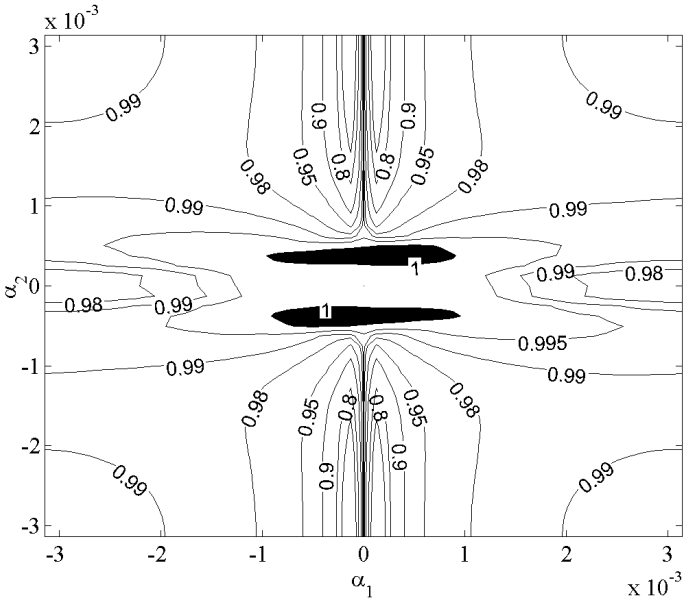

Figure 3: Stability field of an unstable test case. The dark areas represent possible instabilities.

is situated with 8 thin dams, see Fig. 4.1. Different horizontal grid sizes are used for this test case: 1000 m, 333 m, 200 m, 143 m and 111 m. Ten equidistant σ-layers are defined.

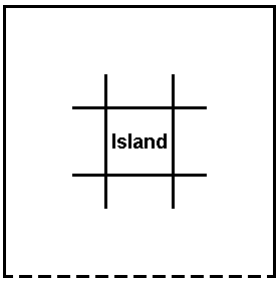

Figure 4: Layout of test case 2.

Test case 3 is a 10 m deep, open ended channel (5 by 19 km). Along one of the open boundaries a tidal condition with an amplitude of 0.5 m is imposed, and on the other one a homogeneous water level condition is imposed. A varying number of σ-layers is used: 1, 5, 10, 20 and 50 σ-layers.

Test case 4, referred to in Table 2.5, is a 2D model of the eastern section of the Westerschelde.

4.2 Absolute error

In addition to the absolute and relative error (see section 2.1) of the separate equations, we can also calculate the absolute difference between single precision and double precision results on whole time steps. We will refer to this as the absolute error for the single precision results.

4.3 Computer architecture

A quick error analysis for test case 1 showed comparable results for three different computer platforms, excluding it as cause for differences in the computed flow data. (*plaatje?*)

4.4 Condition number and row-scaling

As seen in Table 2.5 the condition number for the continuity equation is large. Applying row-scaling leads to a significant improvement of the condition number (Fig. 5). Subsequently, taking Equation 1 into consideration, this leads to a similar reduction in the upper bound for the relative error.

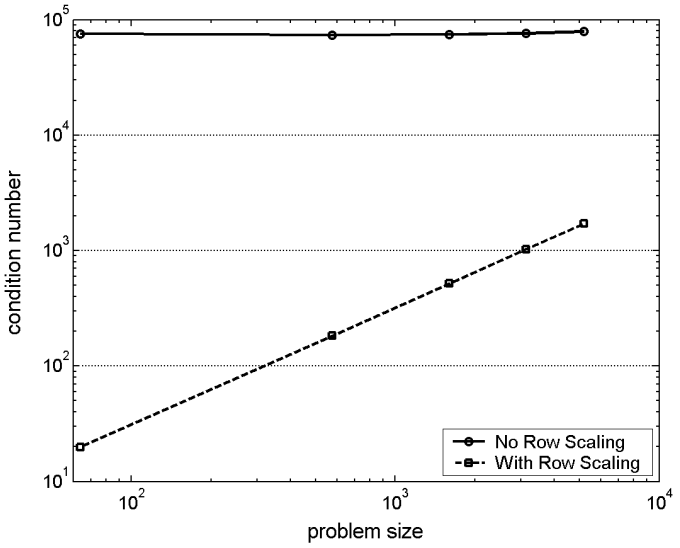

Figure 5: Improvement of the condition number for test case 2 due to row scaling for various problem sizes.

Furthermore, note that without row-scaling the condition number is not dependent in the size of the matrix, while when row-scaling is applied it is.

The extra computation time needed for row-scaling is negligible.

4.5 Size of the model application

Naturally, a larger model implies that more calculations need to be carried out in order to acquire the computed results. It is expected that with increasing model sizes, the errors in the result also increase. Fig. 6 shows that upper bound increases with the number of horizontal grid points.

The number of σ-layers appears to be a significant quantity with respect to the relative error of the momentum equation in x-direction (Fig. 7). The main flow in this test case is in x-direction. In theory the flow velocity in the transverse direction (y-direction) is zero, leading consequently to large *relative* errors. The upper bound on the relative error for the continuity equation is constant with respect to the number of σ-layers.

However, when calculating the absolute error (through comparison with double precision results), the error in the water level does depend on the number of σ-layers (Fig. 8). This is most likely due to the large errors in the main flow velocity.

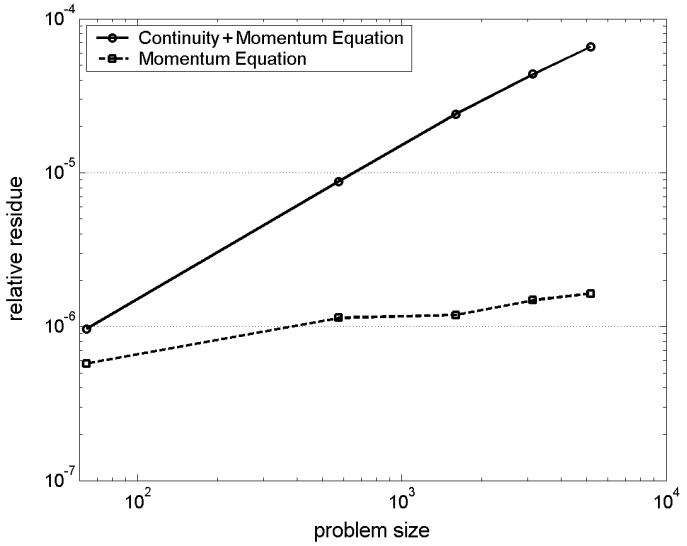

Figure 6: Upper bound for the relative error of the water level against the problem size for test case 2.

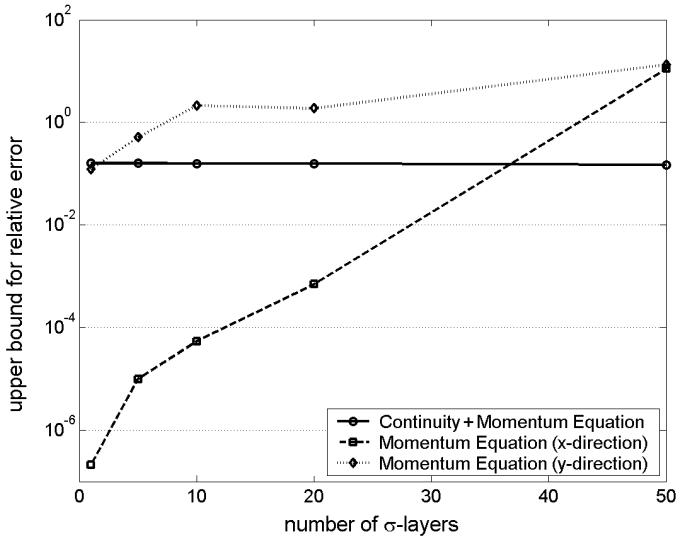

Figure 7: Upper bound for the relative error against the number of σ-layers for test case 3.

4.6 Machine precision

Presently, Delft3D-FLOW is implemented in single precision (6-7 digits accurate for a single operation). However, since the first releases the application have grown considerably in size (up to 100,000 horizontal grid points and 20 σ-layers).

At this moment the absolute error (at whole time steps) of most of the large applications are of the order 10^{-3}, which is just acceptable. In several cases the computed upper bound of the relative error is too high to guarantee sufficiently accurate results. In the future even larger applications might/will show unacceptable results (is dit te sterk?).

Experimental runs with a double precision implementation (14-15 digits) show a decrease of 8 orders in the relative error and the relative error for the continuity and momentum equation, which are computed with a direct method. The other momentum equation, computed with an iterative method, shows hardly any progress due to the stopping criteria.

Test runs show an increase in computation time between 10% - 15%. The necessary memory is almost doubled.

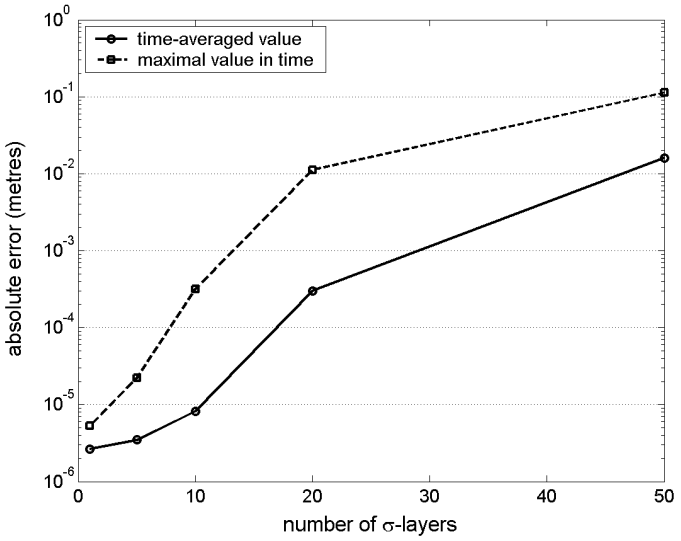

Figure 8: Absolute error of the water level against the number of σ-layers for test case 3.

4.7 Applications for fully closed basins

While for applications with open boundaries, by far the most, tidal conditions will at one point impose a solution and overrun small errors. With fully closed basins, on the other hand, no tidal or other conditions will neutralise these errors.

Fig. 9 shows that these errors accumulate considerably in time. From extrapolation of the graph we can conclude that differences in the order of centimetres will occur after $10,000$ time steps for 6400 grid points.

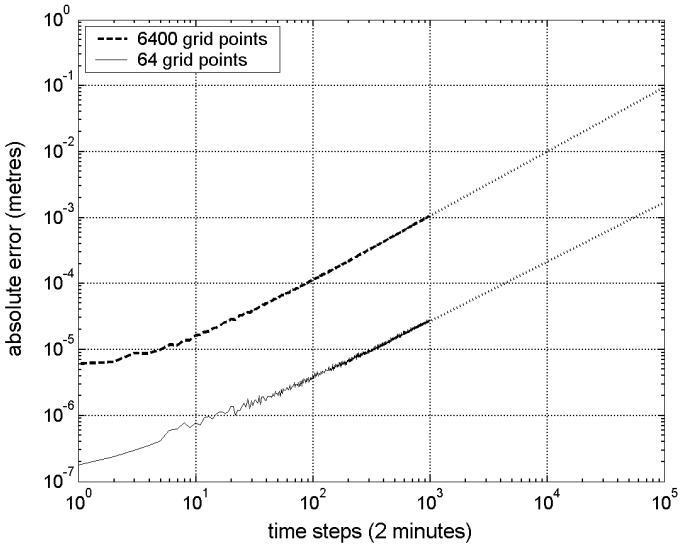

Figure 9: Increase in time of the absolute error of the water level for test case 1. Beyond 1000 time steps the graphs are extrapolated.

4.8 Spin-up time

The simulation time needed for the model to overcome the (klap) of the initial data is called the spin-up time. Short spin-up times are desired in order to lessen computation time. It can be visualised by rerunning the model with a slight perturbation in, for instance, the initial water level and plotting the difference in time. Fig. 10 and Fig. 11 show such a graph for test 2 (where the closed boundaries have been replaced with open boundaries). It can be concluded that row-scaling as well as higher machine precision decreases the spin-up time.

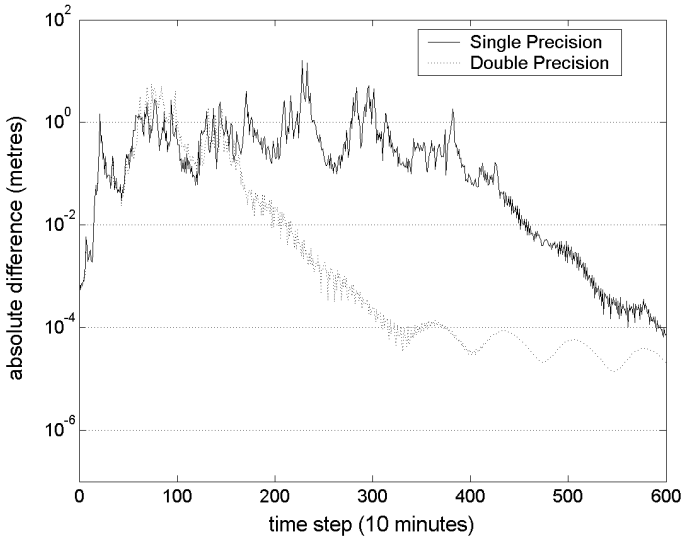

Figure 10: Absolute difference in the water level between the perturbed and the unperturbed system (perturbation: +1 mm on the initial water level) for test case 2.

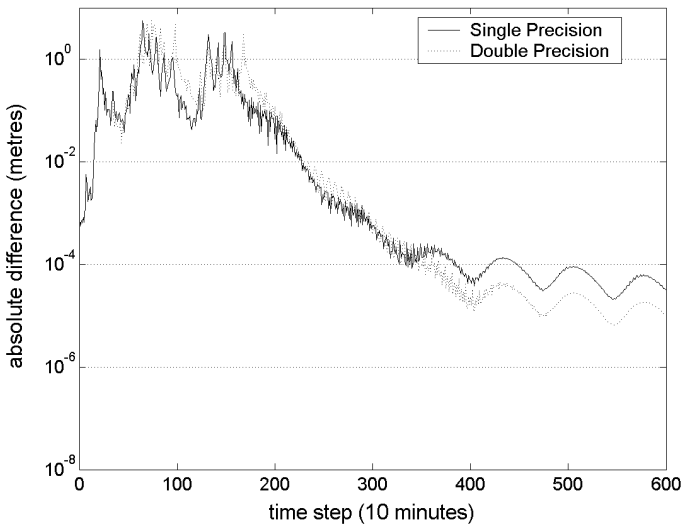

Figure 11: Absolute difference in the water level between the perturbed and the unperturbed system (perturbation: +1 mm on the initial water level) with row-scaling for test case 2.

4.9 Truncation error

The truncation error can be calculated through comparison of the results of the same model run with different grid sizes (Verwijzing). As mentioned before, the (effects of) rounding errors should be smaller than the truncation errors. In Fig. 12 the mean truncation error of test case 2 is shown for two different horizontal grid sizes.

A finer grid hardly contributes to a more accurate result. Furthermore, the order of the truncation error is the same as the order of the absolute error on whole time steps. This implies that grid refinement does not lead to more accurate results.

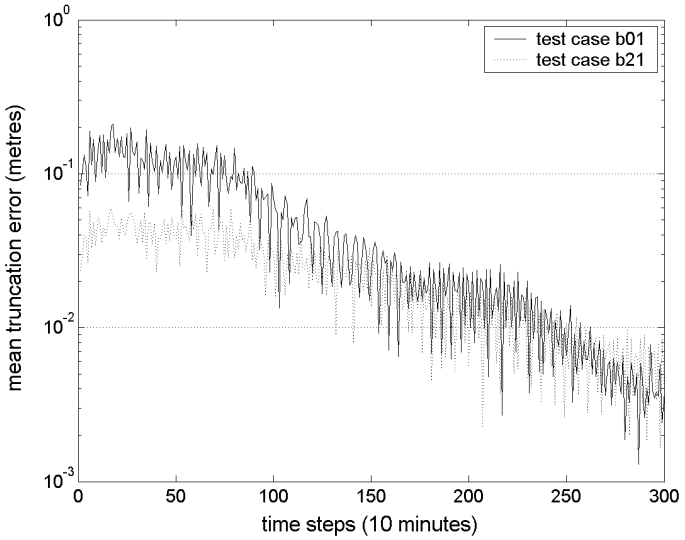

Figure 12: Mean truncation error in the spatial directions for test case 2. Horizontal grid size: b01 - 1000 m; b21 - 333 m.

4.10 Stability

The Matlab program confirms the unconditional stability of the discretised 2D SWE. The stability of the 3D SWE is influenced by several parameters. The program, however, does not predict possible instabilities, as the results do not correspond with the test runs. Therefore we will only analyse the consequences that changes to the parameters have on the stability.

First of all, we look at the flow velocity profile in vertical direction. When this an uniform profile the discretisation is stable. Due to bottom and surface friction this is never the case.

An increase of the horizontal eddy viscosity logically results in a more stable model. However, for the vertical eddy viscosity the opposite appears to be true. (*Ik heb eigenlijk nog steeds geen goede verklaring voor.*)

Furthermore, increasing the number of σ-layers reduces the stability of the model. So, it appears that the embedding (*Is dit goed Engels?*) of the vertical direction is main cause of possible instabilities.

5 CONCLUSIONS

5.1 Rounding errors

The computer architecture is not a source for differences in computed flow data. These differences are mainly of the same order as the effects of the rounding errors.

The two most important quantities which influence the accuracy of the computed results are the machine precision the size of the model. An increase of the first results in a corresponding increase of the accuracy. Increasing the size of the model has a negative effect on the accuracy, especially in the vertical direction.

The condition number for the continuity equation ($10^4 - -10^5$) is large in comparison with the machine precision (10^{-7}). Row-scaling leads to a reduction of the condition number, depending on the size of the model. Larger models benefit less.

Decreasing the grid size does not necessarily results in more accurate results, as the effects of rounding errors may be larger than the truncation error.

Higher accuracy, implementation of double precision and/or row-scaling, leads to shorter spin-up times.

5.2 Stability

It is analytically proved that the discretised 2D SWE (as implemented in Delft3D-FLOW) are unconditionally stable. However, incorporating a third dimension, even when done implicitly, can lead to unstable behaviour.

References

Golub, G. and C. Van Loan (1996). *Matrix Computations* (Third ed.). Baltimore: John Hopkins.

Mitchell, A. and D. Griffiths (1980). *The Finite Difference Method in Partial Differential Equations*, pp. 59–70, 148–154. New York: John Wiley.

Phillips, N. (1957). A coordinate system having some special advantages for numerical forecasting. *Journal of Meteorology 14*, 184–185.

Stelling, G. (1983). *On the construction of computational methods for shallow water flow problems*. Ph. D. thesis, Delft University of Technology, Delft.

Wesseling, P. (2001). *Principles of Computational Fluid Dynamics*. Springer Series in Computational Mathematics, Vol.29. Heidelberg: Springer.