ON THE CONSTRUCTION OF COMPUTATIONAL METHODS FOR SHALLOW WATER FLOW PROBLEMS

ON THE CONSTRUCTION OF COMPUTATIONAL METHODS FOR SHALLOW WATER FLOW PROBLEMS

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR IN DE TECHNISCHE WETENSCHAPPEN AAN DE TECHNISCHE HOGESCHOOL DELFT, OP GEZAG VAN DE RECTOR MAGNIFICUS PROF. B. P. TH. VELTMAN, IN HET OPENBAAR TE VERDEDIGEN TEN OVERSTAAN VAN HET COLLEGE VAN DEKANEN OP DINSDAG 6 DECEMBER 1983 te 14.00 UUR

DOOR

GUSTAAF SJOERD STELLING, WISKUNDIG INGENIEUR

GEBOREN TE ZAANDAM

DRUKKERIJ GRAFIA - PIJNACKER

1983

Dit proefschrift is goedgekeurd
door de promotor
prof. dr. ir. P. Wesseling

aan Joan

Voorwoord

Hierbij wil ik iedereen bedanken die op enigerlei wijze heeft bijgedragen aan de totstandkoming van dit proefschrift.
In het bijzonder wil ik hierbij noemen:
Dr. J.J. Leendertse, A. Staakman, Dr. G.K. Verboom, Prof. P. Wesseling en Ir. J. Willemse.
Tenslotte wil ik de directie van de Dienst Informatie Verwerking van Rijkswaterstaat danken voor de gelegenheid die ik kreeg om dit proefschrift af te ronden.

CONTENTS

CONTENTS (continued)

0 Introduction

The usefulness of mathematical models based on shallow water equations (SWE) is generally recognized for hydraulic problems in civil engineering. Mathematical models based upon SWE are applied not only to estimate water levels but also for the calculation of detailed flow patterns. Not only should the numerical method be accurate but it must also be stable. This stability should not be obtained at the cost of numerical dissipation. Many existing methods produce disappointing results because either instabilities are obtained or numerical dissipation causes very inaccurate results, especially if the flow contains eddies.

This work contains a step-by-step description of the construction of a finite difference method (FDM) for the approximation of SWE. A robust, yet accurate, FDM is constructed which is applicable to a wide range of practical problems of SWE in civil engineering.

The first chapter discusses general notions like stability and convergence. Other aspects such as wave propagation properties and the existence of spurious roots are also important. The understanding of FDM behaviour is enlarged by considering boundary value problems as well as initial value problems. It is sufficient for present purposes to consider only linear equations with one dependent variable.

The second chapter is devoted to special methods for approximating a simple advection equation. These methods can be implemented for the approximation of the advection operator of SWE. FDMs which are accurate with respect to time-dependent problems are not necessarily accurate when they are applied to steady state problems. For practical applications a FDM must be accurate in both cases.

The third chapter describes several FDMs for SWE that are well-known from the literature. To distinguish the differences, it is sufficient to consider only linearized and homogeneous SWE. The important advantages of so-called "staggered grids" will also be explained.

The most efficient existing FDM for practical problems of SWE in civil engineering seems to be the Leendertse method. Moreover, practical experience with this method is extensive. Yet this method has a disadvantage concerning linear

stability with respect to the advection operator. This chapter shows how this method can be stabilized, so that the efficiency is maintained by application of the methods proposed in chapter 2. A linear stability analysis of the resulting scheme is given.
Most of the schemes described in this chapter are of the ADI-type. For practical applications these schemes often yield large inaccuracies for very large timesteps.

In chapter 4 it is shown how one of the FDMs which are proposed in chapter 3 can be extended to an approximation method for nonlinear SWE.
A few aspects of nonlinear FDMs will be illustrated by simple examples.
For practical applications the boundary treatment is very important.
In this chapter the boundary treatment is considered from a practical point of view. For example, water levels or velocities can be prescribed at open boundaries such that almost non-reflective boundary conditions are obtained not requiring so-called "Riemann invariants".
This chapter also contains a description of the numerical treatment of tidal flats.

In chapter 5 a few examples will show that the FDM proposed in chapter 4 is applicable to a wide range of practical problems. The approach is purely numerical, i.e., model adjustment is considered to be beyond the scope of this work. This chapter shows, for example, that for complicated flow patterns the stability of the model is maintained even when the viscosity is very small. It also demonstrates the tidal flat procedure and the sensitivity of the model to variation of several model parameters of non-slip boundary conditions versus perfect slip boundary conditions. A practical steady flow problem is described. All descriptions in this chapter are brief since chapter 5 is meant as an illustration of the applicability of the FDM.

Chapter 6 contains general conclusions.

1 Preliminary Remarks on Linear Equations

1.0 Introduction

Section 1 of this chapter treats basic aspects of the numerical approximation of differential equations such as consistency, convergence and stability. The treatment will be based upon linear equations with a time-space domain that is one-dimensional in space. To explain the relevant concepts however this is not a real limitation.
Section 2 describes a few methods for the determination of stability conditions of finite difference schemes. Examples are given both for ordinary differential equations (ODEs) and for partial differential equations (PDEs).
Section 3 deals with the accuracy of numerical approximations of initial value problems of ODEs and PDEs in terms of phase and amplitude errors.
Section 4 illustrates some aspects of the numerical approximation of boundary conditions concerning the relation between the order of consistency and the order of convergence by means of a simple example.

1.1 Definitions

This section treats the concepts of consistency, convergence, and stability of approximation methods for differential equations. The definitions are based upon the work of Godunov and Ryabenki [4].

A differential equation will be written in the symbolic form given by:

$$L\, u(x,t) = f(x,t),\ x \in \Omega,\ \Omega \in R,\ t \in [o,T] \tag{1.1-1a}$$

with boundary conditions*:

$$1\, u(x,t) = \phi(x,t),\ (x,t) \in \Gamma\ , \tag{1.1-1b}$$

where:
L and 1 denote differential operators and

* Initial conditions are considered as a special type of boundary conditions.

Γ is the boundary of Ω x [0,T]

The system of equations (1.1-1) is assumed to have a unique solution u(x,t) which belongs to a normed linear function space U and for which the following relation holds:

$$\| u(x,t) \|_U \leq C_1 \| f \|_F + C_2 \| \phi \|_\Phi \qquad (1.1\text{-}2)$$

where F and Φ are normed linear function spaces, C_1 and C_2 are constants and $\| \cdot \|_U$, $\| \cdot \|_F$ and $\| \cdot \|_\Phi$ denote norms in the spaces U, F and Φ.
From this relation it follows that (1.1-1) is a well-posed problem in the sense of the definitions given by Kreiss [12].
Instead of solving (1.1-1) exactly, we want to approximate this equation by a "finite difference scheme". For this purpose we define a grid. The grid consists of the Cartesian product of a spatial grid Ω_Δ that is composed of a set of points with coordinates $x=m\Delta$, $m \in Z$ and a time grid T_Δ that is composed of a set of points with coordinates $t=k\tau$, $k \in Z$.

We suppose that τ is a function of Δ such that:

$$\lim_{\Delta \to 0} \tau(\Delta) = 0, \quad \tau(0) = 0 \qquad (1.1\text{-}3)$$

for example $\tau = r\Delta$ were r is a constant.

The boundary of Ω_Δ x T_Δ is denoted by Γ_Δ.
At this point we introduce normed linear function spaces U_Δ, F_Δ and Φ_Δ with norms $\| \cdot \|_{U_\Delta}$, $\| \cdot \|_{F_\Delta}$ and $\| \cdot \|_{\Phi_\Delta}$. The elements of the spaces U_Δ and F_Δ are defined on Ω_Δ x T_Δ while the elements of Φ_Δ are defined on Γ_Δ. These elements are denoted by $u^{(\Delta)}$, $f^{(\Delta)}$, and $\phi^{(\Delta)}$. They will be called "grid functions".

The equations which approximate (1.1-1a) are denoted by:

$$L_\Delta u^{(\Delta)} = f^{(\Delta)} \qquad (1.1\text{-}4a)$$

where L_Δ is called a "finite difference operator" with domain U_Δ and range F_Δ.

The boundary conditions given by (1.1-1b) are approximated by:

$$l_\Delta u^{(\Delta)} = \phi^{(\Delta)} \qquad (1.1\text{-}4b)$$

The domain and range of l_Δ are given by U_Δ and Φ_Δ respectively.

In order to obtain a useful approximation of (1.1-1) by (1.1-4) the approximation must be:

(i) consistent (of order 1 at least)

(ii) convergent (of order 1 at least)

(iii) stable.

For the definition of these concepts we follow Godunov and Ryabenki [4].

Definition (1.1-1) (consistency)

The finite difference scheme given by (1.1-4a) is a consistent approximation of order n of (1.1-1a) if the following relation holds:

$$\| L_\Delta [u]_\Delta - \{Lu\}_\Delta \|_{F_\Delta} + \| f^{(\Delta)} - \{f\}_\Delta \|_{F_\Delta} \leq C_3 \Delta^n$$

where C_3 denotes some constant,

$[.]_\Delta$ is an operator that relates elements of the space U to elements of the space U_Δ, and $\{.\}_\Delta$ is an operator that relates elements of F to elements of F_Δ.

Definition (1.1-2) (convergence)

The finite difference scheme given by (1.1-4) is convergent of order n if the following relation holds:

$$\| [u]_\Delta - u^{(\Delta)} \|_{U_\Delta} \leq C_4 \Delta^n$$

where C_4 denotes some constant not depending on Δ and $[.]_\Delta$ is defined by definition (1.1-1).

Definition (1.1-3) (G-R stability)

The finite difference scheme given by (1.1-4) is said to be G-R stable if the following relation holds:

$$\| u^{(\Delta)} \|_{U_\Delta} \leq K_3 \| f^{(\Delta)} \|_{F_\Delta} + K_4 \| \phi^{(\Delta)} \|_{\Phi_\Delta}, \quad \forall \Delta$$

where K_3 and K_4 are numbers not depending on Δ.

This definition is equivalent to the assumption that the inverse operator related to the finite difference scheme given by (1.1-4) is uniformly bounded as $\Delta \to 0$, cf. Godunov and Ryabenki [4], p. 105.

It is further to be noted that also for the approximation of the boundary conditions consistency can be defined according to definition (1.1-1), see Godunov and Ryabenki [4].

The stability definition given by definition (1.1-3) is by no means the only possible stability definition. A survey of stability definitions is given by Van der Houwen [8]. For hyperbolic problems, see also the stability definition as given by Kreiss et al. [13].

From a practical point of view, the so-called B-H-K stability as treated by Van der Houwen [8] deals with an important aspect of the behaviour of finite difference schemes. To explain this type of stability definition and its difference with respect to G-R stability, we rewrite the finite difference scheme given by (1.1-4) as:

$$u^{k+1} = R_\Delta u^k + \tau \rho^k \qquad (1.1\text{-}5)$$

where R_Δ is an operator with its domain and range in U_Δ. This operator relates to the values of $u^{(\Delta)}$ at time level $t=k\tau$ the values of $u^{(\Delta)}$ at time level $t=(k+1)\tau$, cf. Godunov and Ryabenki [4]. The values of ρ^k depend on the boundary conditions and the right-hand side of the finite difference equation.

Under the conditions that

(i) $\| \rho^k \| \leq L_1 \| f^{(\Delta)} \|_{F_\Delta} + L_2 \| \phi^{(\Delta)} \|_{\Phi_\Delta}$

and

(ii) $\| u^o \| \leq L_3 \| f^{(\Delta)} \|_{F_\Delta} + L_4 \| \phi^{(\Delta)} \|_{\Phi_\Delta}$

stability in the sense of definition (1.1-3) is equivalent to the relation:

$$\| R_\Delta^k \| \leq L_o, \quad k=1,\ldots,T/\tau, \quad \forall\Delta$$

where L_o, L_1, ..., L_4 are constants not depending on Δ.

In other words G-R stability implies that $R_\Delta^{T/\tau}$ is a uniformly bounded operator if $\Delta \to 0$ and thereby $\tau \to 0$ with T constant.

B-H-K stability as defined by Van der Houwen [8] concerns what happens with $\| R_\Delta^k \|$ if $T \to \infty$ while Δ, and thereby τ, are kept constant. In a somewhat simplified form B-H-K stability is defined by:

Definition (1.1-4) (B-H-K stability)

The finite difference scheme given by (1.1-5) is B-H-K stable if

$$\lim_{T\to\infty} \| R_\Delta^{T/\tau} \| \leq M_o$$

while Δ and τ are kept constant.

M_o denotes a constant not depending on T.

From a practical point of view it is convenient if a finite difference scheme is stable in the sense of both stability definitions of this section, although the latter stability definition is in fact meaningful only if the relation given by (1.1-2) also holds if $T\to\infty$. For tidal problems, the final goal of this work, this seems to be a reasonable assumption, however.

Finally we would like to mention that G-R stability implies convergence for consistent finite difference schemes as is proven by Godunov and Ryabenki [4].

1.2 Tools for the determination of stability conditions

To be useful in practice, finite difference schemes need to be consistent, stable, and convergent.
Consistency is generally easy to check by means of Taylor series expansions. Stability, however, and thereby convergence, often implies a complicated analysis. For a practical simulation model, such an analysis is a necessity because ignorance of the stability conditions could lead to meaningless results. This section describes a few tools for the determination of the stability condition of a finite difference scheme.

In general we consider as very important two aspects of a method for the determination of stability conditions:
(i) The method should be fairly simple to apply.
(ii) It should not grossly overestimate the "true" stability conditions. This might lead to very small timesteps, which, in practical applications imply costly and thereby not very competitive computing codes.

In view of this last remark we generally prefer methods that yield only necessary conditions to methods that produce only sufficient conditions. Conditions that are both necessary and sufficient are of course always preferable but are usually hard to come by. The test problem for which the stability study is carried out is almost always a simplification of the problem that is to be solved in reality. Therefore stability always has to be verified by computer calculations with the real model.

Three methods are often applied for the calculation of stability conditions:

(i) The energy method
After the choice of a suitable norm the condition given by definition (1.1-3), or a similar definition, is checked by direct estimation of $\|u^{(\Delta)}\|_{U_\Delta}$. The tools for this method are the triangle inequality, summation by parts etc.
In general only sufficient conditions are found by this method. Boundary conditions are included, and nonlinear problems can also be studied. For nonlinear problems, nonhomogeneous boundary conditions are difficult to study. Examples of this method are given by Richtmyer and Morton [21],

Cuvelier [1], and Temam [27]. A simple example of this method is also given in section 4.1. The disadvantage is that, while providing only sufficient conditions, this method often causes complicated analytical problems for simple finite difference schemes, see also the discussion of Roache [22], p. 48.

(ii) Spectral method

For this method stability is considered as the boundedness of the operator R_Δ^k where R_Δ is defined by (1.1-5).
As already mentioned, stability in the sense of definition (1.1-3) is equivalent to:

$$\| R_\Delta^k \| \leq L_o, \quad k=1,\ldots T/\tau(\Delta), \quad \forall\, \Delta \tag{1.2-1}$$

where T is constant and $\lim_{\Delta\to 0} \tau(\Delta) = 0$, $\tau(o) = 0$

For PDEs the behaviour of R_Δ^k is studied by the definition and the calculation of the spectrum of a family of operators $\{R_\Delta\}$, see Godunov and Ryabenki [4], p. 188.

(iii) Heuristic stability theory

This stability analysis is treated by Hirt [7] and in a somewhat modified form by Warming and Hyett [26]. The method is based upon the following idea:
let the finite difference scheme given by (1.2-1) be an n^{th} order consistent approximation of the following equation.

$$Lu = f, \tag{1.2-2a}$$

with boundary conditions:

$$lu = \phi, \tag{1.2-2b}$$

and a $n+1^{th}$ order consistent approximation of another equation

$$L'u = f', \tag{1.2-3a}$$

with boundary conditions:

$$l'u = \phi'. \tag{1.2-3b}$$

Then for the stability of (1.1-5) not only should (1.2-2) represent a well-posed problem, i.e. its solution should fulfil the condition given by (1.1-2), but (1.2-3) should also represent a well-posed problem.
There is no formal theory for the justification of this method, but the simplicity of its application, especially for semi-discrete problems, makes it an attractive method.

In this section we first treat the calculation of stability conditions for ODEs and then for PDEs.

a. Ordinary differential equations

For the numerical approximation of linear differential equations often linear finite difference equations of order ℓ in the following form are constructed (linear multistep methods):

$$\gamma^{\ell} u^{k+\ell} + \gamma^{\ell-1} u^{k+\ell-1} + \ldots + \gamma^{o} u^{k} = \phi^{k}, \quad k = k_o, k_o+1, \ldots \tag{1.2-4}$$

where γ^j, $j = 0, 1, \ldots, \ell$ are constants independent of k and $\gamma^{\ell} \neq 0$, $\gamma^{o} \neq 0$.

As is well known, see, e.g., Lambert [15], the general solution of (1.2-4) is given by:

$$u^k = \sum_{i=1}^{L} P_i(k)\, r_i^k + \xi^k \tag{1.2-5}$$

where ξ^k denotes a particular solution of (1.2-4) and $P_i(k)$ are polynomials in k. The degree of $P_i(k)$ is μ_{i-1}, with μ_i the multiplicity of the corresponding r_i, while r_i are the roots of the corresponding "characteristic equation", see e.g. Lambert [15], given by:

$$\gamma^{\ell} r^{\ell} + \gamma^{\ell-1} r^{\ell-1} + \ldots + \gamma^{o} = 0 \tag{1.2-6}$$

L denotes the number of roots of (1.2-6), from which it follows that:

$$\sum_{i=1}^{L} \mu_i = \ell \tag{1.2-7}$$

The coefficients of the polynomials $P_i(k)$ are determined by initial or boundary conditions.

(1.2-4) can be denoted also as:

$$[u^{k+\ell},\ldots,u^{k+1}]^T = R[u^{k-1+\ell},\ldots,u^{k}]^T + [\phi^k,0,\ldots 0]^T \tag{1.2-8}$$

where R is given by:

$$R=\begin{bmatrix} -\gamma^{\ell-1}/\gamma^{\ell} & -\gamma^{\ell-2}/\gamma^{\ell} & \cdot & \cdot & \cdot & -\gamma^{o}/\gamma^{\ell} \\ 1 & 0 & & & & \\ & \cdot & \cdot & & & \\ & & \cdot & \cdot & & \\ & & & \cdot & \cdot & \\ & & & & 1 & 0 \end{bmatrix}$$

The eigenvalues of R are determined by the characteristic equation given by (1.2-6).

The general solution of the homogeneous part of (1.2-4) or (1.2-8) can be obtained by superposition of terms $\overline{u}^k$, given by

$$\overline{u}^k = k^{\kappa}\, r_i^{\,k} \tag{1.2-9}$$

where κ is a natural number such that $0 \leq \kappa < \mu_i$; these terms are usually referred to as the "normal modes" of (1.2-3) or (1.2-8).

The stability definitions as defined in the theory of approximate solutions of ODEs relate to the eigenvalues of the operator R_τ that is obtained after application of the approximation method under consideration to a simple "test problem", given by:

$$u_t = \lambda u \tag{1.2-10}$$

The general formulation of a linear multistep method of order ℓ for the approximation of (1.2-10) is given by:

$$\sum_{j=0}^{\ell} \alpha^j u^{k+j} = \tau \lambda \sum_{j=0}^{\ell} \beta^j u^{k+j} \tag{1.2-11}$$

As for (1.2-4), the general solution of (1.2-11) is given by:

$$u^k = \sum_{i=1}^{L} P_i(k) \, r_i^k, \tag{1.2-12}$$

where r_i, $i=1,\ldots,L$, $L \leq \ell$, are the roots of the characteristic equation given by:

$$\pi(r,\lambda\tau) = \rho(r) - \lambda\tau \, \sigma(r) = 0, \tag{1.2-13}$$

where:

$$\rho(r) = \sum_{j=0}^{\ell} \alpha_j r^j$$

$$\sigma(r) = \sum_{j=0}^{\ell} \beta_j r^j.$$

The polynomials $\rho(r)$ and $\sigma(r)$ are referred to as the first and second characteristic polynomials, see, e.g., Lambert [15].

Definition (1.2-1) (zero-stability)

The linear multistep method (1.2-11) is said to be "zero-stable" if no root of the equation $\rho(r) = 0$ has a modulus greater than 1, and if every root with modulus 1 is simple.

Zero stability is a necessary condition for G-R stability because G-R stability of consistent linear finite different schemes ensures convergence, see Van der Houwen [8] p. 12, while zero stability of consistent linear multistep methods is equivalent to convergence, see Lambert [15] p. 33.

Definition (1.2-2) (absolute stability)

The linear multistep method (1.2-11) is said to be "absolutely stable" for a given $\tau\lambda$ if, for that $\tau\lambda$, all roots r_i of (1.2-13) satisfy:
$|r_i| < 1$, i=1, ..., L.

Absolute stability is a sufficient condition for B-H-K stability (def. 1.1-4). Note that if the definition had required $|r_i| \leq 1$ it would have been a necessary condition for B-H-K stability. This shows the close relation between absolute stability and B-H-K stability.

In general, absolute stability does not imply zero stability and vice versa; therefore they will both have to be checked. For linear ODEs their verification is comparatively simple.

The treatment given here of linear multistep methods has been rather limited. More detailed information on these subjects, including Runge Kutta methods, is given by Lambert [15], Gear [3], Lapidus and Seinfeld [17], Henrici, [6] and, Van der Houwen [9].

b. Partial differential equations

So far we have studied the stability of finite difference schemes by establishing the eigenvalues of an operator R as defined, for example, by (1.2-8). We have only studied a scalar ODE, for systems of ODEs however, the analysis is not essentially different, cf. Lambert [15].
The analysis of the stability of finite difference schemes for PDEs, however, involves a fundamentally different aspect from the stability analysis of finite difference schemes for ODEs.

To show this we consider a simple hyperbolic PDE given by:

$$u_t + u_x = 0, \qquad x \in [0,1],\ t \in [0,T] \tag{1.2-14a}$$

$$u(0,t) = 1,\ u(x,0) = 0,\ x > 0 \tag{1.2-14b}$$

For the approximation of this equation we define a spatial grid consisting of M+1 points x_m, $x_m = m\Delta x$, $\Delta x = 1/M$.
For the approximation of the spatial derivatives we use first order "upwind" differencing. The approximation of the derivatives in time is postponed. This yields a so-called "semi-discrete" set of ordinary differential equations given by:

$$(u_m)_t + (u_m - u_{m-1})/\Delta x = 0, \quad m = 1, \ldots, M \tag{1.2-15a}$$

$$u_o(t) = 1, \quad u_m(0) = 0, \quad m = 1, \ldots, M \tag{1.2-15b}$$

This method does not need a special boundary scheme at m = M.

Equation (1.2-17a) can be rewritten as follows:

$$\underline{u}_t = A\,\underline{u}\,, \quad \underline{u}(0) = \underline{u}^o \tag{1.2-16}$$

where

$$\underline{u} = [u_o, u_1, \ldots, u_M]^T$$

$$A = \begin{bmatrix} 0 & & & & \\ +1/\Delta x & -1/\Delta x & & & \\ & +1/\Delta x & -1/\Delta x & & \\ & & \ddots & \ddots & \\ & & & +1/\Delta x & -1/\Delta x \end{bmatrix}$$

$$\underline{u}^o = [1, 0, \ldots, 0]^T$$

The eigenvalues of A are given by 0 and $-1/\Delta x$.

The approximate integration in time is accomplished by the explicit first order Euler method. For the definition of the time grid, in the interval [0,T], we choose K+1 equidistantial points such that $\tau = T/K$, and define the following finite difference scheme:

$$u_m^{k+1} = u_m^k - r\,(u_m^k - u_{m-1}^k),\quad k = 0,\ \ldots, K-1,\quad m = 1,\ldots,\ M \tag{1.2-17a}$$

$$u_o^k = 1,\ k = 0,\ldots,K,\quad u_m^o = 0,\ m = 1,\ldots,\ M, \tag{1.2-17b}$$

where $r = \tau/\Delta x$

Looking at the first characteristic polynomial of (1.2-17a) we immediately see that the method is unconditionally zero stable.
This process, in which discretization in space is followed by numerical integration in time, is generally referred to as the "method of lines".

First we assume (1.2-17) to be an approximation of (1.2-16) and we study the limiting case $\tau \to 0$ while we keep Δx constant.
Second we assume (1.2-17) to be an approximation of (1.2-14) and we study the limiting case $\Delta x \to 0$ while we keep r constant.

If we construct the operator R_τ in the sense of (1.2-1) under the first assumption, then we obtain:

$$R_\tau = \begin{bmatrix} 0 & & \\ r(\tau) & 1 - r(\tau) & \\ & r(\tau) & 1 - r(\tau) \end{bmatrix} \tag{1.2-18}$$

The eigenvalues of this operator are given by $0, 1 - r(\tau)$. Absolute stability is ensured if $0 < r(\tau) < 2$ or:

$$\tau/\Delta x < 2 \tag{1.2-19}$$

If we construct the operator $R_{\Delta x}$ for the second assumption then we obtain:

$$R_{\Delta x} = \begin{bmatrix} 0 & & \\ r & 1-r & \\ & r & 1-r \end{bmatrix} \tag{1.2-20}$$

The eigenvalues are given by 0, 1-r, which means that B-H-K stability is ensured if the condition given by (1.2-19) is fulfilled. For G-R stability, however, the true stability condition, see Godunov and Ryabenki [4] p.190, is given by:

$$r < 1 \qquad (1.2\text{-}21)$$

This condition can be found also by the "Von Neumann condition", see Richtmyer and Morton [21], p. 152.
The reason for the differences between the conditions given by (1.2-19) and (1.2-21) is that for ODEs R_τ has a fixed size and for B-H-K stability the size of R_τ is assumed to be fixed while for G-R stability the size of $R_{\Delta x}$ tends to infinity, and if $r > 1$ some elements of $R^k_{\Delta x}$ tend to infinity as well.

From this example, taken from Godunov and Ryabenki [4], it follows that the "matrix method" as described by Mitchell and Griffiths [19] and used for example by Praagman [20] does not always yield sufficient conditions for G-R stability. The matrix method does not include the Von Neumann method; in fact, by the matrix method B-H-K stability is verified.

Observations of this kind have led Godunov and Ryabenki [4] to introduce the concept of a spectrum of a family of operators denoted by $\{R_{\Delta x}\}$. This is the aggregate of operators $R_{\Delta x}$ formed by letting Δx assume all possible values of the grid size. It is assumed that Δx can be arbitrarily small. The spectrum of $\{R_{\Delta x}\}$ is defined as:

Definition (1.2-4)

The complex number λ denotes a point in the spectrum of the family of operators $\{R_{\Delta x}\}$ if, for any positive ε, we may choose Δx_o ($\Delta x_o > 0$) such that for any Δx, $0 < \Delta x < \Delta x_o$, there exists a vector u (from the appropriate space $U_{\Delta x}$) which satisfies the inequality:

$$\| R_{\Delta x} u - \lambda u \| < \varepsilon \| u \|$$

The aggregate of all such numbers λ forms the spectrum of the family of operators $\{R_{\Delta x}\}$.

After this definition Godunov and Ryabenki [4] prove a very important assertion:

Let one point λ_o in the spectrum of the family of operators $\{R_{\Delta x}\}$ lie outside the unit disk in the complex plane ($|\lambda_o| > 1$). It is then impossible to choose a constant C such that for all Δx the inequality

$$\| R^k_{\Delta x} \| < C$$

is satisfied, where k assumes all integer values from 0 to $k_o(\Delta x)$. It is assumed that $k_o(\Delta x) \to \infty$ if $\Delta x \to 0$.

It is also shown by Godunov and Ryabenki [4] that this spectrum is not equivalent to the aggregate of eigenvalues of each operator $R_{\Delta x}$ by means of an example similar to the one already given in this section.

We will now demonstrate the calculation of the spectrum of $\{R_{\Delta x}\}$ as given by (1.2-20).

The numerical boundary-initial-value problem is split into a number of subproblems.

The first one, the so-called Cauchy problem, see Kreiss [10], assumes the spatial domain to be infinite, i.e. $-\infty < x < \infty$, or:

$$u^{k+1}_m = u^k_m - r\,(u^k_m - u^k_{m-1}), \quad m = 0, \pm 1, \pm 2, \ldots \tag{1.2-22a}$$

This problem is usually referred to as a half plane problem and is a purely initial value problem.

The other two problems are quarter space problems given by:

$$u^{k+1}_m = u^k_m - r\,(u^k_m - u^k_{m-1}), \quad m = 1, 2, \ldots, \quad (\text{i.e. } 0 < x < \infty) \tag{1.2-22b}$$

$$u^k_o = 0,$$

and:

$$u^{k+1}_m = u^k_m - r\,(u^k_m - u^k_{m-1}), \quad m = \ldots, -1, 0, 1, \ldots, M \quad (\text{i.e. } -\infty < x < 1) \tag{1.2-22c}$$

The spectrum of $\{R_{\Delta x}\}$ for this problem consists of the union of the spectra of each of the families of operators that belongs to each subproblem given by (1.2-22 a,b,c). The proof of this is given by Godunov and Ryabenki [4].

To calculate the eigenvalues of each of the three subproblems we assume a solution of (1.2-22) to be of the following form.

$$\tilde{u}_m^k = \lambda^k \hat{u}_m \tag{1.2-23}$$

The Godunov-Ryabenki condition is said to be satisfied if none of the three subproblems has nontrivial solutions in the form (1.2-23) for which $|\lambda| > 1$, see Kreiss et al [13].
This is a necessary condition for G-R stability.

To verify this condition (1.2-23) is substituted into (1.2-22).
This yields the so-called "resolvent equation" given by:

$$\lambda \hat{u}_m = \hat{u}_m - r(\hat{u}_m - \hat{u}_{m-1}) \tag{1.2-24}$$

This resolvent equation is also a finite difference equation for which the general solution is given by:

$$\hat{u}_m = \alpha z^m \tag{1.2-25}$$

where α is a constant depending on boundary conditions and z is the root of the characteristic equation of (1.2-24) given by:

$$\lambda z = z - r(z-1) \tag{1.2-26}$$

By substitution of (1.2-25) into (1.2-23) we obtain:

$$\tilde{u}_m^k = \alpha \lambda^k z^m \tag{1.2-27}$$

Solutions in this form are called the "normal modes" of (1.2-22).

Because the possible solutions of each of the three subproblems of (1.2-22) are bounded at t = 0 it follows that for the verification of the G-R condition the following restrictions for z are to be taken into account:

For (1.2-22a): $|z| = 1$, i.e. $z = e^{i\sigma}$, $0 \leq \sigma < 2\pi$ (1.2-28a)

For (1.2-22b): $|z| \leq 1$ (1.2-28b)

For (1.2-22c): $|z| \geq 1$ (1.2-28c)

If these restriction of $|z|$ are taken into account, the verification of the G-R condition yields (1.2-21) as a necessary condition for stability (by stability we mean G-R stability unless another kind of stability is specified).

In Kreiss [10], Kreiss [11] and Kreiss, Gustafsson, and Sundström [13] a similar stability analysis is presented for hyperbolic problems in general with variable coefficients. In these articles not only are necessary conditions derived but also sufficient conditions. Strikwerda [24] gives a similar analysis for the method of lines.

In fact, this type of stability analysis is always concerned with how to reduce the stability problem to a problem that is simple to deal with. It is done by the construction of the "resolvent equation", which is obtained by substitution of a normal mode into the finite difference equation. The original problem is split into several quarter space problems and one purely initial-value problem; the so-called Cauchy problem, which can be treated with Von Neumann analysis. Even though they are meant to be simple, the quarter space problems are often very difficult to deal with. They have various analytical problems with high order, complex polynomials. For simple problems the insight into the possible behaviour of a numerical solution can be greatly enlarged by posing a purely initial-value problem and several boundary value problems. We will illustrate this later on by an example.

The treatment given in this section on the stability of approximate solutions of initial boundary value problems has been very brief and incomplete.

Because our main purpose is the construction of a stable method for the approximation of the shallow water equations by which practical problems can be solved, a thorough treatment is well beyond the scope of this work. The interested reader should read the articles by Godunov and Ryabenki and Kreiss mentioned in the references.

1.3 Dissipation and dispersion properties of finite difference schemes for initial-value problems

Up to this point we have dealt only with stability problems. Stability, however, is only one of the properties that a good finite difference scheme should have. To make a final choice for a numerical method for the approximation of a differential equation, it is necessary to study other properties. One of the important aspects of a finite difference scheme is accuracy, which concerns the amount of computational labour that is needed to obtain a numerical solution within prescribed error bounds; or, equivalently, it concerns the size of the error that can be expected at a prescribed amount of computational labour.

Qualitative expressions for the accuracy of a numerical method are the order of consistency or the order of convergence, more quantitative expressions however describe the global error of a numerical method in terms of phase and amplitude errors. For the numerical solution of hyperbolic PDEs this is a general approach, cf. Roberts and Weiss [23], Kreiss and Oliger [14] or Roache [22]. Also in the case of numerical methods for ODEs this approach is sometimes adopted, cf. Lambert [16].

In this section we study the phase and amplitude errors for numerical approximation methods that are constructed by discretizing a PDE separately in space and time. For this type of method, the phase and amplitude errors of the spatial and time discretization can be studied separately; in this section phase and amplitude errors of numerical integration methods in time are studied first and then these errors are studied for spatial discretizations. The study will be based upon the same normal mode analysis that has been applied to study stability. Other properties that are important for the final result of a numerical method such as spurious modes, relative stability, and stiff stability will be described briefly.

a. Ordinary Differential Equations

The phase and amplitude errors of a numerical method are defined by application of this method to a simple test problem. The one we use for the ODEs is given by:

$$u_t = \lambda u \tag{1.3-1a}$$

with initial condition given by:

$$u(0) = u_o \tag{1.3-1b}$$

The solution of (1.3) is:

$$u(t) = u_o e^{\lambda t} \tag{1.3-2}$$

If we assume that (1.3) is approximated by a linear one-step method then we obtain a numerical solution given by:

$$u^k = u_o \, r_1^{\,k} \tag{1.3-3}$$

Instead of restricting the continuous solution to a discrete function space, as for the definitions of consistency and convergence, we extend the discrete solution to a continuous function space such that we obtain a function $u'(t) = u_o \, e^{\lambda' t}$, for which the following relation holds:

$$u'(k\tau) = u_o e^{k \lambda' \tau} = u^k \tag{1.3-4}$$

Obviously $\lambda'\tau$ is given by:

$$\lambda'\tau = \ln r_1 \tag{1.3-5}$$

At this point we define the so-called propagation factor, cf. Leendertse [18], by:

$$P(\lambda, t) = e^{(\lambda' - \lambda)t} \tag{1.3-6}$$

The amplitude of this factor is called the amplitude factor and the phase is called the phase error. If, for Im $\lambda > 0$, Im $(\lambda'-\lambda) < 0$, one speaks of a lagging phase error; Im $(\lambda'-\lambda) > 0$ implies a leading phase error.
For later reference the figures (1-1), (1-2) and (1-3) show contour plots of amplitude factors and phase errors per timestep, plotted within their regions of stability in the $\tau\lambda$ plane of the following methods:

Figure (1-1) Euler explicit : $r_1 = 1 + \tau\lambda$
Figure (1-2) Trapezoidal rule : $r_1 = (1 + \tau\lambda/2)/(1-\tau\lambda/2)$
Figure (1-3) Predictor-Corrector: $r_1 = 1 + \tau\lambda + \frac{1}{2}(\tau\lambda)^2$

These methods are often used for the approximation of the time derivative of partial differential equations. From figure (1-2) it follows that the trapezoidal rule has no amplitude errors at the imaginary axis, and the phase errors are lagging. For wave problems the eigenvalues are quite often near the imaginary axis.

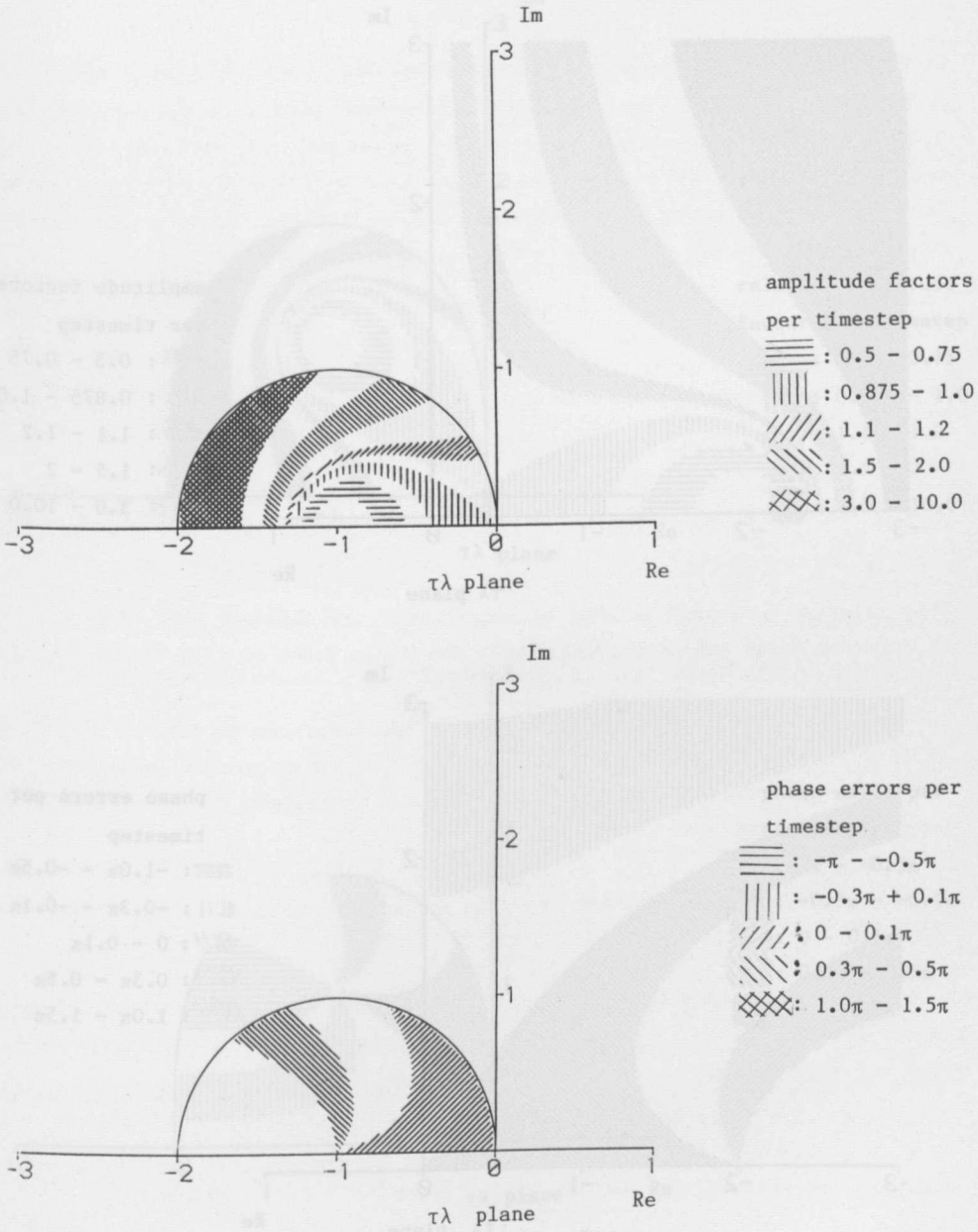


Figure (1-1) Phase errors and amplitude factors of Eulers explicit method

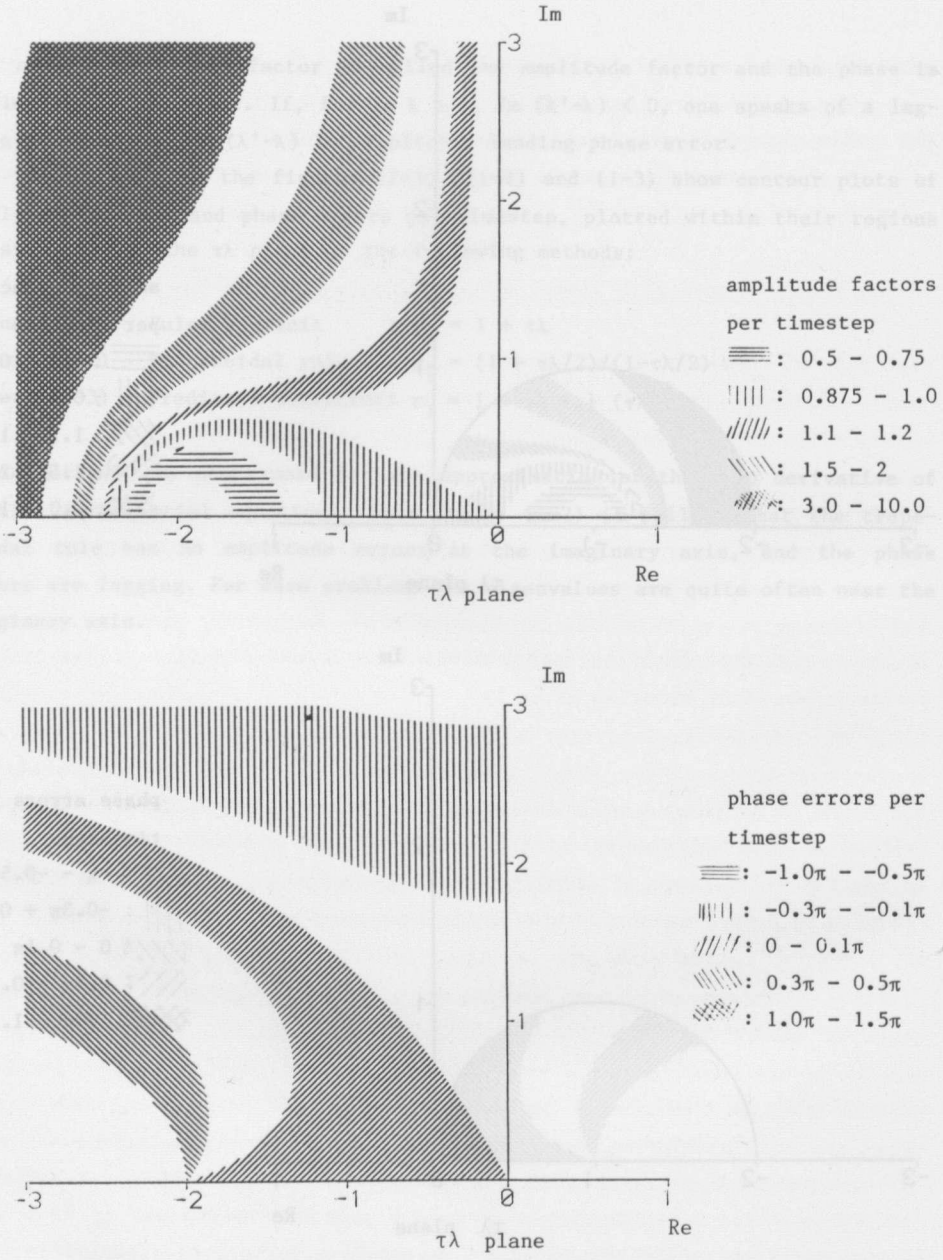


Figure (1-2) Phase errors and amplitude factors of the trapezoidal rule

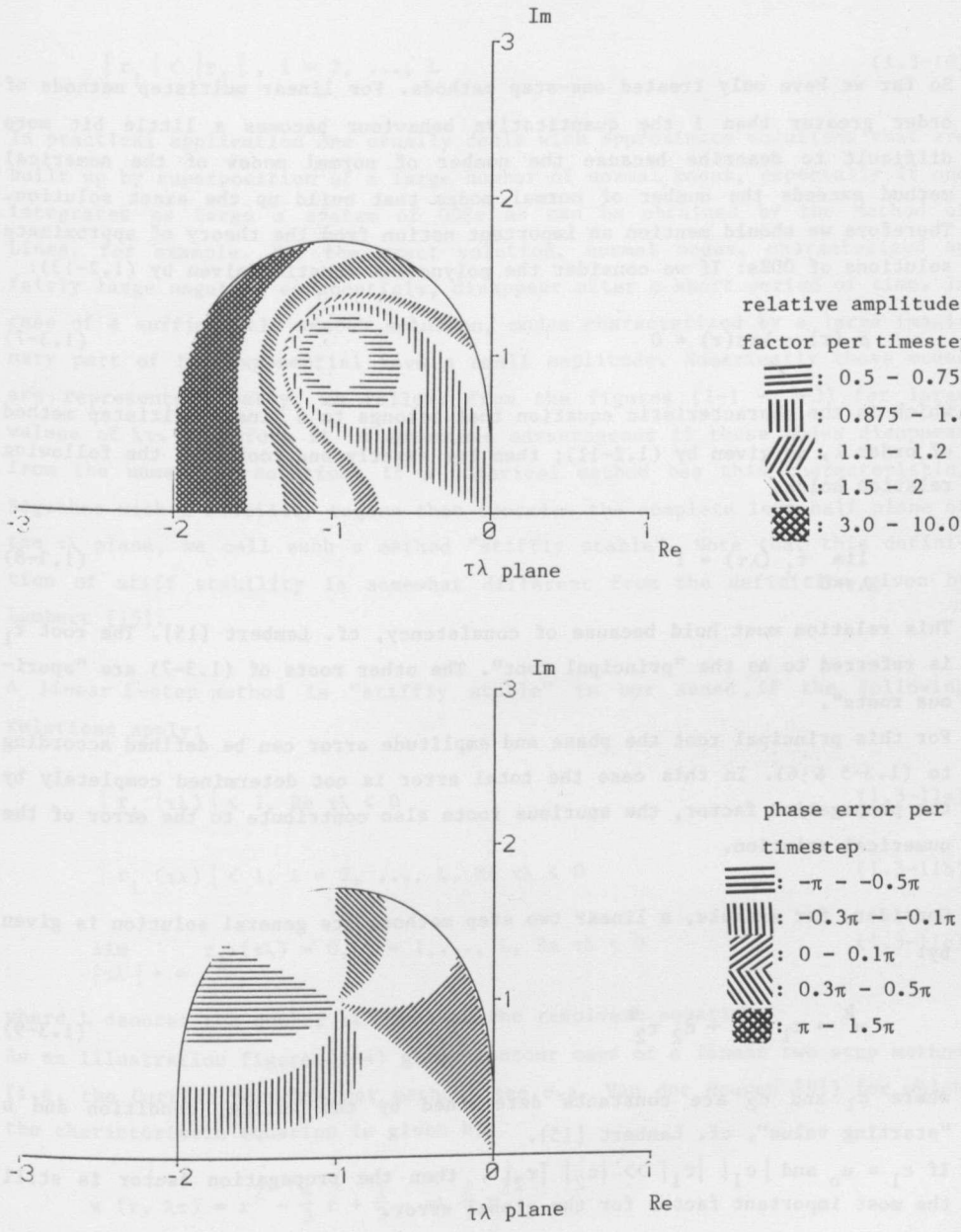


Figure (1-3) Phase errors and amplitude factors of the second order predictor corrector method

So far we have only treated one-step methods. For linear multistep methods of order greater than 1 the quantitative behaviour becomes a little bit more difficult to describe because the number of normal modes of the numerical method exceeds the number of normal modes that build up the exact solution. Therefore we should mention an important notion from the theory of approximate solutions of ODEs: If we consider the polynomial equation given by (1.2-13):

$$\rho(r) - \lambda\tau\sigma(r) = 0 \tag{1.3-7}$$

which is the characteristic equation that belongs to a linear multistep method of order ℓ, as given by (1.2-11); then for exactly one root, r_1, the following relation holds:

$$\lim_{\lambda\tau \to 0} r_1(\lambda\tau) = 1 \tag{1.3-8}$$

This relation must hold because of consistency, cf. Lambert [15]. The root r_1 is referred to as the "principal root". The other roots of (1.3-7) are "spurious roots".
For this principal root the phase and amplitude error can be defined according to (1.3-5 & 6). In this case the total error is not determined completely by the propagation factor, the spurious roots also contribute to the error of the numerical solution.

Consider, for example, a linear two step method. Its general solution is given by:

$$u^k = c_1 r_1^k + c_2 r_2^k \tag{1.3-9}$$

where c_1 and c_2 are constants determined by the initial condition and a "starting value", cf. Lambert [15].
If $c_1 \approx u_o$ and $|c_1|\,|r_1| >> |c_2|\,|r_2|$, then the propagation factor is still the most important factor for the global error.

Another important concept in the theory of the approximate solutions of ODEs is referred to as "relative stability". This means that for a linear multistep method with stepsize ℓ for all spurious roots, $r_2, \ldots, r_L$, $L < \ell$, the following relation holds:

$$| r_i | < | r_1 |, \quad i = 2, \ldots, L \tag{1.3-10}$$

In practical application one usually deals with approximate solutions that are built up by superposition of a large number of normal modes, especially if one integrates as large a system of ODEs as can be obtained by the Method of Lines, for example. For the exact solution, normal modes, characterized by fairly large negative exponentials, disappear after a short period of time. In case of a sufficiently smooth solution, modes characterized by a large imaginary part of the exponential have a small amplitude. Numerically these modes are represented poorly, as follows from the figures (1-1 - 1-3) for large values of $\lambda\tau$. Therefore it is sometimes advantageous if these modes disappear from the numerical solution. If a numerical method has this characteristic, together with a stability region that contains the complete left half plane of the $\tau\lambda$ plane, we call such a method "stiffly stable". Note that this definition of stiff stability is somewhat different from the definition given by Lambert [15].

A linear ℓ-step method is "stiffly stable" in our sense if the following relations apply:

$$| r_1 (\tau\lambda) | \leq 1, \quad \mathrm{Re}\ \tau\lambda \leq 0 \tag{1.3-11a}$$

$$| r_i (\tau\lambda) | < 1, \quad i = 2, \ldots, L, \quad \mathrm{Re}\ \tau\lambda \leq 0 \tag{1.3-11b}$$

$$\lim_{|\tau\lambda| \to \infty} r_i (\tau\lambda) = 0, \quad i = 1, \ldots, L, \quad \mathrm{Re}\ \tau\lambda < 0 \tag{1.3-11c}$$

where L denotes the number of roots of the resolvent equation.
As an illustration figure (1-4) gives contour maps of a linear two step method (i.e. the Curtiss Hirschfelder method, see e.g. Van der Houwen [9]) for which the characteristic equation is given by:

$$\pi (r, \lambda\tau) = r^2 - \frac{4}{3} r + \frac{1}{3} - \tau\lambda \frac{2}{3} r^2$$

Depending on the application amplitude errors are sometimes minimized or phase errors are forced to be as small as possible by special "exponential fitting" or "frequency fitting" methods, see, e.g., Lambert [15], [16].

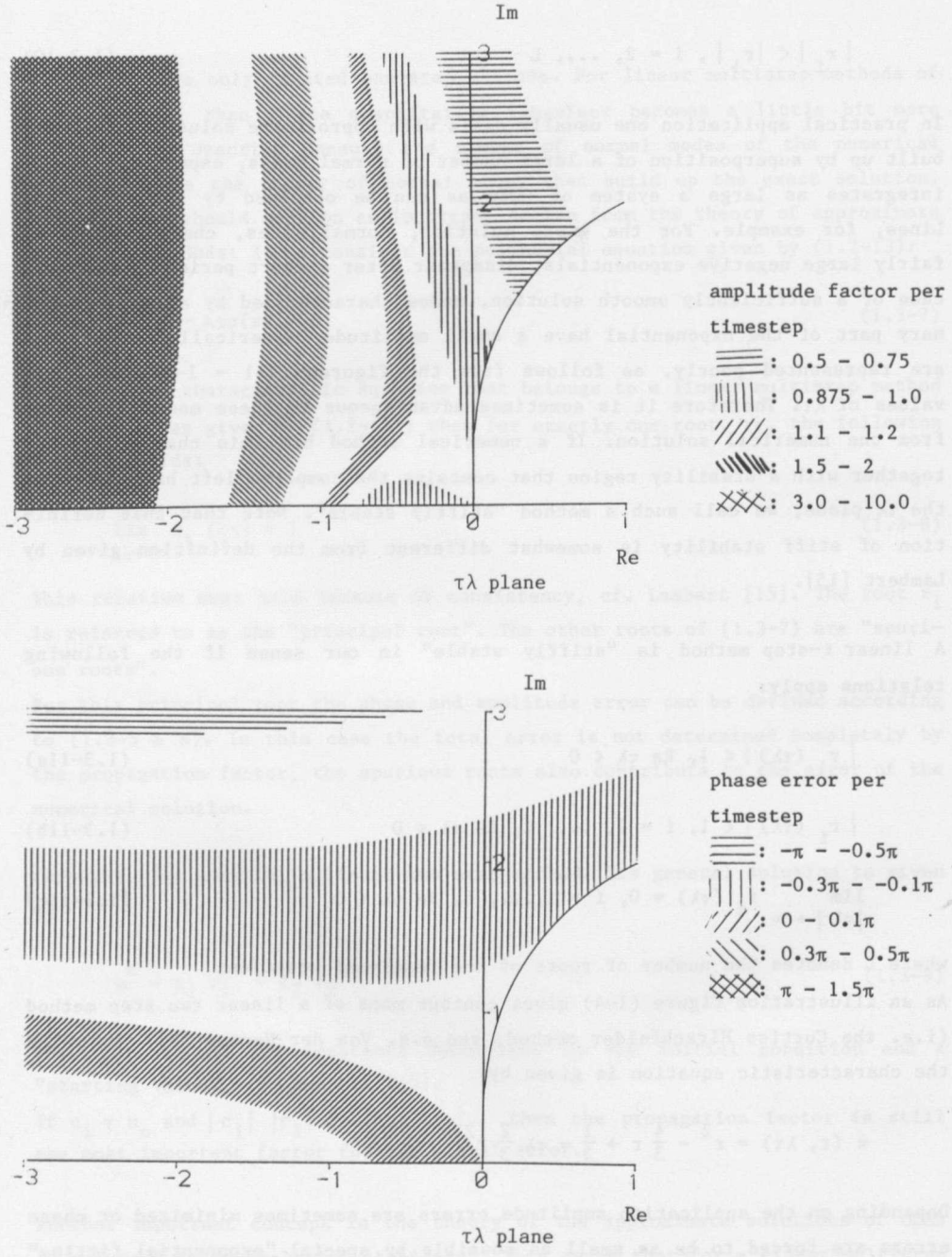


Figure (1-4) Phase errors and amplitude factors of the Curtiss-Hirschfelder method

b. Partial Differential Equations.

This subsection treats the propagation properties of approximate solutions of PDEs. As in the case of ODEs we study this behaviour by means of a simple test problem, which is a hyperbolic initial-value problem given by:

$$u_t + u_x = 0, \quad -\infty < x < \infty, \quad t > 0 \tag{1.3-12a}$$

with initial condition:

$$u(x,0) = e^{i\sigma x} \tag{1.3-12b}$$

The exact solution of this equation is given by:

$$u(x,t) = e^{i\sigma(x-t)} \tag{1.3-13}$$

For the approximation of (1.3-12) we use a semi-discrete system of ODEs given by:

$$(u_m)_t + Du_m = 0, \quad m = 0, \pm 1, \pm 2, \ldots, \tag{1.3-14a}$$

$$u_m(0) = e^{i\sigma m\Delta x} \tag{1.3-14b}$$

where D denotes some spatial difference operator, and Δx is the constant grid size.

On substituting:

$$u_m(t) = \tilde{u}(t)\, e^{i\sigma m\Delta x} \tag{1.3-15}$$

the system of ODEs is reduced to a scalar ODE given by:

$$\tilde{u}_t + \tilde{D}\,\tilde{u} = 0, \quad \tilde{u}(0) = 1 \tag{1.3-16}$$

where $\tilde{D}$ denotes the Fourier transform of D.
The solution of (1.3-16) is given by:

$$\tilde{u}(t) = e^{-\tilde{D}t} \tag{1.3-17}$$

Like Roberts and Weiss [23] we define the "relative wave speed" α by:

$$\alpha = \mathrm{Im}\,\tilde{D}/\sigma \tag{1.3-18}$$

We can define a " propagation factor" as well by:

$$P(\sigma,t) = e^{(-\tilde{D} + i\sigma)\,t} \tag{1.3-19}$$

The phase and amplitude error are determined by this factor.

Figure (1-5) shows the amplitude factor for one wave period, i.e. $t\sigma = 2\pi$, and relative wave speeds as function of the number of points per wavelength for various spatial discretizations.
The number of points per wave length M_p is defined by:

$$M_p = 2\pi/\sigma\Delta x \tag{1.3-20}$$

The following spatial discretizations are used for figure (1-5):

i : $(u_m)_t + (u_m - u_{m-1})/\Delta x = 0$, 1^{th} order upwind differencing

ii : $(u_m)_t + (u_{m+1} - u_{m-1})/2\Delta x = 0$, 2^{nd} order central differencing

iii : $\frac{1}{2}(u_{m+1} + u_m)_t + (u_{m+1} - u_m)/\Delta x = 0$, 2^{nd} order box scheme

iv : $(u_m)_t + (3u_m - 4u_{m-1} + u_{m-2})/2\Delta x = 0$, 2^{nd} order upwind differencing

v : $(u_m)_t + (u_{m+2} + 4u_{m+1} + 18u_m - 28u_{m-1} + 5u_{m-2})/24\Delta x = 0$, 3^{rd} order upwind differencing

Figure (1-5) shows that the methods 1, 2 and 5 have lagging phase errors. Method 4 has a phase error that is partly lagging and partly leading. Method 3 has a leading phase error because of the "mass matrix", i.e. the averaging operator for the time derivatives. By means of such a matrix one can construct rather compact yet accurate methods, see, e.g., Stone and Brian [25].

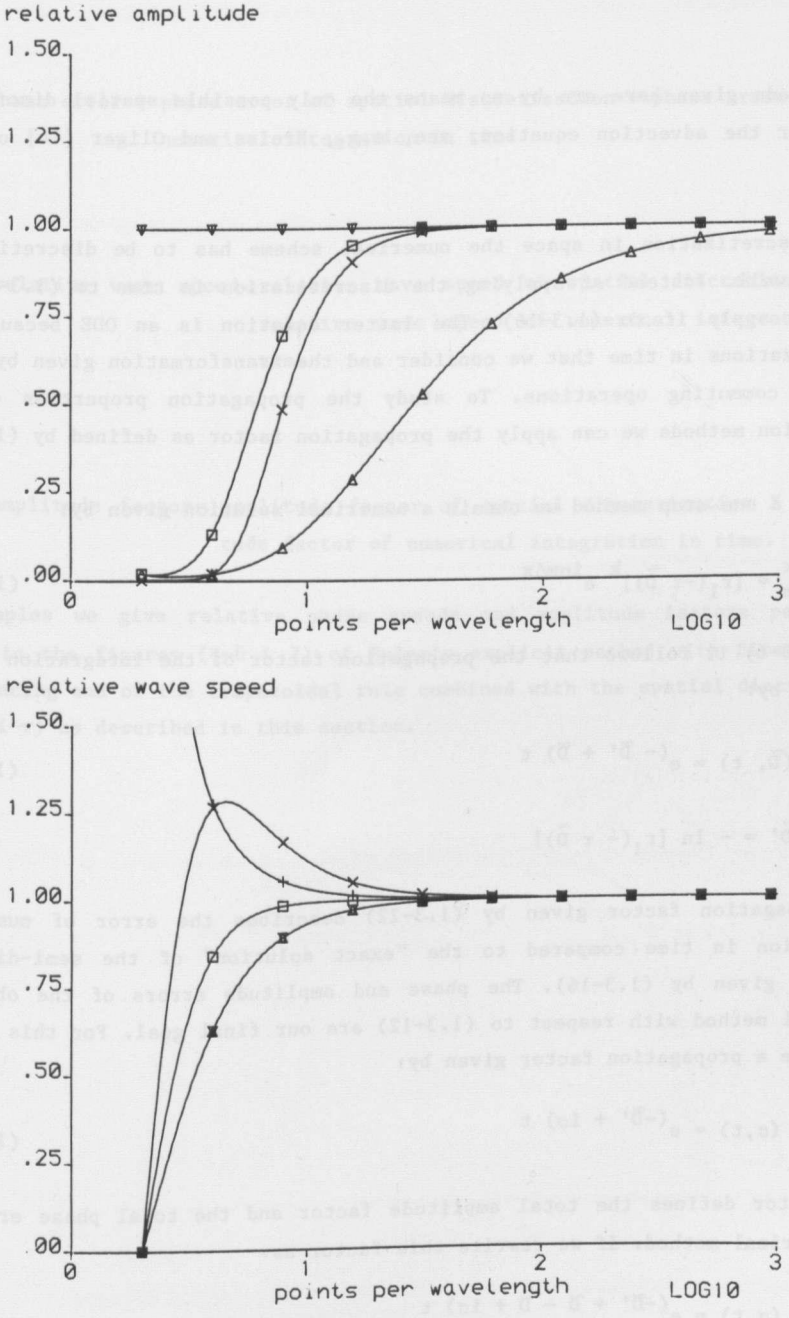


Figure (1-5) Relative amplitude and wave speed of five spatial discretizations

Δ: first order upwind
+: second order box
□: third order upwind
▽: second order central
x: second order upwind

The methods given here are by no means the only possible spatial discretizations for the advection equation, see, e.g., Kreiss and Oliger [14] or Gary [2].

After discretization in space the numerical scheme has to be discretized in time as well. Instead of applying the discretization in time to (1.3-14) we can also apply it to (1.3-16). The latter equation is an ODE because the discretizations in time that we consider and the transformation given by (1.3-15) are commuting operations. To study the propagation properties of the integration methods we can apply the propagation factor as defined by (1.3-6).

On using a one-step method we obtain a numerical solution given by:

$$u_m^k = [r_1(-\tau\ \tilde{D})]^k\ e^{i\sigma m\Delta x} \tag{1.3-21}$$

From (1.3-6) it follows that the propagation factor of the integration method is given by:

$$P(\tilde{D},\ t) = e^{(-\ \tilde{D}' + \tilde{D})\ t} \tag{1.3-22}$$

where $\tau\ \tilde{D}' = -\ \ln\ [r_1(-\ \tau\ \tilde{D})]$

The propagation factor given by (1.3-22) describes the error of numerical integration in time compared to the "exact solution" of the semi-discrete equation given by (1.3-16). The phase and amplitude errors of the observed numerical method with respect to (1.3-12) are our final goal. For this reason we define a propagation factor given by:

$$P\ (\sigma,t) = e^{(-\tilde{D}' + i\sigma)\ t} \tag{1.3-23}$$

This factor defines the total amplitude factor and the total phase error of the numerical method. If we rewrite this factor as:

$$P\ (\sigma,t) = e^{(-\tilde{D}' + \tilde{D} - \tilde{D} + i\sigma)\ t} \tag{1.3-24}$$

it follows that:

Total phase error = phase error of spatial discretization + phase error of the numerical integration in time

or:

Total relative wave speed= relative wave speed of spatial discretization X relative wave speed of numerical integration in time

and:

Total amplitude factor= amplitude factor of spatial discretization X amplitude factor of numerical integration in time.

As examples we give relative phase speeds and amplitude factors per wave period in the figures (1-6 & 7) of Euler's explicit method with first order differencing and of the trapezoidal rule combined with the spatial discretizations (i-v) as described in this section.

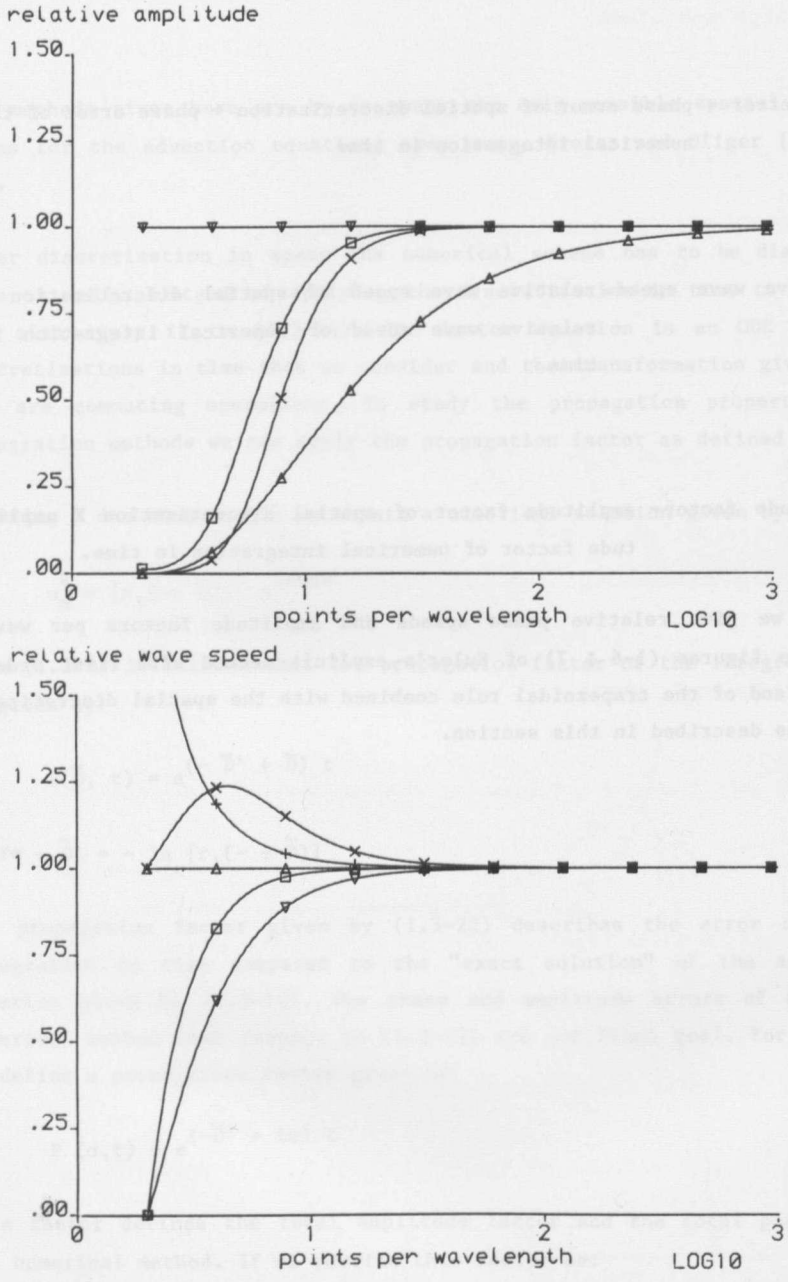


Figure (1-6) Relative wave speed and amplitude factor of several integration schemes, Cf = 0.5

Δ: first order upwind + first order Euler
∇: second order central + trapezoidal
+: second order box + trapezoidal
x: second order upwind + trapezoidal
□: third order upwind + trapezoidal

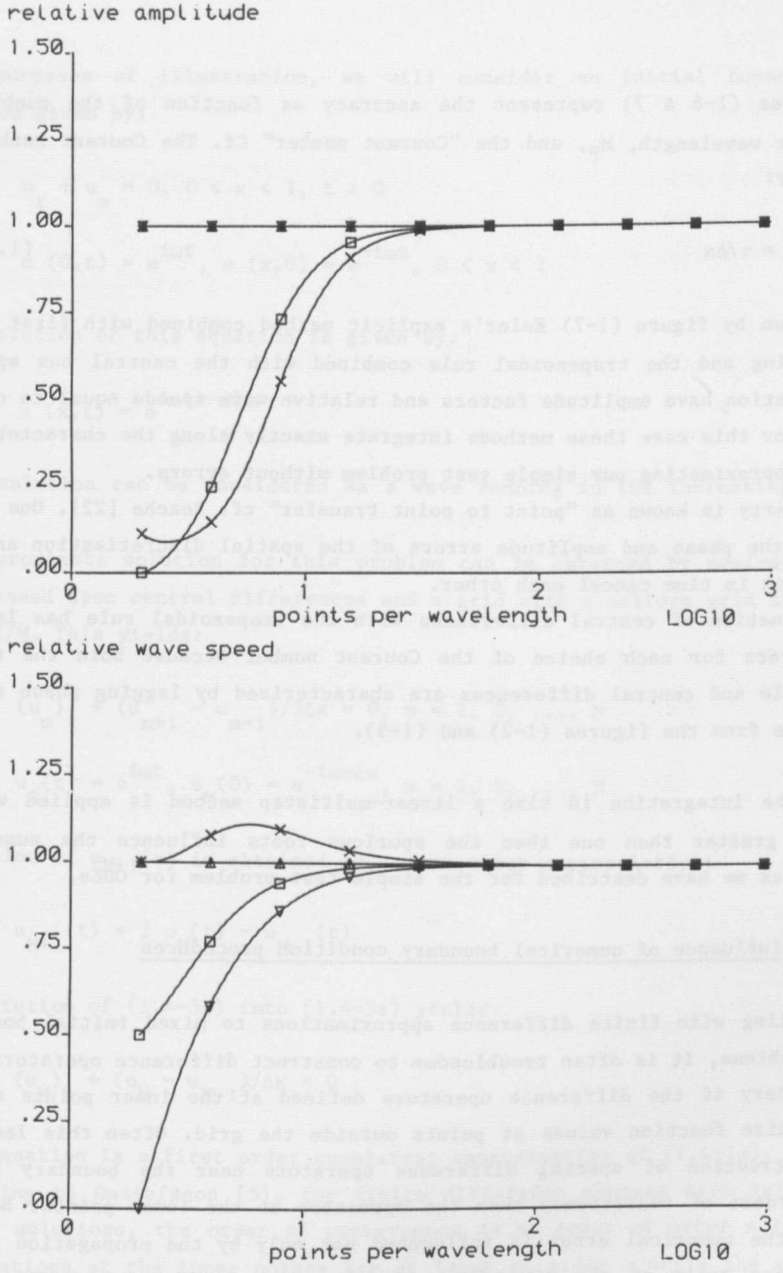


Figure (1-7) Relative wave speed and amplitude factor of several integration schemes, Cf = 1.0.

Δ: first order upwind + first order Euler
∇: second order central + trapezoidal
+: second order box + trapezoidal
x: second order upwind + trapezoidal
□: third order upwind + trapezoidal

The figures (1-6 & 7) represent the accuracy as function of the number of points per wavelength, M_p, and the "Courant number" Cf. The Courant number is defined by:

$$Cf = \tau/\Delta x \qquad (1.3\text{-}25)$$

As is shown by figure (1-7) Euler's explicit method combined with first order differencing and the trapezoidal rule combined with the central box spatial discretization have amplitude factors and relative wave speeds equal to one if Cf = 1. For this case these methods integrate exactly along the characteristic thereby approximating our simple test problem without errors.
This property is known as "point to point transfer" cf. Roache [22]. One could say that the phase and amplitude errors of the spatial discretization and the integration in time cancel each other.
The combination of central differences with the trapezoidal rule has lagging phase errors for each choice of the Courant number because both the trapezoidal rule and central differences are characterized by lagging phase errors as follows from the figures (1-2) and (1-5).

If for the integration in time a linear-multistep method is applied with a stepsize greater than one then the spurious roots influence the numerical solution as we have described for the simple test problem for ODEs.

1.4 The influence of numerical boundary condition procedures

When dealing with finite difference approximations to mixed initial boundary value problems, it is often troublesome to construct difference operators near the boundary if the difference operators defined at the inner points of the grid require function values at points outside the grid. Often this leads to the construction of special difference operators near the boundary having lesser orders of consistency than the operators at the inner points. Because of this the numerical error is influenced not only by the propagation factor or spurious roots of the time discretization but also by numerical reflections, which are due to spurious roots of the spatial discretizations at the inner points combined with a different spatial discretization near the boundary.
An important article on the subject of numerical boundary condition procedures is by Gustafsson [5].

For purposes of illustration, we will consider an initial boundary value problem given by:

$$u_t + u_x = 0, \quad 0 \leq x \leq 1, \quad t > 0 \tag{1.4-1a}$$

$$u(0,t) = e^{i\omega t}, \quad u(x,0) = e^{-i\omega x}, \quad 0 < x \leq 1 \tag{1.4-1b}$$

The solution of this equation is given by:

$$u(x,t) = e^{i\omega(t-x)} \tag{1.4-2}$$

This solution can be considered as a wave running in the increasing x direction.

An approximate solution for this problem can be obtained by semi-discretization based upon central differences and a grid with a uniform grid size, $\Delta x = 1/M$. This yields:

$$(u_m)_t + (u_{m+1} - u_{m-1})/2\Delta x = 0, \quad m = 1, 2, \ldots, M \tag{1.4-3a}$$

$$u_o(t) = e^{i\omega t}, \quad u_m(0) = e^{-i\omega m\Delta x}, \quad m = 1, 2, \ldots, M \tag{1.4-3b}$$

The value of $u_{M+1}(t)$ is obtained by second order extrapolation:

$$u_{M+1}(t) = 2\,u_M(t) - u_{M-1}(t) \tag{1.4-3c}$$

Substitution of (1.4-3c) into (1.4-3a) yields:

$$(u_M)_t + (u_M - u_{M-1})/\Delta x = 0 \tag{1.4-4}$$

This equation is a first order consistent approximation of (1.4-1a).

According to Gustafsson [5], for finite difference schemes with sufficiently smooth solutions, the order of convergence is at least of order m if the approximations at the inner points are at least of order m while the approximations near the boundary are at least of order m-1. We will illustrate this important observation and the existence of numerical reflection by comparison of (1.4-2) with the "exact solution" of (1.4-3).

For the construction of the solution of (1.4-3) we rewrite this equation as:

$$\underline{u}_t = A\,\underline{u} + B \tag{1.4-5a}$$

with initial conditions

$$\underline{u}(0) = \underline{u}^o \tag{1.4-5b}$$

where:

$$A = \begin{bmatrix} 0 & -1/2\Delta x & & & \\ -1/2\Delta x & 0 & -1/2\Delta x & & \\ & -1/2\Delta x & 0 & -1/2\Delta x & \\ & & 1/\Delta x & -1/\Delta x & \end{bmatrix},$$

$$B = [e^{i\omega t}/2\Delta x,\ 0,\ \ldots,\ 0]^T,$$

$$\underline{u}(t) = [u_1(t),\ \ldots,\ u_M(t)]^T,$$

$$\underline{u}^o = [e^{-i\omega\Delta x},\ \ldots,\ e^{-i\omega m\Delta x},\ \ldots,\ e^{-i\omega M\Delta x}]^T.$$

The solution of (1.4-5) can be considered as composed of two parts, denoted as:

$$\underline{u}(t) = \underline{u}^P(t) + \underline{u}^H(t)\ , \tag{1.4-6}$$

where $\underline{u}^P(t) = [u_1^P(t),\ \ldots,\ u_m^P(t),\ \ldots,\ u_M^P(t)]^T$

and $\underline{u}^H(t) = [u_1^H(t),\ \ldots,\ u_m^H(t),\ \ldots,\ u_M^H(t)]^T$,

$\underline{u}^P(t)$ denotes a particular solution of (1.4-5) and $\underline{u}^H(t)$ denotes the homogeneous solution of (1.4-5).

For the construction of $\underline{u}^P(t)$ we substitute a normal mode given by:

$$\tilde{u}_m^P(t) = e^{st}\, r^m \tag{1.4-7}$$

According to Strikwerda [24] this yields the following characteristic equation:

$$2\,Sr + r^2 - 1 = 0 \tag{1.4-8}$$

where $S = s\Delta x$

To fulfil the boundary conditions we pose:

$$s = i\omega \tag{1.4-9}$$

This yields:

$$r_{1,2} = -i\omega\Delta x \pm [-(\omega\Delta x)^2 + 1]^{\frac{1}{2}} \tag{1.4-10}$$

The particular solution $u_m^P(t)$ is given by:

$$u_m^P(t) = a\, e^{i\omega t} r_1^m + b\, e^{i\omega t} r_2^m = e^{i\omega t}(a u_m^1 + b u_m^2), \tag{1.4-11}$$

where $\underline{u}^{1,2} = [u_1^{1,2}, \ldots, u_m^{1,2}, \ldots, u_M^{1,2}]^T$

The constants a and b are given by the boundary conditions:

$$a + b = 1 \tag{1.4-12a}$$

$$a\, r_1^M (r_1 - 2 + 1/r_1) + b r_2^M (r_2 - 2 + 1/r_2) = 0 \tag{1.4-12b}$$

The homogeneous solution $u^H(t)$ is given by:

$$\underline{u}^H(t) = [\underline{u}^o - \underline{u}^P(0)]\, e^{At} = \underline{u}^3\, e^{At} \tag{1.4-13}$$

where $\underline{u}^3 = [u_1^3, \ldots, u_M^3]^T$ or $u_m^3 = u_m^o - u_m^P(0)$

For the eigenvalues λ_A of A the following relation holds:

$$R_e \lambda_A \leq 0, \forall \lambda_A \tag{1.4-14}$$

Proof: define A^* by $A^* = G^{-1}AG$ where G denotes a similarity transformation given by:

$$G = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & \sqrt{2} \end{bmatrix}$$

For A^* the following relation holds: $R_e (A^* u,u) = - u_M{}^2/\Delta x \leq 0$
$\forall$ u denoted as $[u_1, \ldots, u_M]^T \rightarrow R_e \lambda_{A^*} \leq 0 \rightarrow R_e \lambda_A \leq 0 \; \forall \lambda_A$
(end of proof)

From (1.4-14) it follows that the solution $\underline{u}(t)$ remains bounded if $t \rightarrow \infty$.

If $|\omega\Delta x| < 1$ then $r_{1,2}$ can be denoted as

$$r_1 = e^{-i\omega'\Delta x}, \; r_2 = -e^{i\omega'\Delta x} \tag{1.4-15}$$

where $\omega'\Delta x = \arcsin\omega\Delta x$.

At this point we define a spatial propagation factor given by:

$$P(\omega,x) = e^{-i(\omega'-\omega)x}$$

From this, it follows that:

$$u_m^1 = P(\omega, m\Delta x)e^{-i\omega m\Delta x} \tag{1.4-16a}$$

$$u_m^2 = -e^{i\omega' m\Delta x} \tag{1.4-16b}$$

$$u_m^3 = [1 - P(\omega,m\Delta x)]\, e^{-i\omega m\Delta x} + 2\, b \cos \omega' m\Delta x \tag{1.4-16c}$$

$$a = 1/1+\alpha \tag{1.4-16d}$$

$$b = \alpha/1+\alpha \tag{1.4-16e}$$

where u^1, u^2, u^3, a and b have been defined in (1.4-11), (1.4-12) and (1.4-13) and

$$\alpha = (-1)^M e^{-2iM\omega'\Delta x} (\cos \omega'\Delta x - 1)/(\cos \omega'\Delta x + 1).$$

Since it is easy to verify that (i) $au^1_m = [1 + 0(\Delta x^2)]e^{-i\omega m\Delta x}$

or $u^1_m e^{i\omega t} = (1 + 0(\Delta x^2))\, u(m\Delta x, t)$, (ii) $bu^2_m = 0(\Delta x^2)$ and (iii) $u^3_m = 0(\Delta x^2)$

it follows from the comparison of the numerical solution given by (1.4-6) with the exact solution (1.4-2) of (1.4-1) that the convergence is of order Δx^2.

Because $\underline{u}^1$ and $\underline{u}^2$ can be considered as waves propagating in opposite directions and the latter is a spurious numerical wave, it follows that α can be considered as a "numerical reflection coefficient".

Note that for $|\omega\Delta x| > 1$ the numerical solution becomes totally erroneous while the modulus of the spatial propagation factor is no longer equal to one.

A possible first order extrapolation formula at the outflow boundary is given by:

$$u_{M+1} = u_M \tag{1.4-17}$$

Substitution into (1.4-3a) yields:

$$(u_M)_t + (u_M - u_{M-1})/2\Delta x = 0 \tag{1.4-18}$$

This approximation is not consistent with (1.4-1a).
According to Gustafsson [5], however, this boundary procedure does not destroy convergence as can be checked easily in the case of our example. The order of convergence is reduced to first order, because the numerical reflection factor α is for that case of order Δx and is given by:

$$\alpha = (-1)^M e^{-2i\, M\omega'\Delta x} (\cos 2\omega'\Delta x - 1 - i \sin 2\omega'\Delta x)/(1 + \cos\omega'\Delta x) \tag{1.4-19}$$

The propagation factor is unaltered. The matrix A is changed but the eigenvalues still do not have a positive real part.
The results of Gustafsson [5] allow several possible boundary procedures such that convergence is not destroyed. Even inconsistent, or zero order consistent, procedures are possible, as this example illustrates.

Yet there are limitations. Zero order extrapolation or overspecification of boundary conditions can destroy convergence and cause unbounded numerical solutions if $t \to \infty$.

To illustrate this we consider an initial boundary value problem given by:

$$u_t + u_x = 0, \quad 0 < x \leq 1, \quad t \geq 0 \tag{1.4-20a}$$

$$u(0,t) = 1, \quad u(x,0) = 1-x, \quad 0 < x \leq 1 \tag{1.4-20b}$$

The solution of this equation is given by:

$$\begin{aligned} u(x,t) &= 1 + t - x, \quad t - x \leq 0 \\ u(x,t) &= 1, \quad t - x > 0 \end{aligned} \tag{1.4-21}$$

A semi-discrete numerical approximation based upon 2nd order central differences is given by:

$$(u_m)_t + (u_{m+1} - u_{m-1})/2\Delta x = 0 \tag{1.4-22a}$$

$$u_o(t) = 1, \quad u_m(0) = 1 - m\Delta x, \quad m = 1, 2, \ldots, M \tag{1.4-22b}$$

Again this equation is not complete, and we impose an extra boundary condition given by:

$$u_{M+1}(t) = 0 \tag{1.4-22c}$$

Substitution of this "zero order" extrapolation formula into (1.4-22a) shows inconsistency.

For the construction of the solution of (1.4-22) we rewrite this equation as:

$$\underline{u}_t = A\,\underline{u} + B,\ \underline{u}(0) = \underline{u}^o \tag{1.4-23}$$

where:

$$A = \begin{bmatrix} 0 & -1/2\Delta x & & & \\ -1/2\Delta x & 0 & -1/2\Delta x & & \\ & -1/2\Delta x & 0 & -1/2\Delta x & \\ & & -1/2\Delta x & 0 & \end{bmatrix},$$

$B = [1/2\Delta x,\ 0,\ \ldots,\ 0]^T$,

$\underline{u}(t) = [u_1(t),\ \ldots,\ u_M(t)]^T$ and

$\underline{u}^o = [1 - \Delta x,\ \ldots,\ 1-m\Delta x,\ \ldots,\ 0]^T$.

If we write $\underline{u}(t)$ as:

$$\underline{u}(t) = \underline{u}^P(t) + \underline{u}^H(t) \tag{1.4-24}$$

then $\underline{u}^P(t)$, the particular solution, and $\underline{u}^H(t)$, the homogeneous solution are given by:

$$u_m^P(t) = \delta(M)\,\delta(m) + \delta(M+1)[\delta(m)\,(1-\tfrac{m}{M+1}) + \delta(m+1)\,\tfrac{M}{M+1}\,t] \tag{1.4-25}$$

$$\underline{u}^H(t) = [\underline{u}(0) - \underline{u}^P(0)]\,e^{At} \tag{1.4-26}$$

where:

$$\delta(j) = \tfrac{1}{2}\,[1 + (-1)^j].$$

It follows from this solution that convergence is impossible because $u_m^P(t)$ will always oscillate despite the values of M, and independently of Δx ($\Delta x = 1/M$); for odd values of M the solution will grow unboundedly if

$t \to \infty$, and the exact solution on the interval $0 \leq x \leq 1$ will be equal to one in this case.

For boundary procedures as given by (1.4-3c) or (1.4-17), convergence is ensured.

1.5 Concluding remarks

When dealing with stability problems of finite difference methods one must realize in which way stability has been defined.

For PDEs necessary conditions for G-R stability are usually not too difficult to derive by verification of the G-R condition, although for some problems this could lead to complicated analytical problems.

The Von Neumann condition is a necessary condition to fulfil the G-R condition.

Application of the matrix method means verification of B-H-K stability. In the case of implicit methods for PDEs this could be very difficult due to the inversion of large matrices. Moreover if B-H-K stability is established, convergence is not a necessary consequence.

Because stability is almost always established for simplified problems, additional practical experience with the contemplated numerical model remains a necessity.

The tools for the verification of the G-R condition can also be used to estimate the propagation properties of a numerical model.

The fact that first order extrapolation methods near boundaries are sufficient to maintain convergence permits several possible numerical boundary condition procedures. However, overspecification could well lead to instability or completely erroneous results.

References to Chapter I

1 CUVELIER, C.,
Perturbation, approximation et controle optimal d'un système gouverné par les équations de Navier-Stokes couplées à celle de la chaleur.
Thesis, TH Delft, 1976.

2. GARY, J.,
The Method of Lines applied to a Simple Hyperbolic Equation,
Journal of Computational Physics, 22, 1976, pp. 131-149.

3. GEAR, C.W.,
Numerical Initial Value Problems in Ordinary Differential Equations,
Prentice Hall, Englewood Cliffs, N.J., 1971.

4. GODUNOV, S.K. and V.S. RYABENKI,
Theory of Difference Schemes,
North-Holland Publishing Company, Amsterdam 1964.

5. GUSTAFSSON, B.,
The Convergence Rate for Difference Approximations to Mixed Initial Boundary value problems,
Mathematics of Computation, V26, 1975, pp 396-406.

6. HENRICI, P.,
Discrete Variable Methods in Ordinary Differential Equations,
John Wiley & Sons, New York, London, 1962.

7. HIRT, C.W.,
Heuristic Stability Theory for Finite Difference Equations,
Journal of Computational Physics, V2, pp.339-355, 1968.

8. HOUWEN, P.J. van der,
Finite Difference Methods for solving Partial Differential Equations,
Mathematical Centre Tracts, No. 20, Mathematical Centre, Amsterdam, 1968.

9. HOUWEN, P.J. van der,
Construction of Integration Formulas for Initial Value Problems,
North-Holland Publishing Company, Amsterdam, 1977.

10. KREISS, H.O.,
Difference Approximations for the Initial Boundary Value Problem for Hyperbolic Differential Equations,
Proc. Adv. Symp., Madison, Wis, 1966, Wiley, New York, 1966, pp. 141-166.

References (continued)

11. KREISS, H.O.,
Stability Theory for Difference Approximations of Mixed Initial Boundary Value Problems, I,
Mathematics of Computation, V22, 1968, pp 703-714.

12. KREISS, H.O.,
Initial Boundary Value Problems for Hyperbolic Systems,
Comm. on Pure and Applied Math., V23, 1970, pp.272-298.

13. KREISS, H.O., B. GUSTAFSSON and A. SUNDSTROM,
Stability Theory of Difference Approximations for Mixed Initial Boundary Value Problems, II,
Mathematics of Computations, V26, 1972, pp. 649-686.

14. KREISS, H.O., and J. OLIGER
Methods for the Approximate Solution of the Time Dependent Problems,
GARP Publication Series, no. 10, Geneva, 1973.

15. LAMBERT, J.D.,
Computational Methods in Ordinary Differential Equations,
Wiley, London-New York, 1973.

16. LAMBERT, J.D.,
Numerical Methods for Phase Plane Problems in Ordinary Differential Equations,
Procs. of the Dundee Biennial Conference on Numerical Analysis 1979,
Springer, 1980.

17. LAPIDUS, L. and J.H. SEINFELD,
Numerical Solution of Ordinary Differential Equations,
Academic Press, New York, 1971.

18. LEENDERTSE, J.J.,
Aspects of a Computational Model for Long-Period Water-Wave Propagation,
Rand Corporation, Memorandum RM-5294-PR, Santa Monica, 1967.

19. MITCHELL, A.R. and D.F. GRIFFITHS,
The Finite Difference Method in Partial Difference Equations,
Wiley, New York, 1980.

20. PRAAGMAN, N.,
Numerical Solution of the Shallow Water Equations by a Finite Element Method,
Thesis, TH Delft, 1979

References (continued)

21. RICHTMYER, R.D. and K.W. MORTON,
Difference Methods for Initial Value Problems,
Interscience Publishers, Wiley, New York, London, 1967.

22. ROACHE, P.J.,
Computational Fluid Dynamics,
Hermosa Publishers, Albuquerque, 1972.

23. ROBERTS, K.W. and N.O. WEISS,
Convective Difference Schemes,
Mathematics of Computation, No. 2, 1966, pp. 272-299.

24. STRIKWERDA, J.C.,
Initial Boundary Value Problems for the Method of Lines,
Journal of Computational Physics, V34, 1980, pp. 94-107.

25. STONE, H.J. and P.L.T. BRIAN,
Numerical Solution of Convective Transport Problems,
A.I.Ch.E. Journal, V9, 1963, pp. 681-688.

26. WARMING, R.F. and B.J. HYETT,
The Modified Equation Approach to the Stability and Accuracy Analysis of Finite Difference Methods,
Journal of Computational Physics, V14, 1974, pp. 159-179.

27. TEMAM., R.,
Navier Stokes Equations, Theory and Numerical Analysis,
North-Holland Publishing Company, Amsterdam, 1977.

2 Efficient Integration Methods for the Advection Equation

2.0 Introduction

This chapter will describe some efficient and unconditionally stable methods for the integration of a simple hyperbolic equation. This equation is given by:

$$u_t + v(x)\, u_x = f\,(x,t) \qquad (2.0\text{-}1)$$

This equation will be referred to as the nonhomogeneous advection equation. The main motivation for our interest in approximation methods for advection equations is the possible application of these methods for the advection operator of the SWE, which should be chosen carefully; otherwise instabilities are likely to be induced, see e.g. Weare [22]. This integration method, however, should be efficient from a computational point of view. It should be possible to solve not only time-dependent problems but also steady state problems, see, e.g., Vreugdenhil and Wijbenga [21].
Since many integration methods for SWE are defined on a fixed grid with an equidistant grid size the method that we are looking for should be defined on such a grid as well. The following demands are to be satisfied:

(a) Second order consistency at least.
(b) Computational efficiency.
(c) Suitability not only for time-dependent problems but also for steady state problems.
(d) Unconditional stability.
(e) Easy to implement for the approximation of the advection operator of SWE without a significant reduction of the overall efficiency.

Many methods for the integration of advection equations are mentioned in the literature. These can be divided into classes according to the way in which they are constructed:

(i) methods based upon separate approximation of each derivative by finite differences
(ii) characteristic interpolation methods

(iii) characteristic methods
(iv) finite element methods
(v) spectral methods

This chapter deals primarily with methods of class (i). Some authors, e.g. Benqué et al [2], claim that methods of class (ii) are very efficient for the approximation of the advective part of SWE. Therefore we will treat this class as well. The other three classes are not treated here because they are based upon a grid structure different from the one we use for the SWE. For a treatise on characteristic methods see Abbott [1]. Finite element methods for the advection equation are treated by Morton [15]. A monograph on spectral methods has been written by Gottlieb and Orszag [8]. This method is not widely used for the approximation of SWE.

Section 1 will deal with efficient time-spitting methods for the integration of a homogeneous advection equation with a frozen coefficient. Section 2 deals with the stability analysis of these methods. Section 3 shows how the methods of section 1 can be extended to an advection equation with variable coefficients. Section 4 deals with characteristic interpolation methods that are unconditionally stable. Section 5 describes a few simple test problems to compare the methods treated.

2.1 Efficient time splitting methods for the frozen coefficient equation

Consider the following advection equation:

$$u_t + Vu_x = 0,\ 0 \leq x < 1,\ V > 0,\ V = \text{constant ("Frozen")} \qquad (2.1\text{-}1a)$$

The initial and boundary conditions are given by:

$$u(x,0) = 0,\ u(0,t) = g(t) \qquad (2.1\text{-}1b)$$

Let (2.1-1) be approximated by a consistent system of ODEs denoted as:

$$\underline{u}_t = A\,\underline{u} + B \qquad (2.1\text{-}2)$$

where $\underline{u}(t) = [u_1, \ldots, u_M]^T$, the element u_m are grid functions defined on an equidistant grid with grid size Δx, $\Delta x = 1/M$. A denotes a MxM matrix and B a vector with M elements.

After discretization in space (2.1-2) has to be integrated in time. Each of the integration methods that we consider can be written in the form:

Stage 1:

$$(\underline{u}^* - \underline{u}^k)/\tfrac{1}{2}\tau = A_1 \underline{u}^k + B^k \qquad (2.1\text{-}3a)$$

Stage 2:

$$(\underline{u}^{k+1} - \underline{u}^*)/\tfrac{1}{2}\tau = A_2 \underline{u}^{k+1} + B^{k+1} \qquad (2.1\text{-}3b)$$

where $\frac{1}{2}(A_1 + A_2) = A$.

If an integration method can be denoted in the form (2.1-3), which we will call a "two stage split method", and if both (2.1-3a) and (2.1-3b) are consistent approximations of (2.1-1) then it is well structured to be implemented as the advective part of an ADI type of numerical scheme for the SWE. In chapter 3 this will become apparent. It is the reason why in this section we only consider methods of type (2.1-3). Three examples of this type will be described. The first two are well-known: the Crank-Nicolson method and the Angled-Derivative method as proposed by Roberts and Weiss [19]. We will show that these methods are of type (2.1-3). The third method, that we propose has a reduced phase error compared with the other two methods.

a. Crank-Nicolson scheme

This scheme is based upon a spatial discretization given by:

$$(u_m)_t + V(u_{m+1} - u_{m-1})/2\Delta x = 0, \quad m = 1, \ldots, M \qquad (2.1\text{-}4a)$$

$$u_{M+1} = 2u_M - u_{M-1}$$

$$u_o(t) = g(t) \qquad (2.1\text{-}4b)$$

The integration in time is written in form (2.1-3) as follows:

Stage 1:

$$(u_m^* - u_m^k)/\tfrac{1}{2}\tau + V\,(u_{m+1}^k - u_{m-1}^k)/2\Delta x = 0,\ m = 1, \ldots, M \qquad (2.1\text{-}5a)$$

Stage 2:

$$(u_m^{k+1} - u_m^*)/\tfrac{1}{2}\tau + V\,(u_{m+1}^{k+1} - u_{m-1}^{k+1})/2\Delta x = 0,\ m=1, \ldots, M \qquad (2.1\text{-}5b)$$

$$u_{M+1}^{k+1} = 2\,u_M^{k+1} - u_{M-1}^{k+1}$$

$$u_o^{k+1} = g^{k+1} \qquad (2.1\text{-}5c)$$

The relative wave speed of this method and the amplitude factor per wave period are given by figure (2-1), for various Courant numbers, Cf, where:

$$Cf = V\,\tau/\Delta x \qquad (2.1\text{-}6)$$

The unconditional stability of this method, including the boundary procedure, has been proven by Kreiss et al. [11]; see also Wirz et al. [24].
The second stage of this method implies the solution of a tri-diagonal equation. Although the "double sweep method", see e.g. Godunov and Ryabenki [7] is an efficient method for solving tri-diagonal equations, when it is implemented as part of a numerical SWE procedure, with derivatives in more than one spatial dimension, it could decrease overall efficiency. Therefore we consider another two-stage split method based upon (2.1-4):

b. The Angled Derivative method

This method was proposed by Roberts and Weiss [19]. For its application to the advective part of SWE, see Stelling [20]. This method, which is also based upon the spatial discretization given by (2.1-4), is denoted in the form (2.1-3) as follows:

Stage 1:

$$(u_m^* - u_m^k)/\tfrac{1}{2}\tau + V\,(u_{m+1}^k - u_m^k)/\Delta x = 0,\quad m = 1, \ldots, M$$

$$u_{M+1}^k = 2\,u_M^k - u_{M-1}^k \qquad (2.1\text{-}7a)$$

Stage 2:

$$(u_m^{k+1} - u_m^*)/\tfrac{1}{2}\tau + V\,(u_m^{k+1} - u_{m-1}^{k+1})/\Delta x = 0,\quad m = 1, \ldots, M \qquad (2.1\text{-}7b)$$

$$u_o^{k+1} = g^{k+1} \qquad (2.1\text{-}7c)$$

Each of the equations (2.1-7a) or (2.1-7b) are consistent with (2.1-1). If we eliminate u^* then we obtain:

$$(u_m^{k+1} - u_m^k)/\tau + V\,(u_{m+1}^k - u_m^k + u_m^{k+1} - u_{m-1}^{k+1})/2\Delta x = 0 \qquad (2.1\text{-}8)$$

This equation is a second order ($O(\Delta x^2, \tau^2)$) consistent approximation of (2.1-1). The relative wave speed and amplitude factor per wave period are given by figure (2-2). The phase errors of this method are larger than the phase errors of the C-N scheme. In section 2.2 unconditional stability is proved.
Equation (2.1-7b) is implicit. If the calculation of u_m starts at m=0 and proceeds in the increasing m direction then the bi-diagonal equations are solved in one sweep. This makes this method just as efficient as an explicit method and more efficient than the Crank-Nicolson method, especially for equations with derivatives in more than one spatial dimension.
It is also possible to construct an Angled-Derivative method that sweeps in the decreasing m direction. If we write this method as a two stage scheme then we obtain:

Stage 1:

$$(u_m^* - u_m^k)/\tfrac{1}{2}\tau + V\,(u_m^k - u_{m-1}^k)/\Delta x = 0,\quad m = 1, \ldots, M \qquad (2.1\text{-}9a)$$

Stage 2:

$$(u_m^{k+1} - u_m^*)/\tfrac{1}{2}\tau + V\,(u_{m+1}^{k+1} - u_m^{k+1})/\Delta x = 0,\ m = 1, \ldots, M \qquad (2.1\text{-}9b)$$

$$u_{M+1}^{k+1} = 2\,u_M^* - u_{M-1}^k$$

$$u_o^{k+1} = g^{k+1} \qquad (2.1\text{-}9c)$$

The relative wave speed is given by figure (2-3). From this figure it follows that for this sweep direction the phase error is much smaller than for the sweep in the increasing m direction. For Cf = 1 the method has zero phase error. The method given by (2.1-7) however is unconditionally stable whereas (2.1-9) is stable only if Cf < 1, see section 2.2. Now we will try to construct a method that (i) is unconditionally stable, (ii) has small phase errors and (iii) is as efficient as the Angled-Derivative method.

c. Reduced phase error, two stage split scheme

This scheme is based upon the following spatial discretization:

$$(u_1)_t + V(u_2 - u_o)/2\Delta x = 0$$

$$(u_m)_t + V(u_{m+2} + 4u_{m+1} + 18u_m - 28u_{m-1} + 5u_{m-2})/24\Delta x = 0,\ m = 2, \ldots, M-2 \qquad (2.1\text{-}10a)$$

$$(u_{M-1})_t + V(u_M + 3u_{M-1} - 5u_{M-2} + u_{M-3})/4\Delta x = 0$$

$$(u_M)_t + V(3u_M - 4u_{M-1} + u_{M-2})/2\Delta x = 0$$

$$u_o(t) = g(t) \qquad (2.1\text{-}10b)$$

At m=2, ..., M-2, the order of consistency is three, at m=1, M-1, M the order of consistency is two. According to Gustafsson [9] this means that the convergence rate is of order three. The reduced phase error of this spatial discretization of the inner points when compared with central differences is represented in figure (1-5).

For the integration in time we propose the following split scheme:

Stage 1:

$$(u_1^* - u_1^k)/\tfrac{1}{2}\tau + V(u_2^k - u_1^k)/\Delta x = 0$$

$$(u_m^* - u_m^k)/\tfrac{1}{2}\tau + V\,(u_{m+2}^k + 4u_{m+1}^k - 4u_{m-1}^k - u_{m-2}^k)/12\Delta x = 0,\ m=2,\ \ldots,\ M-2 \qquad (2.1\text{-}11a)$$

$$(u_{M-1}^* - u_{M-1}^k)/\tfrac{1}{2}\tau + V\,(u_M^k - u_{M-2}^k)/2\Delta x = 0$$

$$(u_M^* - u_M^k)/\tfrac{1}{2}\tau + V\,(3u_M^k - 4u_{M-1}^k + u_{M-2}^k)/2\Delta x = 0$$

Stage 2:

$$(u_1^{k+1} - u_1^*)/\tfrac{1}{2}\tau + V\,(u_1^{k+1} - u_o^{k+1})/\Delta x = 0 \qquad (2.1\text{-}11b)$$

$$(u_m^{k+1} - u_m^*)/\tfrac{1}{2}\tau + V\,(3u_m^{k+1} - 4u_{m-1}^{k+1} + u_{m-2}^{k+1})/2\Delta x = 0,\ m=2,\ \ldots,\ M$$

$$u_o^{k+1} = g^{k+1} \qquad (2.1\text{-}11c)$$

Each of the equations (2.1-11a) and (2.1-11b) is consistent with (2.1-1). If we eliminate u_m^* then the resulting equation is an $O(\tau^2, \Delta x^3)$ consistent approximation of (2.1-1) at the inner points and an $O(\tau^2, \Delta x^2)$ consistent approximation of (2.1-1) near the boundary.

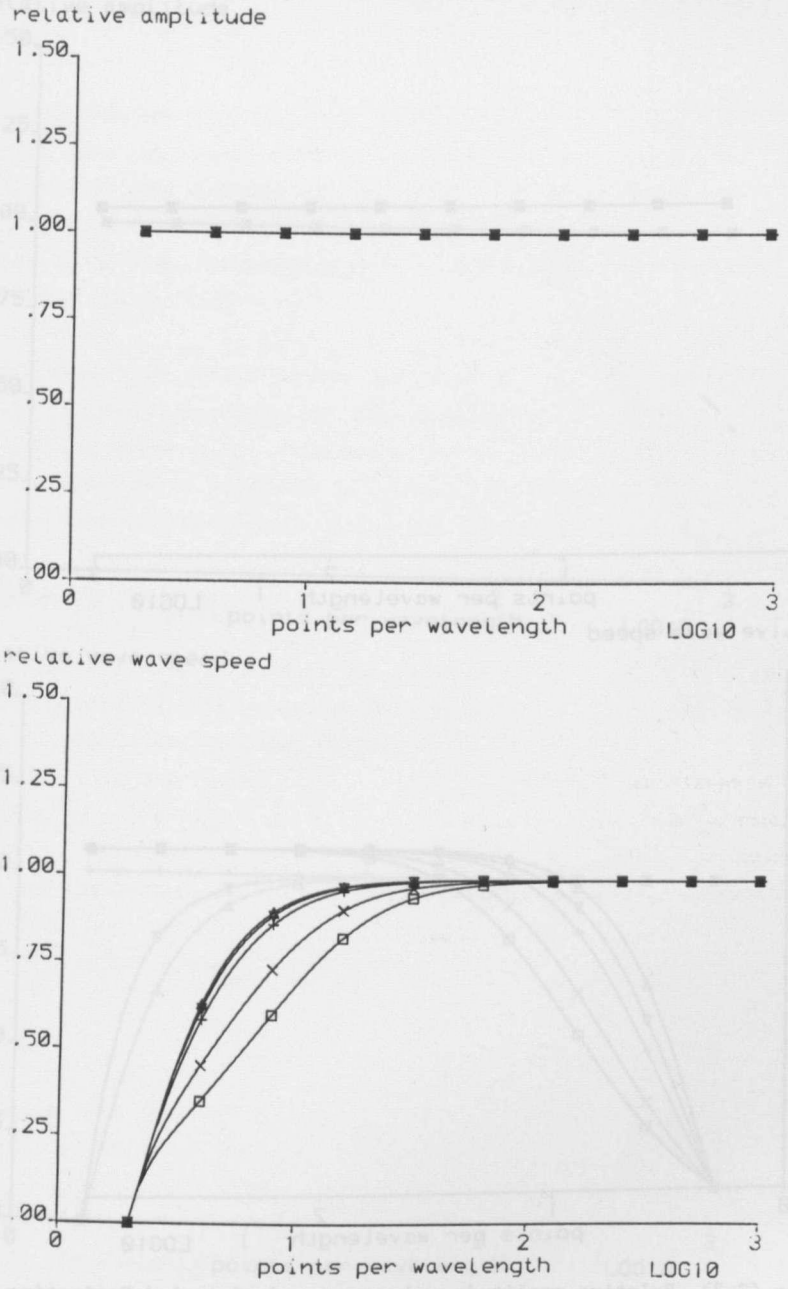


Figure (2-1) Relative amplitude and wave speed of Crank-Nicolson scheme
Δ: Cf = 0.1, ∇: Cf = 0.5, +: Cf = 1.0,
X: Cf = 2.5, □ : Cf = 4.0.

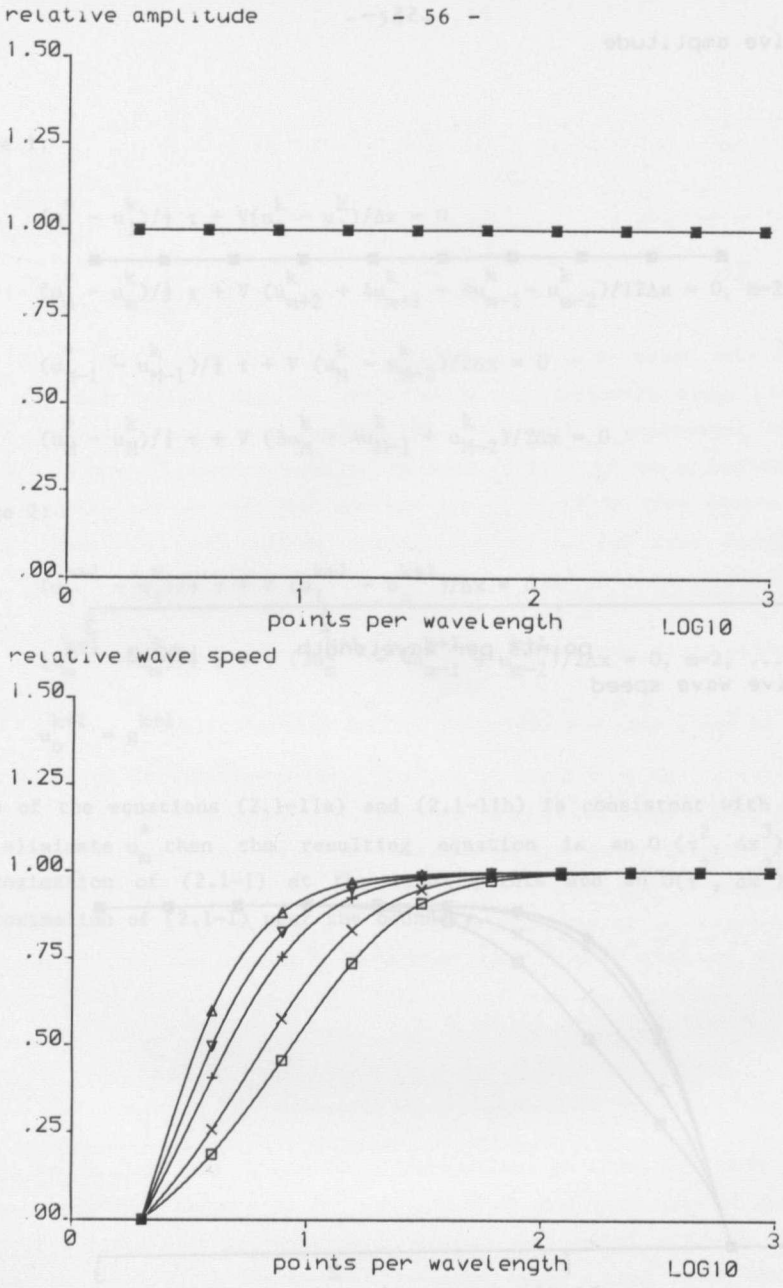


Figure (2-2) Relative amplitude and wave speed of Angled-Derivative method with sweep in increasing m direction
Δ: Cf = 0.1, ∇: Cf = 0.5, +: Cf = 1.0,
X: Cf = 2.5, □ : Cf = 4.0.

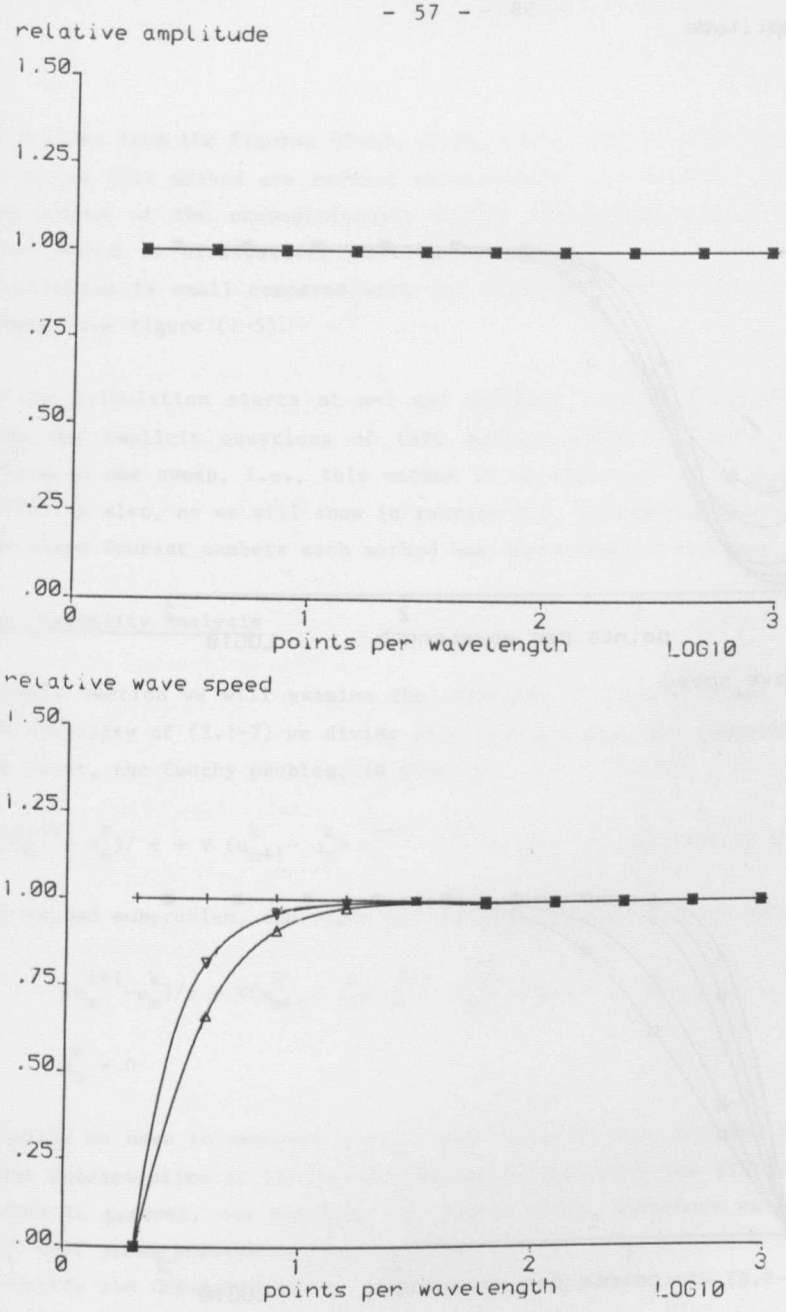


Figure (2-3) Relative amplitude and wave speed of Angled-Derivative method with sweep in decreasing m direction
Δ: Cf = 0.1, ∇: Cf = 0.5, +: Cf = 1.0.

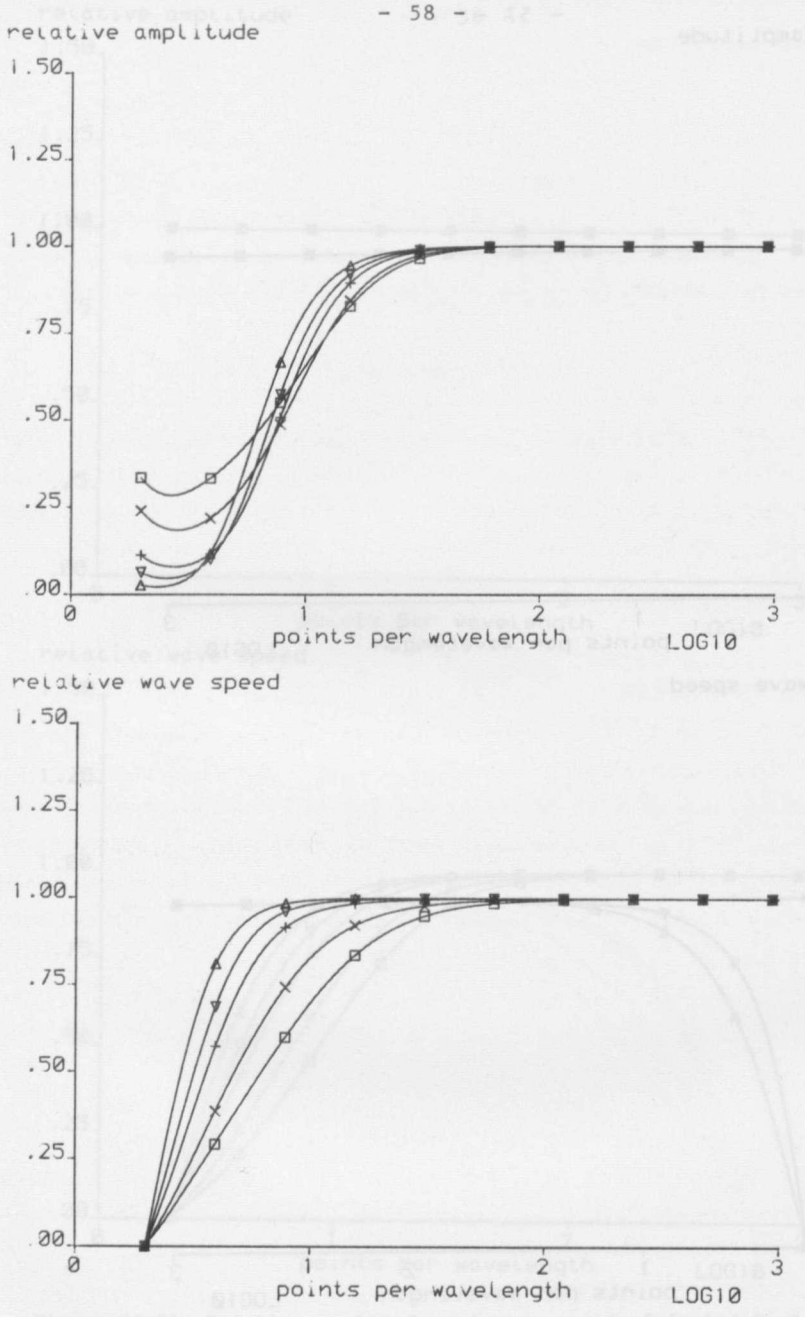


Figure (2-4) Relative amplitude and wave speed of reduced phase error split scheme

Δ: Cf = 0.1, ∇: Cf = 0.5, +: Cf = 1.0,

X: Cf = 2.5, □ : Cf = 4.0.

As follows from the figures (2-1), (2-2), (2-4), (2-10), and (2-11), the phase errors of this method are reduced considerably compared with the Crank-Nicolson scheme or the unconditionally stable Angled-Derivative method. However this method is dissipative, and the other methods are non-dissipative. The dissipation is small compared with the dissipation of the first order upwind scheme, see figure (1-5).

If the calculation starts at m=1 and proceeds in the increasing m direction then the implicit equations of this method, which are lower diagonal, are solved in one sweep, i.e., this method is as efficient as an explicit method, and it is also, as we will show in section 2.2, unconditionally stable.
For large Courant numbers each method has approximately the same phase error.

2.2 Stability Analysis

In this section we will examine the stability of (2.1-7, 9 and 11). To study G-R stability of (2.1-7) we divide this equation into two subproblems:
The first, the Cauchy problem, is given by:

$$(u_m^{k+1} - u_m^k)/\tau + V\,(u_{m+1}^k - u_m^k + u_m^{k+1} - u_{m-1}^{k+1})/2\Delta x = 0,\quad m=0, \pm 1, \pm 2,\ldots \qquad (2.2\text{-}1)$$

The second subproblem, the right half plane problem, is given by:

$$(u_m^{k+1} - u_m^k)/\tau + V(u_{m+1}^k - u_m^k + u_m^{k+1} - u_{m-1}^{k+1})/2\Delta x = 0,\quad m=1,\ 2,\ \ldots \qquad (2.2\text{-}2a)$$

$$u_o^k = 0 \qquad (2.2\text{-}2b)$$

Formally we have to consider a left half plane problem too, but we use second order extrapolation at the outflow boundary from which the stability has been proven in general, see Goldberg and Tadmor [5,6]. Therefore we refer for the left half plane problem to these authors.
To verify the Godunov-Ryabenki condition we substitute into (2.2-1) an eigensolution given by:

$$\tilde{u}_m^k = \lambda^k \hat{u}_m \qquad (2.2\text{-}3)$$

This yields the resolvent equation given by:

$$\lambda\,[\hat{u}_m + \tfrac{r}{2}(\hat{u}_m - \hat{u}_{m-1})] - [\hat{u}_m - \tfrac{r}{2}(\hat{u}_{m+1} - \hat{u}_m)] = 0 \qquad (2.2\text{-}4)$$

where $r = Cf = V\tau/\Delta x$

For the Cauchy problem we pose $\hat{u}_m = e^{im\sigma\Delta x}$. This yields $|\lambda| = 1\ \forall\ r$, i.e., the Cauchy problem is unconditionally stable. The G-R condition is satisfied for (2.2-2) if (2.2-4) combined with:

$$\hat{u}_o = 0,\ \lim_{m\to\infty} \hat{u}_m < C \qquad (2.2\text{-}5)$$

where C denotes an arbitrary constant, for $|\lambda| > 1$, has only the trivial solution $\hat{u}_m = 0$, see Kreiss et al [11].

The general solution of (2.2-4) is given by:

$$\hat{u}_m = a\,z_1^{\,m} + b\,z_2^{\,m} \qquad (2.2\text{-}6)$$

where z_1 and z_2 are the roots of a characteristic equation given by:

$$z^2 + 2z\,(\lambda-1)\,(1+\tfrac{r}{2})/r - \lambda = 0 \qquad (2.2\text{-}7)$$

From this equation it follows that $|z_1|.|z_2| = |\lambda|$ or if $|\lambda| > 1$ then at least one root of (2.2-7) has a root with a modulus larger then one, i.e. either a=0 or b=0. From $\hat{u}_o = 0$ it follows that a = b = 0 or the G-R condition is always satisfied.

If we apply the same procedure for (2.1-9) then we find an unconditionally stable Cauchy problem. To verify the G-R condition for the right half plane problem, we follow the same procedure as described above, again we check if $\hat{u}_m = 0$. This solution is also given by (2.2-6) but here $z_{1,2}$ are roots of the following characteristic equation:

$$z^2 + 2z\,(\lambda - 1)\,(1 - \tfrac{r}{2})/(r\lambda) - 1/\lambda = 0 \qquad (2.2\text{-}8)$$

Following Miller [14] the roots of (2.2-8) $z_{1,2}$ have the property $|z_{1,2}| \leq 1$ if and only if (i) $|\lambda| > 1$ and (ii) if the roots of the "reduced polynomial", see Miller [14], given by:

$$z + (1-r/2)/(r/2) = 0 \qquad (2.2\text{-}9)$$

have a modulus less than one. From this it follows that if:

$$r > 1 \qquad (2.2\text{-}10)$$

then $|z_{1,2}| < 1$ with $|\lambda| > 1$. In that case for $a = -b$ there is a non-trivial solution that fulfils (2.2-5), i.e., in that case the G-R condition is not fulfilled which means instability. The following condition is therefore necessary for stability:

$$Cf < 1 \quad (Cf = r) \qquad (2.2\text{-}11)$$

This is the well-known C-F-L condition, see Courant, Friedrichs and Lewy [3]. Robert and Weiss [19] have derived this condition by means of the so-called "spatial amplification" factor, which in this case yields (2.2-11) as well.

For the calculation of the "spatial amplification factor", S_p, we first eliminate u_m^* from (2.1-7) to obtain the following form:

$$u_M^{k+1} = u_M^k - V\tau\,(u_M^k - u_{M-1}^k)/\Delta x \qquad (2.2\text{-}12a)$$

$$u_m^{k+1} = S_p\,u_{m-1}^k + u_m^k - S_p\,u_{m+1}^{k+1}\,,\quad m = M-1,\ M-2,\ \ldots,\ 1 \qquad (2.2\text{-}12b)$$

$$u_o^{k+1} = g^{k+1} \qquad (2.2\text{-}12c)$$

where $S_p = (r/2)/(1-r/2)$

Equation (2.2-12b) can be written in the form:

$$u_m^{k+1} = S_p u_{m-1}^k + \sum_{j=0}^{M-2-m} (1-S_p^2)S_p^j(-1)^j u_{m+j}^k + S_p^{M-1}(-1^{M-1}u_{M-1}^k + S_p^M\,(-1)^M u_M^{k+1} \qquad (2.2\text{-}13)$$

A necessary condition for u_m^{k+1} to be bounded as $M \to \infty$ is:

$$S_p < 1 \qquad (2.2\text{-}14)$$

This condition is fulfilled for condition (2.2-11).

Finally we will show that application of the C-F-L condition, as introduced by the famous article of Courant, Friedrichs, and Lewy [3], also yields this condition as a necessary condition for convergence. According to this article the numerical domain of influence must contain the characteristic of the approximated PDE. For the sweep in increasing m direction, the domain of influence is the shaded area of figure (2-5); this area will always contain the characteristic if it has a positive angle with the X-axis. For the sweep in the decreasing m direction, however, the characteristic will belong to the numerical domain of dependence only if (2.2-11) is satisfied.

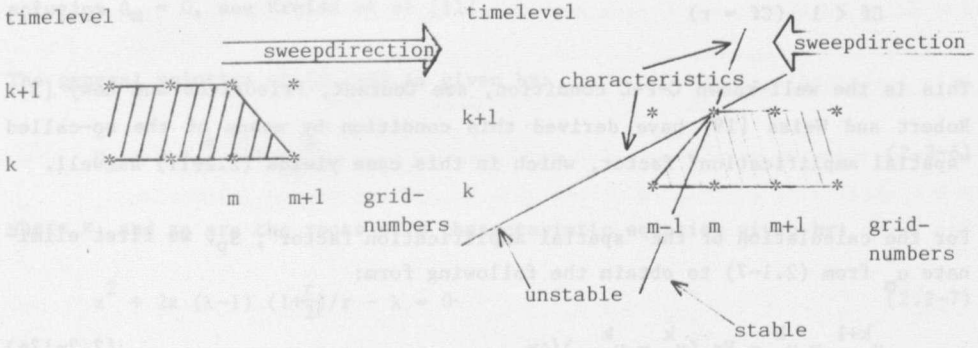


Figure 2-5 Numerical domains of dependence for Angled-Derivative method

Next we study the stability of (2.1-11). If u^* is eliminated from (2.1-11) for the inner points we then obtain:

$$u_m^{k+1} - u_m^k + r[(3u_m^{k+1} - 4u_{m-1}^{k+1} + u_{m-2}^{k+1})/4 + (u_{m+2}^k + 4u_{m+1}^k - 4u_{m-1}^k - u_{m-2}^k)/24] = 0$$

If we substitute $u_m^k = \lambda^k e^{im\phi}$, $\phi = \sigma\Delta x$, then we obtain the propagation factor:

$$P(\sigma,t) = \frac{1 - r/6\ i\ \sin\phi\ (2 + \cos\phi)}{1 + r/2\ [(1 - \cos\phi)^2 + i\sin\phi\ (2 - \cos\phi)]} \qquad (2.2\text{-}15)$$

Because $|P(\sigma,t)| < 1 \quad \forall\ r,\ 0 < \phi < \pi$, the stability of the Cauchy problem is established.

The stability of the boundary schemes is difficult to study due to the degree of the complex resolvent equation. If we apply the heuristic theory of Trapp

and Ramshaw [16], then (2.1-11) is stable. (They assume that an approximation is stable if the boundary schemes when applied to all grid points, $-\infty < m < \infty$, are stable, combined of course with stability of the scheme at the inner points.)

If at m = M-1 we apply the same scheme as at m=M then, following Goldberg and Tadmor [5,6], the outflow part is stable. Finally, practical experience did not show instabilities.

2.3 Time splitting methods for the advection equation with variable coefficients

In this section we will construct methods for the advection equation with a variable velocity given by:

$$u_t + v(x)\, u_x = 0 \tag{2.3-1}$$

The methods that we will describe are based only upon (2.1-8), (2.1-9), and (2.1-11), because the extension of these methods to the variable coefficient case involves an aspect which does not exist for the constant coefficient case.

The straightforward application of the Crank-Nicolson scheme will not be described because there are no principal differences from the constant coefficient case.

Application of (2.1-8) for this equation gives:

$$(u_m^{k+1} - u_m^k)/\tau + v_m\, (u_{m+1}^k - u_m^k + u_m^{k+1} - u_{m-1}^{k+1})/2\Delta x = 0 \tag{2.3-2}$$

where $v_m = v\,(m\Delta x)$.

If we regard v as constant ("freeze the coefficient") the stability analysis of the preceding section is applicable, yielding as local stability condition:

$$v_m \tau/\Delta x > -1 \tag{2.3-3}$$

Application of (2.1-9) gives:

$$(u_m^{k+1} - u_m^k)/\tau + v_m\,(u_{m+1}^{k+1} - u_m^{k+1} + u_m^k - u_{m-1}^k)/2\Delta x = 0 \qquad (2.3\text{-}4)$$

For this scheme the local stability condition is given by:

$$v_m \tau/\Delta x < 1 \qquad (2.3\text{-}5)$$

Obviously if the sign of v_m varies, then neither (2.3-2) nor (2.3-4) represents an unconditionally stable scheme. However by using (2.3-2) if $v_m < 0$ and (2.3-4) if $v_m \geq 0$ an unconditionally stable scheme is obtained. If we introduce the intermediate value u^* this scheme is given by:

Stage 1:

$$(u_m^* - u_m^k)/\tfrac{1}{2}\,\tau + S_{-x}\,(v_m,\, u_m^k) = 0 \qquad (2.3\text{-}6a)$$

Stage 2:

$$(u_m^{k+1} - u_m^*)/\tfrac{1}{2}\,\tau + S_{+x}(v_m,\, u_m^{k+1}) = 0 \qquad (2.3\text{-}6b)$$

where:

$$S_{-x}\,(v_m,\, u_m) = \begin{cases} v_m\,(u_{m+1} - u_m)/\Delta x \text{ for } v_m > 0 \\ v_m\,(u_m - u_{m-1})/\Delta x \text{ for } v_m \leq 0 \end{cases}$$

$$S_{+x}\,(v_m,\, u_m) = \begin{cases} v_m\,(u_m - u_{m-1})/\Delta x \text{ for } v_m > 0 \\ v_m\,(u_{m+1} - u_m)/\Delta x \text{ for } v_m \leq 0 \end{cases}$$

The second stage is the implicit part of the method. The matrix structure of the system of equations for u_m^{k+1} depends on the behaviour of the sign of v_m. If the sign is constant then the structure is bi-diagonal and the equations can be solved in one sweep. If the sign is changing then the structure is tri-diagonal and a double sweep method is one of the possibilities for solving the equations. Instead of a double sweep method we propose an iterative solution method. For that purpose we write (2.3-6b) in the form of a predictor-corrector method given by:

$$(u_m^{k[p]} - u^*)/\tfrac{1}{2}\tau + S_x[v_m, u_m^{k[p]}, \delta(p+p')] = 0 \qquad (2.3\text{-}7a)$$

$$u^{k+1} = u^{k[P]} \qquad (2.3\text{-}7b)$$

where $p = 1, \ldots, P$ and

$$S_x[v_m, u_m^{k[p]}, 0] = \begin{cases} v_m(u_m^{k[p]} - u_{m-1}^{k[p]})/\Delta x, & \text{for } v_m > 0 \\ v_m(u_{m+1}^{k[p-1]} - u_m^{k[p-1]})/\Delta x, & \text{for } v_m \leq 0 \end{cases}$$

$$S_x[v_m, u_m^{k[p]}, 1] = \begin{cases} v_m(u_m^{k[p-1]} - u_{m-1}^{k[p-1]})/\Delta x, & \text{for } v_m > 0 \\ v_m(u_{m+1}^{k[p]} - u_m^{k[p]})/\Delta x, & \text{for } v_m \leq 0 \end{cases}$$

$$\delta(p+p') = \tfrac{1}{2}[1 + (-1)^{p+p'}],$$

$$p' = \begin{cases} 0, & \text{if } \sum_{m=1}^{M} v_m > 0 \\ 1, & \text{if } \sum_{m=1}^{M} v_m \leq 0 \end{cases}$$

Note that if $\delta(p+p') = 0$ then (2.3-7a) proceeds in the increasing m direction, for $\delta(p+p') = 1$ (2.3-7a) proceeds in the decreasing m direction.
In general the predictor-corrector method as described above converges very fast as was found by practical experience. Usually two steps are enough. This means that the cost of this method is approximately the same as when the implicit set of equations is solved by a double sweep method. The principle of the method described here can be extended more easily to multi-dimensional problems.

To illustrate the fast convergence rate of (2.3-7) we give four examples of possible matrix structures depending on the signs of v_m:

Example I:

sign of v_m: + + + + ... $v_m > 0 \; \forall \; m$

main diagonal

Matrix structure:

```
*   *
    *   *
        *   *
            etc.
```

For this case the exact solution of (2.3-6) is reached by (2.3-7) in one step and needs no iterations.

Example II:

sign of v_m: - - - - ... $v_m < 0 \; \forall \; m$

main diagonal

Matrix structure

```
*   *
    *   *
        *   *
            etc.
```

For this case the exact solution of (2.3-6) is reached by (2.3-7) in one step and needs no iterations.

Example III:

sign of v_m: + + + $v_m = 0$ - - - etc.

main diagonal

Matrix structure:

```
*   *
    *   *
        1
            *   *
                *   *
                    etc.
```

For this example the exact solution of (2.3-6) is reached by (2.3-7) in two steps.

Example IV:

sign of v_m $\quad -\quad -\quad v_m = 0\quad +\quad +\quad$ etc.

main diagonal

Matrix structure:

```
*   *
    *   *
        1
        *   *
            *   *
                etc.
```

For this example the exact solution of (2.3-6) is reached by (2.3-7) in two steps.

The method given by (2.1-11) can be treated in the same way, in the case where v is variable and changes sign. We obtain:

$$(u_m^* - u_m^k)/\tfrac{1}{2}\tau + v_m(u_{m+2}^k + 4u_{m+1}^k - 4u_{m-1}^k - u_{m-2}^k)/12\Delta x = 0, \qquad (2.3\text{-}8a)$$

$$(u_m^{k+1} - u_m^*)/\tfrac{1}{2}\tau + S_{+x}(v_m, u_m^{k+1}) = 0 \qquad (2.3\text{-}8b)$$

where:

$$S_{+x}(v_m, u_m) = \begin{cases} v_m\,(3u_m - 4u_{m-1} + u_{m-2})/2\Delta x, & \text{for } v_m > 0 \\ v_m\,(-3u_m + 4u_{m+1} - u_{m+2})/2\Delta x, & \text{for } v_m \leqslant 0 \end{cases}$$

Again the second equation is solved iteratively by a predictor-corrector method:

$$(u_m^{k[p]} - u_m^*)/\tfrac{1}{2}\tau + S_{+x}[v_m, u_m^{k[p]}, \delta(p+p')] = 0 \qquad (2.3\text{-}9a)$$

$$u_m^{k+1} = u_m^{k[P]} \qquad (2.3\text{-}9b)$$

where: p=1, ..., P and

$$S_{+x}[v_m, u_m^{k[p]}, \delta] = \begin{cases} v_m\,(3u_m^{k[p-\delta]} - 4u_{m-1}^{k[p-\delta]} + u_{m-2}^{k[p-\delta]})/2\Delta x, & \text{for } v_m > 0 \\ v_m\,(-3u_m^{k[p-1+\delta]} + 4u_{m+1}^{k[p-1+\delta]} - u_{m+2}^{k[p-1+\delta]})/2\Delta x, & \text{for } v_m \leq 0 \end{cases}$$

$$\delta\,(p+p') = \tfrac{1}{2}\,[1 + (-1)^{p+p'}]$$

$$p' = \begin{cases} 0, & \text{if } \sum_{m=1}^{M} v_m > 0 \\ 1, & \text{if } \sum_{m=1}^{M} v_m < 0 \end{cases}$$

If $\delta\,(p+p') = 0$ then (2.3-9a) proceeds in the increasing m direction; for $\delta\,(p+p') = 1$ (2.3-9a) proceeds in the decreasing m direction.

Each stage of (2.3-6) is an $O(\tau, \Delta x)$, consistent approximation of (2.3-1), the overall order of consistency is $O\,(\tau^2, \Delta x^2)$.
The order of consistency of (2.3-8) is $O\,(\tau, \Delta x^2)$, and the overall consistency is $O\,(\tau^2, \Delta x^3)$.

2.4 Characteristic interpolation methods

Some authors, e.g., Benqué et al. [2], claim that very efficient approximation methods for SWE can be constructed by using "operator splitting". This means that the differential operator is split into several parts. Each part is approximated by a special method. For the advection operator Benqué et al apply an unconditionally stable characteristic interpolation method. In this section we describe the principle of this method for the following simple advection equation:

$$u_t + V\,u_x = 0 \qquad (2.4\text{-}1)$$

The general solution is given by $u(x,t) = f(x-Vt)$. This means that the solution of (2.4-1) is constant along the line $x = Vt$ which is called a characteristic. For characteristic interpolation methods this characteristic is used for the construction of approximate solutions. For the exact solution, the following relation holds:

$$u[(k+1)\tau,\ m\Delta x] = u(k\tau,\ m\Delta x - V\tau) \tag{2.4-2}$$

The point $[(k+1)\tau,\ m\Delta x]$ is a grid point but in general the point $(k\tau,\ m\Delta x - V\tau)$ does not coincide with a grid point, see also figure (2-6).

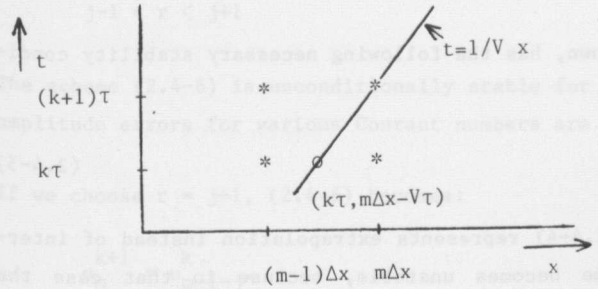


Figure (2-6) x,t space of (2.4-1)
* grid points

An approximation for $u(k\tau, m\Delta x - V\tau)$ can be obtained by Lagrangian interpolation. This gives the following numerical scheme:

$$u_m^k = \mathcal{L}_p\,(m\Delta x - V\tau,\ \underline{u}^{k-1}),\quad m=1,\ \ldots,\ M \tag{2.4-3}$$

where $\mathcal{L}_p\,(x,\ \underline{u})$ denotes a Lagrangian interpolation polynomial of order p which is based upon $[u_0,\ \ldots,\ u_M]^T$.

This well-known principle forms the basis of many numerical methods for the advection equation, see e.g. Fromm [4], Roache [18], and Wesseling [23].
Both Lagrangian polynomials and Hermitian polynomials are used, see, e.g., Holly and Preissmann [10].

In general, characteristic interpolation methods are explicit and not unconditionally stable, although there are exceptions, for example, "Carlson's scheme", see Richtmyer and Morton [17]. This scheme tries to follow the characteristic as close as possible while keeping unconditional stability.

A similar principle is treated by Holly and Preissmann [10]. We treat this principle by means of a second order Leith scheme, which is equivalent to the second order Lax-Wendroff scheme. It is given by:

$$u_m^{k+1} = u_m^k - \tfrac{1}{2} r\,(u_{m+1}^k - u_{m-1}^k) + \tfrac{1}{2} r^2\,(u_{m-1}^k - 2u_m^k + u_{m+1}^k) \qquad (2.4\text{-}4)$$

where $r = V\tau/\Delta x$ ($r = Cf$, the "Courant number")

This scheme, which is well known, has the following necessary stability condition:

$$|r| \leq 1 \qquad (2.4\text{-}5)$$

This means that as soon as (2.4-4) represents extrapolation instead of interpolation, the numerical scheme becomes unstable, because in that case the numerical domain of dependence does not contain the characteristic of (2.4-1), see figure (2-7).

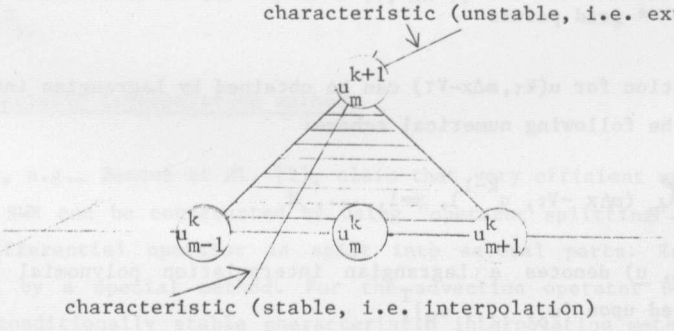


Figure (2-7) Numerical domain of dependence of 2nd order Leith method.

By choosing the basis points of the interpolation polynomial such that it always represents an interpolation formula, an explicit unconditionally stable

method can be constructed.

Based upon second order Lagrangian interpolation such a scheme becomes:

$$u_m^{k+1} = \tfrac{1}{2}\, j(j-1)\, u_{m-1-j}^k + (1-j^2)\, u_{m-j}^k - \tfrac{1}{2}\, j(j+1)\, u_{m+1-j}^k$$

$$- r\,[\tfrac{1}{2}\,(2j-1)\, u_{m-1-j}^k - 2j\, u_{m-j}^k + \tfrac{1}{2}\,(2j+1)\, u_{m+1-j}^k]$$

$$+ \tfrac{1}{2}\, r^2\, (u_{m-1-j}^k - 2u_{m-j}^k + u_{m+1-j}^k) \qquad (2.4\text{-}6)$$

where j is an integer number for which the following relation holds:

$$j-1 < r < j+1 \qquad (2.4\text{-}7)$$

The scheme (2.4-6) is unconditionally stable for the Cauchy problem. Phase and amplitude errors for various Courant numbers are given by figure (2-8).

If we choose r = j+1, (2.4-6) becomes:

$$u_m^{k+1} = u_{m-1-j}^k. \qquad (2.4\text{-}8)$$

In that case, the method yields the exact solution, and the scheme is said to have the "point to point transfer property". This is common to all characteristic interpolation methods.

If a method constructed by the method of lines has this property then it can be written in the form of a characteristic interpolation method. As an example we treat the "box scheme", see Lam and Simpson (12). For (2.4-1) this scheme is given by:

$$\tfrac{1}{2}\,(u_{m+1}^{k+1} + u_m^{k+1} - u_{m+1}^k - u_m^k)/\tau + \tfrac{1}{2}\, V\, (u_{m+1}^{k+1} - u_m^{k+1} + u_{m+1}^k - u_m^k)/\Delta x = 0 \qquad (2.4\text{-}9)$$

which is a combination of the trapezoidal rule, see Lambert [13], for the integration in time, and the box spatial discretization treated in chapter I. By rewriting (2.4-9) we obtain:

$$u_{m+1}^{k+1} = -\frac{1-r}{1+r} u_m^{k+1} + u_m^k + \frac{1-r}{1+r} u_{m+1}^k \qquad (2.4\text{-}10$$

This formulation can be considered as Lagrangian interpolation based upon

$$u_m^{k+1},\ u_m^k \text{ and } u_{m+1}^k,$$

see figure (2-10).

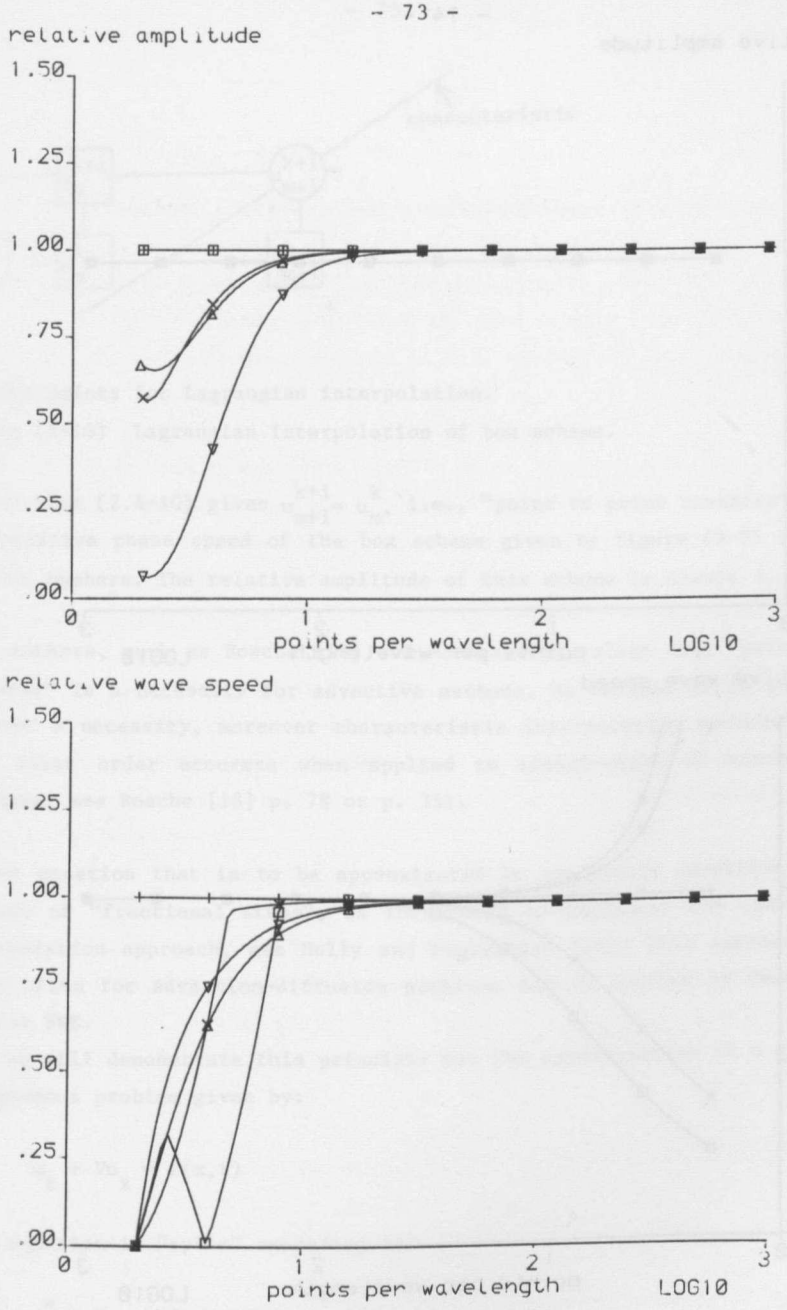


Figure (2-8) Relative amplitude and wavespeed of 2nd order stabilized Leith scheme

Δ: Cf = 0.1, ∇: Cf + 0.5, +: Cf = 1.0,

X: Cf = 2.5, □ : Cf = 4.0.

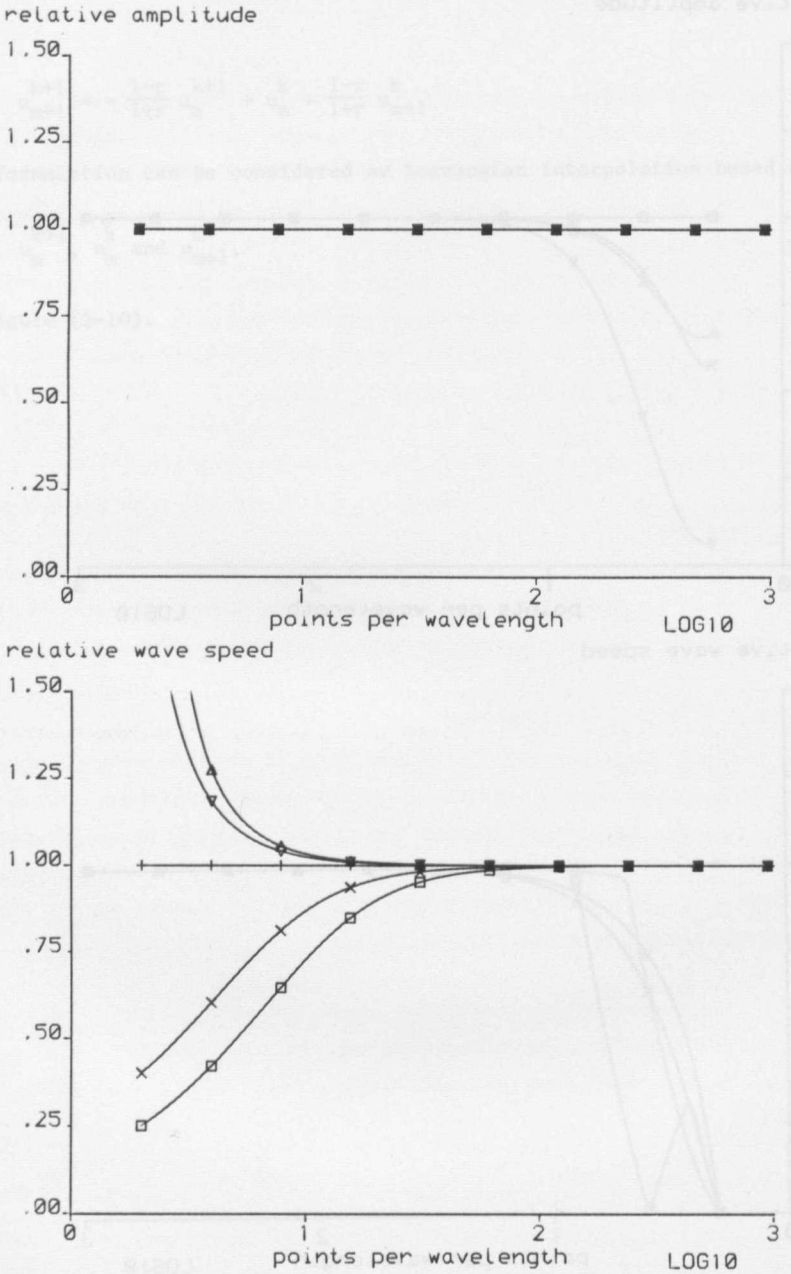
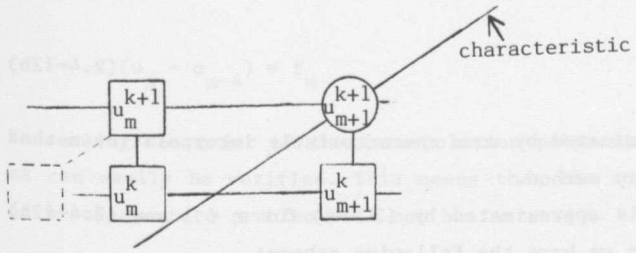


Figure (2-9) Relative amplitude and wavespeed of box scheme
Δ: Cf = 0.1, ∇: Cf = 0.5, +: Cf = 1.0,
X: Cf = 2.5, □ : Cf = 4.0.



▫ basis points for Lagrangian interpolation.

Figure (2-10) Lagrangian interpolation of box scheme.

If r=1 then (2.4-10) gives $u_{m+1}^{k+1} = u_m^k$, i.e., "point to point transfer". See also the relative phase speed of the box scheme given by figure (2-9) for various Courant numbers. The relative amplitude of this scheme is always 1.

Some authors, such as Roache [18] and Morton [15], claim that "point to point transfer" is a necessity for advective methods. We believe it is an advantage but not a necessity, moreover characteristic interpolation methods are often only first order accurate when applied to steady-state or non-homogeneous problems, see Roache [18] p. 78 or p. 351.

If the equation that is to be approximated is not purely advective, then the concept of "fractional steps", is introduced to implement the characteristic interpolation approach, see Holly and Preissmann [16]. This approach is used quite often for advection-diffusion problems and is applied by Benqué et al. [2] for SWE.
Here we will demonstrate this principle for the approximation of a simple non-homogeneous problem given by:

$$u_t + Vu_x = f(x,t) \qquad (2.4\text{-}11)$$

This equation is "split" according to:

$$u_t^* + Vu_x = 0 \qquad (2.4\text{-}12a)$$

and:

$$u_t = f(x,t) \tag{2.4-12b}$$

Equation (2.4-12a) is approximated by some characteristic interpolation method and (2.4-12b) by an arbitrary method.

For example, if (2.4-12a) is approximated by (2.4-6) for $r < 1$ and (2.4-12b) by the trapezoidal rule then we have the following scheme:

Stage 1:

$$u_m^* = u_m^k - \tfrac{1}{2} r\,(u_{m+1}^k - u_{m-1}^k) + \tfrac{1}{2} r^2\,(u_{m-1}^k - 2u_m^k + u_{m+1}^k) \tag{2.4-13a}$$

Stage 2:

$$(u_m^{k+1} - u_m^*)/\tau = \tfrac{1}{2} f_m^{k+1} + \tfrac{1}{2} f_m^k \tag{2.4-13b}$$

where $f_m^k = f(m\Delta x, k\tau)$.

Neither (2.4-13a) nor (2.4-13b) is consistent with (2.4-11) but by eliminating u_m^* we obtain:

$$(u_m^{k+1} - u_m^k)/\tau + V(u_{m+1}^k - u_{m-1}^k)/2\Delta x - \tfrac{1}{2}\tau V^2 (u_{m-1}^k - 2u_m^k + u_{m+1}^k)/\Delta x^2 = \tfrac{1}{2}(f_m^{k+1} + f_m^k) \tag{2.4-14}$$

This equation is only first order consistent with (2.4-11). Let us assume that (2.4-14) is applied as an iterative method for the approximation of a steady-state equation given by:

$$Vu_x = f(x) \tag{2.4-15}$$

then for r=1 (2.4-14) gives an approximation given by:

$$V\,(u_m - u_{m-1}) = f_m \tag{2.4-16}$$

This is only a first order accurate approximation of (2.4-15). By increasing the Courant number the approximation becomes less accurate. For example, r=4 gives:

$$V\,(u_m - u_{m-4}) = f_m \qquad (2.4\text{-}17)$$

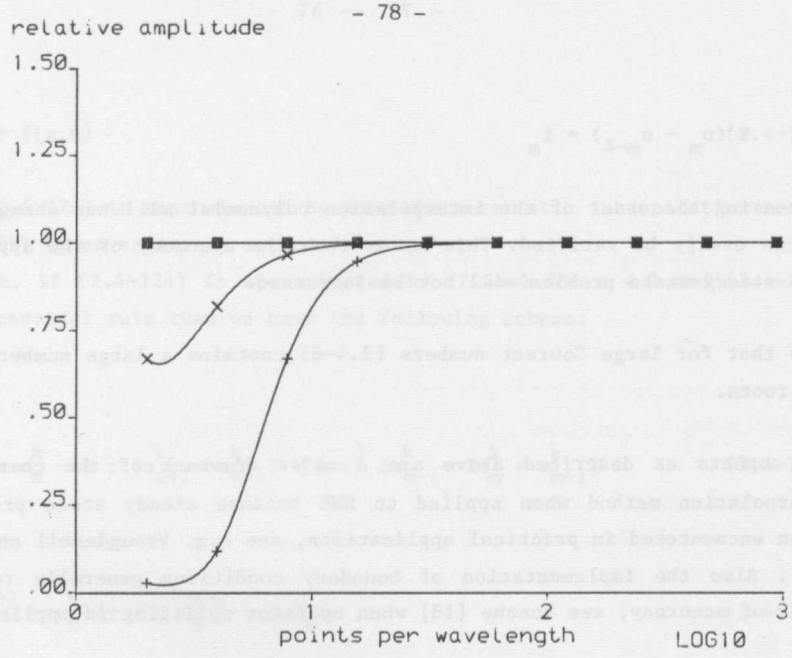
Increasing the order of the interpolation polynomial will not change (2.4-17) as can easily be verified. This means that the accuracy of the approximation of a steady-state problem will not be increased.

Note that for large Courant numbers (2.4-6) contains a large number of spurious roots.

The aspects as described above are a major drawback of the characteristic interpolation method when applied to SWE because steady state problems are often encountered in practical applications, see e.g. Vreugdenhil and Wijbenga [21]. Also the implementation of boundary conditions generally reduces the order of accuracy, see Roache [18] when operator splitting is applied.

Methods of class i, as defined in the introduction, of which the box scheme (2.4-9) is an example too, give no problems when non-homogeneous terms are to be implemented. For these methods the ultimate result, when steady-state problems are approximated, does not depend on τ. Only the convergence rate depends on the timesteps.

Finally we compare the propagation properties of the unconditional stable methods mentioned in this chapter:
(i) Crank Nicolson scheme, (ii) the unconditional stable angled derivative method, (iii) the reduced phase error split scheme, (iv) the stabilized Leith scheme, and (v) the box scheme in the figures (2-11) and (2-12).



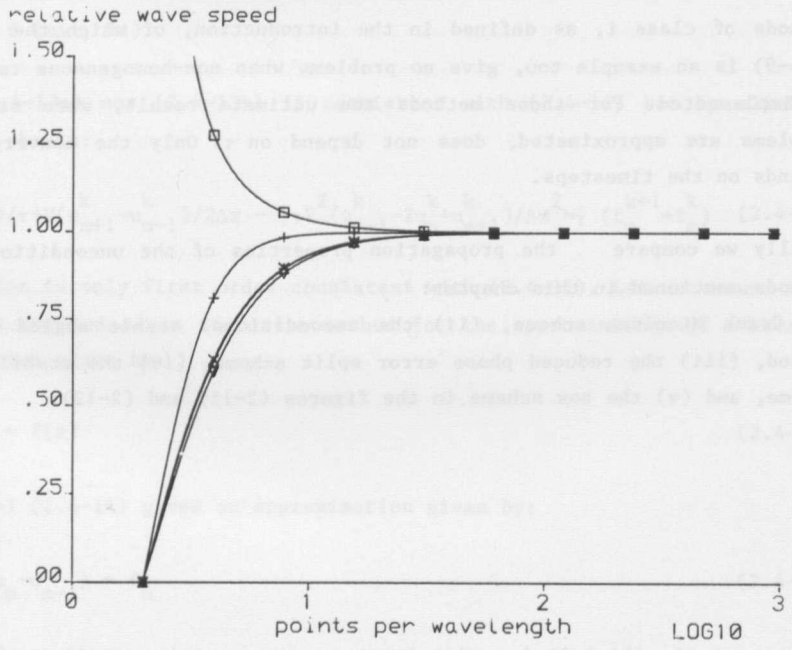


Figure (2-11) Comparison of propagation properties, Cf = 0.1.

Δ: Crank Nicolson

∇: unconditional stable angled derivative

+: reduced phase error split scheme

x: stabilized Leith scheme

□: box scheme

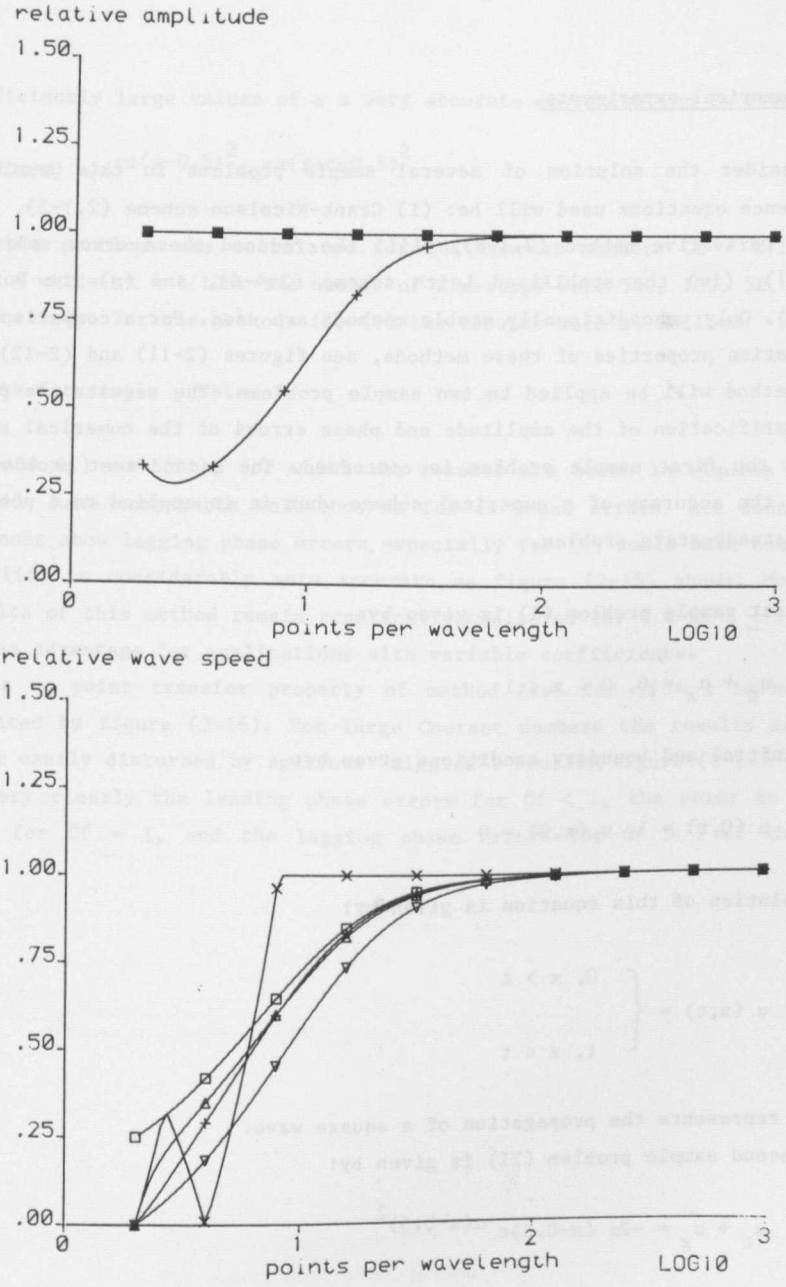


Figure (2-12) Comparison of propagation properties, Cf = 4.0

Δ: Crank Nicolson

∇: unconditional stable angled derivative

+: reduced phase error split scheme

x: stabilized Leith scheme

□: box scheme

2.5 Numerical experiments

We consider the solution of several sample problems in this section. The difference equations used will be: (i) Crank-Nicolson scheme (2.1-5), (ii) the Angled Derivative method (2.1-8), (iii) the reduced phase error split scheme (2.1-11), (iv) the stabilized Leith scheme (2.4-6), and (v) the box scheme (2.4-9). Only unconditionally stable methods are used. For a comparison of the propagation properties of these methods, see figures (2-11) and (2-12).
Each method will be applied to two sample problems. The results are presented as a verification of the amplitude and phase errors of the numerical scheme as far as the first sample problem is concerned. The second test problem illustrates the accuracy of a numerical scheme when it is applied to a non-homogeneous steady-state problem.

The first sample problem (I) is given by:

$$u_t + u_x = 0, \quad 0 \leq x \leq 1 \tag{2.5-1a}$$

with initial and boundary conditions given by:

$$u(0,t) = 1, \quad u(x,0) = 0 \tag{2.5-1b}$$

The solution of this equation is given by:

$$u(x,t) = \begin{cases} 0, & x > t \\ 1, & x \leq t \end{cases} \tag{2.5-2}$$

which represents the propagation of a square wave.
The second sample problem (II) is given by:

$$u_t + u_x = -2\alpha\,(x-0.5)e^{-\alpha(x-0.5)^2} \tag{2.5-3a}$$

with initial and boundary condition:

$$u(0,t) = 0, \quad u(x,0) = 0 \tag{2.5-3b}$$

For sufficiently large values of α a very accurate approximation is given by:

$$u(x,t) = e^{-\alpha(x-0.5)^2} - e^{-\alpha(x-t-0.5)^2} \qquad (2.5\text{-}4)$$

Both sample problems are calculated for the Courant numbers Cf = 0.1, 0.5, 2.5, and 4 while $\Delta x = 1/100$. The number of timesteps were 500, 100, 50, 20 and 13 respectively. For the methods (i)-(v) the results were as follows:

a. Results for sample problem I

The methods (i) and (ii), for which the results are shown in figures (2-13) and (2-14), have comparable solutions as far as phase errors are concerned. Both schemes show lagging phase errors, especially for the small wave numbers. Method (iii) is considerably more accurate, as figure (2-15) shows. Moreover, the results of this method remain practically unaltered for $0 \leq Cf \leq \pm 1.5$. This is an advantage for applications with variable coefficients.
The point to point transfer property of method (iv) for Cf = 1 is clearly demonstrated by figure (2-16). For large Courant numbers the results seem to be rather easily disturbed by spurious "wiggles". Finally, figure (2-17) illustrates very clearly the leading phase errors for $Cf < 1$, the point to point property for Cf = 1, and the lagging phase errors for $Cf > 1$ of the box scheme.

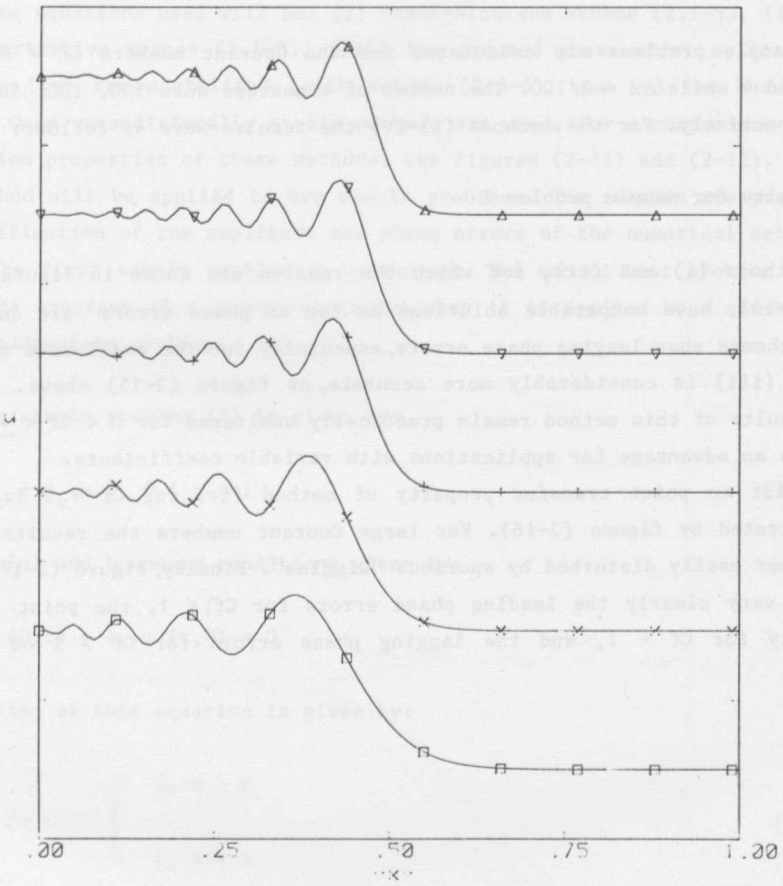


Figure (2-13) Crank-Nicolson scheme

Δ: Cf = 0.1, number of timesteps = 500

∇: Cf = 0.5, number of timesteps = 100

+: Cf = 1.0, number of timesteps = 50

X: Cf = 2.5, number of timesteps = 20

□: Cf = 4.0, number of timesteps = 13

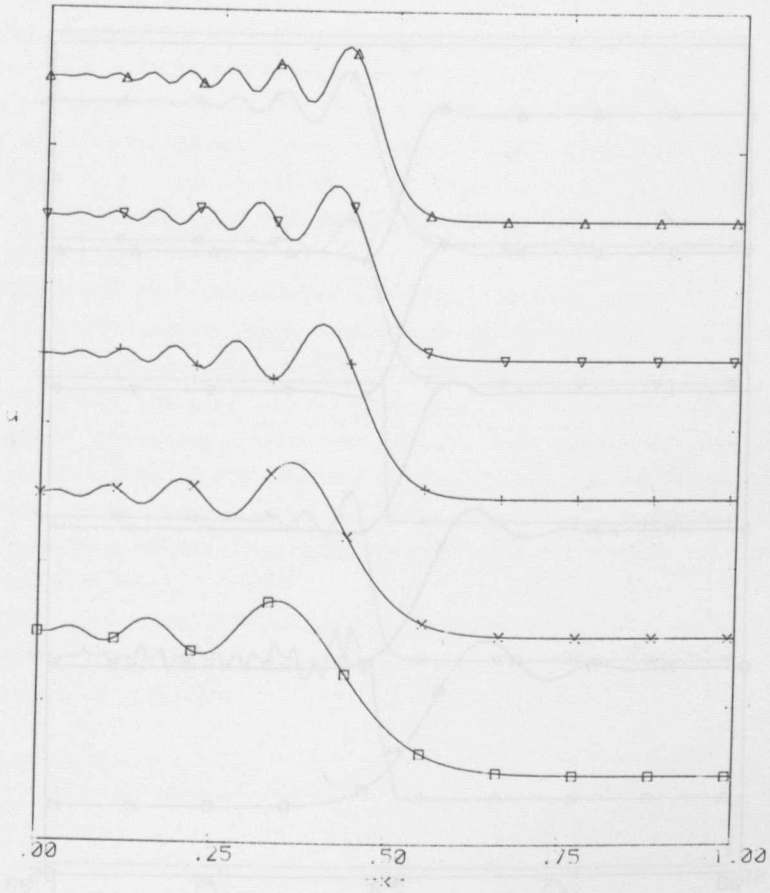


Figure (2-14) Angled-Derivative method

Δ: Cf = 0.1, number of timesteps = 500

∇: Cf = 0.5, number of timesteps = 100

+: Cf = 1.0, number of timesteps = 50

X: Cf = 2.5, number of timesteps = 20

□: Cf = 4.0, number of timesteps = 13

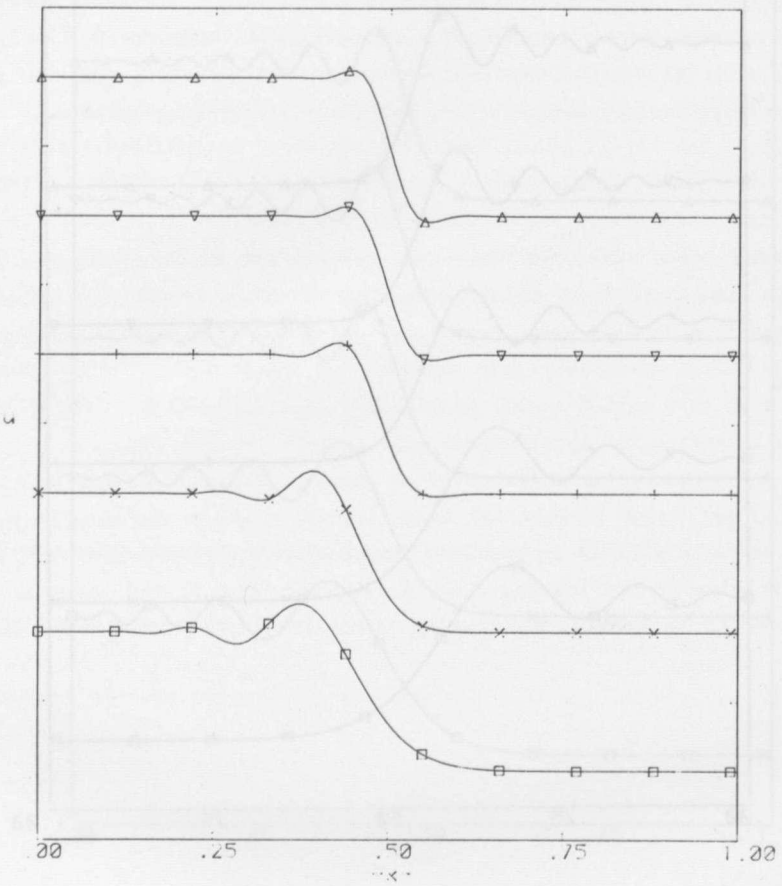


Figure (2-15) Reduced phase error split scheme

Δ: Cf = 0.1, number of timesteps = 500

∇: Cf = 0.5, number of timesteps = 100

+: Cf = 1.0, number of timesteps = 50

X: Cf = 2.5, number of timesteps = 20

□ : Cf = 4.0, number of timesteps = 13

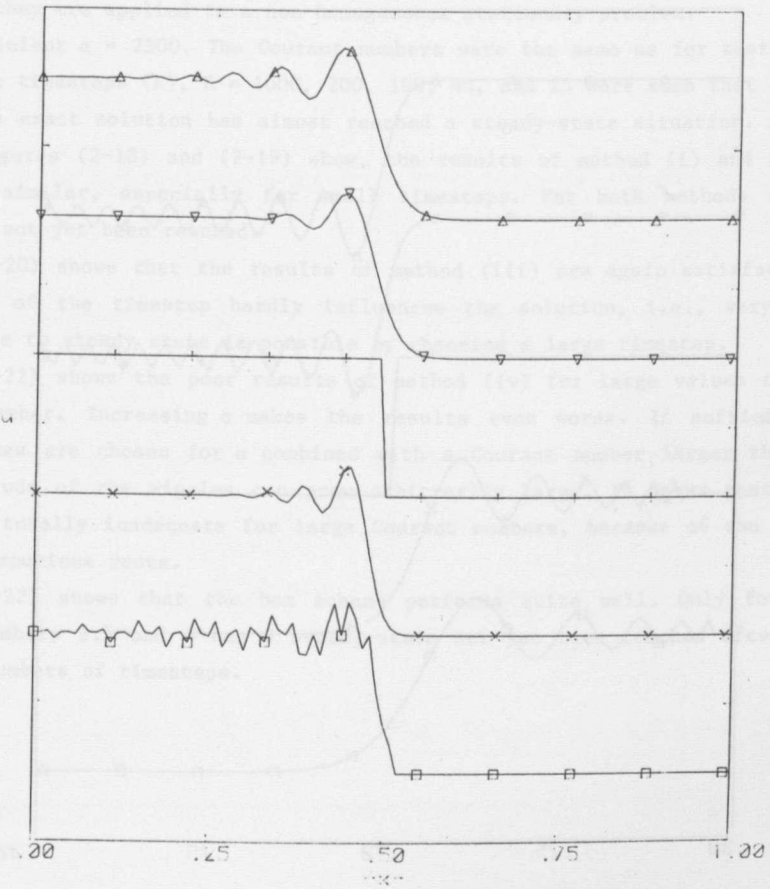


Figure (2-16) Stabilized 2nd order Leith scheme

Δ: Cf = 0.1, number of timesteps = 500

∇: Cf = 0.5, number of timesteps = 100

+: Cf = 1.0, number of timesteps = 50

X: Cf = 2.5, number of timesteps = 20

□: Cf = 4.0, number of timesteps = 13

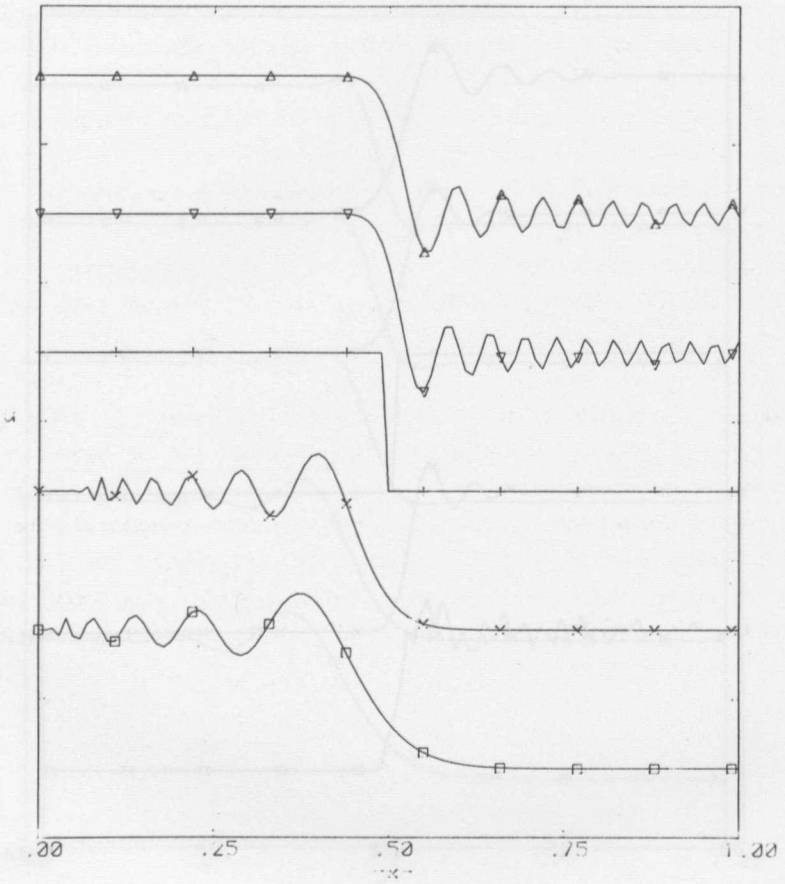


Figure (2-17) box-scheme

Δ: Cf = 0.1, number of timesteps = 500

∇: Cf = 0.5, number of timesteps = 100

+: Cf = 1.0, number of timesteps = 50

X: Cf = 2.5, number of timesteps = 20

□: Cf = 4.0, number of timesteps = 13

b. Results for sample problem II

This test problem is used to illustrate the effectivity of the methods (i) - (v) when they are applied to a non-homogeneous stationary problem.
The coefficient α = 2500. The Courant numbers were the same as for test problem I. The timesteps (K), K = 1000, 200, 100, 40, and 25 were such that at $t = K\tau$ the exact solution has almost reached a steady-state situation. Again, as the figures (2-18) and (2-19) show, the results of method (i) and method (ii) are similar, especially for small timesteps. For both methods steady state has not yet been reached.
Figure (2-20) shows that the results of method (iii) are again satisfactory. The value of the timestep hardly influences the solution, i.e., very fast convergence to steady state is possible by choosing a large timestep.
Figure (2-21) shows the poor results of method (iv) for large values of the Courant number. Increasing α makes the results even worse. If sufficiently large values are chosen for α combined with a Courant number larger than 2, the amplitude of the wiggles can grow arbitrarily large. It shows that this method is totally inadequate for large Courant numbers, because of the large number of spurious roots.
Figure (2-22) shows that the box scheme performs quite well. Only for the Courant numbers 2.5 and 4 has a steady-state not yet been reached after the observed numbers of timesteps.

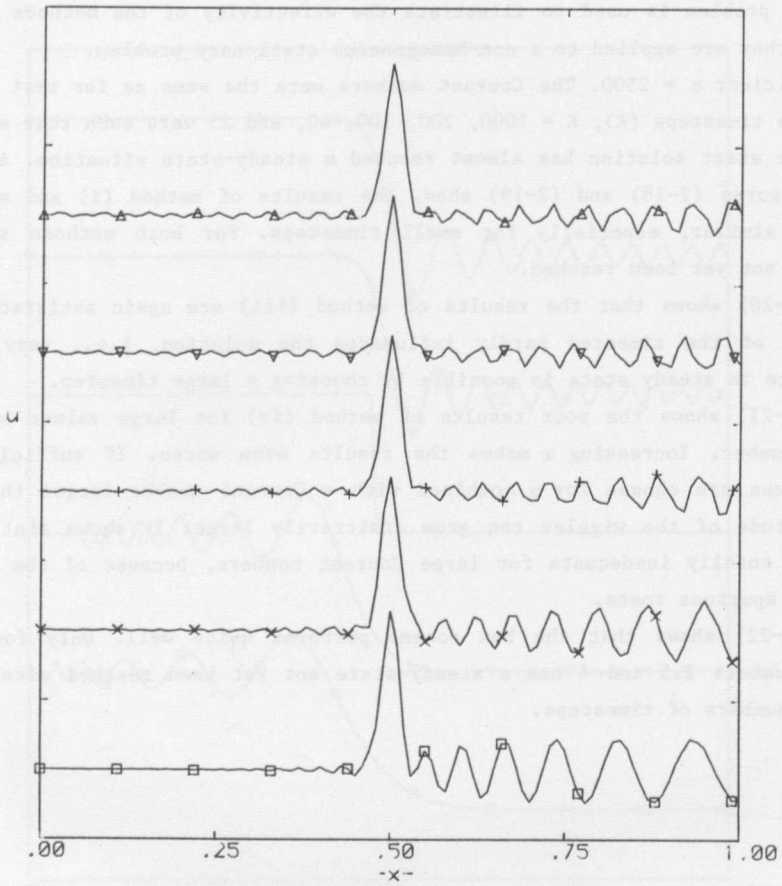


Figure (2-18) Crank-Nicolson scheme

Δ: Cf = 0.1, number of timesteps = 1000

∇: Cf = 0.5, number of timesteps = 200

+: Cf = 1.0, number of timesteps = 100

X: Cf = 2.5, number of timesteps = 40

□: Cf = 4.0, number of timesteps = 25

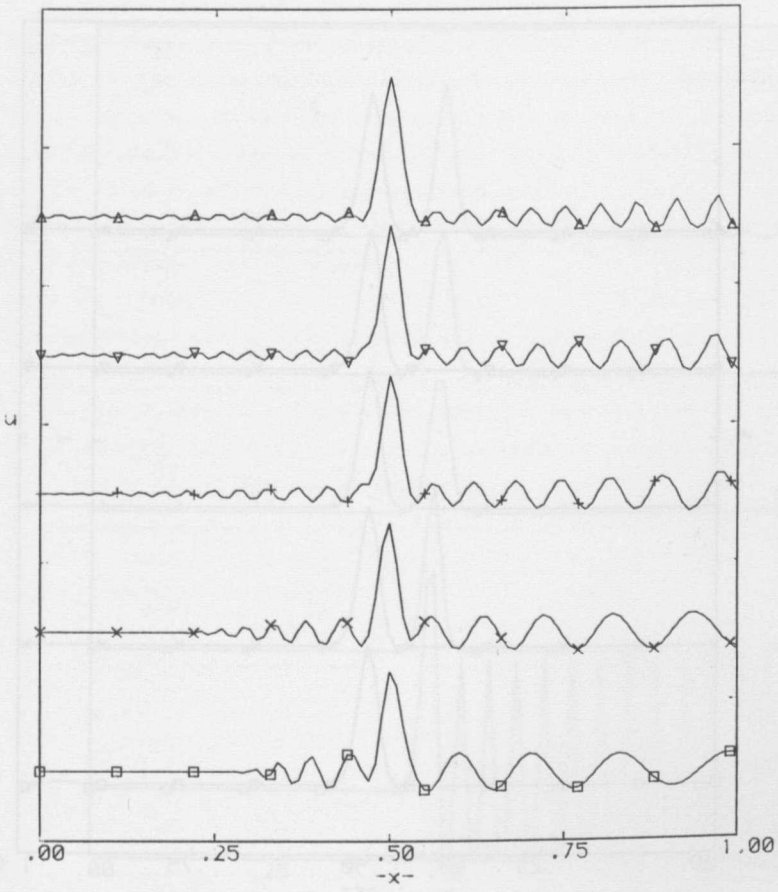


Figure (2-19) Angled-Derivative method

Δ: Cf = 0.1, number of timesteps = 1000

∇: Cf = 0.5, number of timesteps = 200

+: Cf = 1.0, number of timesteps = 100

X: Cf = 2.5, number of timesteps = 40

□ : Cf = 4.0, number of timesteps = 25

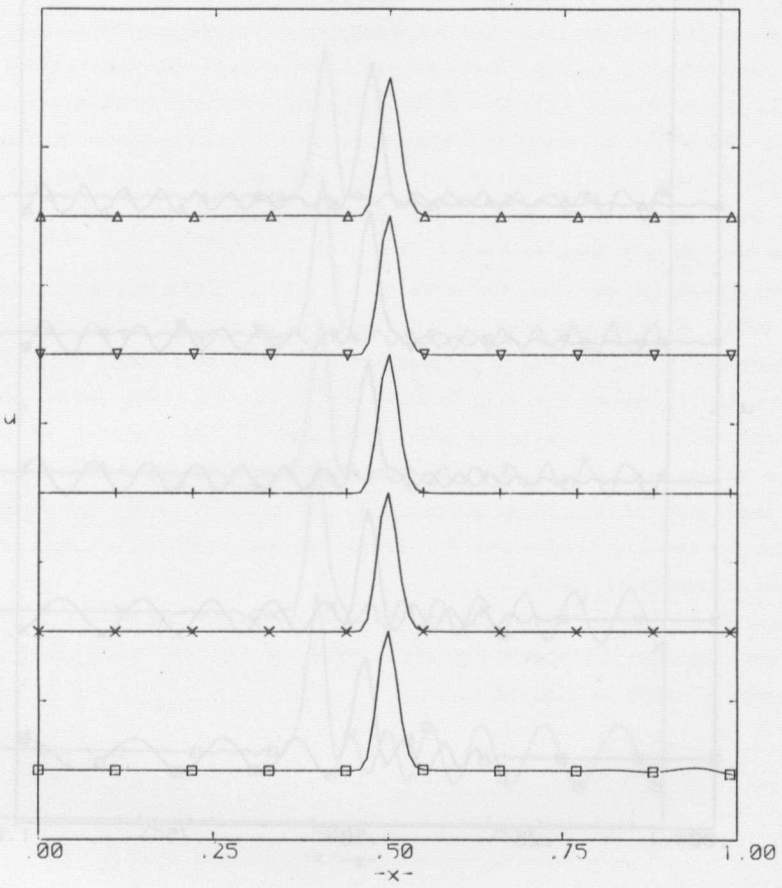


Figure (2-20) Reduced phase error split scheme

Δ: Cf = 0.1, number of timesteps = 1000

∇: Cf = 0.5, number of timesteps = 200

+: Cf = 1.0, number of timesteps = 100

X: Cf = 2.5, number of timesteps = 40

□: Cf = 4.0, number of timesteps = 25

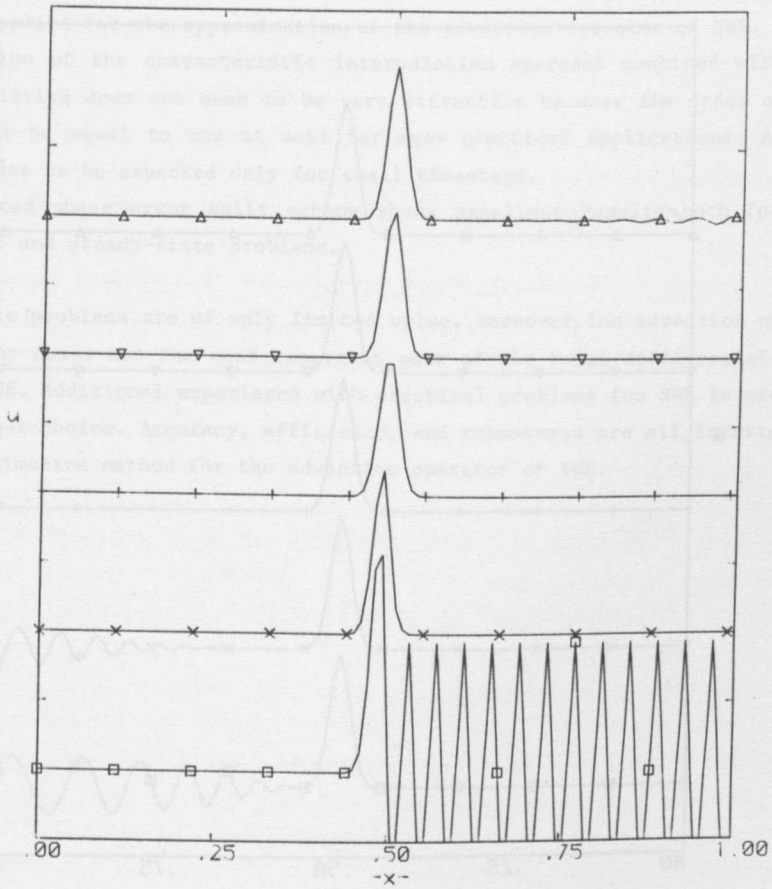


Figure (2-21) Stabilized Leith scheme (2^{nd} order)

Δ: Cf = 0.1, number of timesteps = 1000

∇: Cf = 0.5, number of timesteps = 200

+: Cf = 1.0, number of timesteps = 100

X: Cf = 2.5, number of timesteps = 40

□ : Cf = 4.0, number of timesteps = 25

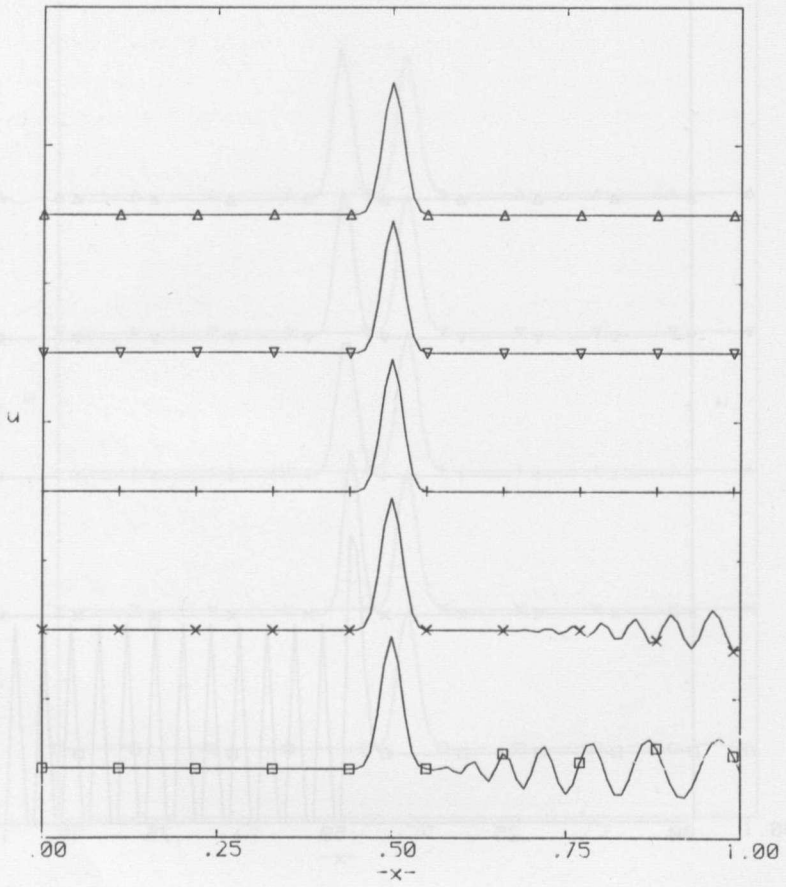


Figure (2-22) box scheme

Δ: Cf = 0.1, number of timesteps = 1000

∇: Cf = 0.5, number of timesteps = 200

+: Cf = 1.0, number of timesteps = 100

X: Cf = 2.5, number of timesteps = 40

□: Cf = 4.0, number of timesteps = 25

2.6 Concluding Remarks

Several numerical methods for the advection equation were compared that can all be applied for the approximation of the advection operator of SWE.
Application of the characteristic interpolation approach combined with operator splitting does not seem to be very attractive because the order of accuracy will be equal to one at most for many practical applications. Adequate results are to be expected only for small timesteps.
The reduced phase error split scheme shows excellent results both for time-dependent and steady-state problems.

The sample problems are of only limited value. Moreover, the advection operator is in many cases not the most important part of the total differential operator of SWE. Additional experience with practical problems for SWE is necessary for a final choice. Accuracy, efficiency, and robustness are all important for an approximation method for the advection operator of SWE.

REFERENCES TO CHAPTER 2

1. ABBOTT, M.B.,
The method of characteristics, in Unsteady Flow in Open Channels edited by V. Mahmood and V. Yevjevich, Water Resources Publications, Fort Collins, 1975.

2. BENQUE, J.P., J.A. CUNGE, J. FEUILLET, A. HAUGUEL and F.M. HOLLY,
New Method for Tidal Current Computation,
Journal of the Waterway, Port, Coastal and Ocean Division, ASCE, 1982, pp 396-417.

3. COURANT, R., K. FRIEDRICHS and H. LEWY,
Uber die partiellen Differenzen-Gleichungen der mathematischen Physik,
Mathematische Annalen, 100, 1928, pp. 32-74.
(Engl. translation by Ph. Fox, On the Partial Difference Equations of Mathematical Physics, IBM-Journal, 1967, pp. 215-234).

4. FROMM, J.E.,
A Method for Reducing Dispersion in Convective Difference Schemes.
Journal of Computational Physics, No. 3, 1968, pp. 176-189.

5. GOLDBERG, M. and E. TADMOR,
Scheme Independent Stability for Difference Approximations of Hyperbolic Initial Boundary Value Problems, I.
Mathematics of Computation, V.32, 1978, pp. 1097-1107.

6. GOLDBERG, M., and E. TADMOR,
Scheme Independent Stability Criteria for Difference Approximations of Hyperbolic Initial Boundary Value Problems, II.
Mathematics of Computation, V.36, 1981, pp. 603-626.

7. GODUNOV, S.K. and V.S. RYABENKI,
Theory of Difference Schemes,
North-Holland Publishing Company, Amsterdam, 1964.

8. GOTTLIEB, D. and S.A. ORSZAG,
Numerical Analysis of Spectral Methods, Theory and Applications.
Society for Industrial and Applied Mathematics, Philadelphia, 1977.

9. GUSTAFSSON, B.,
The Convergence Rate for Difference Approximations to Mixed Initial Boundary Value Problems.
Mathematics of Computation, V.26, 1975, pp. 396-406.

REFERENCES (continued)

10. HOLLY, F.M. and A. PREISSMANN,
Accurate Calculation of Transport in Two Dimensions,
Journal of the Hydraulics Division, ASCE, V.103, 1977, pp. 1259-1277.

11. KREISS, H.O., B. GUSTAFSSON and A. SUNDSTROM,
Stability Theory of Difference Approximations for Mixed Initial Boundary Value Problems, II.
Mathematics of Computation, V.26, 1972, pp. 649-686.

12. LAM, D.C.L. and R.B. SIMPSON,
Centered Differencing and the Box Scheme for Diffusion Convection Problems.
Journal of Computational Physics, No. 22, 1976, pp. 48o-500.

13. LAMBERT, J.D.,
Computational Methods in Ordinary Differential Equations,
Wiley London-New York, 1973.

14. MILLER, J.J.H.,
On the Location of Zeros of Certain Classes of Polynomials with Applications to Numerical Analysis.
Journal Institute of Mathematics and Its Applications, No. 8, 1971, pp. 397-406.

15. MORTON, K.W.,
Petrov Galerkin methods for non-self-adjoint problems,
Procs. of the 8th Biennel Conference on Numerical Analysis, Dundee, 1979, Springer, 1980.

16. TRAPP, J.A. and J.D. RAMSHAW,
A simple and Heuristic Method for Analyzing the Effect of Boundary Conditions on Numerical Stability.
Journal of Computational Physics, V.20, pp.238-242, 1976.

17. RICHTMYER R.D. and K.W. MORTON,
Difference Methods for Initial-Value Problems,
Interscience Publishers, Wiley, New York-London, 1967.

18. ROACHE, P.J.,
Computational Fluid Dynamics,
Hermosa Publishers, Albuquerque, N.M., 1972.

REFERENCES (continued)

19. ROBERTS, K.W. and N.O. WEISS,
Convective Difference Schemes,
Mathematics of Computation, No. 2, 1966, pp. 272-299.

20. STELLING, G.S.,
Improved Stability of Dronkers Tidal Schemes,
Journal of the Hydraulics Division, ASCE, V106, 1980, pp. 1365-1379.

21. VREUGDENHIL, C.B. and J.H.A. WIJBENGA,
Computation of Flow Patterns in Rivers.
Journal of the Hydraulics Division, ASCE, V108, 1982, pp. 1296-1310.

22. WEARE, T.J.,
Instability in Tidal Flow Computational Schemes,
Journal of the Hydraulics Division, ASCE, V102, 1976, pp. 569-580.

23. WESSELING, P.,
On the Construction of Accurate Difference Schemes for Hyperbolic Partial Differential Equations.
Journal of Engineering Mathematics, V7, 1973, pp. 19-31.

24. WIRZ, H.J., F. DE SCHUTTER and A. TURI,
An Implicit, Compact, Finite Difference Method to Solve Hyperbolic Equations,
Mathematics and Computers in Simulation, 1977, pp. 241-261.

3 Implicit finite difference schemes for the linearized shallow-water equations

3.0 Introduction

In this chapter we will treat several numerical schemes for the numerical approximation of the so-called "frozen coefficient" shallow-water equations (SWE).

These linear equations are given by:

$$u_t + Uu_x + Vu_y + g\,\zeta_x = 0\ , \tag{3.0-1a}$$

$$v_t + Vv_y + Uv_x + g\,\zeta_y = 0\ , \tag{3.0-1b}$$

$$\zeta_t + U\zeta_x + V\zeta_y + Hu_x + Hv_y = 0\ . \tag{3.0-1c}$$

where:

$u\,(x,y,t)$ = depth averaged velocity in x direction,

$v\,(x,y,t)$ = depth averaged velocity in y direction,

$\zeta\,(x,y,t)$ = water elevation above some plane of reference,

H = constant averaged depth,

g = acceleration due to gravity and

U,V are constants such that $U^2 + V^2 < gH$, assuming subcritical flow.

In section 1 we introduce the concept of "grid staggering" for semi-discrete approximations of (3.0-1) without advection terms.

In section 2 we will treat several implicit finite difference schemes for the approximation of (3.0-1) that are well-known from the literature. For some of these schemes the advection terms are approximated differently than the numerical treatments of the other first order derivatives, i.e., these schemes are a composition of several numerical methods. The resulting schemes will be called "composite schemes". Each numerical method of this section will be discussed briefly. Each scheme has disadvantages that hamper their application to practical problems in civil engineering.

In section 3 we propose two implicit composite finite difference schemes for the approximation of (3.0-1) of which the advection operator is approximated by methods proposed in chapter 2. The first scheme is very efficient but

conditionally stable. The second scheme is unconditionally stable but requires more operations per timestep than the first method. The approximations in this chapter are based upon the method proposed in chapter 2.
The stability analysis of these schemes is given in section 4. In section 5 the limitations of ADI schemes with respect to accuracy are discussed.

3.1 Grid staggering

If we neglect the advection terms of (3.0-1), i.e. the terms given by Uu_x, Vu_y, Vv_y, Uv_x, $U\zeta_x$, $V\zeta_y$, we obtain the following equations:

$$u_t + g\zeta_x = 0, \tag{3.1-1a}$$

$$v_t + g\zeta_y = 0, \tag{3.1-1b}$$

$$\zeta_t + Hu_x + Hv_y = 0. \tag{3.1-1c}$$

Suppose (3.1-1) is to be approximated numerically. To this end first the spatial grid of figure (3-1) is defined.

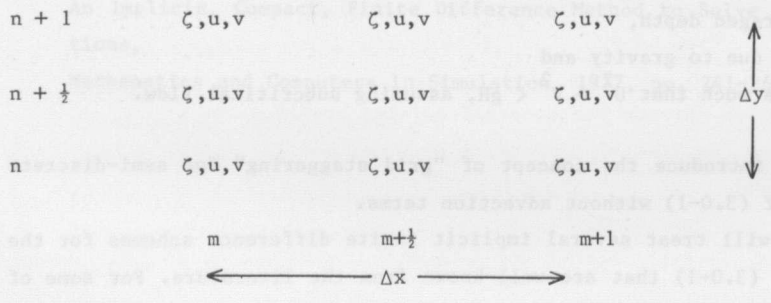


Figure (3-1) Spatial non-staggered grid

In this section we consider only spatial discretizations.
For (3.1-1) a semi-discrete approximation is given by:

$$u_t + g\zeta_{ox} = 0, \text{ at } m,n \tag{3.1-2a}$$

$$v_t + g\zeta_{oy} = 0, \text{ at } m,n \qquad (3.1\text{-}2b)$$

$$\zeta_t + Hu_{ox} + Hv_{oy} = 0, \text{ at } m,n \qquad (3.1\text{-}2c)$$

where $m = \ldots, \frac{1}{2}, 1, 1\frac{1}{2}, 2, \ldots,$

$n = \ldots, \frac{1}{2}, 1, 1\frac{1}{2}, 2, \ldots,$

u_t at m,n denotes $\frac{d}{dt} u_{m,n}(t)$, v_t and ξ_t are defined accordingly

ζ_{ox} at m,n denotes $(\zeta_{m+\frac{1}{2},n} - \zeta_{m-\frac{1}{2},n})/\Delta x$,

u_{ox} and v_{ox} are defined similar to ζ_{ox} ,

ζ_{oy} at m,n denotes $(\zeta_{m,n+\frac{1}{2}} - \zeta_{m,n-\frac{1}{2}})/\Delta y$,

u_{oy} and v_{oy} are defined similar to ζ_{oy} .

If boundary conditions are not taken into account, then, because of the special structure of (3.1-1) combined with the use of central differences in (3.1-2), this latter equation consists of four sets of independent equations. The four grids, each of which relates to one independent set, are given by figure (3-2).

If one chooses just one of the four possible grids, which are all equivalent, the numerical solution that belongs to that grid is just as accurate as the numerical solution that belongs to the original grid. The accuracy is maintained by staggering the grid according to figure (3-2). However, the number of computational values is decreased by a factor of 4. This explains the extensive use of staggered grids for the approximation of the SWE ever since Hansen [9].

n + 1	ζ	u	ζ		v	
n + ½	v		v	u	ζ	u
n	ζ	u	ζ		v	
	m	m+½	m+1	m	m+½	m+1

n + 1	v		v	u	ζ	u
n + ½	ζ	u	ζ		v	
n	v		v	u	ζ	u
	m	m+½	m+1	m	m+½	m+1

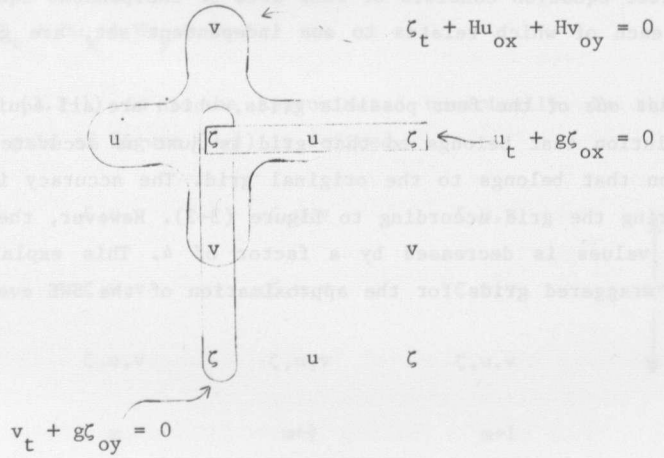


Figure (3-2) Staggered grids

Another advantage of grid staggering is the simplicity of the implementation of boundary conditions; if a "u boundary condition", i.e. at the boundary the velocity in the x direction is given, is located in a u point of figure (3-2), a "v boundary" at a v point or a ζ boundary at a ζ point, then special boundary schemes are not needed. From this it follows that for stability, at least for 1D problems, only the Cauchy problem needs to be investigated, see Kreiss [10]. This means that for a staggered grid a certain class of boundary condi-

tions does not introduce additional stability problems.

The last advantage of staggered grids that we describe in this section is the exclusion of spurious oscillations. Consider a linearized one-dimensional SWE:

$$u_t + g\,\zeta_x = 0, \tag{3.1-3a}$$

$$\zeta_t + H u_x = 0, \tag{3.1-3b}$$

where $0 \leq x \leq 1$ and boundary conditions are given by:

$$\zeta\,(o,t) = e^{st},\ u\,(1,t) = 0 \tag{3.1-3c}$$

The semi-discrete approximation of (3.1-3) similar to (3.1-2) is:

$$u_t + g\,\zeta_{ox} = 0, \text{ at } m,\ m = 0, \tfrac{1}{2}, \ldots, M - \tfrac{1}{2} \tag{3.1-3a}$$

$$\zeta_{-\frac{1}{2}} = 2\zeta_o - \zeta_{\frac{1}{2}}$$

$$\zeta_t + H u_{ox} = 0, \text{ at } m,\ m = \tfrac{1}{2}, 1, \ldots, M \tag{3.1-3b}$$

$$u_{M+\frac{1}{2}} = 2\,u_M - u_{M-\frac{1}{2}}$$

$$\zeta_o(t) = e^{st},\ u_M(t) = 0 \tag{3.1-3c}$$

where $\Delta x = 1/M$

The grid of this scheme is defined by figure (3-3).

u, ζ	u, ζ	u, ζ
m	m+½	m+1

Figure (3-3) One-dimensional grid

In a search for normal modes of the form:

$$[u_m(t), \zeta_m(t)]^T = [\tilde{u}, \tilde{\zeta}]^T e^{st} z^m \tag{3.1-4}$$

we substitute into (3.1-3)

$$[u_m(t), \zeta_m(t)]^T = [\hat{u}_m, \hat{\zeta}_m]^T e^{st} \tag{3.1-5}$$

This yields the following resolvent equation of (3.1-3):

$$s\hat{u} + g\,\hat{\zeta}_{ox} = 0, \text{ at m, } m=0,\tfrac{1}{2},\ldots,M-\tfrac{1}{2}$$

$$\hat{\zeta}_{-\frac{1}{2}} = 2\hat{\zeta}_o - \zeta_{\frac{1}{2}} \tag{3.1-6a}$$

$$s\hat{\zeta} + H\,\hat{u}_{ox} = 0, \text{ at m, } m=\tfrac{1}{2},1,\ldots,M$$

$$\hat{u}_{M+\frac{1}{2}} = 2\hat{u}_M - \hat{u}_{M-\frac{1}{2}} \tag{3.1-6b}$$

$$\hat{\xi}_o = 1, \quad \hat{u}_M = 0. \tag{3.1-6c}$$

The solution of (3.1-6) is given by:

$$\begin{bmatrix} \hat{u}_m \\ \hat{\zeta}_m \end{bmatrix} = \begin{bmatrix} 1 \\ \sqrt{\frac{H}{g}} \end{bmatrix} [\alpha_1 z_1^{2m} + \beta_1 (-z_1)^{2m}] + \begin{bmatrix} 1 \\ -\sqrt{\frac{H}{g}} \end{bmatrix} [\alpha_2 z_2^{2m} + \beta_2 (-z_2)^{2m}] \tag{3.1-7}$$

where $z_{1,2}$ are roots of the characteristic equation of (3.1-6) given by:

$$\begin{vmatrix} \Delta xsz & g(z^2-1) \\ H(z^2-1) & \Delta xsz \end{vmatrix} = 0 \tag{3.1-8}$$

and $\alpha_{1,2}$ and $\beta_{1,2}$ are constants determined by boundary conditions and boundary schemes

From (3.1-7) it follows that if m is an integer, then (3.1-5) can be written as:

$$\begin{bmatrix} u_m(t) \\ \xi_m(t) \end{bmatrix} = e^{st}\left\{\begin{bmatrix} 1 \\ \sqrt{\frac{H}{g}} \end{bmatrix} (\alpha_1 + \beta_1) z_1^{2m} + \begin{bmatrix} 1 \\ -\sqrt{\frac{H}{g}} \end{bmatrix} (\alpha_2 + \beta_2) z_2^{2m}\right\} \tag{3.1-9a}$$

while if m is an odd multiple of $\frac{1}{2}$ then (3.1-5) becomes:

$$\begin{bmatrix} u_m(t) \\ \zeta_m(t) \end{bmatrix} = e^{st} \left\{ \begin{bmatrix} 1 \\ \sqrt{\frac{H}{g}} \end{bmatrix} (\alpha_1 - \beta_1)\, z_1^{2m} + \begin{bmatrix} 1 \\ -\sqrt{\frac{H}{g}} \end{bmatrix} (\alpha_2 - \beta_2)\, z_2^{2m} \right\} \qquad (3.1\text{-}9b)$$

In general $\beta_{1,2} \neq 0$, and we see that a spurious "2Δx wave" will be present in the "non-staggered" grid. The amplitude of this wave depends on the boundary conditions and the value for s, which is determined by the wave periods of the boundary conditions.

Summarizing, for the simple SWE (3.1-1) two independent sets of equations result in the 1-dimensional case and four independent sets in the 2-dimensional case. The differences between the solutions may become apparent in the result of the calculations as spurious "2Δx waves", to an extent determined by the boundary conditions.

If only one of the sets is chosen, these spurious solutions are impossible. This follows immediately from a general solution for a staggered grid which is given by:

$$\begin{bmatrix} u_{m+\frac{1}{2}}(t) \\ \zeta_m(t) \end{bmatrix} = e^{st} \left\{ \begin{bmatrix} \sqrt{z_a} \\ \sqrt{\frac{H}{g}} \end{bmatrix} \gamma_a z_a^{\,m} + \begin{bmatrix} \sqrt{z_b} \\ \sqrt{\frac{H}{g}} \end{bmatrix} \gamma_b\, z_b^{\,m} \right\} \qquad (3.1\text{-}10)$$

where $z_a = z_1^{\,2}$, $z_b = z_2^{\,2}$ and $\gamma_{a,b}$ are constants.

Obviously this general solution does not contain spurious oscillations.

3.2 Review of existing implicit methods

In this section we describe several FDMs for the approximation of SWE that are well-known from the literature. In the original papers these FDMs are described for nonlinear SWE with various dependent variables, for example, $(u, v, \zeta)^T$, $(uh, vh, \zeta)^T$ or $(u, v, 2\sqrt{gh})^T$, where h denotes the total depth of the fluid. To explain the relevant aspects of the methods it is sufficient to describe the form they take when they are applied to (3.0-1). We rewrite (3.0-1) in the following form:

$$\underline{w}_t + U\underline{w}_x + V\underline{w}_y + A\underline{w}_x + B\underline{w}_y = 0 \qquad (3.2\text{-}1)$$

where $\underline{w} = (u, v, \zeta)^T$,

$$A = \begin{bmatrix} 0 & 0 & g \\ 0 & 0 & 0 \\ H & 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & g \\ 0 & H & 0 \end{bmatrix}$$

We will describe the FDMs proposed by Gustafsson [8], Elvius and Sundström [3], Leendertse [14] and Benqué et al. [1].

a. Gustafsson's method

This method is based upon a non-staggered grid as given by figure (3-1). When applied to (3.2-1) this method is a straightforward ADI method consisting of two stages which are given by:

Stage 1:

$$(\underline{w}^{k+\frac{1}{2}} - \underline{w}^{k})/\tfrac{1}{2}\tau + U\underline{w}_{ox}^{k+\frac{1}{2}} + V\underline{w}_{oy}^{k} + A\underline{w}_{ox}^{k+\frac{1}{2}} + B\underline{w}_{oy}^{k} = 0, \text{ at } m,n \qquad (3.2\text{-}2a)$$

Stage 2:

$$(\underline{w}^{k+1} - \underline{w}^{k+\frac{1}{2}})/\tfrac{1}{2}\tau + U\underline{w}_{ox}^{k+\frac{1}{2}} + V\underline{w}_{oy}^{k+1} + A\underline{w}_{ox}^{k+\frac{1}{2}} + B\underline{w}_{oy}^{k+1} = 0, \text{ at } m,n \qquad (3.2\text{-}2b)$$

where $\underline{w}$ at m,n denotes $[u_{m,n}, v_{m,n}, \zeta_{m,n}]^T$,

$\underline{w}_{ox}$ at m,n denotes $(\underline{w}_{m+\frac{1}{2},n} - \underline{w}_{m-\frac{1}{2},n})/\Delta x$,

and $\underline{w}_{oy}$ at m,n denotes $(\underline{w}_{m,n+\frac{1}{2}} - \underline{w}_{m,n-\frac{1}{2}})/\Delta y$.

For nonlinear SWE another method has been proposed by Fairweather and Navon [4]. If this method is applied to (3.2-1) then it also yields (3.2-2).
A similar method for a non-staggered grid has been described by Gerritsen [5].
The methods of Gustafsson [8] and Fairweather and Navon [4] are unconditionally stable when they are applied to (3.2-1).

The FDM (3.2-2) is unconditionally stable but is not very efficient because of the non-staggered grid. The solution contains spurious roots as explained in section 3.1, and the structure of the implicit equations is such that the method for the solution of linear equations should be chosen carefully in

order to prevent the amplification of rounding errors.

If $U^2 + V^2 < gH$ the advection terms given by $U\underline{w}_x$ and $V\underline{w}_y$ need not be approximated by the same method as $A\underline{w}_x$ and $B\underline{w}_y$. Practically the same accuracy can be obtained by a slightly different approximation of the advection operator. In that case a more efficient staggered grid can be applied. If the time-discretization of the advection operator is different from the time-discretization of the other part of the differential operator, the resulting linear equations can be solved very efficiently. An example of such a composite FDM has been proposed by Elvius and Sundström [3].

b. Elvius and Sundström's method

This method is based upon the grid of figure (3-3).

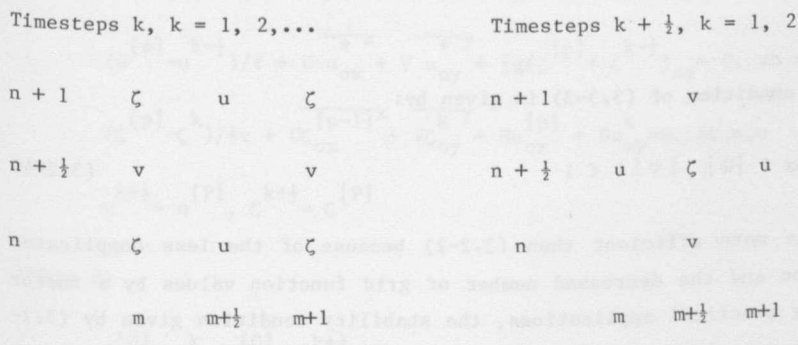


Figure (3-3) Elvius and Sundström grid.

For this grid the number of grid function values is half as much as for the non-staggered grid of figure (3-1).
The Elvius and Sundström method is a combination of the mid-point rule, see Lambert [12], and the trapezoidal rule. If the ADI pertubation of the trapezoidal rule is not taken into account their method is given by:

$$u_{ot} + U\overline{u}^{y}_{ox} + V\overline{u}^{x}_{oy} + g\zeta^{t}_{ox} = 0, \text{ at } m+\tfrac{1}{2}, n, k+\tfrac{1}{2} \text{ and } (m, n+\tfrac{1}{2}, k) \qquad (3.2\text{-}3a)$$

$$v_{ot} + V\overline{v}^{x}_{oy} + U\overline{v}^{y}_{ox} + g\zeta^{t}_{oy} = 0, \text{ at } m, n+\tfrac{1}{2}, k^{+\frac{1}{2}} \text{ and } (m+\tfrac{1}{2}, n, k) \qquad (3.2\text{-}3b)$$

$$\zeta_{ot} + U\overline{\zeta}^{y}_{ox} + V\overline{\zeta}^{x}_{oy} + H\overline{u}^{t}_{ox} + H\overline{v}^{t}_{oy} = 0, \text{ at } m,\ n,\ k \text{ and } m+\tfrac{1}{2},\ n+\tfrac{1}{2},\ k+\tfrac{1}{2} \qquad (3.2\text{-}3c)$$

where: $\overline{u}^{t}$ at m,n,k denotes $\frac{1}{2}(u^{k+\frac{1}{2}}_{m,n} + u^{k-\frac{1}{2}}_{m,n})$

$\overline{v}^{t}$ and $\overline{\zeta}^{t}$ are defined in the same way as $\overline{u}^{t}$

u_{ot} at m,n,k denotes $(u^{k+\frac{1}{2}}_{m,n} - u^{k-\frac{1}{2}}_{m,n})/\tau$

V_{ot} and ζ_{ot} are defined in the same way as u_{ot}

$\overline{u}^{x}$ at m,n,t denotes $\frac{1}{2}(u^{k}_{m+\frac{1}{2},n} + u^{k}_{m-\frac{1}{2},n})$

$\overline{u}^{y}$ at m,n,k denotes $\frac{1}{2}(u^{k}_{m,n+\frac{1}{2}} + u^{k}_{m,n-\frac{1}{2}})$ and

$\overline{v}^{x}$ $\overline{v}^{y}$, $\overline{\zeta}^{x}$ and $\overline{\zeta}^{y}$ are defined in the same way as $\overline{u}^{x}$ and $\overline{u}^{y}$

The stability condition of (3.3-3) is given by:

$$\tau/\Delta x \max (|U|, |V|) < 1 \qquad (3.2\text{-}4)$$

This method is more efficient than (3.2-2) because of the less complicated linear equation and the decreased number of grid function values by a factor of 2; for many practical applications, the stability condition given by (3.2-4) is not too restrictive. Yet a fully staggered grid will allow even more efficient methods. A very well-known example is the following method:

c. The method of Leendertse

The Leendertse [14] method is probably the most widely used method in civil engineering applications, many of which have been reported. See Leendertse et al. [15]. It is part of a computer modelling system described by Leendertse et al. [16].

The method is based upon a fully staggered grid as given by figure (3-4).

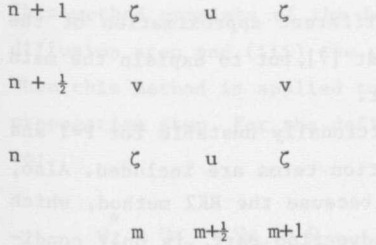


Figure (3-4) Grid of the Leendertse method

Application of this method to (3.2-1) yields:

Stage 1:

$$u^{[0]} = u^{k-\frac{1}{2}},\ \zeta^{[0]} = \zeta^{k};\ \text{for } p = 1, \ldots, P:$$

$$(u^{[p]} - u^{k-\frac{1}{2}})/\tau + U\,\overline{u^{*}}^{x}_{ox} + V\,\overline{u^{*}}^{y}_{oy} + \tfrac{1}{2}g(\zeta^{[p]} + \zeta^{k-\frac{1}{2}})_{ox} = 0,\ \text{at } m+\tfrac{1}{2},n \qquad (3.2\text{-}5a)$$

$$(\zeta^{[p]} - \zeta^{k})/\tfrac{1}{2}\tau + U\overline{\zeta^{[p-1]}}^{x}_{ox} + V\overline{\zeta^{k}}^{y}_{oy} + Hu^{[p]}_{ox} + Hv^{k}_{oy} = 0,\ \text{at } m,n \qquad (3.2\text{-}5b)$$

$$u^{k+\frac{1}{2}} = u^{[P]},\ \zeta^{k+\frac{1}{2}} = \zeta^{[P]}$$

Stage 2:

$$v^{[0]} = v^{k},\ \zeta^{[0]} = \zeta^{k+\frac{1}{2}};\ \text{for } p = 1, \ldots, P:$$

$$(v^{[p]} - v^{k})/\tau + \overline{v^{+}}^{y}_{oy} + \overline{v^{+}}^{x}_{ox} + \tfrac{1}{2}g(\zeta^{[p]} + \zeta^{k})_{ox} = 0,\ \text{at } m,n+\tfrac{1}{2} \qquad (3.2\text{-}5c)$$

$$(\zeta^{[p]} - \zeta^{k+\frac{1}{2}})/\tfrac{1}{2}\tau + U\overline{\zeta^{k+\frac{1}{2}}}^{x}_{ox} + V\overline{\zeta^{[p-1]}}^{y}_{oy} + Hu^{k+\frac{1}{2}}_{ox} + Hv^{[p]}_{oy} = 0,\ \text{at } m,n \qquad (3.2\text{-}5d)$$

$$v^{k+1} = v^{[P]},\ \zeta^{k+1} = \zeta^{[P]}.$$

where $u^{*} = \tfrac{1}{2}(u^{[p-1]} + u^{k-\frac{1}{2}})$ and

$v^{+} = \tfrac{1}{2}(v^{[p-1]} + v^{k})$.

Note that Leendertse [14] reports a slightly different approximation of the advection part based upon the work of Grammeltvedt [7], but to explain the main ideas of Leendertse's method this is not relevant.
Weare [20] has shown that this scheme is unconditionally unstable for P=1 and can therefore be used only when stabilizing friction terms are included. Also, for P=2 this scheme is not likely to be stable because the RK2 method, which is applied for the time discretization of the advection part, is only conditionally stable and its stability region does not contain any part of the imaginary axis, see figure (1-3). The scheme is likely to be stable only for the case without stabilizing friction terms, if it is "corrected to convergence", see Lambert [12], p86. This may entail a large number of iterations, which decreases its effectivity.

Benqué et al. [1] claim that the ADI structure could decrease the accuracy considerably when large timesteps are applied. They propose a completely implicit method based upon operator splitting, of which we will give a very brief review:

d. The operator splitting method of Benqué et al.

The grid employed by this method is given by figure (3-5).

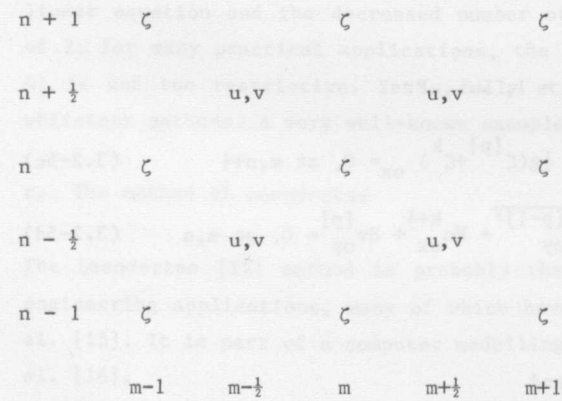


Figure (3-5) Grid for method of Benqué et al. [1]

This method consists of the following steps: (i) the advection step, (ii) the diffusion step and (iii) the propagation step.
When this method is applied to (3.2-1) we only need the advection step and the propagation step. For the definition of these steps (3.2-1) is split according to:

$$w^*_t + Uw_x + Vw_y = 0 \,, \tag{3.2-6a}$$

$$w_t + Aw^*_x + Bw^*_y = 0 \,. \tag{3.2-6b}$$

After this the advection step is performed by approximating (3.2-6a) by a stabilized characteristic interpolation method as described in chapter 2. During this step the grid function values are calculated at the u,v points of figure (3-5).

The propagation step implies the approximation of (3.2-6b) at the ζ points of figure (3-5). If values needed for the evaluation of the FDM are missing, they are calculated by interpolation of grid function values at adjacent grid-points.

By elimination of the unknown values for u and v an implicit set of equations results where the values for ζ are the unknowns. The method is implicit and because an ADI type of method is not used the implicit equations involve all grid points. The equations are solved by an iterative method based upon conjugate gradients.

After the calculation of the ζ values, values for u and v at the "u,v" points are calculated. Again missing values are calculated by interpolation based upon values at adjacent grid points.

Formally the stability of this method is unconditional. However, we believe the advection step imposes a practical timestep limit given by:

$$\tau \ \mathrm{Max} \left(|U| / \Delta x, |V| / \Delta y \right) < 2 \tag{3.2-7}$$

The reason for this timestep limit was discussed in section 2.4. For steady-state problems the timestep has to be chosen much smaller than (3.2-7) allows

in order to avoid τ dependent numerical viscosity, see Roache [19] or section 2.4.

In view of (3.2-7) or the much more severe accuracy limit for steady state problems the method does not seem to be very effective (although excellent results for practical problems were reported by Benqué et al) especially if one takes into account the computational work needed to solve the fully implicit propagation step. An approximation method for SWE completely based upon characteristics has been given by Daubert et al. [2].

Of the methods just considered, the Leendertse method seems to be the most efficient, because of the fully staggered grid and the simple implicit equations. Another advantage is the widespread experience with this method for practical applications. Disadvantages are (i) formal instability because of the approximation of the advection operator, (ii) second order accuracy is obtained only if $P > 2$, (iii) stable results could imply a large number of iterations, which decrease the efficiency.

3.3 On the stabilization of Leendertse's method

In this section we propose two possible stabilizations of the Leendertse method, called M1 and M2. The first method, M1, is based upon the same spatial discretization as has been used for (3.2-5) and is given by:

$$u_t + U\overline{u}^x_{ox} + V\overline{u}^y_{oy} + g\zeta_{ox} = 0, \qquad \text{at } m+\tfrac{1}{2},\ n \qquad (3.3\text{-}1a)$$

$$v_t + V\overline{v}^y_{oy} + U\overline{v}^x_{ox} + g\zeta_{oy} = 0, \qquad \text{at } m,\ n+\tfrac{1}{2} \qquad (3.3\text{-}1b)$$

$$\zeta_t + U\overline{\zeta}^x_{ox} + V\overline{\zeta}^y_{oy} + Hu_{ox} + Hv_{oy} = 0, \qquad \text{at } m,\ n \qquad (3.3\text{-}1c)$$

For the approximation of the advection operator we propose the two stage Angled-Derivative method given by (2.1-7). This yields the following approximation method for (3.0-1):

Method M1:

Stage 1:

$$(u^{k+\frac{1}{2}}-u^{k})/\tfrac{1}{2}\tau+Uu^{k}_{+x}+Vu^{k+\frac{1}{2}}_{-y}+g\zeta^{k+\frac{1}{2}}_{ox}=0,\ \text{at } m+\tfrac{1}{2},n \tag{3.3-2a}$$

$$(v^{k+\frac{1}{2}}-v^{k})/\tfrac{1}{2}\tau+Vv^{k+\frac{1}{2}}_{-y}+Uv^{k}_{+x}+g\zeta^{k}_{oy}=0,\ \text{at } m,n+\tfrac{1}{2} \tag{3.3-2b}$$

$$(\zeta^{k+\frac{1}{2}}-\zeta^{k})/\tfrac{1}{2}\tau+U\zeta^{k}_{+x}+V\zeta^{k+\frac{1}{2}}_{-y}+Hu^{k+\frac{1}{2}}_{ox}+Hv^{k}_{oy}=0,\ \text{at } m,n \tag{3.3-2c}$$

Stage 2:

$$(u^{k+1}-u^{k+\frac{1}{2}})/\tfrac{1}{2}\tau+Uu^{k+1}_{-x}+Vu^{k+\frac{1}{2}}_{+y}+g\zeta^{k+\frac{1}{2}}_{ox}=0,\ \text{at } m+\tfrac{1}{2},n \tag{3.3-2d}$$

$$(v^{k+1}-v^{k+\frac{1}{2}})/\tfrac{1}{2}\tau+Vv^{k+\frac{1}{2}}_{+y}+Uv^{k+1}_{-x}+g\zeta^{k+1}_{oy}=0,\ \text{at } m,n+\tfrac{1}{2} \tag{3.3-2e}$$

$$(\zeta^{k+1}-\zeta^{k+\frac{1}{2}})/\tfrac{1}{2}\tau+U\zeta^{k+1}_{-x}+V\zeta^{k+\frac{1}{2}}_{+y}+Hu^{k+\frac{1}{2}}_{ox}+Hv^{k+1}_{oy}=0,\ \text{at } m,n \tag{3.3-2f}$$

where: u_{+x} at m,n denotes $(u_{m+1,n}-u_{m,n})/\Delta x$,
u_{-x} at m,n denotes $(u_{m,n}-u_{m-1,n})/\Delta x$,
u_{+y} at m,n denotes $(u_{m,n+1}-u_{m,n})/\Delta y$,
u_{-y} at m,n denotes $(u_{m,n}-u_{m,n-1})/\Delta y$,
and v_{+x}, v_{-x}, ζ_{+x}, ζ_{-x}, v_{+y}, v_{-y}, ζ_{+y} and ζ_{-y} are defined accordingly.

The equations (3.3-2a) and (3.3-2c) are coupled implicitly. If the evaluation of the grid function values concerning these equations starts at the row with the lowest number of n, the first stage of the Angled Derivative method is effectively explicit as the calculation proceeds in the increasing n direction.

For the same reason the evaluation of (3.3-2b) has to start at the lowest number for m, of (3.3-2e) and (3.3-2f) at the lowest number for m, and (3.3-2d) at the lowest number for n. If these simple rules are implemented, (3.3-2) constitutes an effectively partially explicit, and partially implicit, hence very efficient FDM. The stability conditions for (3.3-2) are given by:

$$U\tau/\Delta x > -1 \qquad V\tau/\Delta y > -1 \tag{3.3-3}$$

This is due to the CFL condition for the advection operator as has been explained in section 2.2.

Method M1 is second order accurate; the accuracy is comparable to the Elvius and Sundström method and the stability condition (3.3-3) is equivalent to (3.2-4). The efficiency is doubled, however, because (3.3-2) needs half as much grid function values as (3.2-3) and the tri-diagonal structure of the implicit equations is similar for both methods.

Method M1 is an improvement over the Leendertse method given by (3.2-5) as well because iterations of the implicit equations to increase the stability are not necessary.

In section 3.4 it will be argued that the Cauchy-problem for (3.3-2) is unconditionally stable.
As will also be shown in section (3.4) that this method has eigenvalues only on the unit circle. Hence, the method is non-dissipative. This might cause spurious wiggles because of the advection operator approximation. Also (3.3-3) could be too restrictive for several possible applications. The advection FDMs of chapter 2 allow a large number of possible advection operator approximations, however.
The second stabilization of the Leendertse scheme that we propose (method M2) is based upon approximation of the advection operator by the Crank-Nicolson scheme given by (2.1-5) and the dissipative reduced phase error scheme given by (2.1-11). To keep the numerical dissipation as small as possible, only Vu_y and Uv_x will be approximated by this last method, although (2.1-11) introduces only a fourth order dissipative term. The resulting scheme is given by:

Method M2:
Stage 1:

$$(u^{k+\frac{1}{2}} - u^{k})/\tfrac{1}{2}\tau + \overline{Uu^{k}_{ox}}^{x} + S_{oy}(V, u^{k}) + g\zeta^{k+\frac{1}{2}}_{ox} = 0, \text{ at } m+\tfrac{1}{2}, n \qquad (3.3\text{-}4a)$$

$$(v^{k+\frac{1}{2}} - v^{k})/\tfrac{1}{2}\tau + \overline{Vv^{k+\frac{1}{2}}_{oy}}^{y} + S_{+x}(U, v^{k+\frac{1}{2}}) + g\zeta^{k}_{oy} = 0, \text{ at } m, n+\tfrac{1}{2} \qquad (3.3\text{-}4b)$$

$$(\zeta^{k+\frac{1}{2}} - \zeta^{k})/\tfrac{1}{2}\tau + \overline{U\zeta^{k+\frac{1}{2}}_{ox}}^{x} + \overline{V\zeta^{k}_{oy}}^{y} + Hu^{k+\frac{1}{2}}_{ox} + Hv^{k}_{oy} = 0, \text{ at } m, n \qquad (3.3\text{-}4c)$$

Stage 2:

$$(u^{k+1}-u^{k+\frac{1}{2}})/\tfrac{1}{2}\tau+U\overline{u_{ox}^{k+1}}^{x} + S_{+y}(V,u^{k+1}) + g\zeta_{ox}^{k+\frac{1}{2}} = 0, \text{ at } m+\tfrac{1}{2},n \qquad (3.3\text{-}4d)$$

$$(v^{k+1}-v^{k+\frac{1}{2}})/\tfrac{1}{2}\tau + V\overline{v_{oy}^{k+\frac{1}{2}}}^{y} + S_{ox}(U,v^{k+\frac{1}{2}}) + g\zeta_{oy}^{k+1} = 0, \text{ at } m,n+\tfrac{1}{2} \qquad (3.3\text{-}4e)$$

$$(\zeta^{k+1}-\zeta^{k+\frac{1}{2}})/\tfrac{1}{2}\tau + U\overline{\zeta_{oy}^{k+\frac{1}{2}}}^{x} + V\overline{\zeta_{oy}^{k+1}}^{y} + Hu_{ox}^{k+\frac{1}{2}} + Hv_{oy}^{k+1} = 0, \text{ at } m,n \qquad (3.3\text{-}4f)$$

where:

$S_{oy}(V,u)$ at m,n denotes $V(u_{m,n+2}+4u_{m,n+1}-4u_{m,n-1}-u_{m,n-2})/12\Delta y$ and

$$S_{+y}(V,u) \text{ at m,n denotes} \begin{cases} V(3u_{m,n}-4u_{m,n-1}+u_{m,n-2})/2\Delta y \text{ if } V > 0 \\ V(-3u_{m,n}+4u_{m,n+1}-u_{m,n+2})/2\Delta y \text{ if } V \leq 0 \end{cases}$$

The functions $S_{ox}(U,v)$ and $S_{+x}(U,v)$ are defined accordingly.

The implicit equations that result from (3.3-4) are all tri-diagonal if they start at the proper row or column number depending on the sign of U or V.
We will deal with the boundary treatment in chapter 4 where (3.3-4) will be extended to an approximation method of the nonlinear SWE.

3.4 Stability analysis of stabilized versions of Leendertse's method

In this section we study the G-R stability of (3.3-2) and (3.3-4). The stability of the Cauchy problem will be studied, in fact only the Von Neumann condition.
Verification of the G-R condition for initial-boundary value problems is probably not only very complex but also not defined. The theory of Godunov and Ryabenki [6] or of Kreiss et al. [11] has been developed only for problems with one spatial dimension. A recent paper of Michelson [17] is perhaps the first to extend this theory to multidimensional problems.

For implicit methods we usually try to prove unconditional stability. In order to simplify the stability analysis we first give a few definitions and lemmas.

Definition (3.4-1):

If A is a matrix with complex-valued entries a_{ij}, then its adjoint A^H is defined by $a^H_{ij} = \overline{a}_{ji}$, where the overbar denotes the complex conjugate.

Definition (3.4-2):

If A is a matrix which has the property that it commutes with its adjoint, i.e. $A^HA = AA^H$, then A is called a normal matrix.

Definition (3.4-3):

If A is a matrix for which $A^HA = AA^H = I$, where I denotes the identity matrix, then A is called a unitary matrix.

Lemma (3.4-1):

If A is a normal matrix then A^HA^{-1} is a unitary matrix.

proof: $(A^HA^{-1})(A^HA^{-1})^H = A^HA^{-1}A^{-H}A = A^H(A^HA)^{-1}A = A^H(AA^H)^{-1}A$
$= A^HA^{-H}A^{-1}A = I.$

Lemma (3.4-2):

If the matrices A and B are unitary matrices then the matrix AB is also a unitary matrix.

proof: $(AB)(AB)^H = ABB^HA^H = I$

Lemma (3.4-3):

The eigenvalues of unitary matrices are on the unit circle.

proof: $(A\underline{w}, A\underline{w}) = (A^HA\underline{w}, \underline{w}) = (\underline{w}, \underline{w})$ $\forall \underline{w}$, where $\underline{w}$ denotes an arbitrary vector

$\rightarrow |\lambda| = 1$

The lemmas given above are simple and given in almost any introductory monograph on linear algebra.

To study the stability of the Cauchy problem for (3.3-2) we look at solutions of (3.3-2) which have the form:

$$[\tilde{u}^k_{m,n}, \tilde{v}^k_{m,n}, \tilde{\zeta}^k_{m,n}]^T = [\hat{u}^k, \hat{v}^k, \hat{\zeta}^k]^T e^{i(\sigma_1 m\Delta x + \sigma_2 n\Delta y)} \qquad (3.4\text{-}1)$$

where σ_1, σ_2 are real numbers.

Substitution of (3.4-1) into (3.3-2) leads to:

$$\underline{\hat{w}}^{k+1} = A^{-1} B C^{-1} D \hat{w}^k = G\hat{w}^k \qquad (3.4\text{-}2)$$

where $\underline{\hat{w}} = [\hat{u}, \hat{v}, \hat{\zeta}]^T$ and $G = A^{-1} B C^{-1} D$ is the well known "amplification matrix", see Richtmyer and Morton [18].

The matrices A, B, C and D are given by:

$$A = \begin{bmatrix} 1+a & 0 & 0 \\ 0 & 1+\hat{a} & g\frac{\tau}{2}\hat{D}_{oy} \\ 0 & H\frac{\tau}{2}\hat{D}_{oy} & 1+\hat{a} \end{bmatrix}, \quad B = \begin{bmatrix} \bar{\hat{a}}+b & 0 & -g\frac{\tau}{2}\hat{D}_{ox} \\ 0 & \bar{\hat{a}}+b & 0 \\ -H\frac{\tau}{2}\hat{D}_{ox} & 0 & \bar{\hat{a}}+b \end{bmatrix}$$

$$C = \begin{bmatrix} 1+\hat{b} & 0 & g\frac{\tau}{2}\hat{D}_{ox} \\ 0 & 1+\hat{b} & 0 \\ H\frac{\tau}{2}\hat{D}_{ox} & 0 & 1+\hat{b} \end{bmatrix}, \quad D = \begin{bmatrix} \bar{\hat{a}}+a & 0 & 0 \\ 0 & \bar{\hat{a}}+a & -g\frac{\tau}{2}\hat{D}_{oy} \\ 0 & -H\frac{\tau}{2}\hat{D}_{oy} & \bar{\hat{a}}+a \end{bmatrix}$$

where: $\hat{a} = \frac{\tau U}{2\Delta x}\,[1-\cos(\sigma_1 \Delta x) + i\sin(\sigma_1 \Delta x)]$,

$\hat{b} = \frac{\tau}{2\Delta y} V\,[1 - \cos(\sigma_2 \Delta y) + i\sin(\sigma_2 \Delta y)]$,

$\hat{D}_{ox} = i\sin(\sigma_1 \tfrac{1}{2}\Delta x) / (\tfrac{1}{2}\Delta x)$,

$\hat{D}_{oy} = i\sin(\sigma_2 \tfrac{1}{2}\Delta y) / (\tfrac{1}{2}\Delta y)$ and

$\bar{\hat{a}}$, $\bar{\hat{b}}$ denote the complex conjugates of $\hat{a}$ and $\hat{b}$.

Stability of the Cauchy problem is ensured if $\| G^k \|$ is bounded $\forall$ k.
To verify this we write G in the following form:

$$G = \Lambda\Lambda^{-1}A^{-1}\ \Lambda\Lambda^{-1}\ B\ \Lambda\Lambda^{-1}\ C^{-1}\Lambda\Lambda^{-1}\ D\Lambda\Lambda^{-1} =$$

$$= \Lambda(\Lambda^{-1}A\Lambda)^{-1}\ (\Lambda^{-1}B\Lambda)\ (\Lambda^{-1}C\Lambda)^{-1}\ (\Lambda^{-1}D\Lambda)\Lambda^{-1} =$$

$$= \Lambda A'^{-1}\ B'\ C'^{-1}\ D'\Lambda^{-1} \qquad (3.4\text{-}3)$$

where $\Lambda = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sqrt{\frac{H}{g}} \end{bmatrix}$ and consequently:

$$A' = \begin{bmatrix} 1+\hat{a} & 0 & 0 \\ 0 & 1+\hat{a} & \frac{\tau}{2}\sqrt{gH}\,\hat{D}_{oy} \\ 0 & \frac{\tau}{2}\sqrt{gH}\,\hat{D}_{oy} & 1+\hat{a} \end{bmatrix},\quad C' = \begin{bmatrix} 1+\hat{b} & 0 & \frac{\tau}{2}\sqrt{gH}\,\hat{D}_{ox} \\ 0 & 1+\hat{b} & 0 \\ \frac{\tau}{2}\sqrt{gH}\,\hat{D}_{ox} & 0 & 1+\hat{b} \end{bmatrix},$$

$D' = A'^{H}$ and $B' = C'^{H}$

Hence G can be written as:

$$G = \Lambda A'^{-1}\ C'^{H}\ C'^{-1}\ A'^{H}\Lambda^{-1} \qquad (3.4\text{-}4)$$

Equation (3.4-4) implies the following relation:

$$\| G^k \| \leq \|\Lambda\ A'^{-1}\|\ \|C'^{H}\ C'^{-1}\|^{k}\ \|A'^{H}\ A'^{-1}\|^{k-1}\ \|A'^{H}\ \Lambda^{-1}\| \qquad (3.4\text{-}5)$$

It is easy to see that A' and C' are normal, hence $C'^{H}\ C'^{-1}$ and $A'^{H}\ A'^{-1}$ are unitary.
From this and from (3.4-5) it follows that:

$$\| G^k \| \leq \|\Lambda\ A'^{-1}\|\,\|A'^{H}\ \Lambda^{-1}\|,\ \forall\ k \qquad (3.4\text{-}6)$$

which proves the stability of the Cauchy problem for (3.3-2).

The eigenvalues of G are on the unit circle, as can be seen as follows:

$$G' = A' \Lambda^{-1} G \Lambda A'^{-1} = C'^{H} C'^{-1} A'^{H} A'^{-1} \qquad (3.4\text{-}7)$$

Because $A' \Lambda^{-1}$ is a similarity transformation it follows that G' and G have the same eigenvalues. G' is a product of two unitary matrices which means, according to lemma (3.4-2), that G' is also a unitary matrix. According to lemma (3.4-3) G' has eigenvalues on the unit circle which means that the eigenvalues of G are also on the unit circle.

The amplification matrix of (3.3-4) can be written in the form of (3.4-3) with:

$$A' = \begin{bmatrix} a' & 0 & 0 \\ 0 & 1 & \frac{\tau}{2}\sqrt{gH}\,\hat{D}_{oy} \\ 0 & \frac{\tau}{2}\sqrt{gH}\,\hat{D}_{oy} & 1+\frac{\tau}{2}V\,\hat{D}_{1y} \end{bmatrix}, \quad B' = \begin{bmatrix} 1 & 0 & -\frac{\tau}{2}\sqrt{gH}\,\hat{D}_{ox} \\ 0 & b' & 0 \\ -\frac{\tau}{2}\sqrt{gH}\,\hat{D}_{ox} & 0 & 1-\frac{\tau}{2}U\,\hat{D}_{1x} \end{bmatrix},$$

$$C' = \begin{bmatrix} 1 & 0 & \frac{\tau}{2}\sqrt{gH}\,\hat{D}_{ox} \\ 0 & c' & 0 \\ \frac{\tau}{2}\sqrt{gH}\,\hat{D}_{ox} & 0 & 1+\frac{\tau}{2}U\,\hat{D}_{1x} \end{bmatrix}, \quad D' = \begin{bmatrix} d' & 0 & 0 \\ 0 & 1 & -\frac{\tau}{2}\sqrt{gH}\,\hat{D}_{oy} \\ 0 & -\frac{\tau}{2}\sqrt{gH}\,\hat{D}_{oy} & 1-\frac{\tau}{2}V\,\hat{D}_{1y} \end{bmatrix}.$$

where $\hat{D}_{1x} = i\sin(\sigma_1 \Delta x)/\Delta x$,

$\hat{D}_{1y} = i\sin(\sigma_2 \Delta y)/\Delta y$,

$a' = 1 + \frac{\tau}{2}U\,\hat{D}_{1x} + \frac{\tau}{2}V\,\hat{S}_{+y}$

$b' = 1 - \frac{\tau}{2}V\,\hat{D}_{1y} - \frac{\tau}{2}U\,\hat{S}_{ox}$,

$c' = 1 + \frac{\tau}{2}V\,\hat{D}_{1y} + \frac{\tau}{2}U\,\hat{S}_{+x}$,

$d' = 1 - \frac{\tau}{2}U\,\hat{D}_{1x} - \frac{\tau}{2}V\,\hat{S}_{oy}$,

$$\hat{S}_{+x} = [(1 - \cos \sigma_1 \Delta x)^2 + i \sin \sigma_1 \Delta x \, (2 - \cos \sigma_1 \Delta x)]/\Delta x,$$

$$\hat{S}_{ox} = i \sin \sigma_1 \Delta x \, (2 + \cos \sigma_1 \Delta x) \, / \, 3\Delta x,$$

$$\hat{S}_{+y} = [(1 - \cos \sigma_2 \Delta y)^2 + i \sin \sigma_2 \Delta y \, (2 - \cos \sigma_2 \Delta y)]/\Delta y \text{ and}$$

$$\hat{S}_{oy} = i \sin \sigma_2 \Delta y \, (2 + \cos \sigma_2 \Delta y)/3\Delta y.$$

Similar to (3.4-5) the following relation holds:

$$\| G^k \| \leq \| \Lambda \, A'^{-1} \| \, \| B' \, C'^{-1} \|^k \| D' \, A'^{-1} \|^{k-1} \| D \, \Lambda^{-1} \| \qquad (3.4\text{-}8)$$

Hence, for stability of (3.3-4) it is sufficient that
$\| B' \, C'^{-1} \| \leq 1$ and $\| D' \, A'^{-1} \| \leq 1$.
To prove this we write D' and A' in the following partitioned form:

$$A' = \begin{bmatrix} a' & 0 \\ 0 & A'_s \end{bmatrix}, \quad D' = \begin{bmatrix} d' & 0 \\ 0 & D'_s \end{bmatrix} \qquad (3.4\text{-}9)$$

where:

$$A'_s = \begin{bmatrix} 1 & \frac{\tau}{2} \sqrt{gH} \, \hat{D}_{oy} \\ \frac{\tau}{2} \sqrt{gH} \, \hat{D}_{oy} & 1 + \frac{\tau}{2} V \hat{D}_{1y} \end{bmatrix}, \quad D'_s = \begin{bmatrix} 1 & -\frac{\tau}{2} \sqrt{gH} \, \hat{D}_{oy} \\ -\frac{\tau}{2} \sqrt{gH} \, \hat{D}_{oy} & 1 - \frac{\tau}{2} V \hat{D}_{1y} \end{bmatrix}$$

It follows that $D'_s = A'^H_s$ or:

$$D' \, A'^{-1} = \begin{bmatrix} d'/a' & 0 \\ 0 & A'^H_s \, A'^{-1}_s \end{bmatrix}$$

With lemma 3.4-1 it follows that

$$\| D' A'^{-1} \| \leq \text{Max} (|d'/a'|, 1) \qquad (3.4\text{-}10)$$

Because $|d'/a'| \leq 1$, as can easily be verified, it follows that $\| D'A'^{-1} \| \leq 1$.

Similarly one can prove that $\| B' C'^{-1} \| \leq 1$, which completes the proof of the stability of (3.3-4) for the Cauchy problem.

3.5 An aspect of the accuracy of ADI schemes for shallow water equations

If advection terms are omitted then every scheme described in this chapter, except (3.2-6), which concerns the method of Benqué et al [1], is an ADI pertubation of the Crank-Nicolson scheme. This means that the wave propagation properties are given by figure (2-1) for $Cf \leq 4$, and by figure (3-6) for $Cf \leq 20$. If these figures are applied to SWE the Courant number Cf is defined by:

$$Cf = \tau \sqrt{gH} / \Delta x \qquad (3.5\text{-}1)$$

Note that for 2-D problems the Courant number is defined as

$$Cf = \tau \sqrt{gH} \left(\frac{1}{\Delta x^2} + \frac{1}{\Delta y^2} \right)^{\frac{1}{2}}.$$

For the calculation of the wave propagation properties a uniform depth and an infinite spatial domain in all directions has been assumed. For practical applications this is never the case. Benqué et al [1] show that ADI schemes for geometries with a non-uniform depth badly represent the flow patterns for very large timesteps. (Cf = 96, see Benqué et al. [1]). We will now give an explanation of this phenomenon for complicated geometries.

Consider for example the geometry of figure (3-7). Suppose that a large Courant number is used, such that from an analytical point of view point P should be contained within the region of influence of Q and vice versa within one timestep. The analytical regions of influence are the characteristic cones of P and Q. The numerical region of influence of point P during one complete timestep, however, is the shaded area of figure (3-7), which does not contain Q. Although this does not cause instabilities, inaccuracies are to be expected.

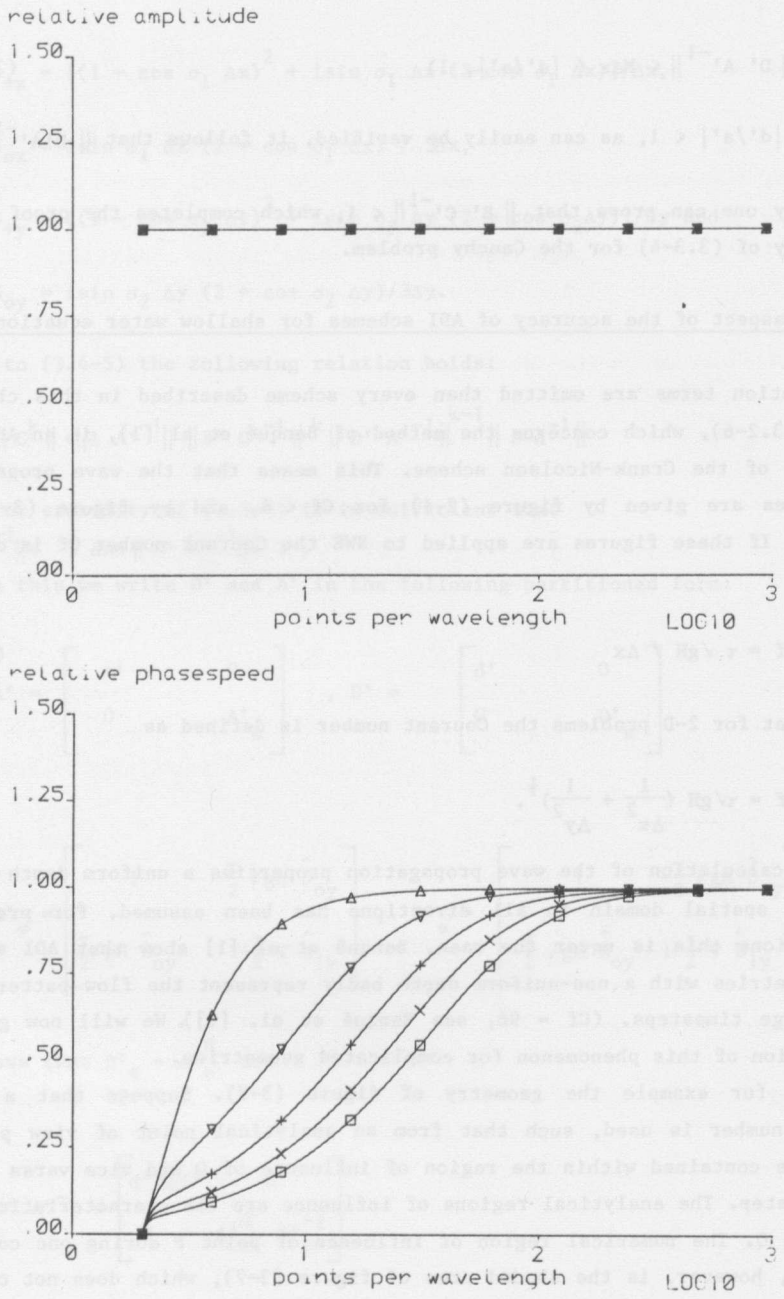


Figure (3-6) Propagation properties of the Crank-Nicolson scheme

Δ: Cf = 0.1 X: Cf = 15.0

∇: Cf = 5.0 □: Cf = 20.0

+: Cf = 10.0

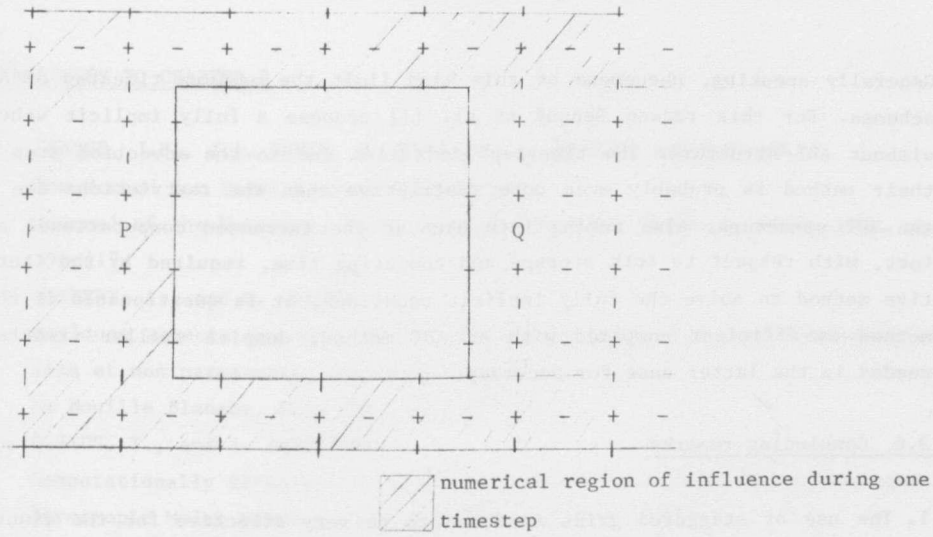


Figure (3-7) Example of complicated geometry.

To increase the accuracy the timestep has to be chosen such that with one timestep, also analytically there is no influence from P onto Q and vice versa. Another example of this possible inaccuracy is the "zig-zag" channel of figure (3-8).

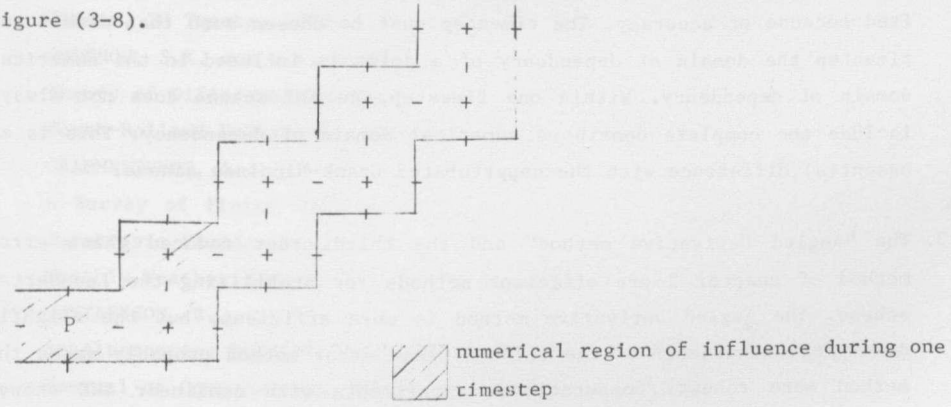


Figure (3-8) "zig-zag" channel

The shaded area of figure (3-8) involves only two grid spacings. This means that for Courant numbers larger than two the numerical solution must be inaccurate for this example.

We believe that the inaccuracies as observed by Benqué et al. [1] are for a similar reason.

Generally speaking, phenomena of this kind limit the maximum timestep of ADI schemes. For this reason Benqué et al. [1] propose a fully implicit scheme without ADI structure. The timestep limitation due to the advection step of their method is probably even more restrictive than the restrictions due to the ADI structure. Also taking into account the increased computational effort, with respect to both storage and computing time, required by the iterative method to solve the fully implicit equations, it is questionable if this method is efficient compared with an ADI method, despite smaller timesteps needed in the latter case for accuracy.

3.6 Concluding remarks

1. The use of staggered grids is found to be very effective for the discretization of the SWE. If advective terms are omitted then the accuracy is the same as for a fully non-staggered grid; the number of grid points however has been reduced by a factor of 4. Moreover, the use of staggered grids reduces the possibility of spurious "$2\Delta x$ waves".

2. For complicated geometries the maximum timestep for an ADI scheme is limited because of accuracy. The timestep must be chosen such that within one timestep the domain of dependency of a point is included in the numerical domain of dependency. Within one timestep, an ADI scheme does not always include the complete domain as numerical domain of dependency. This is an essential difference with the unpertubated Crank-Nicolson scheme.

3. The "Angled Derivative method" and the third order reduced phase error method of chapter 2 are efficient methods for stabilizing the Leendertse scheme. The Angled Derivative method is more efficient, but the slightly dissipative character of the reduced phase error method propably makes the method more robust. Computational experiments with nonlinear SWE showed that the latter method increases the computational overhead less than 6 percent compared with the unconditionally stable Angled Derivative method. In fact with the methods described in chapter 2 it is possible to stabilize the Leendertse method in many ways. Which possibility is most satisfactory for practical problems remains a question that can be answered only by practical experience.

REFERENCES TO CHAPTER 3

1. BENQUE, J.P., J.A. CUNGE, J. FEUILLET, A. HAUGUEL and F.M. HOLLY,
New Method for Tidal Current Computation,
Journal of the Waterway, Port, Coastal and Ocean Division, ASCE, 1982, pp. 396-417.

2. DAUBERT, A., and O. GRAFFE,
Quelques aspects des écoulements presque horizontaux à deux dimensions en plan et non permanents. Applications aux estuaires.
La Houille Blanche, V22, 1967, pp. 847-860.

3. ELVIUS, T., and A. SUNDSTRÖM,
Computationally Efficient Schemes and Boundary Conditions for a Fine-mesh Barotropic Model based on the Shallow-Water Equations,
Tellus XXV, 25, 1973, pp. 132-156.

4. FAIRWEATHER, G. and I.M. NAVON,
A linear ADI Method for the Shallow-Water Equations,
Journal of Computational Physics, 37, 1980, pp. 1-18.

5. GERRITSEN, H.,
Accurate Boundary Treatment in Shallow Water Flow Computations,
Thesis, TH Twente, 1982.

6. GODUNOV, S.K. and V.S. RYABENKI,
Theory of Difference Schemes,
North-Holland Publishing Company, Amsterdam, 1964.

7. GRAMMELTVEDT, A.,
A Survey of Finite Difference Schemes for the primitive Equations for a Barotropic Fluid,
Monthly Weather Review, 97, 1969, pp. 384-404.

8. GUSTAFSSON, B.,
An Alternating Implicit Method for Solving the Shallow Water Equations,
Journal of Computational Physics, 7, 1971, pp. 239-254.

9. HANSEN, W.,
Theorie zur Errechnung des Wasserstandes und der Strömungen in Randmeeren nebst Anwendungen,
Tellus, 8, 1956, pp. 289-300.

10. KREISS, H.O.,
Difference Approximations for the Initial Boundary Value Problem for Hyperbolic Differential Equations,
Proc. Adv. Symp. Madison, Wis, 1966, Wiley New Yrok, 1966, pp.141-166.

REFERENCES (continued)

11. KREISS, H.O., B. GUSTAFSSON and A. SUNDSTROM,
Stability Theory of Difference Approximations for Mixed Initial Boundary Value Problems II,
Mathematics of Computation, 26, 1972, pp. 649-686.

12. LAMBERT, J.D.,
Computational Methods in Ordinary Differential Equations,
Wiley, London-New-York, 1973.

13. LEENDERTSE, J.J.,
Aspects of Computational Model for Long-Period Water-Wave Propagation,
Rand Corporation, Memorandum RM-5294-PR, Santa Monica, 1967.

14. LEENDERTSE, J.J.,
A Water-Quality Simulation Model for Well-Mixed Estuaries and Coastal Seas: Volume I, Principles of Computation,
Rand Corporation, Memorandum RM-6230-RC, Santa Monica, 1970.

15. LEENDERTSE, J.J., A. LANGERAK and M.A.M. DE RAS,
Adjustment and Verification of the Rand Delta II model,
Rand Corporation, P-6247, Santa Monica, 1978.

16. LEENDERTSE, J.J., C.N. JOHNSON, M.C. FUTSISAKI and A.I. NELSON,
Notebook,
Rand Corporation, Santa Monica, 1981.

17. MICHELSON, D.,
Stability Theory of Difference Approximations for Multidimensional Initial-Boundary Value Problems,
Mathematics of Computation, 40, 1983, pp. 1-45.

18. RICHTMYER, R.D. and K.W. MORTON,
Difference Methods for Initial Value Problems,
Interscience Publishers, New York, NY, 1967.

19. ROACHE, P.J.,
Computational Fluid Dynamics,
Hermosa Publishers, Albuquerque, 1972.

20. WEARE, T.J.,
Instability in Tidal Flow Computational Schemes,
Journal of the Hydraulics Division, ASCE, 102, 1976, pp. 569-580.

4 A finite difference method for nonlinear shallow water equations

4.0 Introduction

In this chapter we describe a nonlinear extension of the linear FDMs of the preceding chapters to the nonlinear SWE given by:

$$u_t + uu_x + vu_y - fv + g\zeta_x + gu\,(u^2+v^2)^{\frac{1}{2}}/(C^2H) - \nu(u_{xx}+u_{yy}) = F^{(x)} \qquad (4.0\text{-}1a)$$

$$v_t + vv_y + uv_x + fu + g\zeta_y + gv\,(u^2+v^2)^{\frac{1}{2}}/(C^2H) - \nu\,(v_{xx}+v_{yy}) = F^{(y)} \qquad (4.0\text{-}1b)$$

$$\zeta_t + (Hu)_x + (Hv)_y = 0 \qquad (4.0\text{-}1c)$$

where: u = velocity in x direction,

v = velocity in y direction,

ζ = waterelevation above some plane of reference,

h = waterdepth below some plane of reference,

$H = h + \zeta$ = total waterdepth,

f = coriolis parameter,

g = acceleration due to gravity,

C = Chezy coefficient for bottom roughness,

$F^{(x\ y)}$ = external forcing functions of windstress or barometric pressure

and ν = viscosity coefficient.

In order to make a choice for a FDM, we adopt the following criteria:

1. The numerical solution should be sufficiently accurate. Hence, the method should be consistent to a sufficient order and stable. According to practical experience second order accuracy is satisfactory. It is also necessary that the numerical solution is not greatly influenced by spurious solutions and rounding errors.

2. The method should be robust. In our case this means that the method should be applicable to a wide range of practical 2-D flow problems in civil engineering such as tidal problems in coastal seas and estuaries with tidal flats, model problems in tidal flumes, or steady state problems in rivers.

3. The method should be computationally efficient. Efficiency should not be obtained at the cost of robustness, so robustness has a higher priority.

4. The numerical treatment of the boundary conditions should be such that the overall accuracy and efficiency are not greatly decreased.

Robustness excludes the use of explicit methods. The FDM given by (3.3-2) has moderate stability conditions and is very efficient. Nevertheless we choose a nonlinear extension of (3.3-4) for the numerical approximation of (4.0-1), even though a nonlinear extension of (3.3-2) would be twice as efficient per timestep. This is because of the robustness of (3.3-4), which has been demonstrated by extensive numerical testing.

In the first section we discuss a few general aspects of nonlinear extensions of linear FDMs. The main ideas will be illustrated by means of a simple nonlinear equation.
In section 2 we propose a FDM for the approximation of (4.0-1). This choice is based partly on the results of the investigations described in the preceding chapters and partly on extensive testing with practical problems. The approximation of each term of (4.0-1) is discussed separately. For brevity we describe only the results of these tests.
Sections 3 and 4 deal with the numerical approximations of (4.0-1) near the boundaries. The boundary approximations are based upon a heuristic principle described in section 3. This section also describes the boundary treatment near closed boundaries. Section 4 is devoted to open boundaries.
In section 5 the numerical treatment of tidal flats is discussed. All implicit equations are tri-diagonal, as will be shown in section 6, and can be solved by a simple recursive algorithm. In section 6 this well-known algorithm will be described briefly. It is verified that the structures of the implicit equations are such that rounding errors remain bounded.

4.1 On nonlinear extensions of linear finite difference methods

First we will discuss some aspects of nonlinear extensions of linear FDMs by means of the inviscid Burgers' equation:

$$u_t + (\tfrac{1}{2} u^2)_x = 0, \quad 0 \leq x \leq 1, \quad t > 0 \tag{4.1-1}$$

This equation is "conservative". This means that, for homogeneous boundary conditions at x=0 and x=1, the following relation holds:

$$\frac{d}{dt} \|u\| = 0 \tag{4.1-2}$$

where

$$\|u\|^2 = \int_0^1 u^2 \, dx$$

Consider a semi-discrete numerical approximation of (4.1-1) given by:

$$(u_m)_t + u_m (u_{m+1} - u_{m-1})/2\Delta x = 0, \quad m=1, \ldots, M-1 \tag{4.1-3}$$

This equation can be considered as a nonlinear extension of the following linear finite difference scheme:

$$(u_m)_t + U (u_{m+1} - u_{m-1})/2\Delta x = 0 \tag{4.1-4}$$

where U is a constant.

Equation (4.1-4) is the so-called "frozen coefficient" equation associated with (4.1-3), which is obtained by assuming that the coefficient u_m of (4.1-3) is a constant, $u_m = U$.

The FDM (4.1-3) is by no means the only possible nonlinear extension of (4.1-4). Consider for example the following semi-discrete FDM:

$$(u_m)_t + \frac{1}{3} (u_{m-1} + u_m + u_{m+1}) (u_{m+1} - u_{m-1})/2\Delta x = 0 \tag{4.1-5}$$

If we "freeze" the coefficient $\frac{1}{3} (u_{m-1} + u_m + u_{m+1})$ then we also obtain (4.1-4). This means that the linear properties of (4.1-3) and (4.1-5) are the same. The nonlinear properties, however, are different. The FDM (4.1-5) is conservative, cf. Kreiss and Oliger [14] p. 62. This means that for (4.1-5) the following relation holds for homogeneous boundary conditions:

$$\frac{d}{dt} \|u\|_{\Delta x}^2 = 0 \tag{4.1-6}$$

where $\|u\|_{\Delta x} = \{\sum_{o}^{M-1} u_m^2\}^{\frac{1}{2}} \Delta x$

To obtain a numerical solution, also a discretization in time will have to be defined. For (4.1-3) we consider the following time-discretization:

$$(u^{k+1} - u^k)/\tau + \tfrac{1}{2}u^k u_{ox}^{k+1} + \tfrac{1}{2}u^{k+1} u_{ox}^k = 0, \text{ at } m \qquad (4.1\text{-}7)$$

where u_{ox} is defined as for (3.2-2).
This FDM is locally linear and consequently the solution can be obtained without the application of an iterative method to solve nonlinear equations.

For (4.1-5) we consider the following conservative discretization in time:

$$(u^{k+1} - u^k)/\tau + \frac{1}{6}\{[(u^{k+1})^2]_{ox} + u^{k+1}(u^{k+1})_{ox} + [(u^k)^2]_{ox} + u^k(u^k)_{ox}\} = 0, \text{ at } m \qquad (4.1\text{-}8)$$

Both (4.1-7) and (4.1-8) are nonlinear extensions of the same "frozen coefficient" equation given by:

$$(u_m^{k+1} - u_m^k)/\tau + \tfrac{1}{2} U(u_m^{k+1})_{ox} + \tfrac{1}{2} U (u_m^k)_{ox} = 0 \qquad (4.1\text{-}9)$$

The stability of (4.1-9) is sufficient for the convergence of (4.1-7) and (4.1-8) if the solution of (4.1-1) is sufficiently smooth, cf. Richtmyer and Morton [20], p. 127. In general, however, the stability of the frozen coefficient equation is only a necessary ondition, see Oliger and Sundström [19]. The stability of (4.1-8) can be proven by the energy method. This means that for homogeneous boundary conditions, cf. Richtmyer and Morton [20] p. 142 or Kreiss and Oliger [14] p. 62, the following relation can be proven:

$$\|u^k\|_{\Delta x} = \|u^o\|_{\Delta x} \quad \forall \; k \qquad (4.1\text{-}10)$$

where $\|u^k\|_{\Delta x} = [\sum_m (u_m^k)^2]^{\frac{1}{2}} \Delta x$ and u^o is the initial value of u.

Because of (4.1-10), (4.1-8) seems a safer approximation of (4.1-1) than (4.1-7). The solution of (4.1-8) requires an iterative algorithm, while (4.1-7) can be solved directly. Moreover, the solution of (4.1-8) is not necessarily a more

accurate approximation of (4.1-1) than the solution of (4.1-7) despite the conservation property (4.1-6). To illustrate this point we consider the initial boundary value problem given by (4.1-1) with initial and boundary condition given by:

$$u(x,0) = x,\ u(0,t) = 0,\ t \geq 0 \tag{4.1-11}$$

For this case only one boundary condition at inflow is allowed.
The exact solution is given by:

$$u(x,t) = x/(1+t) \tag{4.1-12}$$

For the numerical approximation of (4.1-1) and (4.1-11) we define a grid with grid points $(k\tau, m\Delta x)$, $k=0,\ldots K$, $m=0,\ldots, M\Delta x = 1/M$.

We consider the numerical approximations (4.1-7) and (4.1-8). For both schemes the initial and boundary conditions are given by:

$$u_m^o = m\Delta x,\ m = 1, \ldots, M,\ u_o^k = 0,\ k = 0, \ldots, k \tag{4.1-13}$$

In order to apply (4.1-7) and (4.1-8) at m=M we define a virtual grid function value by:

$$u_{M+1}^k = 2\, u_M^k - u_{M-1}^k \tag{4.1-14}$$

The order of accuracy of both (4.1-7) and (4.1-8) is determined by substitution into these equations of:

$$u(m\Delta x,\ k\tau) = m\Delta x/(1+k\tau) \tag{4.1-15}$$

From this substitution it follows that (4.1-7) represents the exact solution of (4.1-1) and (4.1-11) without any error, while (4.1-8) is a second order accurate approximation of (4.1-1) and (4.1-11) clearly illustrating that conservative FDMs do not always yield more accurate approximations than non-conservative FDMs, especially if the boundary and initial value conditions are such that the conservation property does not hold.

A practical disadvantage of (4.1-8), despite its conservation property, is that there is no guarantee that the nonlinear equations have real solutions. We will illustrate this point with a simple example. Suppose that for (4.1-13) M is chosen as: M=1. Then by substitution of (4.1-14) into (4.1-8) we obtain, for k=1, the following equation:

$$(u_1^1)^2 + \frac{2}{\tau} u_1^1 + 1 - \frac{2}{\tau} = 0 \qquad (4.1\text{-}16)$$

This simple quadratic equation has real solutions only if the following relation:

$$\tau \leq 1 + \sqrt{2} \qquad (4.1\text{-}17)$$

is satisfied. This means that for $\tau > 1 + \sqrt{2}$ any iterative procedure to solve (4.1-16) that does not account for imaginary solutions does not converge.
It is obvious that (4.1-7) always yields real solutions for any value of τ for this simple example.

Which method is most satisfactory for "real life" applications remains a question that can be answered only by practical experience.

4.2 The finite difference method at the inner points

For the linear FDMs described in chapter 3 the criteria formulated in the introduction of this chapter are satisfied as much as possible by (3.3-4) because of: (i) complete grid staggering for optimal efficiency and minimization of wiggles, (ii) unconditionally linear stabilities for robustness, (iii) second order accuracy, (iv) dissipativity to increase robustness; this dissipativity, however, is small and does not decrease the accuracy.
Therefore, in this section we propose a nonlinear extension of (3.3-4) for the approximation of (4.0-1).
We will only deal with the numerical treatment of the inner points. The treatment near or at the boundary is described in the following sections.
The method of this section has the following general structure:

Stage 1 (S1):

$$\underline{w}^{[0]} = \underline{w}^{k},$$

$$L_1 (\underline{w}^{[p]}, \underline{w}^{[p-1]}, \underline{w}^{k}) = \underline{F}_1, \quad p = 1, 2, \ldots, P, \tag{4.2-1a}$$

$$\underline{w}^{k+\frac{1}{2}} = \underline{w}^{[P]}$$

Stage 2 (S2):

$$\underline{w}^{[0]} = \underline{w}^{k+\frac{1}{2}},$$

$$L_2 (\underline{w}^{[p]}, \underline{w}^{[p-1]}, \underline{w}^{k+\frac{1}{2}}) = \underline{F}_2, \quad p = 1, 2, \ldots, P \tag{4.2-1b}$$

$$\underline{w}^{k+1} = \underline{w}^{[P]}$$

where $\underline{w}$ denotes a vector of grid functions, $L_{1,2}$ finite difference operators and $F_{1,2}$ vector functions.

We will describe each term of our FDM separately and we will elucidate each approximation. For convenience we repeat (4.0-1), omitting the nonhomogeneous part:

$$u_t + uu_x + vu_y - fv + g\zeta_x + gu\,(u^2 + v^2)^{\frac{1}{2}}/(C^2H) - \nu(u_{xx} + u_{yy}) = 0, \tag{4.2-2a}$$

$$v_t + vv_y + uv_x + fu + g\zeta_x + gv\,(u^2 + v^2)^{\frac{1}{2}}/(C^2H) - \nu(v_{xx} + v_{yy}) = 0, \tag{4.2-2b}$$

$$\zeta_t + (Hu)_x + (Hv)_y = 0. \tag{4.2-2c}$$

The staggered grid that is used is defined by figure (4-1) or figure (3-4).

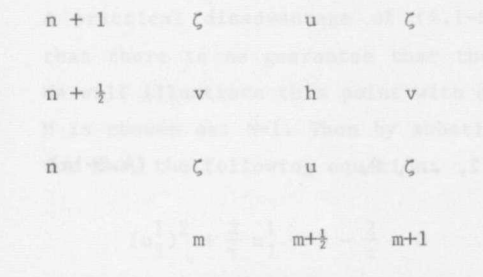


Figure (4-1) Staggered spatial grid

If we do not take into account the iterative or predictor corrector procedures that will be introduced later, then, at the inner points of the grid, each term of (4.2-2) is approximated as follows:

a): u_t, at $m+\frac{1}{2}$, n:

S1: $(u^{k+\frac{1}{2}} - u^{k})/\frac{1}{2}\tau$ at $m + \frac{1}{2}$, n

S2: $(u^{k+1} - u^{k+\frac{1}{2}})/\frac{1}{2}\tau$ at $m + \frac{1}{2}$, n (4.2-3a)

b): v_t; as u_t, but at m, $n+\frac{1}{2}$ (4.2-3b)

c): ζ_t; as u_t, but at m, n (4.2-3c)

d): uu_x; at $m+\frac{1}{2}$, n:

S1: $u^{k+\frac{1}{2}}_{m+\frac{1}{2},n}(u^{k}_{m+1\frac{1}{2},n} - u^{k}_{m-\frac{1}{2},n})/2\Delta x$ (explicit)

S2: $u^{k+\frac{1}{2}}_{m+\frac{1}{2},n}(u^{k+1}_{m+1\frac{1}{2},n} - u^{k+1}_{m-\frac{1}{2},n})/2\Delta x$ (implicit) (4.2-3d)

This discretization has been the subject of numerous numerical experiments carried out by the author. Also such discretizations as:

S1: $u^{k}_{m+\frac{1}{2},n}(u^{k}_{m+1\frac{1}{2},n} - u^{k}_{m-\frac{1}{2},n})/2\Delta x$ (explicit)

S2: $u^{k+1}_{m+\frac{1}{2},n}(u^{k+1}_{m+1\frac{1}{2},n} - u^{k+1}_{m-\frac{1}{2},n})/2\Delta x$ (implicit) (4.2-3d')

were tested. For fairly large Courant numbers (4.2-3d) especially turned out to have better stability properties than (4.2-3d'). Also more conservative approximations have been tested, see section 4.1, but these nonlinear approximations were not less susceptible to instability than (4.2-3d). Moreover, for steady flow problems the conservative approximations introduce an increased amount of numerical dissipation as was found by practical experiments.

e): vv_y, at $m, n+\frac{1}{2}$:

$$\begin{aligned} &S1: \; v^{k}_{m,n+\frac{1}{2}}(v^{k+\frac{1}{2}}_{m,n+1\frac{1}{2}} - v^{k+\frac{1}{2}}_{m,n-\frac{1}{2}})/2\Delta y \qquad \text{(implicit)} \\ &S2: \; v^{k+1}_{m,n+\frac{1}{2}}(v^{k+\frac{1}{2}}_{m,n+1\frac{1}{2}} - v^{k+\frac{1}{2}}_{m,n-\frac{1}{2}})/2\Delta y \qquad \text{(explicit)} \end{aligned} \tag{4.2-3e}$$

This approximation is similar to (4.2-3d).

f): vu_y, at $m+\frac{1}{2},n$:

$$S1: \; \overset{=k+\frac{1}{2}}{v}_{m+\frac{1}{2},n}(u^{k}_{m+\frac{1}{2},n+2}+4u^{k}_{m+\frac{1}{2},n+1}-4u^{k}_{m+\frac{1}{2},n-1}-u^{k}_{m+\frac{1}{2},n-2})/12\Delta y \qquad \text{(explicit)}$$

$$S2: \begin{cases} \overset{=k+\frac{1}{2}}{v}_{m+\frac{1}{2},n}(3u^{k+1}_{m+\frac{1}{2},n}-4u^{k+1}_{m+\frac{1}{2},n-1}+u^{k+1}_{m+\frac{1}{2},n-2})/2\Delta y, \text{ if } \overset{=k+\frac{1}{2}}{v}_{m+\frac{1}{2},n} > 0 \\ \qquad\qquad \text{(implicit)} \\ \overset{=k+\frac{1}{2}}{v}_{m+\frac{1}{2},n}(-3u^{k+1}_{m+\frac{1}{2},n}+4u^{k+1}_{m+\frac{1}{2},n+1}-u^{k+1}_{m+\frac{1}{2},n+2})/2\Delta y, \text{ if } \overset{=k+\frac{1}{2}}{v}_{m+\frac{1}{2},n} \leq 0 \end{cases} \tag{4.2-3f}$$

where $\overset{=k+\frac{1}{2}}{v}_{m+\frac{1}{2},n} = (v^{k+\frac{1}{2}}_{m,n+\frac{1}{2}} + v^{k+\frac{1}{2}}_{m+1,n+\frac{1}{2}} + v^{k+\frac{1}{2}}_{m,n-\frac{1}{2}} + v^{k+\frac{1}{2}}_{m+1,n-\frac{1}{2}})/4$

At stage 2 (4.2-3f) is implicit and is approximately solved with 2 iterations, cf. section 2.3.

g): uv_x, at $m,n+\frac{1}{2}$:

$$S1: \begin{cases} \overset{=k}{u}_{m,n+\frac{1}{2}}(3v^{k+\frac{1}{2}}_{m,n+\frac{1}{2}}-4v^{k+\frac{1}{2}}_{m-1,n+\frac{1}{2}}+v^{k+\frac{1}{2}}_{m-2,n+\frac{1}{2}})/2\Delta x, \text{ if } \overset{=k}{u}_{m,n+\frac{1}{2}} > 0 \\ \qquad\qquad \text{(implicit)} \\ \overset{=k}{u}_{m,n+\frac{1}{2}}(-3v^{k+\frac{1}{2}}_{m,n+\frac{1}{2}}+4v^{k+\frac{1}{2}}_{m+1,n+\frac{1}{2}}-v^{k+\frac{1}{2}}_{m+2,n+\frac{1}{2}})/2\Delta x, \text{ if } \overset{=k}{u}_{m,n+\frac{1}{2}} \leq 0 \end{cases} \tag{4.2-3g}$$

$$S2:\ \overset{=k+1}{u}_{m,n+\frac{1}{2}}(v^{k+\frac{1}{2}}_{m+2,n+\frac{1}{2}}+4v^{k+\frac{1}{2}}_{m+1,n+\frac{1}{2}}-4v^{k+\frac{1}{2}}_{m-1,n+\frac{1}{2}}-v^{k+\frac{1}{2}}_{m-2,n+\frac{1}{2}})/12\Delta x \qquad \text{(explicit)}$$

where $\overset{=k}{u}_{m,n+\frac{1}{2}}=(u^{k}_{m+\frac{1}{2},n}+u^{k}_{m+\frac{1}{2},n+1}+u^{k}_{m-\frac{1}{2},n}+u^{k}_{m-\frac{1}{2},n+1})/4$

This approximation is similar to (4.2-3f) but this time the first stage is implicit.

h): $-fv$ at $m+\frac{1}{2},n$:

$$S1:\ -f\overset{=k+\frac{1}{2}}{v}_{m+\frac{1}{2},n} \qquad \text{(implicit)}$$

S2: as S1, which is explicit at this stage (4.2-3h)

i): fu at $m,n+\frac{1}{2}$:

$$S1:\ f\,\overset{=k}{u}_{m,n+\frac{1}{2}} \qquad \text{(explicit)}$$

$$S2:\ f\,\overset{=k+1}{u}_{m,n+\frac{1}{2}} \qquad \text{(implicit)} \qquad (4.2\text{-}3i)$$

j): $gu\,(u^2+v^2)^{\frac{1}{2}}/[C^2(\zeta+h)]$, at $m+\frac{1}{2},n$

$$S1:\ g\,u^{k+\frac{1}{2}}_{m+\frac{1}{2},n}[(u^{k}_{m+\frac{1}{2},n})^2+(\overset{=k+\frac{1}{2}}{v}_{m+\frac{1}{2},n})^2]^{\frac{1}{2}}/(C^2_{m+\frac{1}{2},n}H^{k}_{m+\frac{1}{2},n})$$

$$S2:\ g\,u^{k+1}_{m+\frac{1}{2},n}[(u^{k+\frac{1}{2}}_{m+\frac{1}{2},n})^2+(\overset{=k+\frac{1}{2}}{v}_{m+\frac{1}{2},n})^2]^{\frac{1}{2}}/(C^2_{m+\frac{1}{2},n}H^{k+\frac{1}{2}}_{m+\frac{1}{2},n}) \qquad (4.2\text{-}3j)$$

where $H^{k}_{m+\frac{1}{2},n}=\frac{1}{2}(\zeta^{k}_{m+1,n}+\zeta^{k}_{m,n}+h_{m+\frac{1}{2},n+\frac{1}{2}}+h_{m+\frac{1}{2},n-\frac{1}{2}})$

This approximation is implicit at both stages.

k): $gv\,(u^2+v^2)^{\frac{1}{2}}/[C(\zeta+h)]$ at $m,n+\frac{1}{2}$:

$$S1:\ g\,v^{k+\frac{1}{2}}_{m,n+\frac{1}{2}}[(v^{k}_{m,n+\frac{1}{2}})^2+(\overset{=k}{u}_{m,n+\frac{1}{2}})^2]^{\frac{1}{2}}/(C^2_{m,n+\frac{1}{2}}H^{k}_{m,n+\frac{1}{2}})$$

$$S2:\ g\,v^{k+1}_{m,n+\frac{1}{2}}[(v^{k+\frac{1}{2}}_{m,n+\frac{1}{2}})^2+(\overset{=k+1}{u}_{m,n+\frac{1}{2}})^2]^{\frac{1}{2}}/(C^2_{m,n+\frac{1}{2}}H^{k+\frac{1}{2}}_{m,n+\frac{1}{2}}) \qquad (4.2\text{-}3k)$$

where $H^{k}_{m,n+\frac{1}{2}}=\frac{1}{2}(\zeta^{k}_{m,n+1}+\zeta^{k}_{m,n}+h_{m+\frac{1}{2},n+\frac{1}{2}}+h_{m-\frac{1}{2},n+\frac{1}{2}})$

This approximation is similar to (4.2-3j)

l): $-\nu\ (u_{xx}+u_{yy})$ at $m+\frac{1}{2},n$:

$$\text{S1: } -\nu[(u^{k}_{m+\frac{1}{2},n})_{oxx}+(u^{k}_{m+\frac{1}{2},n})_{oyy}] \quad \text{(explicit)}$$

$$\text{S2: } -\nu[(u^{k+1}_{m+\frac{1}{2},n})_{oxx}+(u^{k+1}_{m+\frac{1}{2},n})_{oyy}] \quad \text{(implicit)} \qquad (4.2\text{-}3\ell)$$

where:

$$(u^{k}_{m+\frac{1}{2},n})_{oxx} = (u^{k}_{m+1\frac{1}{2},n}-2u^{k}_{m+\frac{1}{2},n}+u^{k}_{m-\frac{1}{2},n})/\Delta x^2$$

$$(u^{k}_{m+\frac{1}{2},n})_{oyy} = (u^{k}_{m+\frac{1}{2},n+1}-2u^{k}_{m+\frac{1}{2},n}+u^{k}_{m+\frac{1}{2},n-1})/\Delta y^2,$$

$(u^{k+1}_{m+\frac{1}{2},n})_{oxx}$ and $(u^{k+1}_{m+\frac{1}{2},n})_{oyy}$ are defined accordingly.

The second stage is implicit. The implicit part in the y direction is solved iteratively, similarly to vu_y. For fairly small values of ν two iterations are enough. If not, it is probably preferable to change this part of the discretization method, which is very possible.

m): $-\nu\ (v_{xx}+v_{yy})$ at $m,n+\frac{1}{2}$:

$$\text{S1: } -\nu[(v^{k+\frac{1}{2}}_{m,n+\frac{1}{2}})_{oxx}+(v^{k+\frac{1}{2}}_{m,n+\frac{1}{2}})_{oyy}] \quad \text{(implicit)} \qquad (4.2\text{-}3m)$$

S2: as S1, which is explicit at this stage

where:

$$(v^{k+\frac{1}{2}}_{m,n+\frac{1}{2}})_{oxx} = (v^{k+\frac{1}{2}}_{m+1,n+\frac{1}{2}}-2v^{k+\frac{1}{2}}_{m,n+\frac{1}{2}}+v^{k+\frac{1}{2}}_{m-1,n+\frac{1}{2}})/\Delta x^2$$

$$(v^{k+\frac{1}{2}}_{m,n+\frac{1}{2}})_{oyy} = (v^{k+\frac{1}{2}}_{m,n+1\frac{1}{2}}-2v^{k+\frac{1}{2}}_{m,n+\frac{1}{2}}+v^{k+\frac{1}{2}}_{m,n-\frac{1}{2}})/\Delta y^2$$

The implicit part in the x-direction is solved iteratively, similar to uv_x.

n): $g\zeta_x$ at $m+\frac{1}{2},n$:

S1: $g\,(\zeta^{k+\frac{1}{2}}_{m+1,n}-\zeta^{k+\frac{1}{2}}_{m,n})/\Delta x$ (implicit)

S2: as stage 1, which is explicit at this stage (4.2-3n)

o): $g\zeta_y$ at $m,n+\frac{1}{2}$:

S1: $g\,(\zeta^{k}_{m,n+1}\ \zeta^{k}_{m,n})/\Delta y$ (explicit)

S2: $g\,(\zeta^{k+1}_{m,n+1}\ \zeta^{k+1}_{m,n})/\Delta y$ (implicit) (4.2-3o)

p): $(Hu)_x$ at m,n:

S1: $(H^{k+\frac{1}{2}}_{m+\frac{1}{2},n}u^{k+\frac{1}{2}}_{m+\frac{1}{2},n}-H^{k+\frac{1}{2}}_{m-\frac{1}{2},n}u^{k+\frac{1}{2}}_{m-\frac{1}{2},n})/\Delta x$ (implicit)

S2: as stage 1, which is explicit at this stage (4.2-3p)

The implicit part here requires an iterative procedure. Locally linear schemes, which are cheaper, were tested as well; but it was found experimentally that this local linearization causes instabilities, especially at very shallow regions with an accidented bottom profile.
For example, the following "local linearization" turned out to be unstable:

S1: $(\overline{h}^{y}\,u^{k+\frac{1}{2}})_{ox}+\zeta^{k}u^{k+\frac{1}{2}}_{ox}+\overline{u^{k}\zeta^{k+\frac{1}{2}}_{ox}}^{x}$, at m,n (effectively implicit)

S2: $(\overline{h}^{y}\,u^{k+\frac{1}{2}})_{ox}+\zeta^{k+1}u^{k+\frac{1}{2}}_{ox}+\overline{u^{k+1}\zeta^{k+\frac{1}{2}}_{ox}}^{x}$ at m,n (effectively explicit)

where it is to be noted that:

$$(Hu)_{ox}=(\overline{h}^{y}\,u)_{ox}+\zeta\,u_{ox}+\overline{u\,\zeta_{ox}}^{x}\ \text{at } m,n$$

For the iterative solution of (4.2-3p) several possibilities were tested. The procedure that turned out to be very efficient is given by (4.2-4).

Usually, two iterations are enough, both for accuracy and stability. Only for very shallow regions with an accidented bottom profile and points changing from dry to wet and vice versa, see section 4.5, more iterations are occasion-

ally necessary for stability. Second order accuracy is obtained for $Q \geq 2$, for Q see 4.2-4.

q): $(Hv)_y$ at m,n:

$$S1: (H^k_{m,n+\frac{1}{2}}\, v^k_{m,n+\frac{1}{2}} - H^k_{m,n-\frac{1}{2}}\, v^k_{m,n-\frac{1}{2}})/\Delta y \quad \text{(explicit)}$$

$$S2: (H^{k+1}_{m,n+\frac{1}{2}}\, v^{k+1}_{m,n+\frac{1}{2}} - H^{k+1}_{m,n-\frac{1}{2}}\, v^{k+1}_{m,n-\frac{1}{2}})/\Delta y \quad \text{(implicit)} \qquad (4.2\text{-}3q)$$

The iterative procedure to solve the implicit part is given by (4.2-4).

If all the iterative procedures to solve implicit equations are taken into account then the discretizations (4.2-3a) up to and including (4.2-3q) yield the following FDM for the approximation of (4.2-2):

Stage 1:

$$u^{[0]} = u^k,\ v^{[0]} = v^k,\ \zeta^{[0]} = \zeta^k$$

For p = 1,2, q = 1,2, ...,Q:

$$(u^{[q]} - u^k)/\tfrac{1}{2}\tau + u^{[q]}\, \overline{u^k_{ox}}^{x} + S_{oy}(\overset{=}{v}^{k+\frac{1}{2}}, u^k) - f\overset{=}{v}^{k+\frac{1}{2}} + g\zeta^{[q]}_{ox}$$

$$+ gu^{[q]}[(\overset{=}{v}^{k+\frac{1}{2}})^2 + (u^k)^2]^{\frac{1}{2}}/(C^2 H^k) - \nu(u^k_{oxx} + u^k_{oyy}) = 0, \text{ at } m+\tfrac{1}{2},n \qquad (4.2\text{-}4a)$$

$$(v^{[p]} - v^k)/\tfrac{1}{2}\tau + v^k\, \overline{v^{[p]}_{oy}}^{y} + S_{+x}[\overset{=}{u}^k, v^{[p]}, \delta(p+p')] + f\overset{=}{u}^k + g\zeta^k_{oy}$$

$$+ gv^{[p]}[(\overset{=}{u}^k)^2 + (v^k)^2]^{\frac{1}{2}}/(C^2 H^k) - \nu(v^{[*]}_{oxx} + v^{[p]}_{oyy}) = 0, \text{ at } m,n+\tfrac{1}{2} \qquad (4.2\text{-}4b)$$

$$(\zeta^{[q]} - \zeta^k)/\tfrac{1}{2}\tau + (\overline{h}^y u^{[q]})_{ox} + \zeta^{[q-1]} u^{[q]}_{ox} + \overline{u^{[q-1]}\zeta^{[q]}_{ox}}^{x} + (H^k v^k)_{oy} = 0, \text{ at } m,n \qquad (4.2\text{-}4c)$$

$$u^{k+\frac{1}{2}} = u^{[Q]},\ v^{k+\frac{1}{2}} = v^{[2]},\ \zeta^{k+\frac{1}{2}} = \zeta^{[Q]}$$

where:

$$S_{oy}(v^{=k+\frac{1}{2}}, u^{k})\text{at } m+\tfrac{1}{2}, n=v^{=k+\frac{1}{2}}_{m+\frac{1}{2},n}(u^{k}_{m+\frac{1}{2},n+2}+4u^{k}_{m+\frac{1}{2},n+1}-4u^{k}_{m+\frac{1}{2},n-1}-u^{k}_{m+\frac{1}{2},n-2})/12\Delta y$$

$$S_{+x}[u^{=k}, v^{[p]}, \delta]\text{at } m, n+\tfrac{1}{2}=\begin{cases} u^{=k}_{m,n+\frac{1}{2}}(3v^{[p-\delta]}_{m,n+\frac{1}{2}}-4v^{[p-\delta]}_{m-1,n+\frac{1}{2}}+v^{[p-\delta]}_{m-2,n+\frac{1}{2}})/2\Delta x \text{ if } u^{=k}_{m,n+\frac{1}{2}}>0 \\ u^{=k}_{m,n+\frac{1}{2}}(-3v^{[p-1+\delta]}_{m,n+\frac{1}{2}}+4v^{[p-1+\delta]}_{m+1,n+\frac{1}{2}}-v^{[p-1+\delta]}_{m+2,n+\frac{1}{2}})/2\Delta x \text{ if } u^{=k}_{m,n+\frac{1}{2}}\leq 0 \end{cases}$$

$$\delta\,(p+p') = \tfrac{1}{2}[1 + (-1)^{p+p'}]$$

$$p' = \begin{cases} 0, \text{ if } \sum_{m,n} u^{k} > 0 \quad (\textstyle\sum u \text{ denotes the sum of } u \text{ over all grid points}) \\ 1, \text{ if } \sum_{m,n} u^{k} \leq 0 \end{cases}$$

$$\text{and } v^{[*]}_{oxx}\text{at } m, n+\tfrac{1}{2} = (v^{[p-1+\delta(p+p')]}_{m+1,n+\frac{1}{2}}-2v^{[p]}_{m,n+\frac{1}{2}}+v^{[p-\delta(p+p')]}_{m-1,n+\frac{1}{2}})/\Delta x^{2}$$

Stage 2:

$$u^{[o]} = u^{k+\frac{1}{2}},\ v^{[o]} = v^{k+\frac{1}{2}},\ \zeta^{[o]} = \zeta^{k+\frac{1}{2}},$$

For p = 1,2 and q = 1, ..., Q:

$$(u^{[p]}-u^{k+\frac{1}{2}})/\tfrac{1}{2}\tau+u^{k+\frac{1}{2}}\overline{u^{[p]}}^{x}_{ox}+S_{+y}[v^{=k+\frac{1}{2}}, u^{[p]}, \delta(p+p')]-fv^{=k+\frac{1}{2}}+g\zeta^{k+\frac{1}{2}}_{ox}$$

$$+gu^{[p]}[(v^{=k+\frac{1}{2}})^{2}+(u^{k+\frac{1}{2}})^{2}]^{\frac{1}{2}}/(C^{2}H^{k+\frac{1}{2}})-\nu(u^{[p]}_{oxx}+u^{[*]}_{oyy})=0 \text{ at } m+\tfrac{1}{2}, n \qquad (4.2\text{-}4d)$$

$$(v^{[q]}-v^{k+\frac{1}{2}})/\tfrac{1}{2}\tau+v^{[q]}\overline{v^{k+\frac{1}{2}}}^{x}_{ox}+S_{ox}(u^{=k+1}, v^{k+\frac{1}{2}})+fu^{=k+1}+g\zeta^{[q]}_{ox}$$

$$+gv^{[q]}[(v^{k+\frac{1}{2}})^{2}+(u^{=k+1})^{2}]^{\frac{1}{2}}/(C^{2}H^{k+\frac{1}{2}})-\nu(v^{k+\frac{1}{2}}_{oxx}+v^{k+\frac{1}{2}}_{oyy})=0 \text{ at } m, n+\tfrac{1}{2} \qquad (4.2\text{-}4e)$$

$$(\zeta^{[q]}-\zeta^{k+\frac{1}{2}})/\tfrac{1}{2}\tau+(H^{k+\frac{1}{2}}\overline{u^{k+\frac{1}{2}}}^{x})_{ox}+(\overline{h}^{x}v^{[q]})_{oy}+\zeta^{[q-1]}v^{[q]}_{oy}+v^{[q-1]}\zeta^{[q]}_{oy}=0, \text{ at } m, n \qquad (4.2\text{-}4f)$$

$$u^{k+1} = u^{[2]},\quad v^{k+1} = v^{[Q]},\quad \zeta^{k+1} = \zeta^{[Q]}$$

where:

$$S_{+y}(v^{=k+\frac{1}{2}}, u^{[p]}, \delta) \text{ at } m+\tfrac{1}{2},n = \begin{cases} v^{=k+\frac{1}{2}}_{m+\frac{1}{2},n}(3u^{[p-\delta]}_{m+\frac{1}{2},n} - 4u^{[p-\delta]}_{m+\frac{1}{2},n-1} + u^{[p-\delta]}_{m+\frac{1}{2},n-2})/2\Delta x, \\ \qquad \text{if } v^{=k+\frac{1}{2}}_{m+\frac{1}{2},n} > 0 \\ v^{=k+\frac{1}{2}}_{m+\frac{1}{2},n}(-3u^{[p-1+\delta]}_{m+\frac{1}{2},n} + 4u^{[p-1+\delta]}_{m+\frac{1}{2},n+1} - u^{[p-1+\delta]}_{m+\frac{1}{2},n+2})/2\Delta x, \\ \qquad \text{if } v^{=k+\frac{1}{2}}_{m+\frac{1}{2},n} \leq 0 \end{cases}$$

$$\delta(p+p') = [1 + (-1)^{p+p'}]$$

$$p' = \begin{cases} 0, \text{ if } \sum_{m,n} v^{k+\frac{1}{2}} > 0 \ (\Sigma \text{ v denotes the sum of v over all grid points}) \\ 1, \text{ if } \sum_{m,n} v^{k+\frac{1}{2}} \leq 0 \end{cases}$$

$$S_{ox}(u^{=k+1}, v^{k+\frac{1}{2}}) \text{ at } m,n+\tfrac{1}{2} = u^{=k+1}_{m,n+\frac{1}{2}}(v^{k+\frac{1}{2}}_{m+2,n+\frac{1}{2}} + 4v^{k+\frac{1}{2}}_{m+1,n+\frac{1}{2}} - 4v^{k+\frac{1}{2}}_{m-1,n+\frac{1}{2}} - v^{k+\frac{1}{2}}_{m-2,n+\frac{1}{2}})/12\Delta x$$

$$\text{and } u^{[*]}_{oyy} \text{ at } m+\tfrac{1}{2},n = (u^{[p-1+\delta(p+p')]}_{m+\frac{1}{2},n+1} - 2u^{[p]}_{m+\frac{1}{2},n} + u^{[p-\delta(p+p')]}_{m+\frac{1}{2},n-1})/\Delta y^2$$

At stage 1 (4.2-4b) is an implicit equation. Because of the definition of p', (4.2-4b) is solved column by column in the dominant flow direction of u. If the sign of u is constant then (4.2-4b) is solved in one iteration, otherwise a second step is necessary. This step proceeds in the opposite direction. After (4.2-4b), (4.2-4a) and (4.2-4c) are solved. These equations are coupled implicitly. By substitution of (4.2-4a) at $m-\frac{1}{2},n$ and at $m+\frac{1}{2},n$ into (4.2-4c) at m,n the implicit equations are tri-diagonal and of the same size as the tri-diagonal equations of (4.2-4b).

At stage 2 (4.2-4d) is implicit and is solved similarly to (4.2-4b). The coupled implicit equations (4.2-4e) and (4.2-4f) are solved according to (4.2-4a) and (4.2-4c).

By solving (4.2-4) in the sequence (4.2-4b), [(4.2-4a) and (4.2-4c)], (4.2-4d) and [(4.2-4e) and (4.2-4f)] the computer implementation needs only one array per dependent variable (u,v,ζ) and one work array of the size of the number of "ζ-points" of the grid. This means that the FDM is very efficient with respect to storage.

4.3 Boundary conditions, closed boundaries

The SWE given by (4.0-1) are of course incomplete without boundary conditions. In general two types of boundary conditions are to be distinguished: closed and open. Closed boundaries are land-water boundaries; they are physical because they relate to a really existing boundary. Open boundaries are mathematical; they are introduced to restrict the size of the domain of the problem. This section describes the numerical treatment of closed boundaries.

At a closed boundary, one boundary condition is to be prescribed if $\nu=0$; then the equations are hyperbolic. If $\nu\neq 0$, two boundary conditions are to be prescribed. In that case the equations are "incompletely parabolic", see Oliger and Sundström [19].
At closed boundaries the following boundary conditions are given:

$$u_{\perp} = 0, \tag{4.3-1a}$$

$$(1-\alpha)\, u_{//} + \alpha \frac{\delta}{\delta n} u_{//} = 0 \tag{4.3-1b}$$

where $u_{\perp}$ denotes the velocity normal to the boundary, $u_{//}$ denotes the velocity parallel to the boundary and $\frac{\partial}{\partial n}$ denotes the derivative normal to the boundary. If $\alpha=1$ then (4.3-1b) describes a "perfect slip" boundary condition but if $\alpha=0$ then (4.3-1b) represents a "no slip" boundary condition. In general $\alpha=1$.

For the FDM (4.2-4) closed boundaries are represented by zero velocities in either the x-direction or the y direction. For some geometries this yields the typical "zig-zag" lines of which an example is figure (4-2).

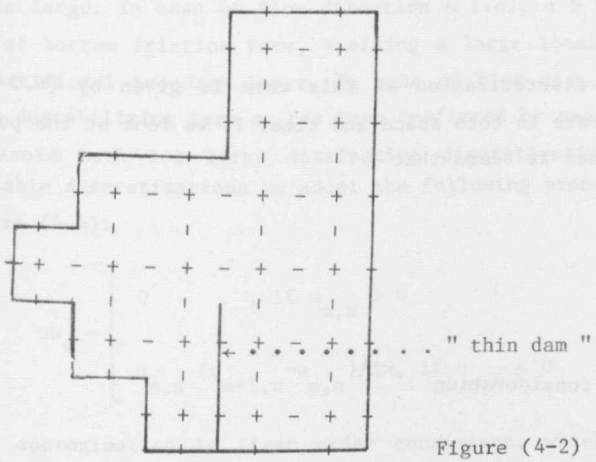


Figure (4-2) Example of closed boundaries as represented by the staggered grid.

The staggered grid allows a simple treatment of the boundary conditions. At a closed boundary only one dependent variable is to be calculated. Because this variable equals zero, no special boundary scheme is needed at the boundary itself. Near closed boundaries, however, the discretizations of section 4.2 cannot always be applied. This problem concerns only the discretizations of the momentum equations. For the discretization of the continuity equation special boundary schemes are not necessary.
We will describe the boundary schemes for the momentum equation in the x-direction. The boundary treatment of the momentum equation in the y-direction is similar.

We will give a separate description of the discretization of the momentum equation for each term, that needs a special boundary treatment. It is to be noted that u_t and $g\zeta_x$ never need special discretization. This is an important advantage because these are in general the most important terms of the momentum equation. In fact the special boundary discretizations are only necessary for the advection terms, which only have limited influence in case of some applications, see, e.g., Verboom [26] or Stelling [21].

The boundary treatment, near closed boundaries, for the advection terms is as follows:

a) uu_x:

For the inner points the discretization of this term is given by (4.2-3d) which is second order accurate in both space and time. If we look at the point depicted in figure (4-3) then it seems that

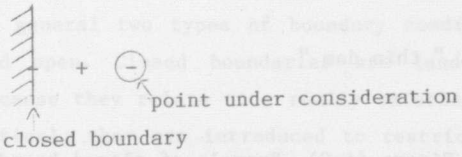


Figure (4-3)

for uu_x a special boundary discretization is not necessary. Theoretically this is true. But for practical applications when (4.2-3d) is applied near boundaries, it can produce instability or artificial boundary layers. Consider for example the situation of figure (4-4).

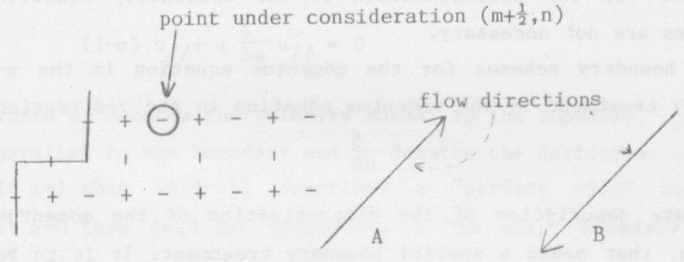


Figure (4-4)

Application of (4.2-3d) for the discretization of the advection term at the point under consideration yields:

$$(uu_x)_{m+\frac{1}{2},n} \approx u_{m+\frac{1}{2},n}(u_{m+1\frac{1}{2},n} - 0)/2\Delta x \qquad (4.3\text{-}3)$$

In practical applications the velocity in the point under consideration can be

quite large. In case of flow direction A i.e., $u > 0$, (4.3-3) act as an artificial bottom friction term, yielding a large local water level gradient and an artificial boundary layer. In case of flow direction B, (4.3-3) might act as a destabilizing term as has been confirmed by numerical experiments.
To avoid both too large dissipative discretizations near the boundary or unstable discretizations we adopt the following procedure for the situation of figure (4-4):

$$uu_x \simeq \begin{cases} 0 & , \text{ if } u_{m,n} > 0 \\ u_{m,n}\,(u_{m+1,n} - u_{m,n})/\Delta x, & \text{if } u_{m,n} \leq 0 \end{cases} \qquad (4.3\text{-}4)$$

This approximation is first order consistent, which does not affect second order convergence because this approximation is applied only near the boundary, see chapter I.
The discretization (4.3-4) turned out to give satisfactory results for a large variety of geometries. By practical experience we found that in general satisfactory discretizations for advection terms near boundaries are generated by taking into account the following principles:

i) Always avoid negative diffusion in the truncation error of the discretization.

ii) If the discretization formula contains the boundary value $u_\perp = 0$ then try to avoid this by using a discretization that needs fewer grid points, or by upwind differencing. If it is not possible to avoid the boundary value $u_\perp = 0$ in the discretization formula, then the advection term has to be approximated by a zero value.

iii) The discretization should be such that if the boundary procedure is applied to a frozen coefficient Cauchy problem, then the resulting scheme is stable, cf. Goldberg and Tadmor [7], [8] or Trapp and Ramshaw [23].

By (i) and (iii) instabilities are avoided while (ii) suppresses artificial boundary layers.

Summarizing, for the discretization of uu_x we adopt the following procedure for the situation of the figures (4-4) and (4-5):

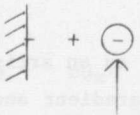


Figure (4-4), $u_{m-\frac{1}{2},n} = 0$

Stage 1 ($u^k_{m-\frac{1}{2},n}=0$, $u^k_{m+1\frac{1}{2},n}\neq 0$):

$$uu_x \simeq \begin{cases} 0, & \text{if } u^k_{m+\frac{1}{2},n} > 0 \\ u^k_{m+\frac{1}{2},n}(u^{k+\frac{1}{2}}_{m+1\frac{1}{2},n}-u^{k+\frac{1}{2}}_{m+\frac{1}{2},n})/\Delta x, & \text{if } u^k_{m+\frac{1}{2},n} < 0 \end{cases} \qquad (4.3\text{-}5a)$$

Stage 2 ($u^{k+\frac{1}{2}}_{m-\frac{1}{2},n}=0$, $u^{k+\frac{1}{2}}_{m+1\frac{1}{2},n}\neq 0$):

$$uu_x \simeq \begin{cases} 0 & \text{if } u^{k+\frac{1}{2}}_{m+\frac{1}{2},n} > 0 \\ u^{k+1}_{m+\frac{1}{2},n}\,(u^{k+\frac{1}{2}}_{m+1\frac{1}{2},n}-u^{k+\frac{1}{2}}_{m+\frac{1}{2},n})/\Delta x & \text{if } u^{k+\frac{1}{2}}_{m+\frac{1}{2},n} \leq 0 \end{cases} \qquad (4.3\text{-}5b)$$

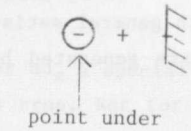


Figure (4-5) $u_{m+1\frac{1}{2},n} = 0$

Stage 1 ($u^k_{m-\frac{1}{2},n}\neq 0$, $u^k_{m+1\frac{1}{2},n}= 0$):

$$uu_x \simeq \begin{cases} u^k_{m+\frac{1}{2},n}(u^{k+\frac{1}{2}}_{m+\frac{1}{2},n}-u^{k+\frac{1}{2}}_{m-\frac{1}{2},n})/\Delta x, & \text{if } u^k_{m+\frac{1}{2},n} > 0 \\ 0, & \text{if } u^k_{m+\frac{1}{2},n} \leq 0 \end{cases} \qquad (4.3\text{-}6a)$$

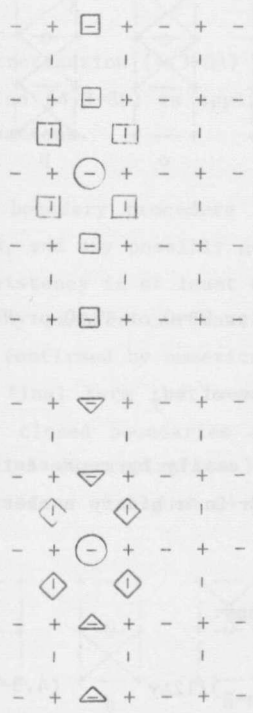
Stage 2 ($u^{k+\frac{1}{2}}_{m-\frac{1}{2},n}\neq 0$, $u^{k+\frac{1}{2}}_{m+1\frac{1}{2},n}=0$):

$$uu_x \simeq \begin{cases} u^{k+1}_{m+\frac{1}{2},n}(u^{k+\frac{1}{2}}_{m+\frac{1}{2},n}-u^{k+\frac{1}{2}}_{m-\frac{1}{2},n})/\Delta x, & \text{if } u^k_{m+\frac{1}{2},n} > 0 \\ o, & \text{if } u^k_{m+\frac{1}{2},n} \leq 0 \end{cases} \qquad (4.3\text{-}6b)$$

Note that if $u_{m-\frac{1}{2},n}= 0$ and $u_{m+\frac{1}{2},n} = 0$ then $uu_x \simeq 0$.
The discretizations of vv_y near boundaries are similar to those of uu_x.

b) vu_y:

For the discretization of vu_y we take into account the same principles that were used for uu_x. Also for this case first order consistency is assumed as sufficient to maintain second order convergence. For vu_y the boundary procedure is more complex, which stems from the fact that the "full" discretization at the inner points as given by (4.2-3fa) and (4.2-3fb) involves more grid points, as is illustrated by figure (4-6).



○ point under consideration $(m+\frac{1}{2},n)$

□ grid points involved in (4.2-3fa)

○ point under consideration $(m+\frac{1}{2},n)$

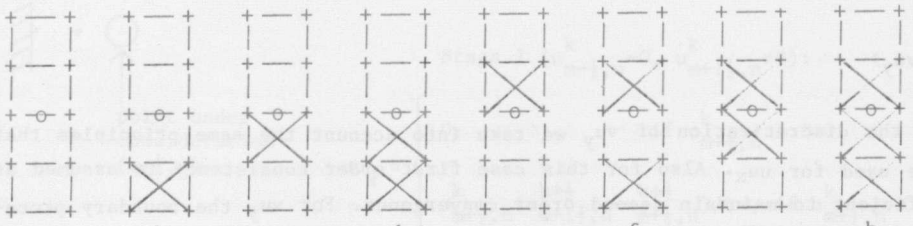
◇ grid points involved in (4.2-3fb) for each sign of $v^{=k+\frac{1}{2}}_{m+\frac{1}{2},n}$

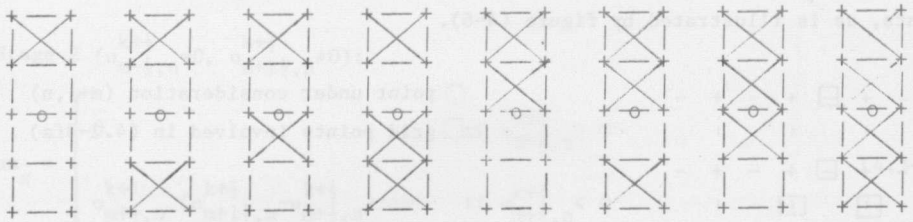
△ grid points involved in (4.2-3fb) if $v^{=k+\frac{1}{2}}_{m+\frac{1}{2},n} > 0$

▽ grid points involved in (4.2-3fb) if $v^{=k+\frac{1}{2}}_{m+\frac{1}{2},n} < 0$

Figure (4-6) Grid points involved in (4.2-3f)

Because of the large number of grid points involved in (4.2-3f) there are many possibilities that might cause a considerable amount of computational overhead. To reduce the overhead we consider only the situations of figure (4-7).





denotes: at least UT=0 or UB=0 or VL=0 or VR=0

Figure (4-7) Possible situations for the discretization of vu_y

The situations of figure (4-7) can be determined very easily by representing the situation in each computational cell as a position in a binary number of four bits.

At stage 1 we consider only the following discretizations:

$$vu_y \simeq \overset{=k+\frac{1}{2}}{v_{m+\frac{1}{2},n}} (u^k_{m+\frac{1}{2},n+2} + 4u^k_{m+\frac{1}{2},n+1} - 4u^k_{m+\frac{1}{2},n-1} - u^k_{m+\frac{1}{2},n-2})/12\Delta y \qquad (4.3\text{-}7a)$$

$$vu_y \simeq \overset{=k+\frac{1}{2}}{v_{m+\frac{1}{2},n}} (u^k_{m+\frac{1}{2},n+1} - u^k_{m+\frac{1}{2},n-1})/2\Delta y \qquad (4.3\text{-}7b)$$

$$vu_y \simeq 0 \qquad (4.3\text{-}7c)$$

Discretization (4.3-7a) is applied only for situation a, (4.3-7b) is applied for situation b, i, and j; and $vu_y \simeq 0$ for the remaining situations.

At stage 2 we consider two cases: , (I) $\overset{=k+\frac{1}{2}}{v}_{m+\frac{1}{2},n} > 0$ and (II) $\overset{=k+\frac{1}{2}}{v}_{m+\frac{1}{2},n} \leq 0$. We consider only case (I) because (II) is treated completely similarly. For case (I) we consider only the following discretizations:

$$vu_y \simeq \overset{=k+\frac{1}{2}}{v}_{m+\frac{1}{2},n}(3u^{k+1}_{m+\frac{1}{2},n} - 4u^{k+1}_{m+\frac{1}{2},n-1} + u^{k+1}_{m+\frac{1}{2},n-2})/2\Delta y \qquad (4.3\text{-}8a)$$

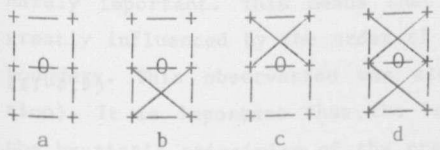
$$vu_y \simeq \overset{=k+\frac{1}{2}}{v}_{m+\frac{1}{2},n}(u^{k+1}_{k+\frac{1}{2},n} - u^{k+1}_{k+\frac{1}{2},n-1})/\Delta y \qquad (4.3\text{-}8b)$$

$$vu_y \simeq 0 \qquad (4.3\text{-}8c)$$

Discretization (4.3-8a) is applied for the situations a, e, i and m. Discretization (4.3-8b) is applied for b, f, j and n while $vu_y \simeq 0$ for the remaining situations.

The boundary procedure as described above causes little computational overhead, and any possibly geometry can be treated by this procedure. The order of consistency is at least one. The stability of the procedures when applied to a Cauchy problem can easily be verified; the stability for practical problems was confirmed by numerical experiments.

The final term that we treat in this section needs a special discretization near closed boundaries and is given by u_{yy}. (Note that u_{xx} does not need a special discretization.) For the boundary procedure of u_{yy} we consider the situations of figure (4-8).



0 point under consideration $(m+\frac{1}{2},n)$

Figure (4-8) Possible situations for the discretization of u_{yy}.

The following discretizations are applied:

for a: $u_{yy} \simeq (u_{m+\frac{1}{2},n+1} - 2u_{m+\frac{1}{2},n} + u_{m+\frac{1}{2},n-1})/\Delta y^2 \qquad (4.3\text{-}9a)$

for b: $u_{yy} \simeq [(1-\alpha)(-3u_{m+\frac{1}{2},n} + u_{m+\frac{1}{2},n+1}) + \alpha(-u_{m+\frac{1}{2},n} + u_{m+\frac{1}{2},n+1})]/\Delta y^2 \qquad (4.3\text{-}9b)$

for c: $u_{yy} \simeq [(1-\alpha)(-3u_{m+\frac{1}{2},n}+u_{m+\frac{1}{2},n-1})+\alpha(-u_{m+\frac{1}{2},n}+u_{m+\frac{1}{2},n-1})]/\Delta y^2$ (4.3-9c)

for d: $u_{yy} \simeq -(1-\alpha)\ 4u_{m+\frac{1}{2},n}/\Delta y^2$ (4.3-9d)

The time discretization is similar to (4.2-3ℓ).
It is easily verified that for $\alpha=1$, the usual case (4.3-9) is completely second order accurate. If $\alpha\neq1$ then (4.3-9a) is second order accurate and the remaining discretizations are first order consistent.

4.4 Boundary conditions, open boundaries

As already mentioned open boundaries are artificial water-water boundaries that have been arbitrarily drawn somewhere across a wider flow field to restrict the domain of the problem. Computational fluid dynamicists, as well as numerical analysts are quite active in the subject of open boundary approximation, see e.g. Kreiss [13], Kreiss and Gustafsson [15], Oliger and Sundström [19], Verboom et al [25], [26] Gerritsen [5], Gottlieb et al [9], Enquist and Majda [4], Kutler [16], Strikwerda [22], Elvius and Sundström [3], Gustafsson [10]; this list is nowhere near complete. A thorough theoretical treatment is beyond the scope of this work.

This section describes the numerical boundary condition procedures for the following boundary conditions at open boundaries.

a) Velocity boundary conditions given by:

$$u_{\perp} = f^{u}(t) \tag{4.4-1a}$$

$$u_{//} = 0 \tag{4.4-1b}$$

$$\frac{\partial}{\partial n} u_{//} = 0, \text{ if } \nu \neq 0 \tag{4.4-1c}$$

b) Water level boundary conditions given by:

$$\zeta = f^{\zeta}(t) \tag{4.4-2a}$$

$$u_{//} = 0 \tag{4.4-2b}$$

$$\frac{\partial}{\partial n} u_{//} = 0, \text{ if } \nu \neq 0 \tag{4.4-2c}$$

where (4.4-1b) and (4.4-2b) are prescribed only at inflow, i.e., if $u_{\perp}$ is directed from outside the domain of the problem to inside. The well-posedness of (4.4-1) and (4.4-2) is treated by Verboom et al [24], [25].

The conditions (4.4-1) and (4.4-2) are often applied to practical flow problems in civil engineering because the prescribed quantities can be measured in nature. However, "absorbing boundary conditions", see Enquist and Majda [4], can also be implemented by the procedures that we describe in this section.

Because of the staggered grid, the special boundary procedures are necessary only because of the advection terms. Various extrapolation procedures were tested as proposed for example by Elvius and Sundström [3] or Gerritsen [5]. When applied to our grid structure these procedures were not satisfactory for inflow boundaries. At inflow, extrapolation procedures are sensitive to instabilities as was found by practical experience. It seems that boundary extrapolation methods are stable only if they are based upon extrapolation of quantities that reach the boundary from inside, like the outgoing "Riemann invariants", see Moretti [18] or Gottlieb et al [9]. For relevant applications in civil engineering it was also found by practical experience that the order of consistency of the advection discretization near the open boundary is hardly important. This means that the solution at the inner points is not greatly influenced by the order of the advection discretization near the open boundary. This observation was also made by Leendertse (private communication). It is important that the boundary procedure be stable. We found that the heuristic principles of the preceding section are also applicable for the construction of open boundary procedures. For the discretization of (4.4-1) and (4.4-2) we propose the following procedures:

a): Velocity boundary conditions:

for velocity boundary conditions we assume a geometry as given by figure (4-9)

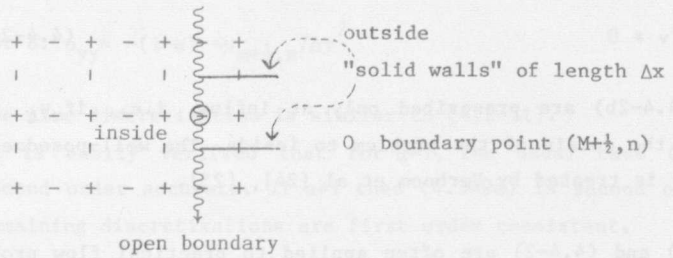


Figure (4-9) Open velocity boundary

The following discretizations are used:

$$u^k = f^u(k\tau), \text{ at } M+\tfrac{1}{2},n \tag{4.4-3a}$$

$$v^k = 0, \text{ at } M+1,\ n+\tfrac{1}{2} \text{ and } M+1,\ n-\tfrac{1}{2} \tag{4.4-3b}$$

$$\zeta^k_{M+1,n} = \zeta^k_{M,n} \tag{4.4-3c}$$

$$\text{at } M-\tfrac{1}{2},n\ uu_x \text{ is approximated by } \begin{cases} uu_{-x} & \text{if } u > 0 \\ uu_{+x} & \text{if } u \leq 0 \end{cases} \tag{4.4-3d}$$

where u_{-x} and u_{+x} are defined according to (3.3-2).
This procedure is consistent with (4.4-1). The condition given by (4.4-3a) is consistent with (4.4-1a). Because of the procedures for the discretizations of vu_x and v_{yy} near closed boundary as described in section 4.3, (4.4-3b) is consistent with (4.4-1b) and (4.4-1c). Because of (4.4-3c) at point M,n of figure (4-9), the continuity equation, is approximated with zero order consistency. According to Gustafsson [10] or Beam et al. [2], this is sufficient to maintain convergence. Experiments with extrapolations of a higher order did not greatly change the results at the inner points or became unstable. Boundary conditions in x and y direction are treated similarly.

b): Water level boundary conditions

For water level boundary conditions we assume a geometry as given by figure (4-10).

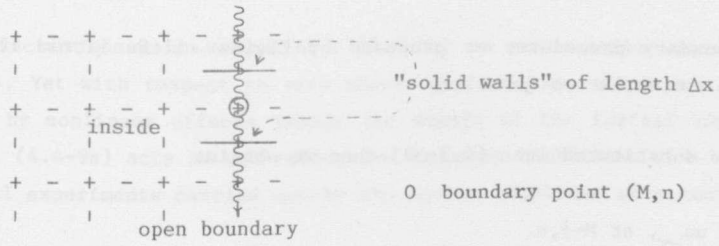


Figure (4-10) Open water level boundary

The following discretizations are proposed:

$$\zeta^k = f^{\zeta}(k\tau), \text{ at } M,n \tag{4.4-4a}$$

$$v^k = 0 \quad , \text{ at } M,n+\tfrac{1}{2} \text{ and } M,n-\tfrac{1}{2} \tag{4.4-4b}$$

$$\text{at } M-\tfrac{1}{2},n \; uu_x \text{ is approximated by} \begin{cases} uu_{-x} & \text{if } u > 0 \\ 0 \quad , & \text{if } u \leq 0 \end{cases} \tag{4.4-4c}$$

$$\text{at } M-1\tfrac{1}{2},n \; uu_x \text{ is approximated by} \begin{cases} uu_{-x} & \text{if } u > 0 \\ uu_{+x} & \text{if } u \leq 0 \end{cases} \tag{4.4-4d}$$

This discretization is a consistent approximation of (4.4-2).

Note that at inflow uu_x is approximated by a zero value at $M-\frac{1}{2},n$. This yields a zero order consistent approximation of the momentum equation at $M-\frac{1}{2},n$. Again this yields only first order convergence from a theoretical point of view. Yet the influence on the inner points of this approximation is negligible as was found by practical experiments. The only noticeable effect of approximations

of a higher order of consistency is that sometimes they became unstable. Consider for example the following extrapolation formula for a "virtual" u value at $M+\frac{1}{2},n$:

$$u^{k}_{M+\frac{1}{2},n} = 2u^{k}_{M-\frac{1}{2},n} - u^{k}_{M-1\frac{1}{2},n} \tag{4.4-5}$$

where the boundary procedures as proposed by Elvius and Sundström are based upon similar extrapolation formulae.

If (4.4-5) is substituted into (4.2-3d) then we obtain:

$$uu_{x} \simeq uu_{-x}, \text{ at } M-\tfrac{1}{2},n \tag{4.4-6}$$

If we calculate the truncation error of the spatial discretization of (4.4-6) then we obtain:

$$u(x,y,t)[u(x,y,t)]_{ox} = u(x,y,t)[u(x,y,t)]_{x} - \frac{\Delta x}{2}u(x,y,t)[u(x,y,t)]_{xx} + O(\Delta x^{2}) \tag{4.4-7}$$

If $u < 0$, then the truncation error of (4.4-6) contains negative diffusion as follows from (4.4-7). This is probably the reason for the observed instabilities.

A stabilizing effect is often experienced as a result of the prescription of Riemann invariants at the open boundaries, see e.g. Oliger and Sundström [19]. In this case we obtain the following boundary conditions:

$$u_{\perp} \pm 2(gH)^{\frac{1}{2}} = f^{R}(t) , \tag{4.4-8a}$$

$$u_{//} = 0 , \tag{4.4-8b}$$

$$\frac{\partial}{\partial n} u_{//} = 0 . \tag{4.4-8c}$$

As mentioned before, Riemann invariants are quantities that are not measured in nature. In order to profit from the stabilizing effect of Riemann invariants while velocities are still prescribed as boundary conditions, we propose the following boundary conditions:

$$u_{\perp} + \varepsilon \frac{\partial}{\partial t} [u \pm 2 (gH)^{\frac{1}{2}}] = f^{u}(t) , \qquad (4.4\text{-}9a)$$

$$u_{//} = 0 , \qquad (4.4\text{-}9b)$$

$$\frac{\partial}{\partial n} u_{//} = 0 . \qquad (4.4\text{-}9c)$$

For sufficiently small values of ε (4.4-9a) is an accurate approximation of (4.4-1a). Yet with respect to very short wave lengths, which are produced for example by nonlinear effects inside the domain of the initial boundary value problem, (4.4-9a) acts as non-reflective boundary condition.
Numerical experiments carried out by the author confirmed this conjecture.

4.5 Tidal flats

In this section we describe the treatment of land-water boundaries for which the location is a function of the water level. The location of these boundaries is implicitly given by the following relation:

$$\zeta + h = 0 \qquad (4.5\text{-}1)$$

If ζ (x,y,t) varies as a function of time then the location of tidal flat boundaries varies as well, depending on the shape of the bottom profile, see e.g. figure (4-11).

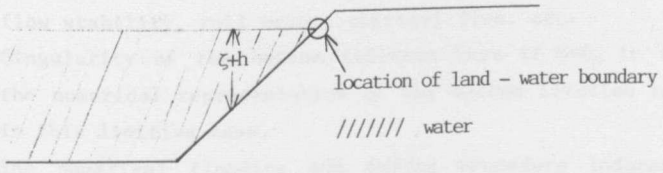


Figure (4-11) 1D Bottom profile with varying land-water boundary

For the fixed 1-dimensional grid of figure (4-12) a continuously varying position of the boundary is not possible. As discretization for the situation of figure (4-11) the bottom profile of figure (4-12) is assumed.

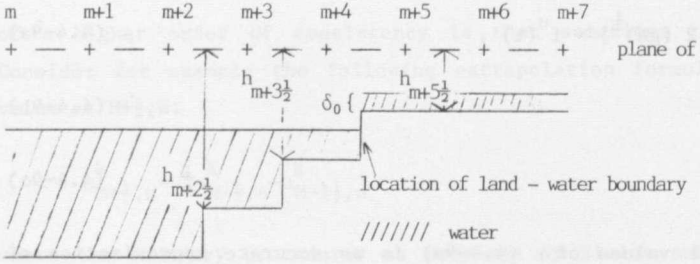


Figure (4-12) Discretized 1-D bottom profile .

For the bottom profile of figure (4-12) the boundary conditions are given by:

$$u_{m+\frac{1}{2}}=0, \text{ if } h_{m+\frac{1}{2}}+\tfrac{1}{2}\,(\zeta_m+\zeta_{m+1}) \leq \delta_o \qquad (4.5\text{-}2)$$

where $h_{m+\frac{1}{2}}$ denotes the bottom depth below some plane of reference.

For the two-dimensional grid of figure (4-13) the boundary conditions are given by:

$$u_{m+\frac{1}{2},n} = 0, \text{ if } H_{m+\frac{1}{2},n} \leq \delta_o \qquad (4.5\text{-}3a)$$

$$v_{m,n+\frac{1}{2}} = 0, \text{ if } H_{m,n+\frac{1}{2}} \leq \delta_o \qquad (4.5\text{-}3b)$$

These equations intrinsically assume a bottom profile as given by figure (4-13).

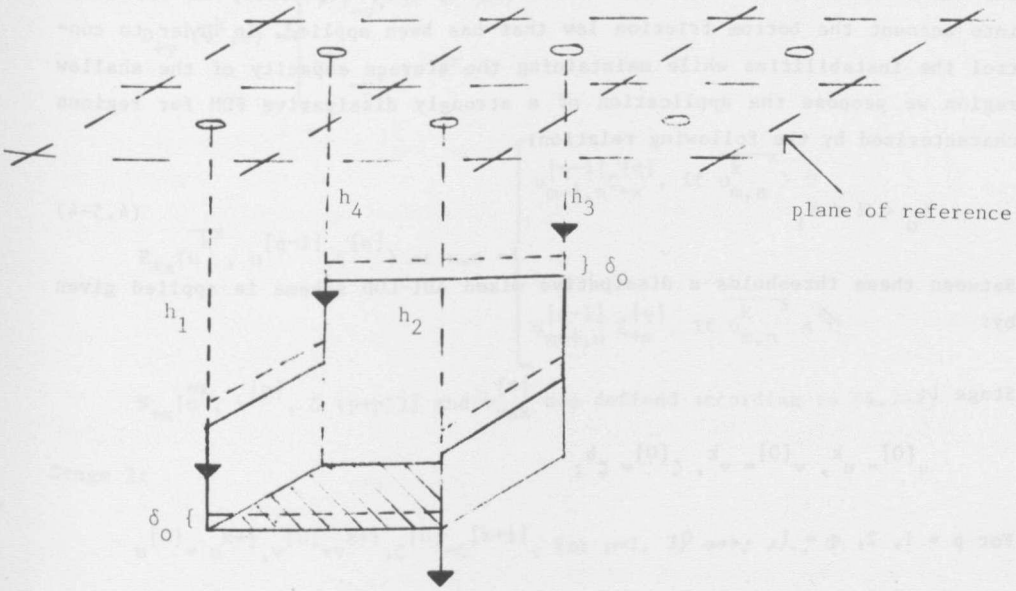


Figure (4-13) Discretized 2-D bottom profile

In very shallow regions there are several sources that might destabilize a FDM. Examples of destabilizing sources are given by:

(i) Singularities due to the intersection of characteristics. See Abbott [1] for a discussion on characteristics and Liggett [17] for a discussion on flow stability, roll waves, critical flow, etc.

(ii) Singularity of the bottom friction term if $H \to 0$. It is questionable if the numerical representation of the bottom friction term remains stable in this limiting case.

(iii) The numerical flooding and drying procedure induces disturbances by suddenly changing the status of a velocity point from flowing, i.e. $u,v \neq 0$, to dry, i.e., $u,v=0$ and vice versa.

In general the flow distribution of the shallow regions is not very important for the flow distribution of the deep regions. Considered as a storage basin, however, the shallow regions may be important. Therefore δ_o should be chosen as small as possible. Very small values of H facilitate the onset of instabilities, however. Moreover it is questionable whether at very shallow regions

the water movement is described accurately by SWE, especially if one takes into account the bottom friction law that has been applied. In order to control the instabilities while maintaining the storage capacity of the shallow region we propose the application of a strongly dissipative FDM for regions characterized by the following relation:

$$\delta_o < H < \delta_1 \tag{4.5-4}$$

Between these thresholds a dissipative mixed ADI-LOD scheme is applied given by:

Stage 1:

$$u^{[0]} = u^k,\ v^{[0]} = v^k,\ \zeta^{[0]} = \zeta^k;$$

For p = 1, 2, q = 1, ..., Q:

$$(u^{[q]} - \tilde{u}^k)/\tfrac{1}{2}\tau - fv^{=k+\frac{1}{2}} + 2g\zeta^{[q]}_{ox} + gu^{[q]}[(v^{=k+\frac{1}{2}})^2 + (\tilde{u}^k)^2]^{\frac{1}{2}}/c^2\delta_1$$

$$-\nu\,(u^k_{oxx} + u^k_{oyy}) = 0,\ \text{at } m+\tfrac{1}{2},\ n \tag{4.5-5a}$$

$$(v^{[p]} - \tilde{v}^k)/\tfrac{1}{2}\tau + 2\,S_{+x}[u^{=k},\ v^{[p]},\ \delta(p+p')] + 2D_{+y}(\tilde{v}^k, v) + f\,u^{=k}$$

$$+ g\,v^{[p]}[(u^{=k})^2 + (\tilde{v}^k)^2]^{\frac{1}{2}}/\delta_1 - \nu(v^{[x]}_{oxx} + v^{[p]}_{oyy}) = 0 \text{ at } m, n+\tfrac{1}{2} \tag{4.5-5b}$$

$$(\zeta^{[q]} - \tilde{\zeta}^k)/\tfrac{1}{2}\tau + (\overline{h}^y u^{[q]})_{ox} + \zeta^{[q-1]} u^{[q]}_{ox} + E_{+x}(\overline{u^k}^x, u^{[q-1]}, \zeta^{[q]}) + (H^k v^k)_{ox} = 0 \text{ at } m, n \tag{4.5-5c}$$

$$u^{k+1} = u^{[Q]},\ v^{k+\frac{1}{2}} = v^{[2]},\ \zeta^{k+\frac{1}{2}} = \zeta^{[Q]}$$

where:

$$\tilde{\zeta}^k_{m,n} = (1-4\gamma)\zeta^k_{m,n} + \gamma(\zeta^k_{m,n+1} + \zeta^k_{m+1,n} + \zeta^k_{m-1,n} + \zeta^k_{m,n-1})$$

$\tilde{u}^k$ en $\tilde{v}^k$ are defined similarly to $\tilde{\zeta}^k$

$$D_{+y}(\tilde{v}^k, v) = \begin{cases} \tilde{v}^k v_{-y}, & \text{if } \tilde{v}^k > 0 \\ \tilde{v}^k v_{-y}, & \text{if } \tilde{v}^k < 0 \end{cases}$$

$$E_{+x}(\overline{u^k}^x, u^{[q-1]}, \zeta^{[q]}) \text{ at } m,n = \begin{cases} u^{[q-1]}_{m-\frac{1}{2},n} \zeta^{[q]}_{-x}, & \text{if } \overline{u^k_{m,n}}^x > 0 \\ u^{[q-1]}_{m+\frac{1}{2},n} \zeta^{[q]}_{+x}, & \text{if } \overline{u^k_{m,n}}^x \leq 0 \end{cases}$$

$S_{+x}[\overline{u}^k, v^{[p]}, \delta(p+p')]$ and $v^{[*]}_{oxx}$ are defined according to (4.2-4)

Stage 2:

$$u^{[0]} = u^{k+\frac{1}{2}}, v^{[0]} = v^{k+\frac{1}{2}}, \zeta^{[0]} = \zeta^{[k+\frac{1}{2}]}, \text{ for } p=1, 2, q=1, \ldots, Q:$$

$$(u^{[p]} - \tilde{u}^{k+\frac{1}{2}})/\tfrac{1}{2}\tau - f\overline{v}^{k+\frac{1}{2}} + 2S_{+y}[\overline{v}^{k+\frac{1}{2}}, u^{[p]}, \delta(p+p')] + 2D_{+x}(\tilde{u}^{k+\frac{1}{2}}, u)$$

$$g u^{[p]} [(\overline{v}^{k+\frac{1}{2}})^2 + (\tilde{u}^{k+\frac{1}{2}})^2]^{\frac{1}{2}}/C^2\delta_1 - \nu(u^{[p]}_{oxx} + u^{[*]}_{oyy}) = 0, \text{ at } m+\tfrac{1}{2}, n \qquad (4.5\text{-}5d)$$

$$(v^{[q]} - \tilde{v}^{k+\frac{1}{2}})/\tfrac{1}{2}\tau + f\overline{u}^{k+1} + 2g\zeta^{[q]}_{oy} + g v^{[q]} [(\overline{u}^{k+1})^2 + (\tilde{v}^{k+\frac{1}{2}})^2]^{\frac{1}{2}}/C^2\delta_1$$

$$-\nu(v^{k+\frac{1}{2}}_{oyy} + v^{k+\frac{1}{2}}_{oxx}) = 0, \text{ at } m, n+\tfrac{1}{2} \qquad (4.5\text{-}5e)$$

$$(\zeta^{[q]} - \zeta^{k+\frac{1}{2}})/\tfrac{1}{2}\tau + (H^{k+\frac{1}{2}} u^{k+\frac{1}{2}})_{ox} + (\overline{h}^x v^{[q]})_{oy} + \zeta^{[q-1]} v^{[q]}_{oy} + E_{+y}(\overline{v^{k+\frac{1}{2}}}^y, v^{[q-1]}, \zeta^{[q]}) = 0$$
$$\text{at } m,n \qquad (4.5\text{-}5f)$$

$$u^{k+1} = u^{[2]}, \quad v^{k+1} = v^{[Q]}, \zeta^{k+1} = \zeta^{[Q]}$$

where $\tilde{u}^{k+\frac{1}{2}}$, $\tilde{v}^{k+\frac{1}{2}}$ and $\tilde{\zeta}^{k+\frac{1}{2}}$ are defined similarly to ζ^k,

$$D_{+x}(\tilde{u}^{k+\frac{1}{2}}, u) = \begin{cases} \tilde{u}^{k+\frac{1}{2}} u_{-x}, & \text{if } \tilde{u}^{k+\frac{1}{2}} > 0 \\ \tilde{u}^{k+\frac{1}{2}} u_{+x}, & \text{if } \tilde{u}^{k+\frac{1}{2}} \leq 0 \end{cases}$$

$$E_{+y}(\overline{v^{k+\frac{1}{2}}}^{y}, v^{[q-1]}, \zeta^{[q]}) \text{ at } m,n = \begin{cases} v^{[q-1]}_{m,n-\frac{1}{2}}\ \zeta^{[q]}_{-y}, & \text{if } \overline{v^{k+\frac{1}{2}}}^{y}_{m,n} > 0 \\ v^{[q-1]}_{m,n+\frac{1}{2}}\ \zeta^{[q]}_{+y}, & \text{if } \overline{v^{k+\frac{1}{2}}}^{y}_{m,n} \leq 0 \end{cases}$$

$S_{+y}\ [v^{=k+\frac{1}{2}}, u^{[p]}, \delta(p+p')]$ and $u^{[*]}_{oyy}$ are defined according to (4.2-4).

The equations (4.5-5c) and (4.5-5f) are conservative with respect to mass because of the following relations:

$$\zeta\ u_{ox} + E_{+x}(U,u,\zeta) \text{ at } m,n = \begin{cases} (\zeta_{m,n}\ u_{m-\frac{1}{2},n})_{ox}, & U > 0 \\ (\zeta_{m,n}\ u_{m+\frac{1}{2},n})_{ox}, & U \leq 0 \end{cases}$$

$$\zeta\ v_{oy} + E_{+y}(V,v,\zeta) \text{ at } m,n = \begin{cases} (\zeta_{m,n}\ v_{m,n-\frac{1}{2}})_{oy}, & V > 0 \\ (\zeta_{m,n}\ v_{m,n+\frac{1}{2}})_{oy}, & V \leq 0 \end{cases}$$

Numerical experiments showed the excellent stability properties of (4.5-5). Note that for the large majority of practical applications (4.5-5) is not necessary.

4.6 On the structure of the implicit equations

For a numerical method not only is stability an important aspect but also the solution methods that are used should not be over-sensitive to rounding errors, see e.g. Wilkinson [28].
The solution method we will use is a simple recursive algorithm to solve equations with a tri-diagonal coefficient matrix. By this method a tri-diagonal system of equations given by:

$$a_m z_{m-1} + b_m z_m + c_m z_{m+1} = Z_m,\ m = 1, 2, \ldots, M \qquad (4.6\text{-}1)$$

where $a_1 = 0$, $c_M = 0$

is reduced to a bi-diagonal system of equations given by:

$$z_m + X_m z_{m+1} = Y_m, \quad m = 1, 2, \ldots M-1$$
$$z_M = Y_M \tag{4.6-2}$$

where the coefficients X_m and Y_m can be calculated by means of simple recursion formulae given by:

$$x_1 = c_1/b_1, \quad Y_1 = Z_1/b_1$$
$$X_m = c_m/(b_m - a_m X_{m-1}), \quad Y_m = (-a_m Y_{m-1}+Z_m)/(b_m-a_m X_{m-1}), \quad m=2,\ldots,M \tag{4.6-3}$$

The final solution of (4.6-1) is obtained by backward substitution of (4.6-3) as follows:

$$z_M = Y_M$$
$$z_m = Y_m - X_m z_{m+1}, \quad m = M-1, M-2, \ldots, 1 \tag{4.6-4}$$

This method is described by many authors, see Godunov and Ryabenki [6] or Isaacson and Keller [11].

To prevent the amplification of rounding errors the following relation must hold:

$$|X_m| \leq 1, \quad m = 1, \ldots, M \tag{4.6-5}$$

A sufficient condition to fulfil (4.6-5) is given by:

$$|b_m| \geq |a_m| + |c_m|, \quad m = 1, 2, \ldots, M \tag{4.6-6}$$

as can be easily verified.

We will study the structure of the equations given by (4.2-4a) and (4.2-4b). It is only necessary to consider the coefficients of the implicit part, which is of the following form:

$$-c_{m-\frac{1}{2}}\,\zeta_{m-1} + u_{m-\frac{1}{2}} + c_{m-\frac{1}{2}}\,\zeta_m = R_{m-\frac{1}{2}}$$

$$-C_{m-\frac{1}{2}}\,\zeta_{m-1} - \bar{B}_m u_{m-\frac{1}{2}} + (1 + C_{m-\frac{1}{2}} - C_{m+\frac{1}{2}})\,\zeta_m + \bar{B}_m^{+} u_{m+\frac{1}{2}} + C_{m+\frac{1}{2}}\,\zeta_{m+1} = R_m \qquad (4.6\text{-}7)$$

$$-c_{m+\frac{1}{2}}\,\zeta_m + u_{m+\frac{1}{2}} + c_{m+\frac{1}{2}}\,\zeta_{m+1} = R_{m+\frac{1}{2}}$$

where:

$$c_m = \frac{\tau}{2\Delta x}\, g/\{1+\tfrac{\tau}{2}(u^2+\bar{v}^2)^{\frac{1}{2}}/[C^2(\bar{h}^y+\bar{\zeta}^x)] + \frac{\tau}{4\Delta x}\,(u_{m+\frac{1}{2}} - u_{m-\frac{1}{2}})\} \text{ at } m,n$$

$$C_m = \frac{\tau}{4\Delta x}\, u_m$$

$$\bar{B}_m = \frac{\tau}{2\Delta x}\,(\zeta_m + \bar{h}^y_{m-\frac{1}{2}}) \quad \text{and}$$

$$\bar{B}_m^{+} = \frac{\tau}{2\Delta x}\,(\zeta_m + \bar{h}^y_{m+\frac{1}{2}}).$$

Equation (4.6-7) is penta-diagonal, but by direct substitution it can be reduced to:

$$-A_m\zeta_{m-1} + (1+A_m+B_m)\,\zeta_m - B_m\zeta_{m+1} = R_m + \bar{B}_m R_{m-\frac{1}{2}} - \bar{B}_m^{+} R_{m+\frac{1}{2}} \qquad (4.6\text{-}8)$$

where:

$$A_m = \frac{\tau}{4\Delta x}\, u_{m-\frac{1}{2}} + (\frac{\tau}{2\Delta x})^2 \frac{g(\zeta_m + \bar{h}^y_{m-\frac{1}{2}})}{1+\frac{\tau}{2}g(u^2+\bar{v}^2)^{\frac{1}{2}}/[C^2(\bar{h}^y+\bar{\zeta}^x)] + \frac{\tau}{4\Delta x}(u_{m+\frac{1}{2}} - u_{m-\frac{1}{2}})}$$

$$B_m = -\frac{\tau}{4\Delta x}\, u_{m+\frac{1}{2}} + (\frac{\tau}{2\Delta x})^2 \frac{g\,(\zeta_m + \bar{h}^y_{m+\frac{1}{2}})}{1+\frac{\tau}{2}g(u^2+\bar{v}^2)^{\frac{1}{2}}/[C^2(\bar{h}^y+\bar{\zeta}^x)] + \frac{\tau}{4\Delta x}\,(u_{m+\frac{1}{2}} - u_{m-\frac{1}{2}})}$$

A sufficient condition to fulfil (4.6-6) is given by:

$$\frac{\tau}{\Delta x}\,|u_{m+\frac{1}{2}}| < 2 \qquad (4.6\text{-}9)$$

where it has been assumed that $\zeta_m + \overline{h}^y_{m+\frac{1}{2}} > 0$.

In stead of verification of (4.6-6), (4.6-5) can be verified directly as well. Suppose that (4.6-8) has been reduced up to the $(m-1)^{th}$ equation and is written in the following form:

$$\zeta_{m-1} + X_{m-1}\,\zeta_m = Y_m \tag{4.6-10}$$

From (4.6-8) and (4.6-10) it follows that:

$$X_m = -B_m/(1+A_m+B_m+A_m X_{m-1}) \tag{4.6-11}$$

Suppose that the boundary condition at $m = 1$ is such that $|X_1| \leq 1$, and assume that $|X_{m-1}| \leq 1$, if for X_m as given by (4.6-11) the relation given by (4.6-5) holds by induction, the existence of this relation has been proven. By this procedure we can prove that the following conditions are sufficient for (4.6-5):

$$\frac{\tau}{\Delta x}\,|u_{m+\frac{1}{2}}| < 4, \text{ if } u_{m-\frac{1}{2}}\,u_{m+\frac{1}{2}} \geq 0 \tag{4.6-12a}$$

$$\frac{\tau}{\Delta x}\,|u_{m+\frac{1}{2}}| < 2, \text{ if } u_{m-\frac{1}{2}}\,u_{m+\frac{1}{2}} < 0 \tag{4.6-12b}$$

The condition given by (4.6-12a) is not as restrictive as (4.6-9) while (4.6-12b) is just as restrictive. Because (4.6-12b) has to be satisfied only if the velocity u changes its sign and hence is small, it seems reasonable to assume that (4.6-12a) imposes the real restriction. By a similar analysis one can prove that in order to control amplification of rounding errors for (4.2-4d), (4.6-12) must hold as well.

If we analyse the FDM for the very shallow regions it follows that for (4.5-5), (4.6-5) is always satisfied.

For equations in the y direction obviously similar conditions must be fulfilled. It follows that in order to control the amplification of rounding errors the following relations are to be satisfied:

$$\text{Max}\left(|u_{m+\frac{1}{2},n}|\,\frac{\tau}{\Delta x},\ |v_{m,n+\frac{1}{2}}|\,\frac{\tau}{\Delta y}\right) \leq 4 \tag{4.6-13}$$

The condition given by (4.6-13) seems to be the only restriction for the time step for the FDM treated in this chapter. This restriction disappears by the

application of partial pivoting, but it increases the computational effort. By application of nonlinear extensions (2.3-7) this restriction can be circumvented. For many applications, however, (4.6-13) is not very restrictive; accuracy considerations, see e.g. section 3.5, will often impose more severe restrictions.

For the boundary schemes described in this section the condition given by (4.6-5) is satisfied. The effects of the discretizations of vu_y, uv_x, $\nu(u_{xx}+u_{yy})$ and $\nu(v_{xx}+v_{yy})$ have been neglected for this analysis. These discretizations only contribute to reducing the restrictions given by (4.6-13), as can easily be verified.

4.7 Concluding Remarks

In this chapter a numerical method has been constructed based upon nonlinear extension of a linear method. The advantage of first considering the linearized equations is that linear stability and efficiency can be studied first and then the nonlinear aspects of a FDM.

There are no arguments to support the opinion that fully nonlinear integration by the trapezoidal rule is more accurate than a locally linearized integration. Both methods are second order accurate. In fact, a simple nonlinear example could be constructed that is integrated exactly by local linearization.

It is important that the boundary treatment is such that the amount of computational control remains bounded in order to minimize the overhead.

At inflow boundaries it seems difficult to construct a stable boundary advection treatment with an order of consistency greater than zero.

For practical applications zero order of consistency for the advection operator near inflow boundaries seems to be sufficiently accurate. Stability is more important near open boundaries.

Application of a dissipative FDM in very shallow regions improves the robustness of a FDM without effecting the overall accuracy.

The matrix structures of the implicit equations are such that in order to control the amplification of rounding errors the maximum timestep might be restricted. For practical applications these restrictions are not severe; moreover, they can be circumvented by application of the unconditionally stable Angled Derivative method or by partial pivoting.

REFERENCES TO CHAPTER 4

1. ABBOTT, M.B.,
Weak Solution of the Equations of Open Channel Flow, in Unsteady Flow in Open Channels, edited by K. Mabwood and V. Yevjevich, Water Resources Publications, Fort Collins, 1975.

2. BEAM, R.M., R.F. WARMING and H.C. YEE,
Stability Analysis of Numerical Boundary Conditions and Implicit Difference, Approximations for Hyperbolic Equations,
Journal of Computational Physics 48, 1982, pp. 200-222.

3. ELVIUS, T. and A. SUNDSTROM,
Computationally Efficient Schemes and Boundary Conditions for a Finemesh Barotropic Model based on the Shallow Water Equations,
Tellus XXV, 25, 1973, pp. 132-156.

4. ENQUIST, B. and A. MAJDA,
Absorbing Boundary Conditions for the Numerical Simulation of Waves,
Mathematics of Computation, 31, 1977, pp. 629-651.

5. GERRITSEN, H.,
Accurate Boundary Treatment in Shallow Water Flow Computations,
Thesis, T.H. Twente, 1982.

6. GODUNOV, S.K. and V.S. RYABENKI,
Theory of Difference Schemes,
North Holland Publishing Company, Amsterdam, 1964.

7. GOLDBERG, M., and E. TADMOR,
Scheme Independent Stability Criteria for Difference Approximations of Hyperbolic Initial Boundary Value Problems I,
Mathematics of Computation, 32, 1978, pp. 1097-1107.

8. GOLDBERG, M. and E. TADMOR,
Scheme Independent Stability Criteria for Difference Approximations of Hyperbolic Initial Boundary Value Problems II,
Mathematics of Computation, 36, 1981, pp. 603-626.

9. GOTTLIEB, D., M. GUNZBURGER and E. TURKEL,
On Numerical Boundary Treatment of Hyperbolic Systems for Finite Difference and Finite Element Methods,
SIAM, J. Numer Anal, 4, 1982, pp. 671-682.

REFERENCES (continued)

10. GUSTAFSSON, B.,
The Convergence Rate for Difference Approximation to Mixed Initial Boundary Value Problems,
Mathematics of Computation, 36, 1975, pp. 396-406.

11. ISAACSON, E., and H.B. KELLER,
Analysis of Numerical Methods,
John Wiley and Sons, London, 1966.

12. JEFFREY, A.,
Quasilinear Hyperbolic Systems and Waves,
Pitman Publishing, London, etc. 1976.

13. KREISS, H.O.,
Initial Boundary Value Problems for Hyperbolic Systems,
Communications on Pure and Applied Mathematics, XXIII, 1970, pp. 277-298.

14. KREISS, H.O. and J. OLIGER,
Methods for the Approximate Solution of Time Dependent Problems,
GARP Publication Series, 10, Geneva 1973.

15. KREISS, H.O. and B. GUSTAFSSON,
Boundary Conditions for Time Dependent Problems with an Artificial Boundary, Journal of Computational Physics, 30, 1979, pp. 333-351.

16. KUTLER, P., (editor)
Numerical Boundary Condition Procedures,
NASA Conference Publications 2201, NASA Ames Research Center, Moffett Field, 198

17. LIGGETT, J.A.,
Stability, in Unsteady Flow in Open Channels, edited by K. Mahmood and V. Yevjevich, Water Recources Publications, Fort Collins, 1975.

18. MORETTI, G.,
A Physical Approach to the Numerical Treatment of Boundaries in Gas Dynamics,
Numerical Boundary Condition Procedures, NASA Conference Publication 2201, 1981.

19. OLIGER, J. and A. SUNDSTROM,
Theoretical and Practical Aspects of some Initial Boundary Value Problems in Fluid Dynamics,
SIAM, Journal Appl. Math., 35, 1978, pp 419-446.

REFERENCES (continued)

20. RICHTMYER, R.D. and K.W. MORTON,
Difference Methods for Initial Value Problems
Interscience Publishers, Wiley, New York-London, 1967.

21. STELLING, G.S.,
Vergelijking van getijberekeningen van de Zuidelijke Noordzee met behulp van verschillende Numerieke Modellen, Delft Hydraulics Laboratory, Information X60, 1979.

22. STRIKWERDA, J.,
Initial Boundary Value Problems for Incompletely Parabolic Systems,
Thesis, Stanford University, 1976.

23. TRAPP, J.A. and J.D. RAMSHAW,
A Simple and Heuristic Method for Analyzing the Effect of Boundary Conditions on Numerical Stability,
Journal of Computational Physics, 20, 1976, pp 238-242.

24. VERBOOM, G.K., G.S. STELLING and M.J. OFFICIER,
Non-reflective Boundary Conditions in Horizontal Flow Models,
Proceedings Int. Conf. Numerical Methods for Coupled Problems, Swansea, 1981.

25. VERBOOM, G.K., G.S. STELLING and M.J. OFFICIER,
Boundary Conditions for the Shallow Water Equations,
In Engineering Applications for Computational Hydraulics, Volume 1,
(M.B. Abbott and J.A. Cunge ed.) Pitman Publishing, 1982.

26. VERBOOM, G.K.,
Een getijberekening voor de zuidelijke Noordzee met verschillende numerieke modellen; vergelijking van de resultaten. Delft Hydraulics Laboratory, Report R1718.

27. VERBOOM, G.K.,
Weakly Reflective Boundary Conditions for the Shallow Water Equations,
Delft Hydraulics Laboratory, Publication no. 266, 1982.

28. WILKINSON, J.H.,
Rounding Errors in Algebraic Processes,
Her Majesty's Stationary Office, London, 1963.

5 Numerical experiments

5.0 Introduction

During the development of the FDM of chapter 4 many numerical experiments were made. These may be subdivided into two classes. The first class concerns numerical experiments with simple geometries. Even so, complicated flow patterns, containing eddies, are induced. The experiments in this class were performed to study the stability properties of the FDM, the effect of the size of the timestep, the sensitivity of the FDM to the variation of the value of viscosity, and the effect of perfect slip boundary conditions versus no slip boundary condition.
The second class concerns the application of the FDM for practical problems. These experiments were performed to study the flooding and drying procedure, the accuracy as function of the timestep, the ability to solve steady state problems, and stability.
The approach followed in this chapter is purely numerical, i.e., aspects concerning calibration with prototype measurements are considered to be beyond the scope of this work.
In section 1 the numerical experiments concerning eddies in simple geometries are described. The importance of advection, viscosity, and boundary conditions will be shown with respect to the resulting flow pattern.
Section 2 contains the description of practical experiments considering an estuary, the tidal inlet of an estuary, and a river section.
The textual explanation of this chapter is brief. The major part consists of figures that illustrate the numerical experiments.

5.1 Simple geometries

In this section the following aspects are treated: (i) stability of the advection discretization, (ii) effects of viscosity, and (iii) effects of perfect slip and no-slip boundary conditions.
These aspects were studied by means of two test problems with a simple geometry and a uniform depth.

a. Flow past a jetty

The geometry for this experiment is given by figure 5-1.

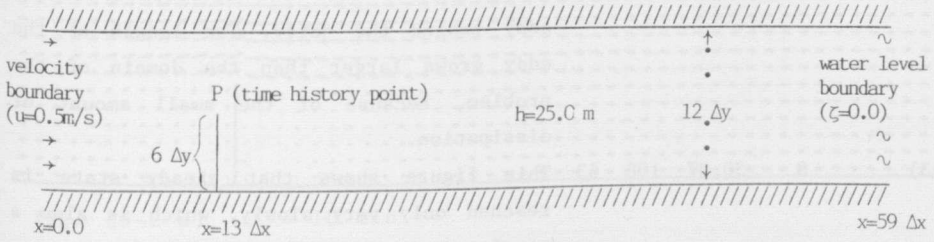


Figure 5-1 Geometry of flow past a jetty

The rectangular basin of figure (5-1) has a length of 59 Δx and a width of 12 Δy where $\Delta x = \Delta y = 25$ m. The depth is uniformly 25 m. A jetty is situated at $x=13\ \Delta x$. The length of the jetty is 6 Δy.
We consider a steady-state problem that is initially time-dependent. For this aim at $x=0$ a uniform velocity $u=0.5$ m/s is given as boundary condition and at $x=59\Delta x$ a uniform water level $\zeta=0$ is given as boundary condition.

Because of the jetty, the influence of advection is very important, as can be seen from comparison of the flow patterns of figure (5-2 g) and figure (5-4).

The numerical values of the parameters for figures (5-2) - (5-10) are given by table (5-1).

Table 5.1

Figure	ν	τ,Cf	ε	C	comment
(5-2 a-g)	0	30,27	100	63	These figures show the development of an eddy behind the jetty. The length of the eddy grows larger than the domain of the problem, because of the small amount of dissipation.
(5-3)	0	30,27	100	63	This figure shows that steady state is reached only very slowly, which is also a result of the small amount of dissipation.
(5-4)	0	30,27	100	63	This flow pattern is obtained if the advection terms are omitted. Comparison with figure (5-2 g) shows the importance of advection.
(5-5)	0	30,27	100	63	Absence of advection gives a real steady state solution, as this time history shows.
(5-6 a-g)	0	10,9	100	63	Development of the eddy for a smaller timestep. The final stage shows two eddies. The influence of the timestep is noticeable.
(5-7)	0	10,9	100	63	Steady state is reached slowly, but faster than for τ=30. Yet a time-dependent disturbance remains noticeable.
(5-8)	0	30,27	0	63	This standing wave superimposed on the numerical solution is obtained if both boundary conditions are purely reflective. Comparison with figure (5-3) shows the important influence of (4.4-9) when $\varepsilon \neq 0$.
(5-9)	10	30,27	100	63	Addition of viscosity shortens the length of the eddy as follows from comparison of figure (5-9) with (5-2-g).
(5-10)	10	30,27	100	63	Addition of dissipation gives very fast convergence to steady state.

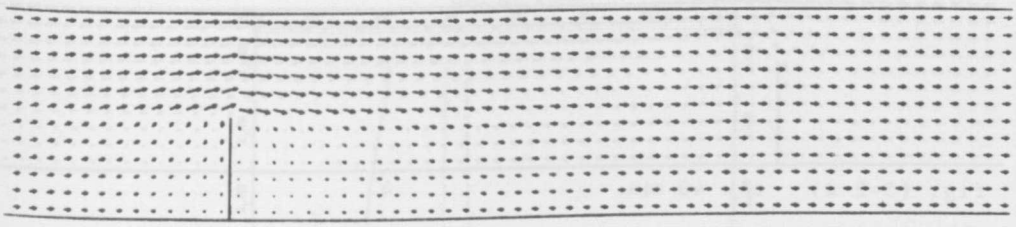


fig. (5-2 a) t = 10 min.

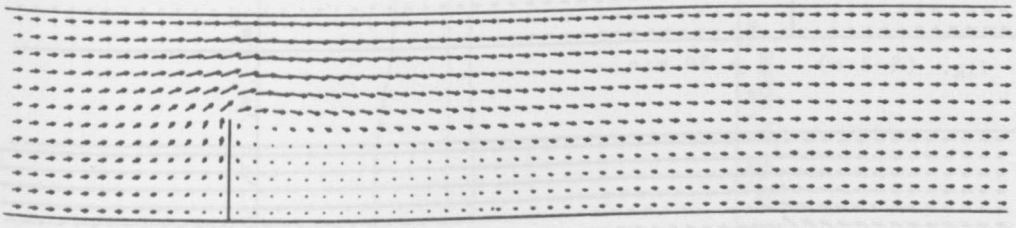


fig. (5-2 b) t = 20 min.

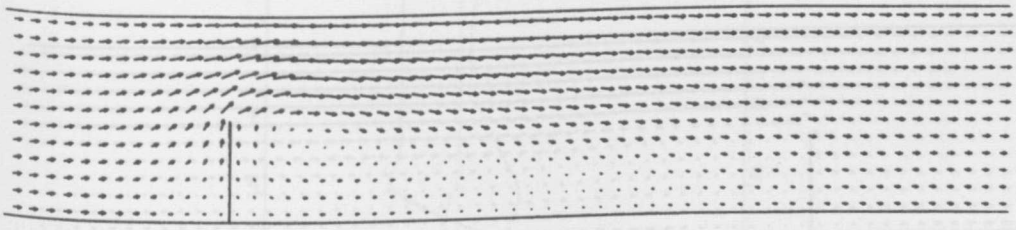


fig. (5-2 c) t = 30 min.

Figure (5-2) Development of eddy behind jetty

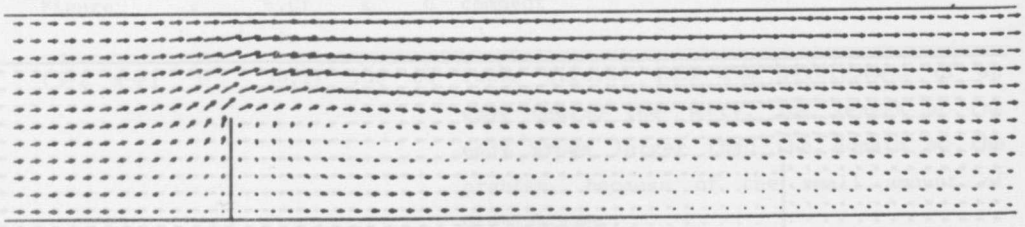


fig. (5-2 d) t = 40 min.

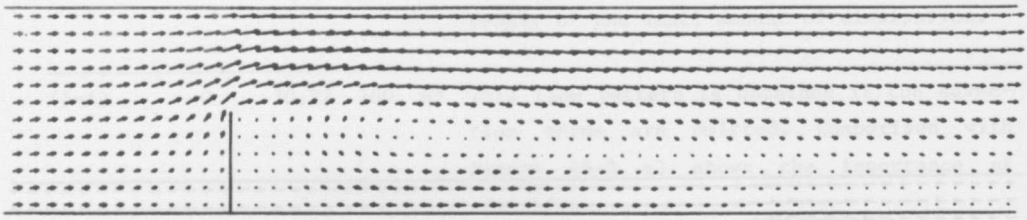


fig. (5-2 e) t = 50 min.

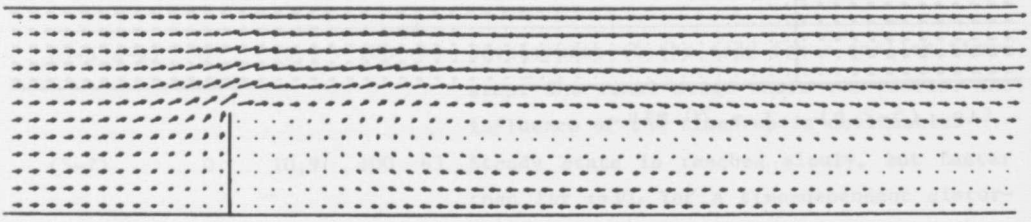


fig. (5-2 f) t = 60 min.

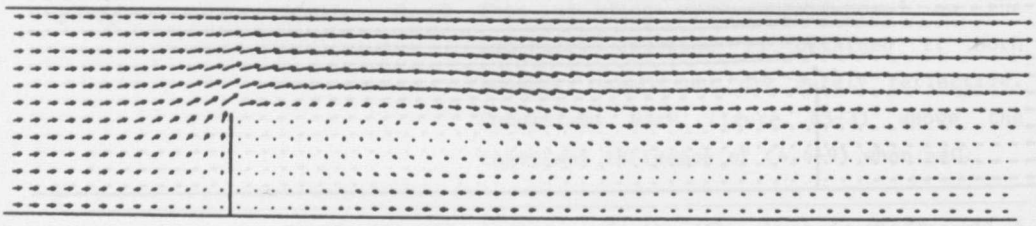


fig. (5-2 g) t = 180 min.

Figure (5-2) Development of eddy behind jetty

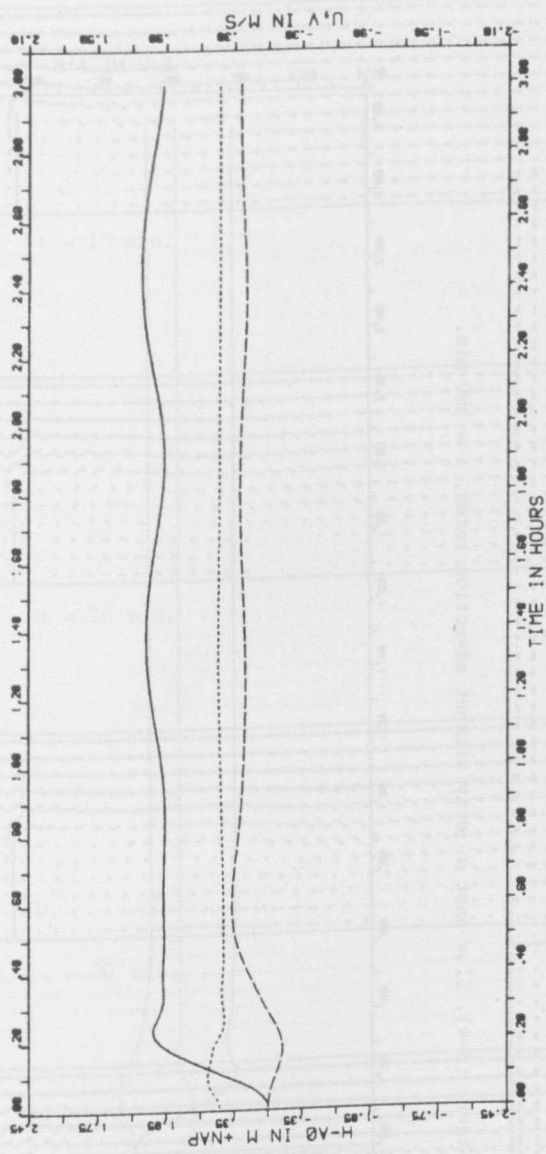


Figure (5-3) Time history at P (see fig. (5-1)) for flow past a jetty, ε=100

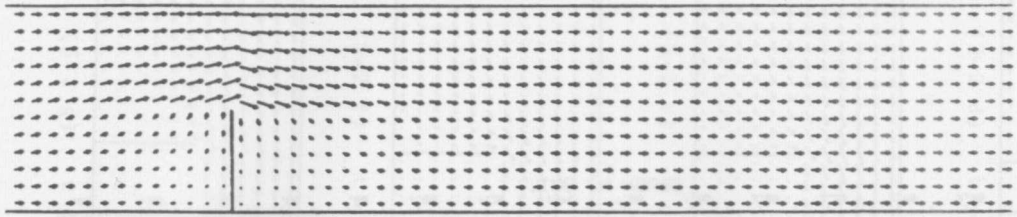


Figure (5-4), flow past a jetty without advection terms, t = 180 min.

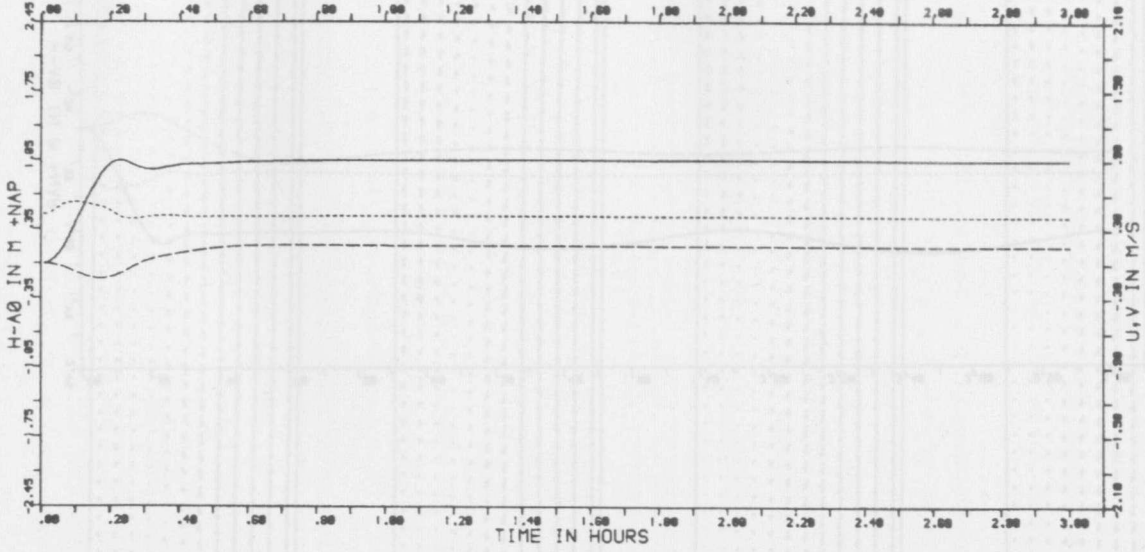


Figure (5-5) Time history at P (see (5-1)) for flow past a jetty without advection terms

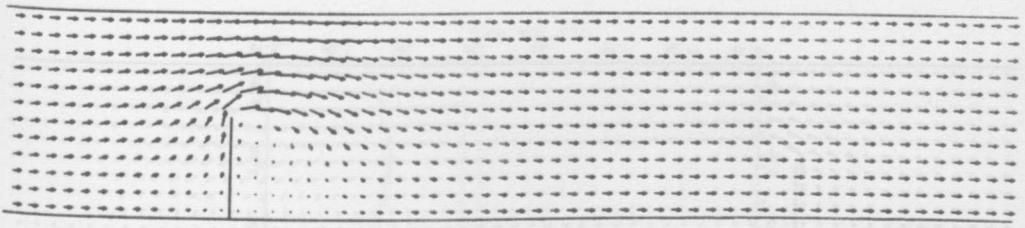


Fig. (5-6 a) t = 10 min.

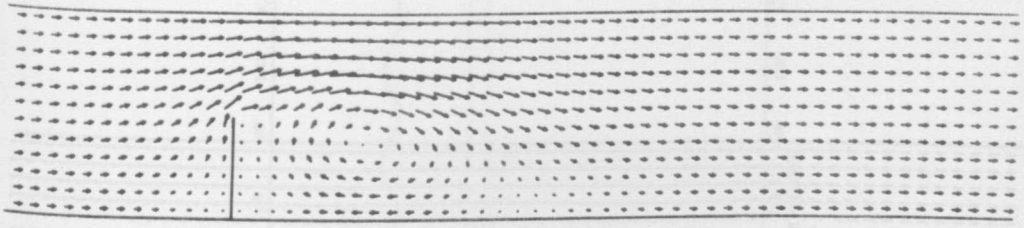


Figure (5-6 b) t = 20 min.

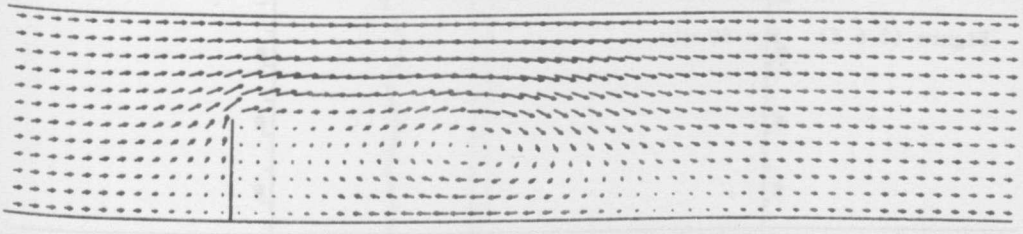


Figure (5-6 c) t = 30 min.

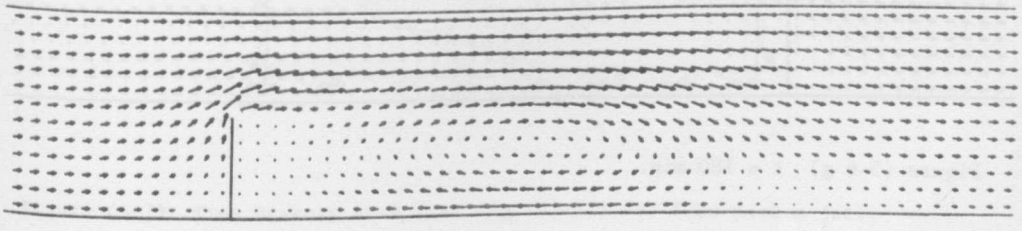


Figure (5-6 d) t = 40 min.

Figure (5-6) Development of eddy behind jetty

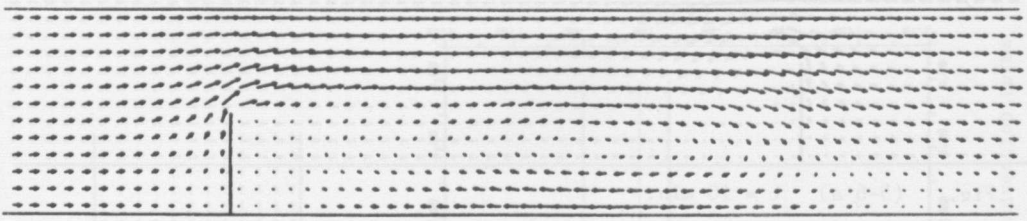


Figure (5-6 e) t = 50 min.

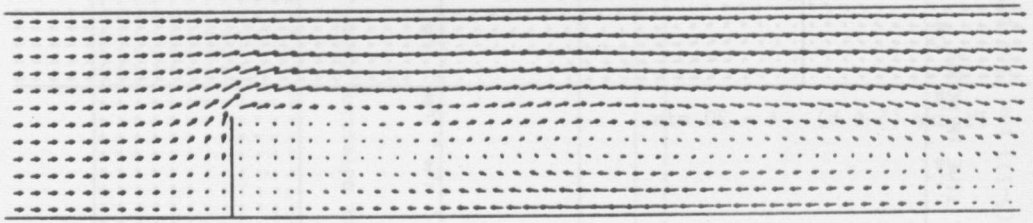


Figure (5-6 f) t = 60 min.

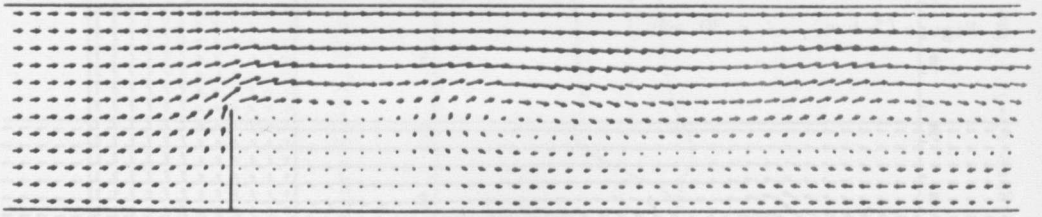


Figure (5-6 g) t = 180 min.

Figure (5-6) Development of eddy behind jetty

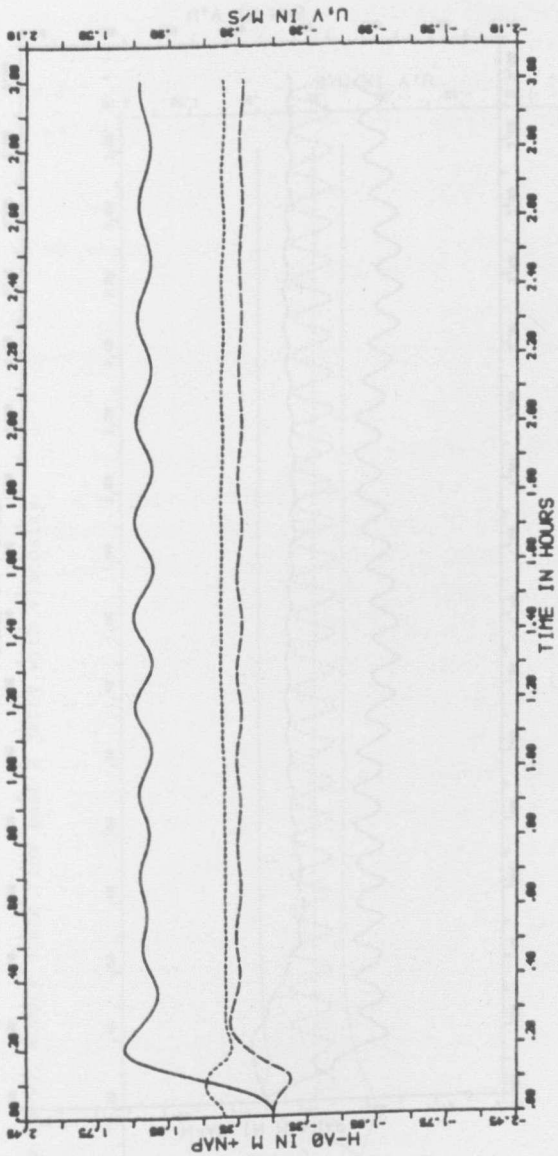


Figure (5-7) Time history at P (see fig. (5-1)) for flow past a jetty

ε = 100

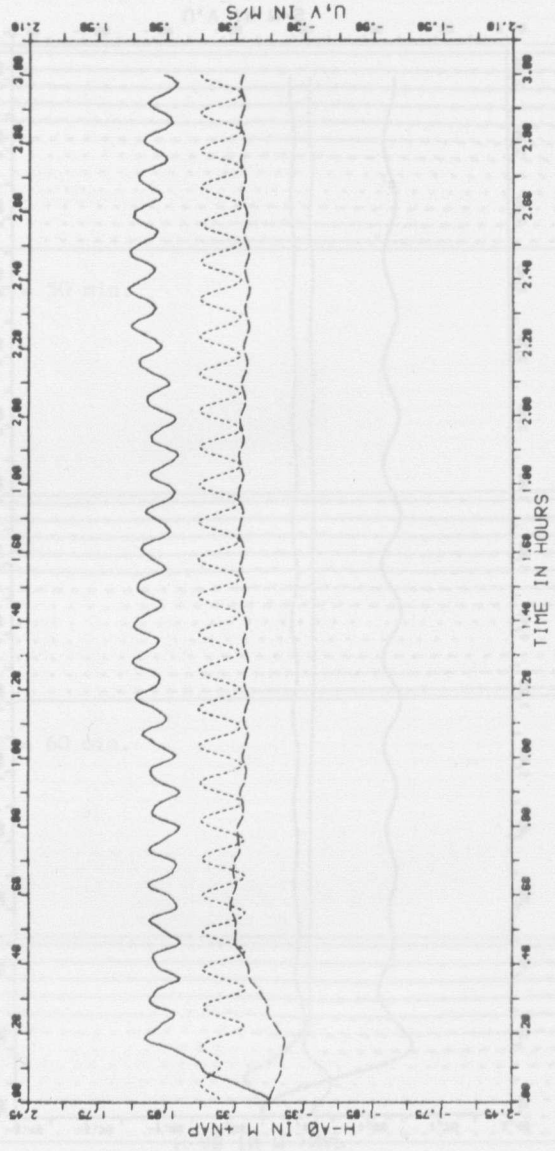


Figure (5-8) Time history at P, (see fig. (5-1)) for flow past a jetty with purely reflective boundary conditions
$\nu = 0$, $\tau = 30$, $\varepsilon = 0$

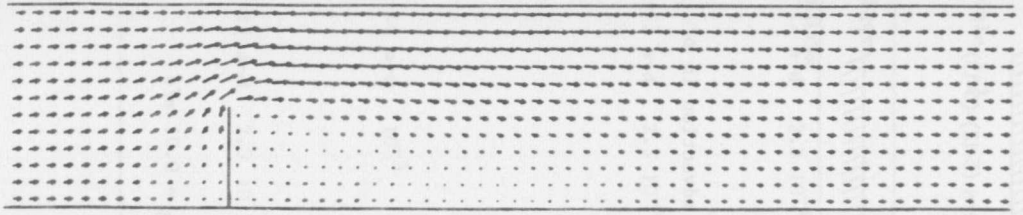


Figure (5-9), steady state flow past a jetty with viscosity

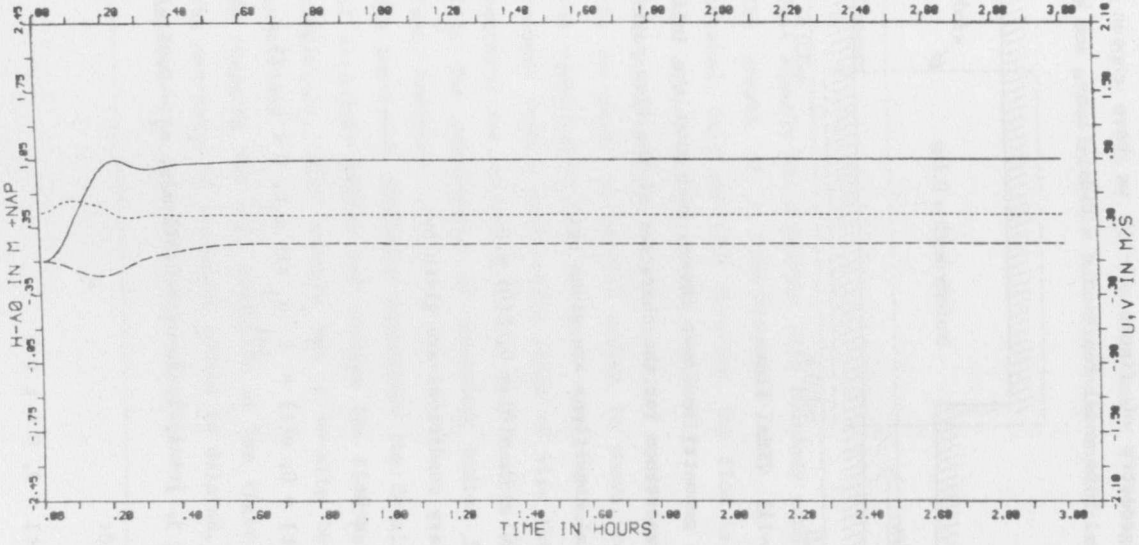


Figure (5-10) Time history at P (see fig. (5-1)) for flow past a jetty with viscosity

b) Flow past a backward step

For the geometry of figure (5-11) we have chosen time-dependent boundary conditions. This model represents a tidal flume, see Wang [4].

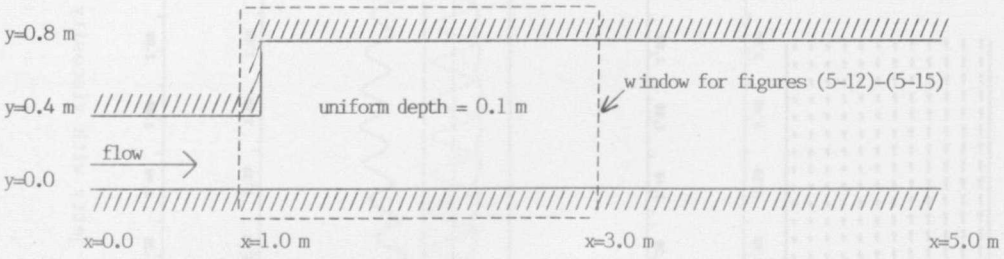


Figure (5-11) Tidal flume

Again the geometry has been chosen such that the influence of advection is of crucial importance for the character of the flow pattern.

The initial conditions are given by:

$$u(0) = 0,\ v(0) = 0,\ \zeta(0) = 0$$

The boundary conditions are given by:

At x=0:

$$v(t) = 0,\ u(t) = \sum_{j=1}^{3} \hat{u}_j \sin \omega_j t,\ 0 \leq t \leq 75 \text{ s}$$

where

$\omega_j = \frac{2\pi}{150} j$, $j = 1, 2, 3$, $\hat{u}_1 = 0.375$ m/s, $\hat{u}_2 = 0.05$ m/s, and $\hat{u}_3 = 0.01$ m/s

At x = 5.0:

$$\zeta(t) = 0,\ 0 \leq t < 5$$

$$\zeta(t) = \sum_{j=1}^{3} \zeta_j \sin \omega_j (t-5)\ 5 \leq t < 75$$

$\hat{\zeta}_1 = 0.021$, $\hat{\zeta}_2 = 0.001$ and $\hat{\zeta}_3 = 0.0005$

For the initial and boundary conditions given above several numerical values for ν and α were used. The results are illustrated by the figures (5-12) to (5-15). The grid size was $\Delta x = \Delta y = 0.025$ m. The timestep was $\tau = 0.125$ s. The numerical values for ν and α are given by table 5.2. The Chezy coefficient was $C = 62.64$.

Table 5.2

Figure	ν	α	Comment
5-12 a-d	$2.3\ 10^{-4}$	1	This example has a perfect slip boundary condition. The growth of a time-dependent eddy is demonstrated. Only one eddy develops. The flow in backward direction follows the rigid walls.
5-13 a-d	$2.3\ 10^{-4}$	0	This example has a no-slip boundary condition. Here the emergence of several eddies is shown. Despite the complicated flow patterns stability was maintained. From a qualitative point of view the flow patterns are according to the measurements of Wang [4]. The development of secondary eddies is too fast, however.
5-14 a-d	10^{-3}	0	The increased viscosity suppresses the development of secondary eddies and changes the flow patterns completely. This example has a no-slip boundary condition.
5-15 a-d	$2.3\ 10^{-4}$	0.1	By changing the slip condition at the rigid walls the emergence of secondary eddies is delayed, which improves the agreement with measurements.

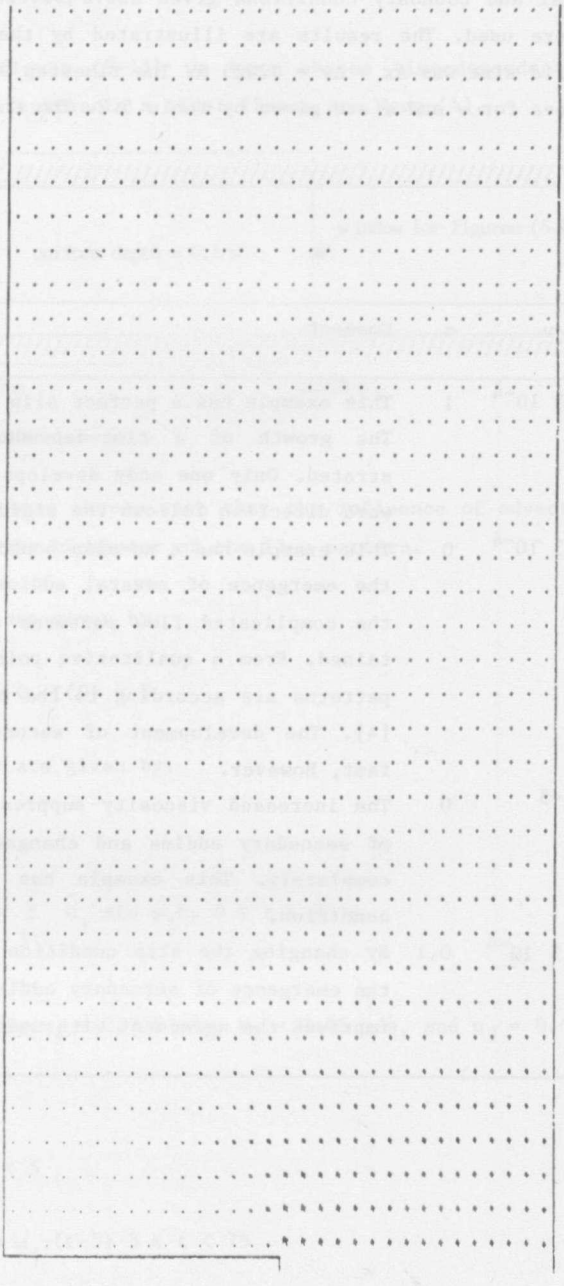


Figure (5-12 a) Development of time-dependent eddy in tidal flume, perfect slip boundary condition, t = 5 s

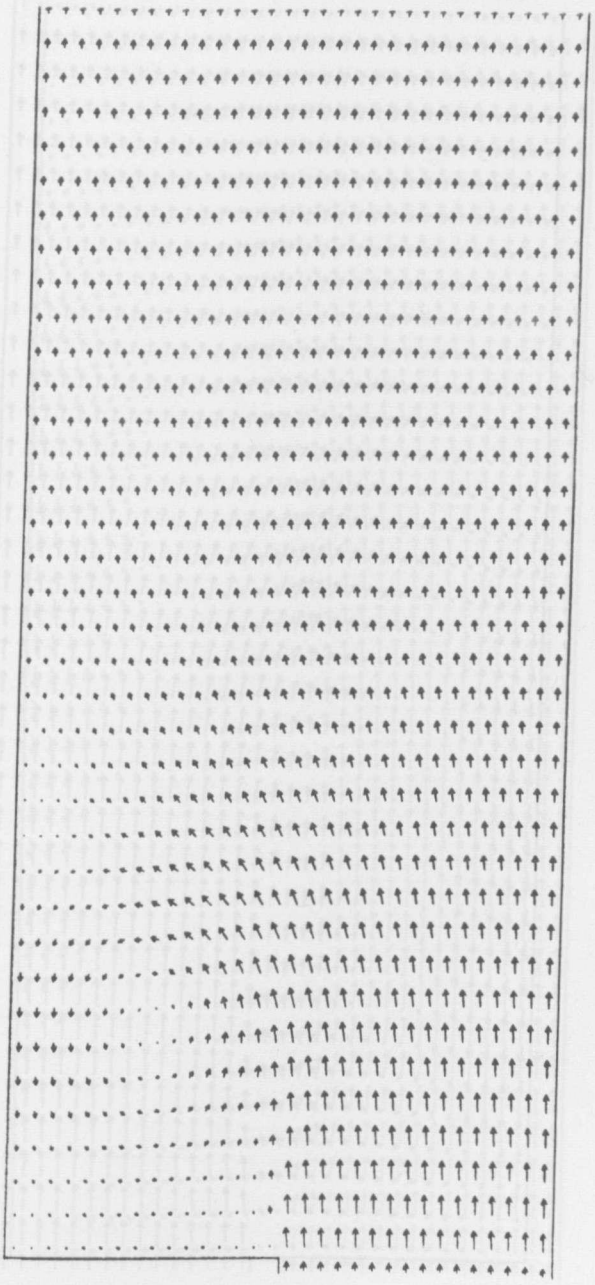


Figure (5-12-b) t = 15 s

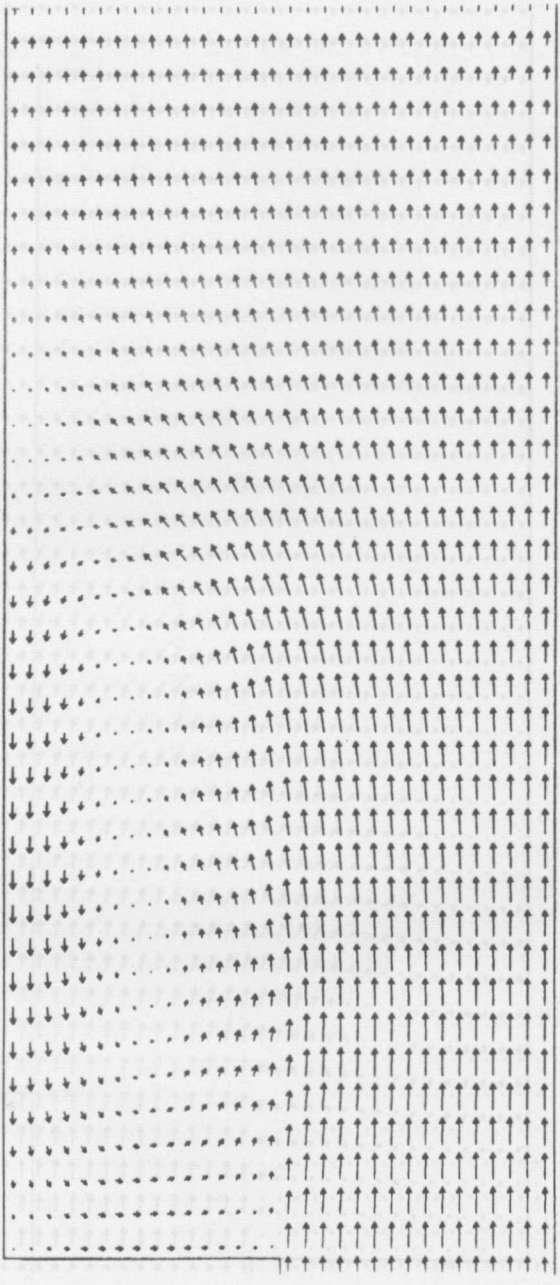


Figure (5-12-c) t = 25 s

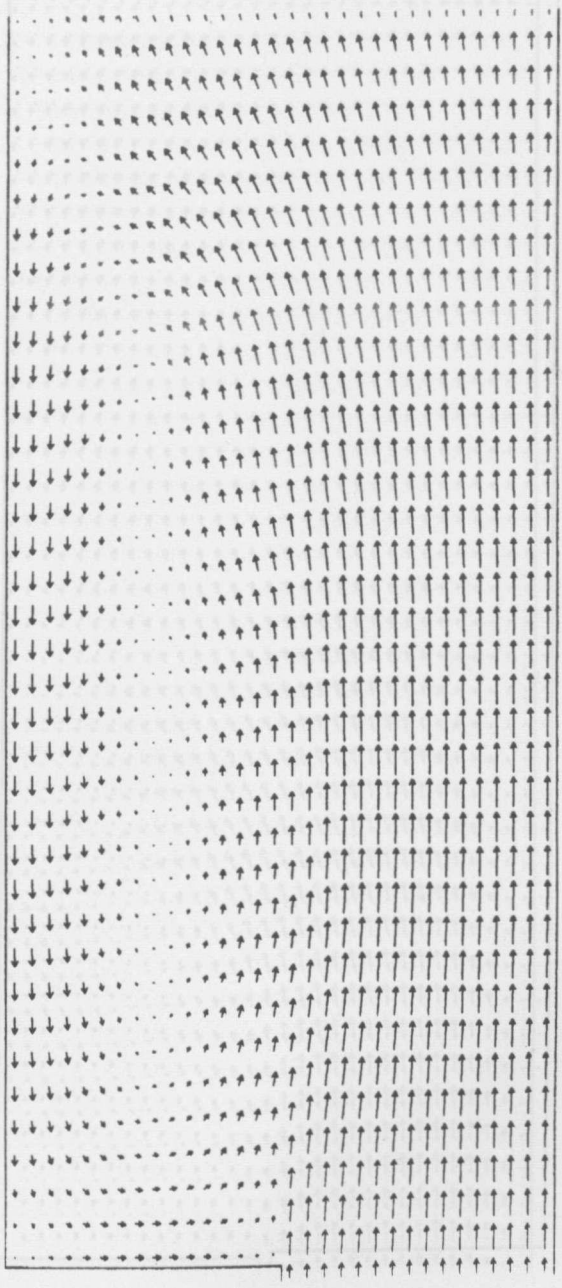


Figure (5-12-d) t = 35 s

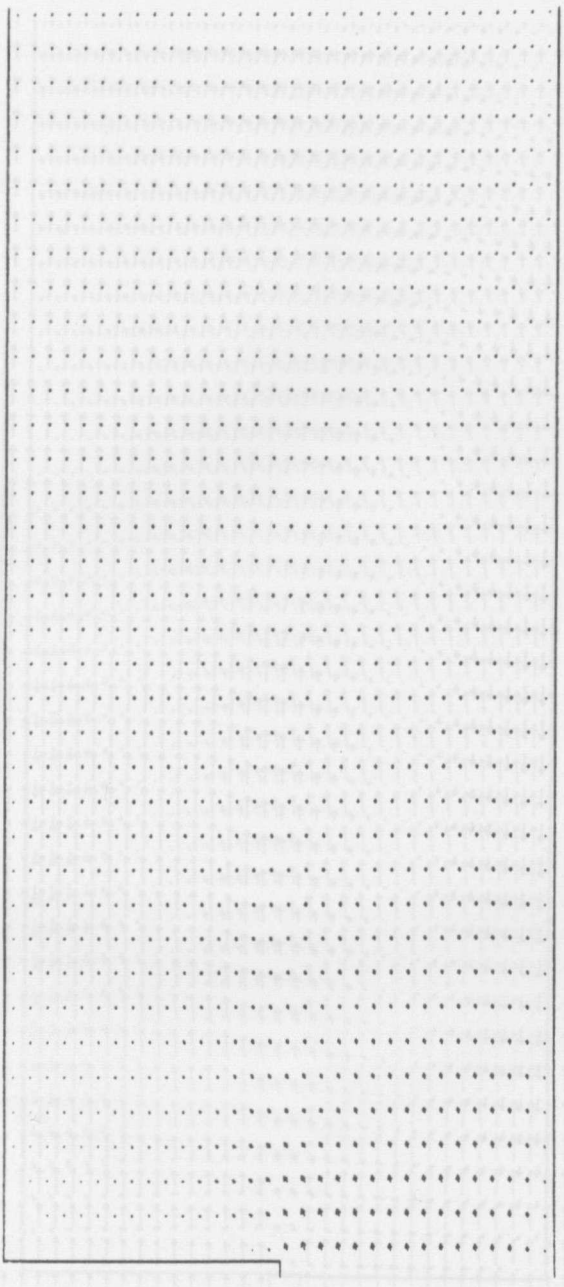


Figure (5-13 a) Development of time-dependent eddies in tidal flume, no-slip boundary condition, t = 5 sec

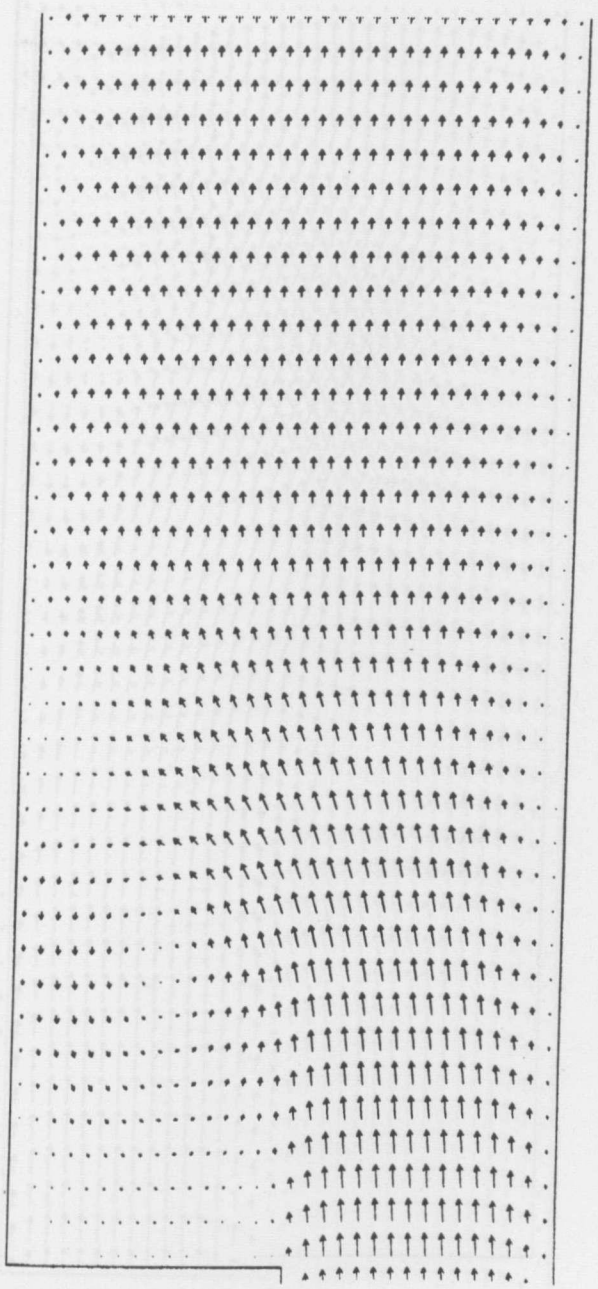


Figure (5-13-b) t = 15 s

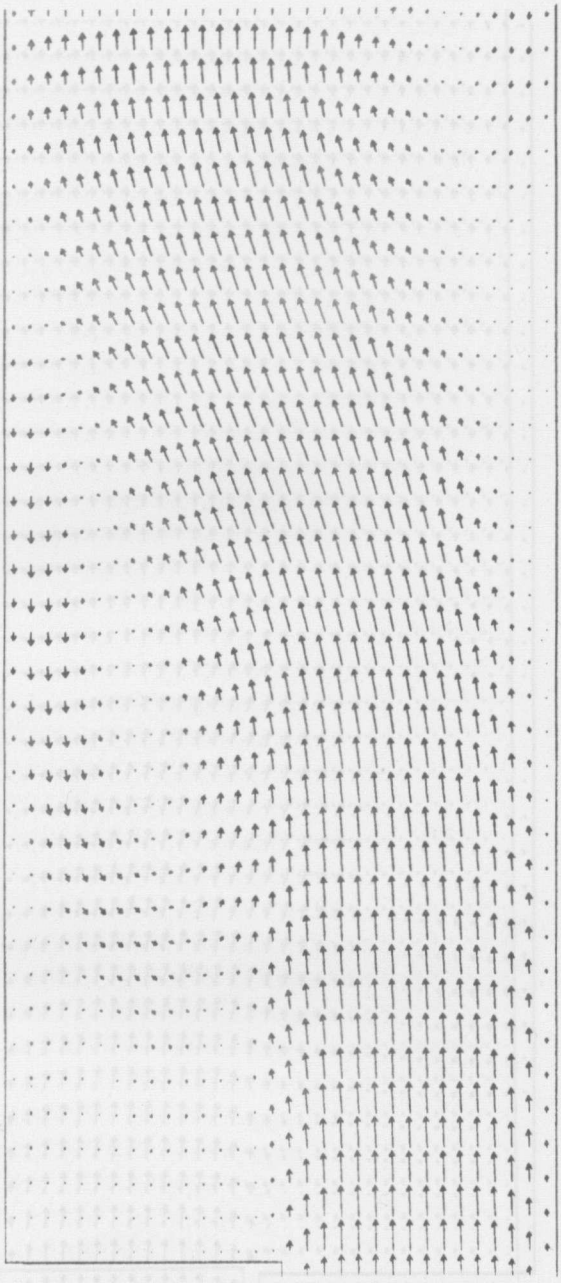


Figure (5-13-c) t = 25 s

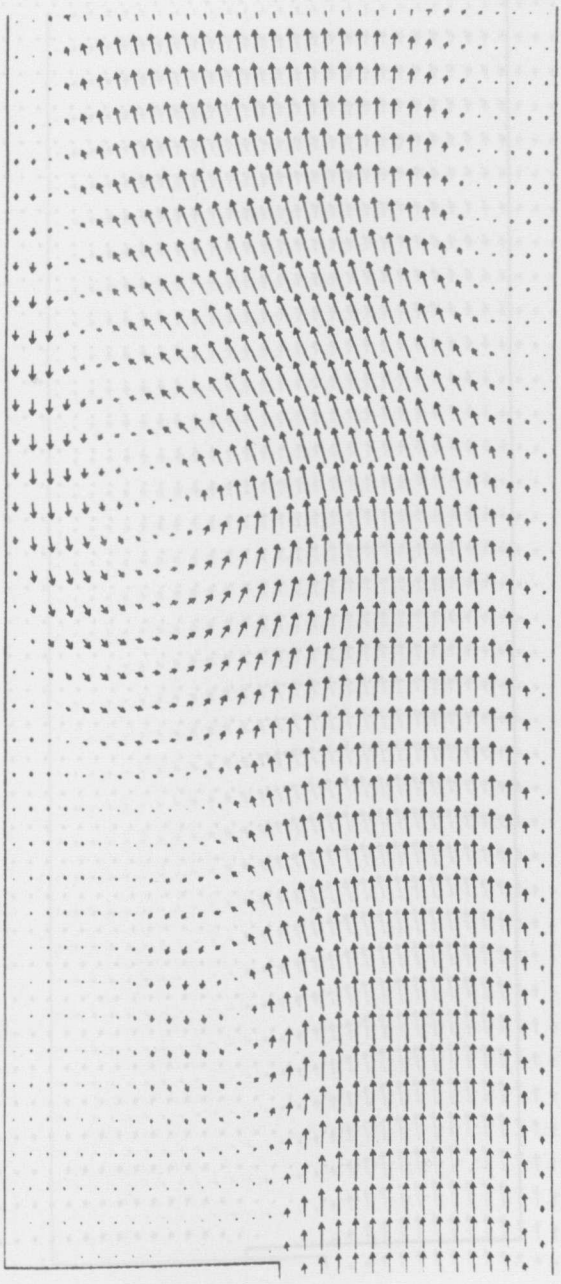


Figure (5-13 d) t = 35 s

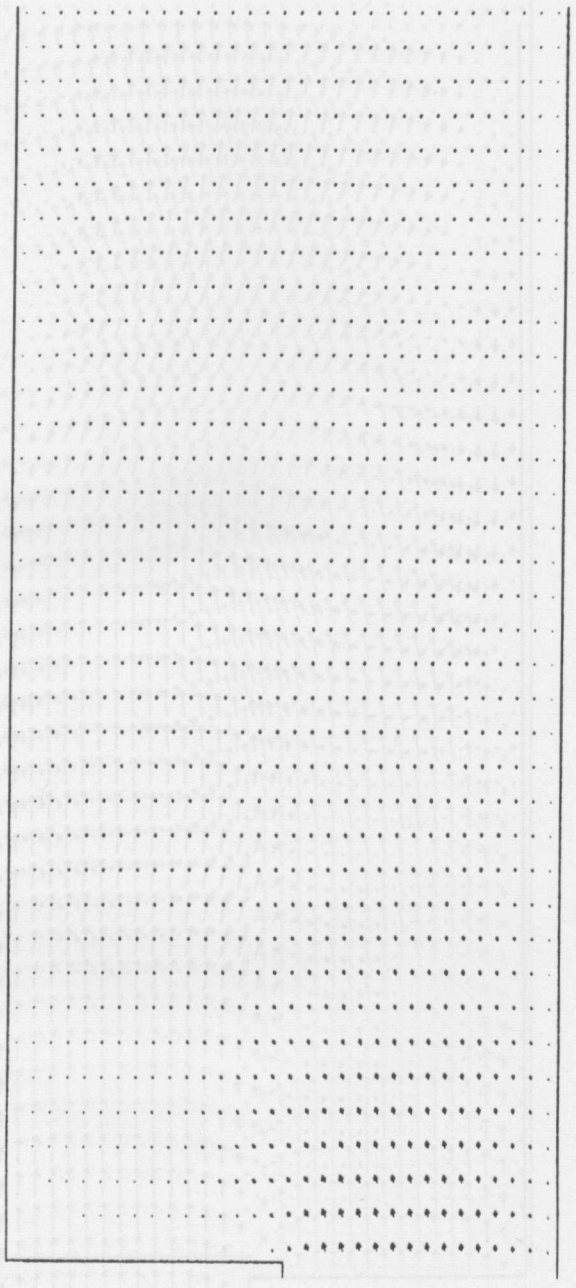


Figure (5-14-a) Development of time-dependent eddy in tidal flume with increased viscosity, t = 5 s

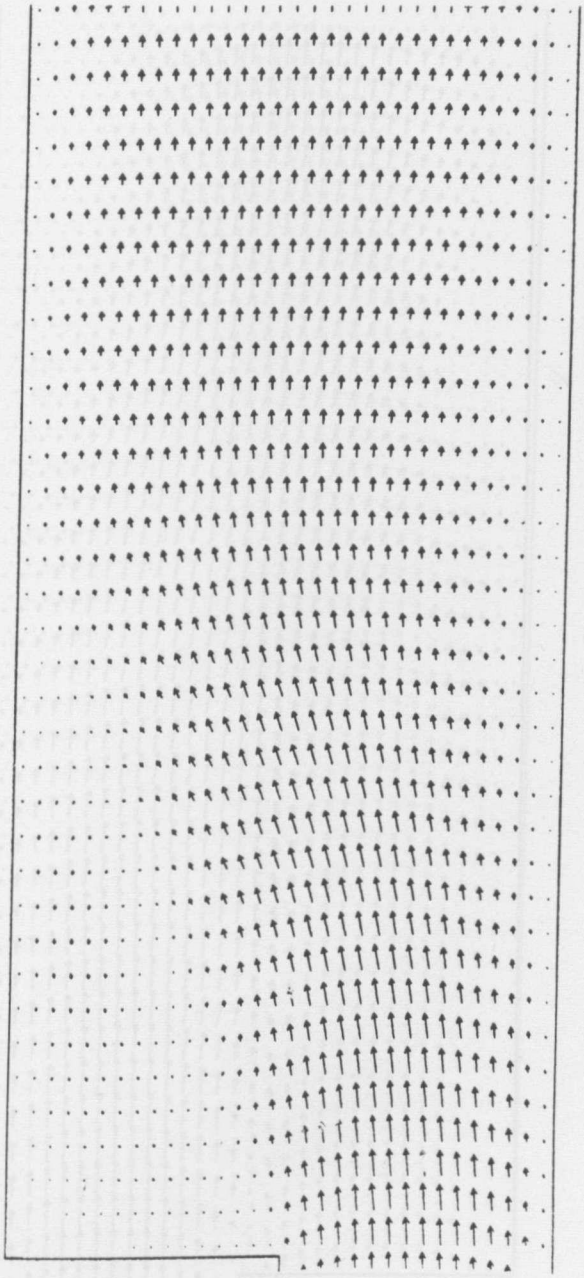


Figure (5-14-b) t = 15 s

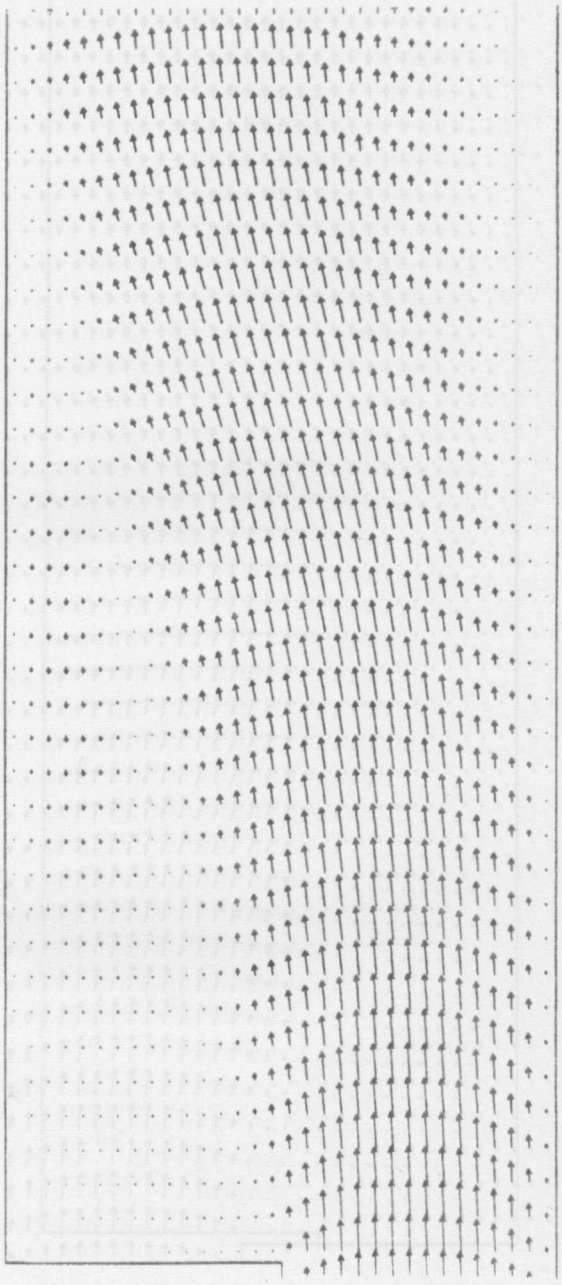


Figure (5-14-c) t = 25 s

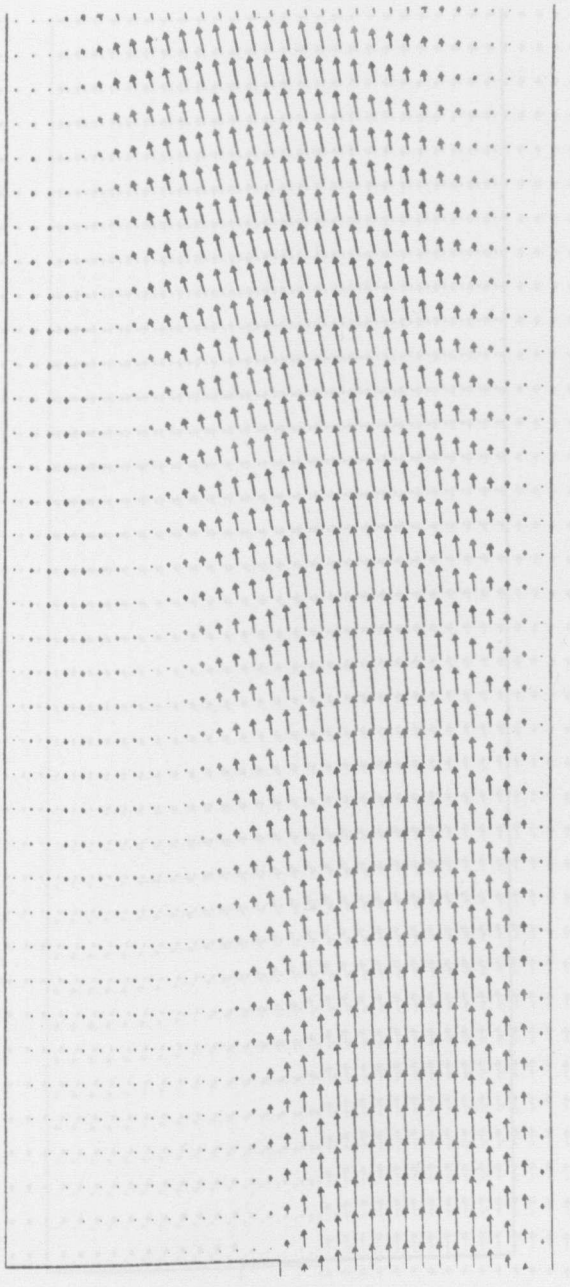


Figure (5-14-d) t = 35 s

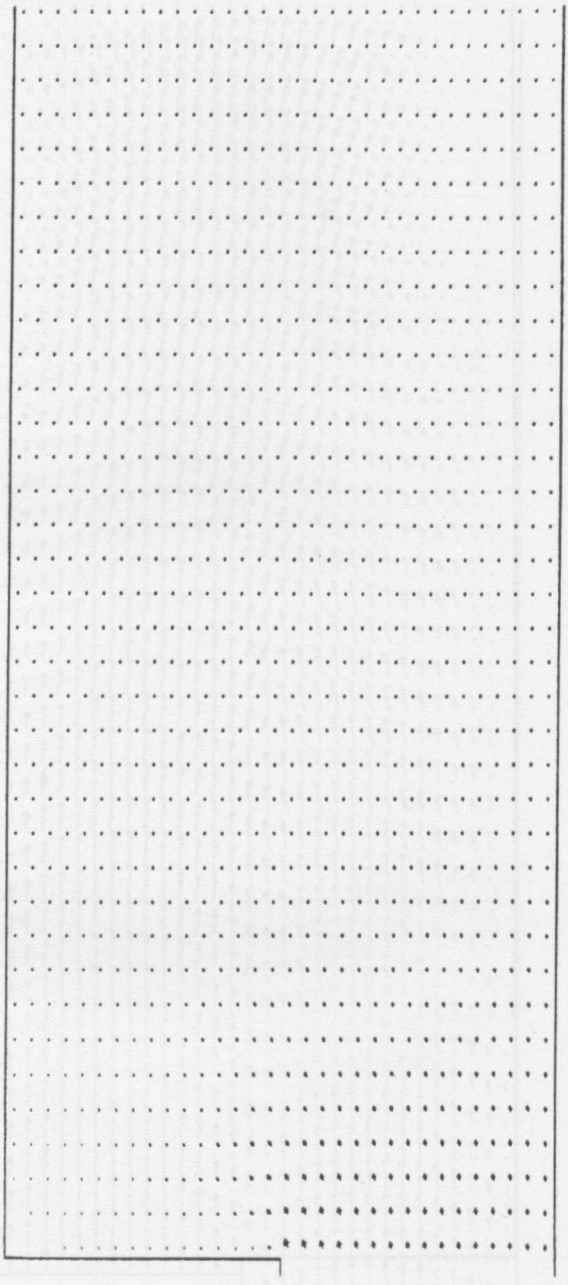


Figure (5-15-a) Development of time-dependent eddies in tidal flume, t = 5 s

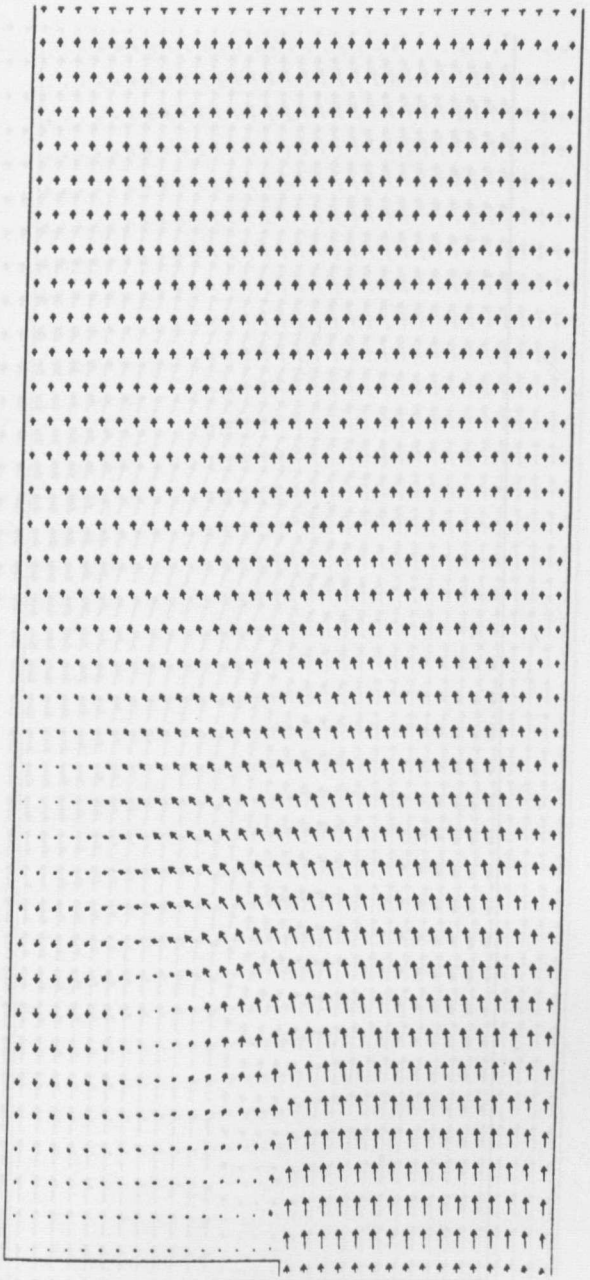


Figure (5-15-b) t = 15 s

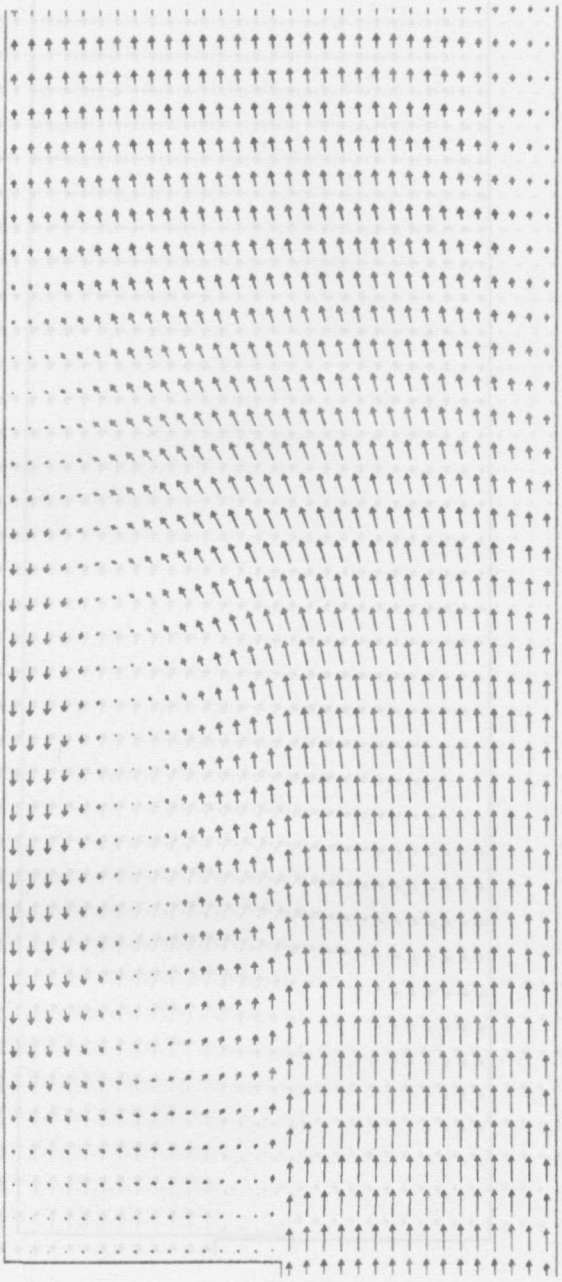


Figure (5-15-c) t = 25 s

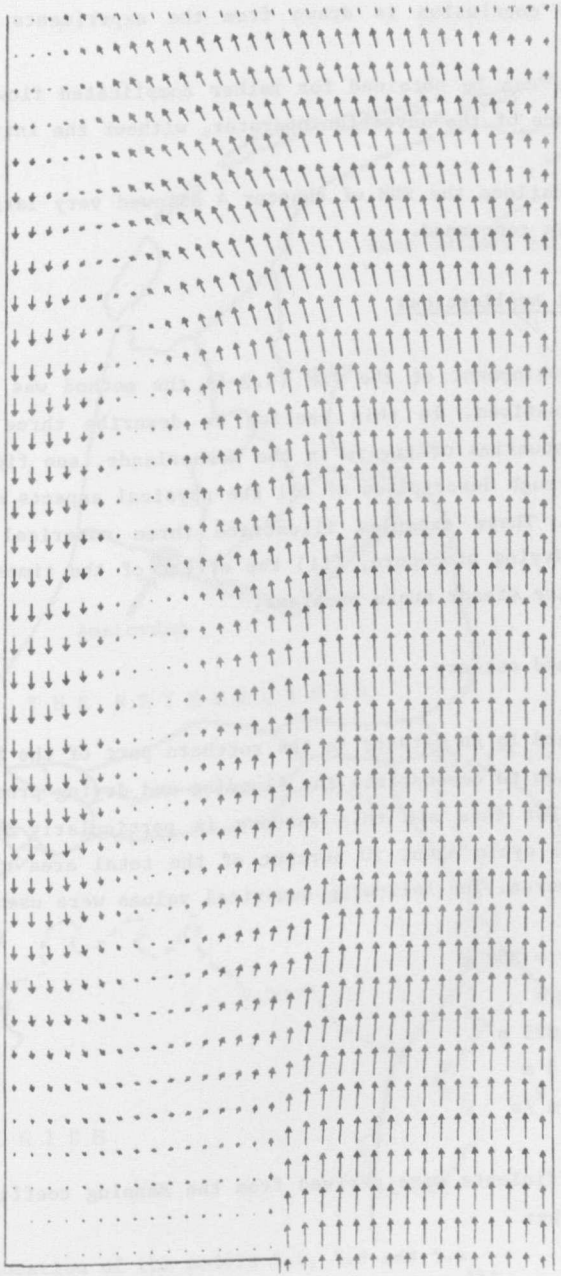


Figure (5-15-d) t = 35 s

The following conclusion is drawn from the experiments described in this section:
stable results can be obtained for rather complicated flow patterns, because of the influence of the advection operator, without the introduction of numerical viscosity.
For these situations the FDM of chapter 4 allowed very large timesteps as far as stability is concerned.

5.2 Practical applications

During the development of the FDM (4.2-4), the method was applied to several practical situations. In this section we describe three applications, all relating to estuaries or rivers in the Netherlands, see figure 5-16, but they are not a thorough description of all the physical aspects of the waters under consideration. These examples illustrate three numerical aspects: (i) the flooding and drying procedure, (ii) the effect of the timestep, and (iii) the ability to solve steady state problems.

a. Eems-Dollard estuary

The Eems-Dollard is an estuary in the northern part of the Netherlands. We use this application to demonstrate the flooding and drying procedure described in section 4.5. For this aim this estuary is particularly well suited because during a tidal cycle about 50 percent of the total area changes from dry to wet and vice versa. The following numerical values were used:

$\Delta x = \Delta y = 300$ m
$\tau = 150$ s
$\delta_o = 0.025$ m
$\delta_1 = 0.3$ m
$\nu = 10$ m^2/s

The Chezy coefficients were derived from the Manning coefficients. This relation is given by:

$$C = 1.49/\eta \; H^{1/6} \qquad (5.2\text{-}1)$$

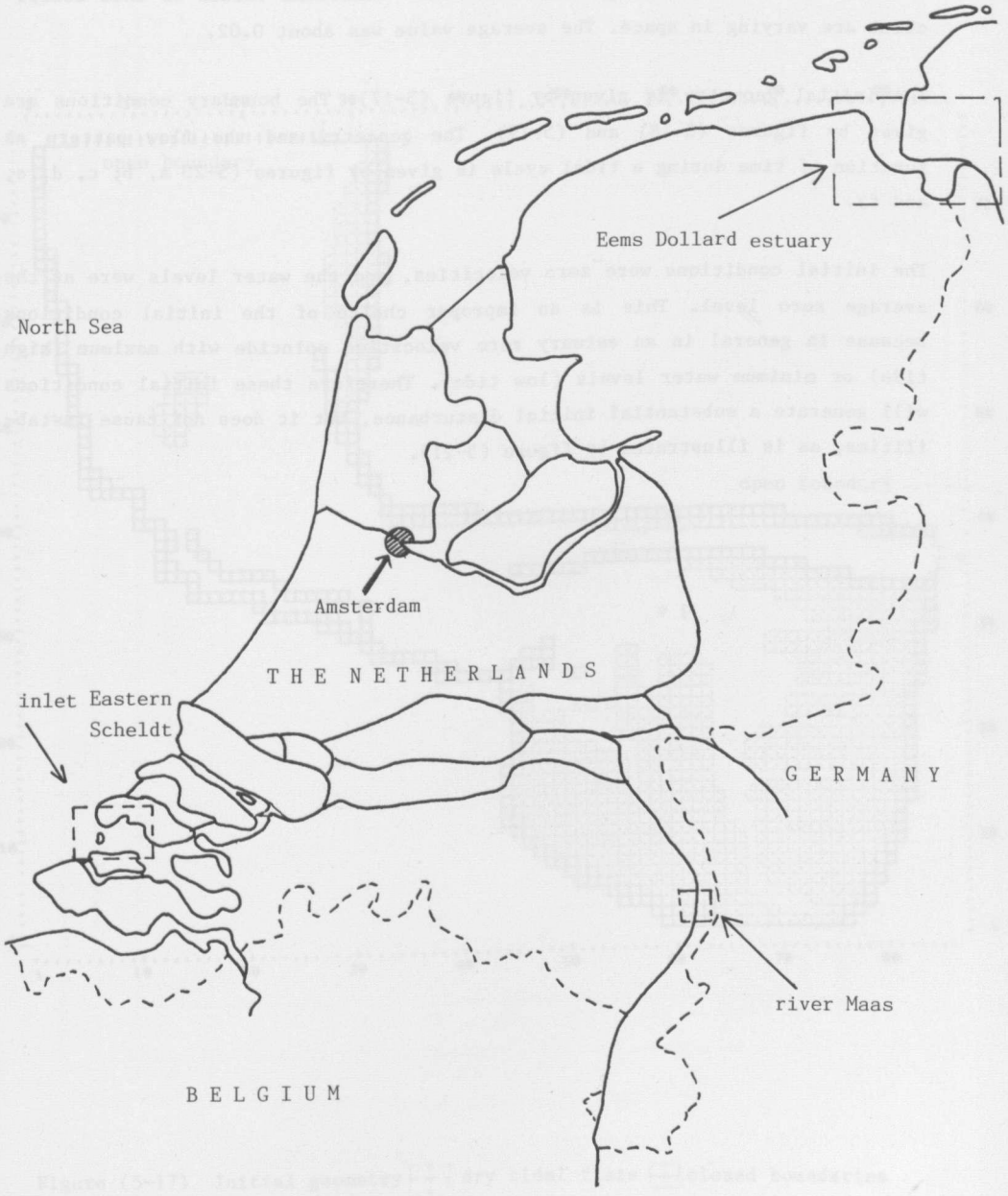


Figure (5-16) Location of the models 5.2, 5.3 and 5.4

where η denotes the Manning coefficient. The numerical values of this coefficient are varying in space. The average value was about 0.02.

The initial geometry is given by figure (5-17). The boundary conditions are given by figures (5-18) and (5-19). The geometry and the flow pattern as function of time during a tidal cycle is given by figures (5-20 a, b, c, d, e, and f).

The initial conditions were zero velocities, and the water levels were at the average zero level. This is an improper choice of the initial conditions because in general in an estuary zero velocities coincide with maximum (high tide) or minimum water levels (low tide). Therefore these initial conditions will generate a substantial initial disturbance, but it does not cause instabilities, as is illustrated by figure (5-21).

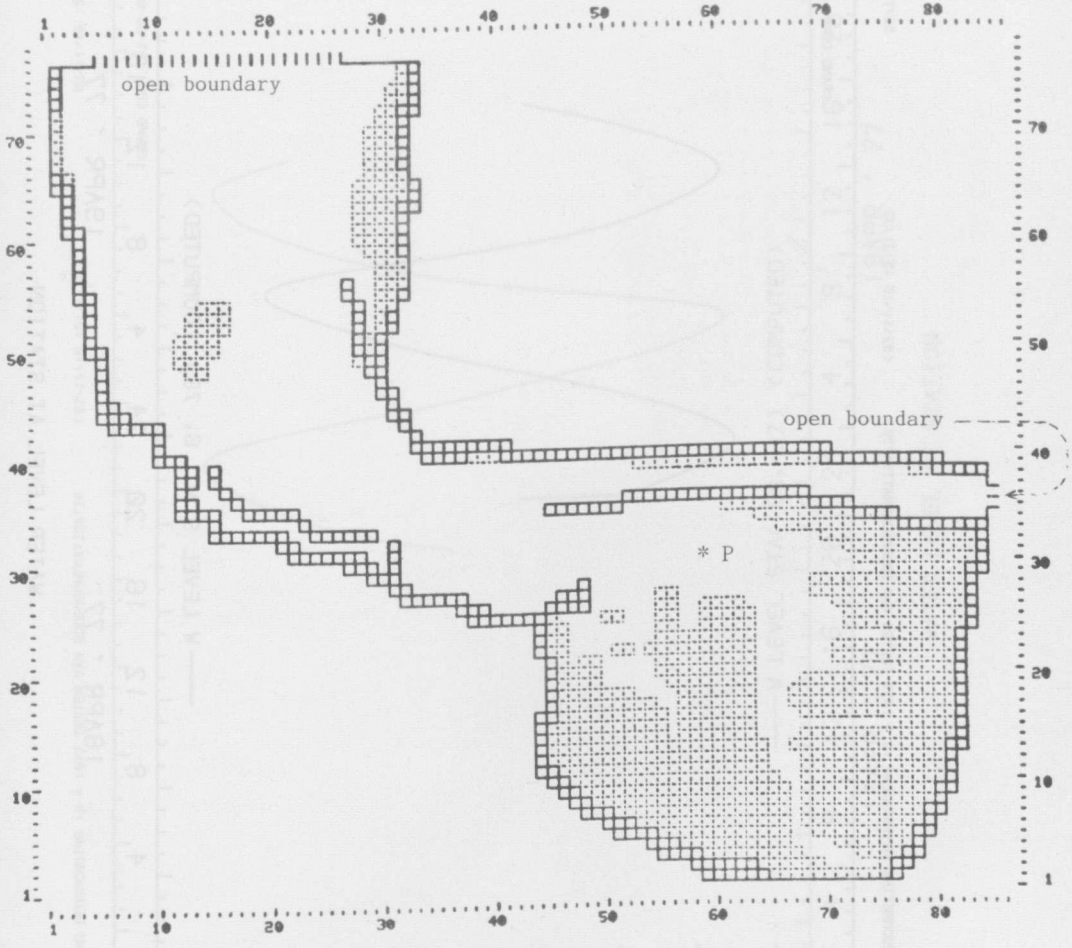


Figure (5-17) Initial geometry dry tidal flats closed boundaries

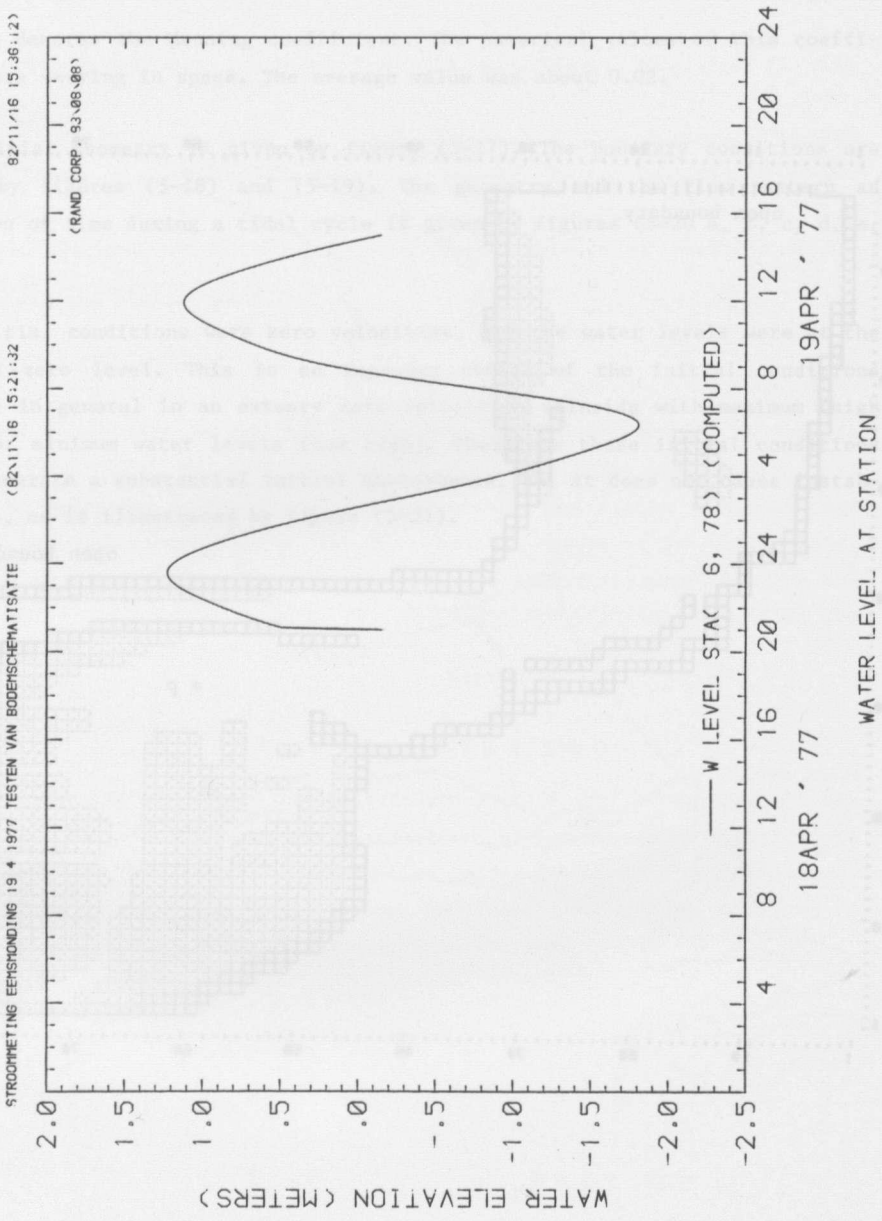


Figure (5-18) Boundary conditions

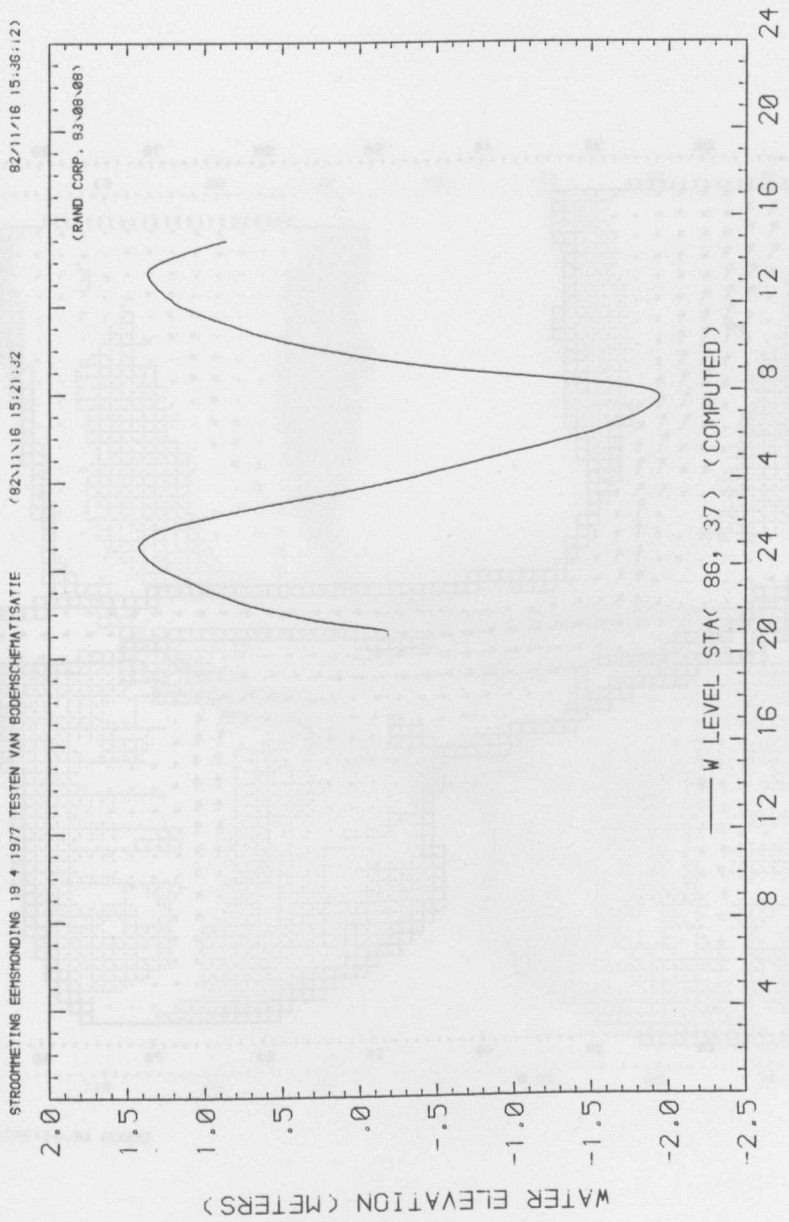


Figure (5-19) Boundary condition

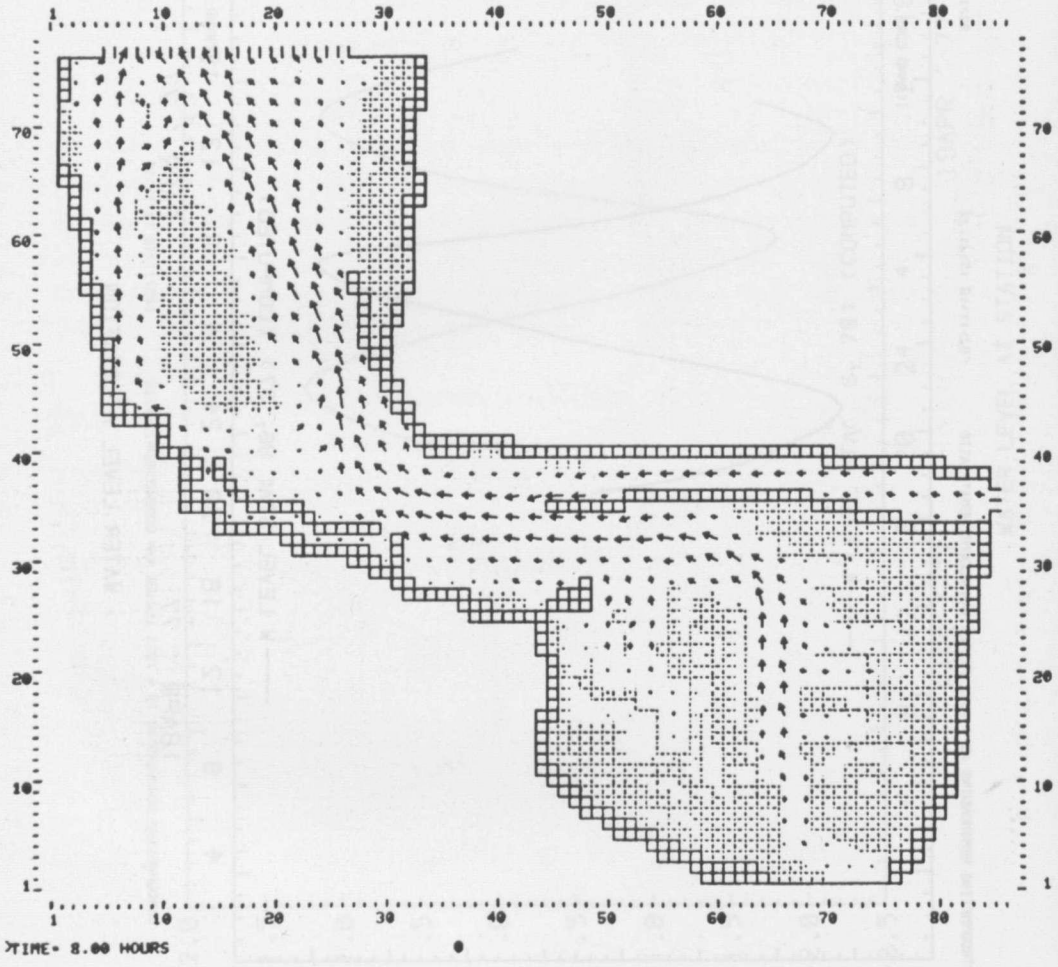


Figure (5-20) a Geometry of tidal flats and flow pattern

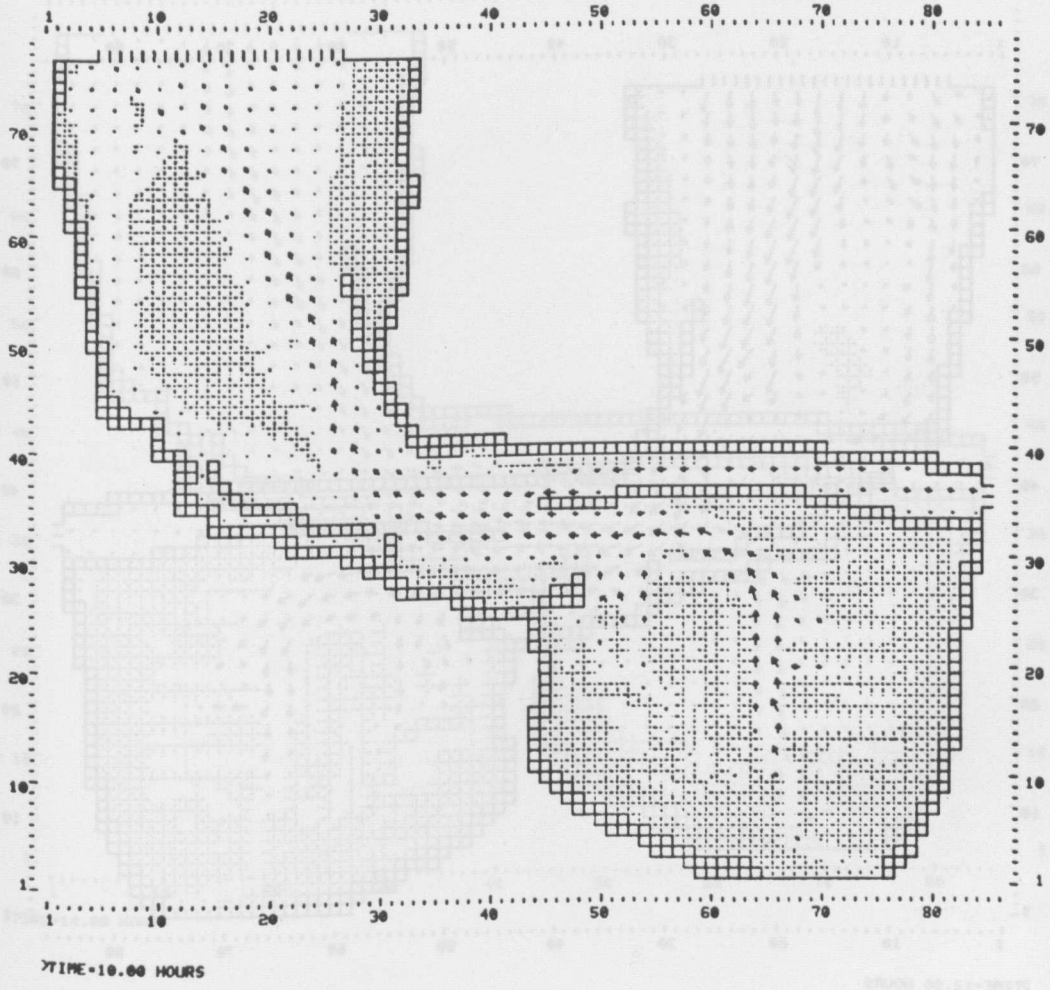


Figure (5-20) b Dry tidal flats

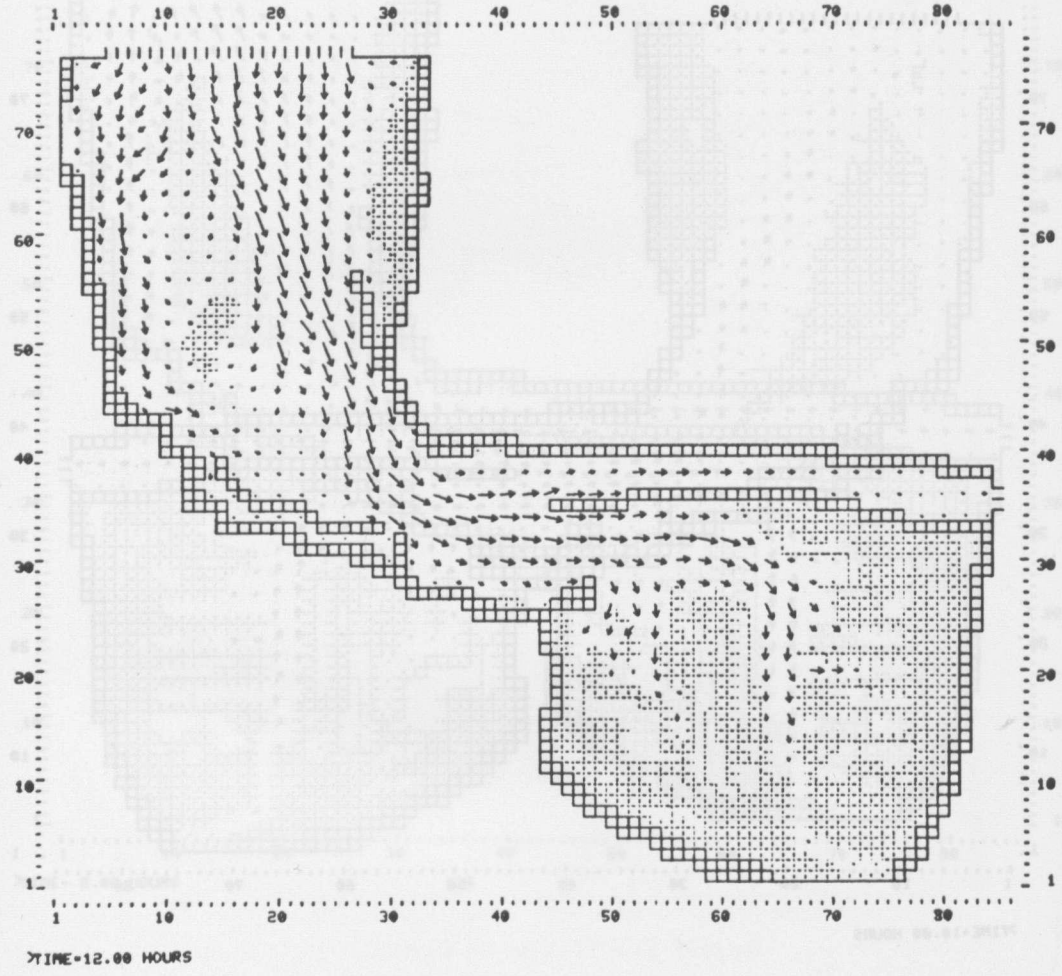


Figure (5-20) c

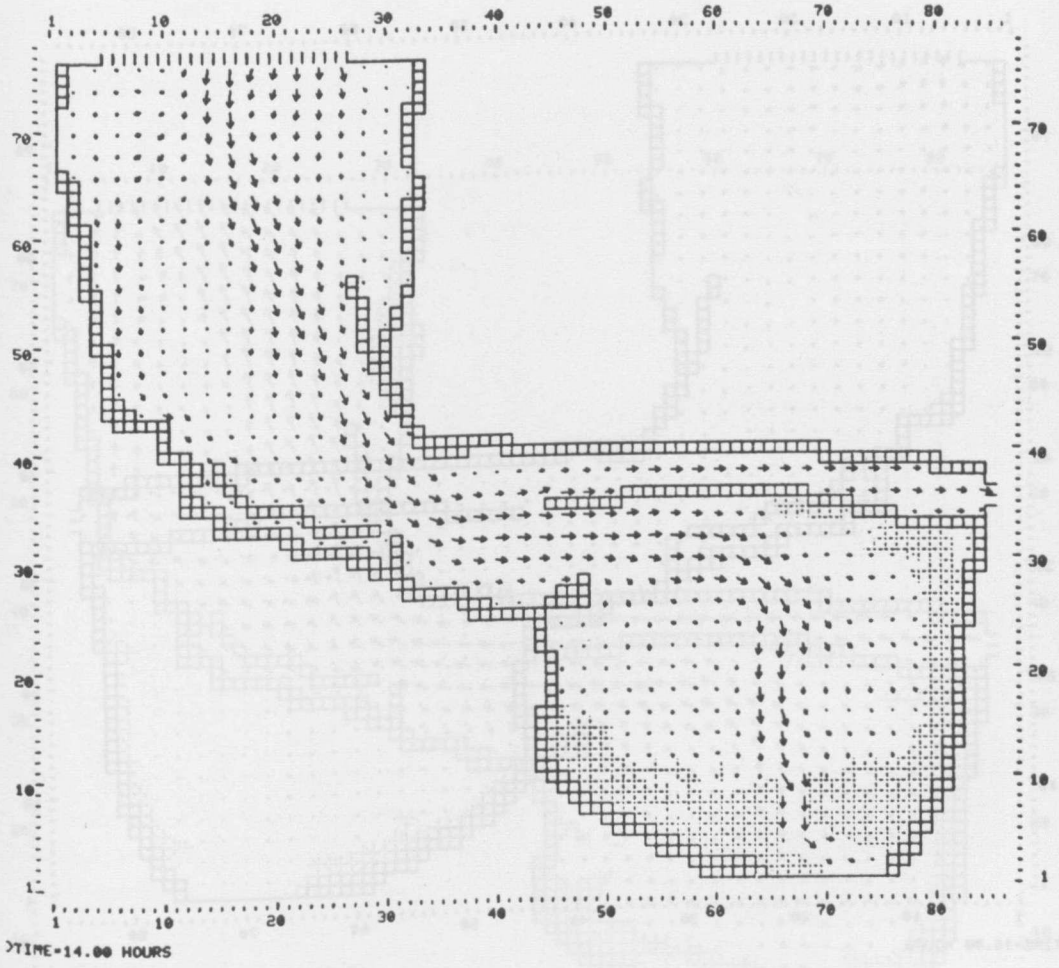


Figure (5-20) d

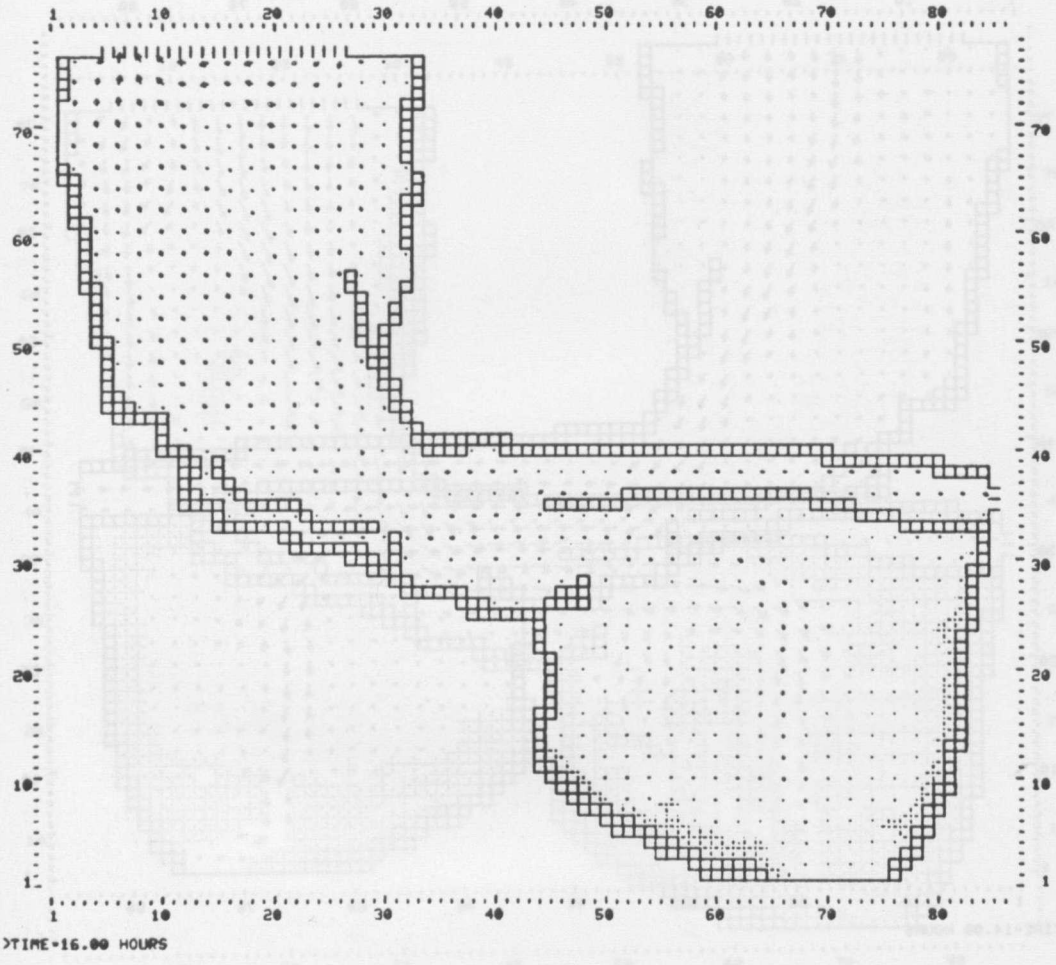


Figure (5-20) e

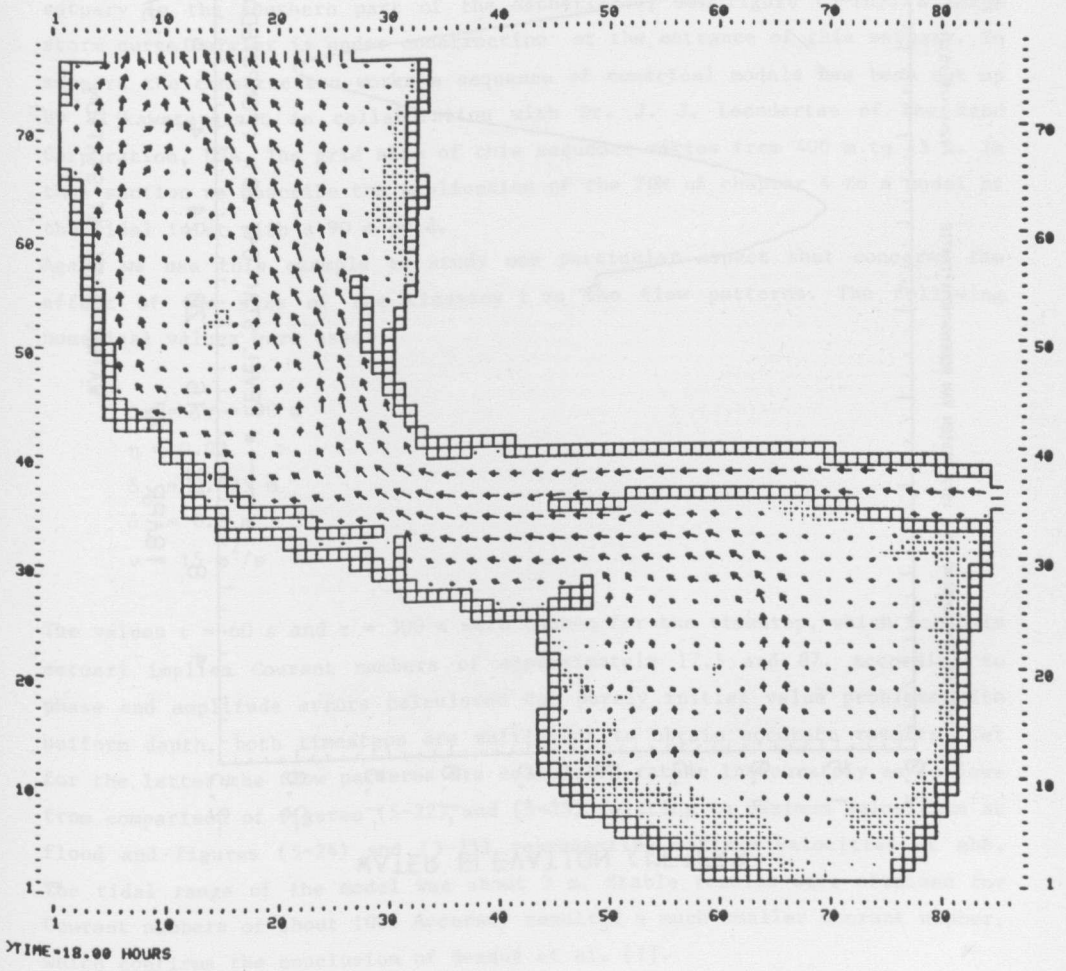


Figure (5-20) f

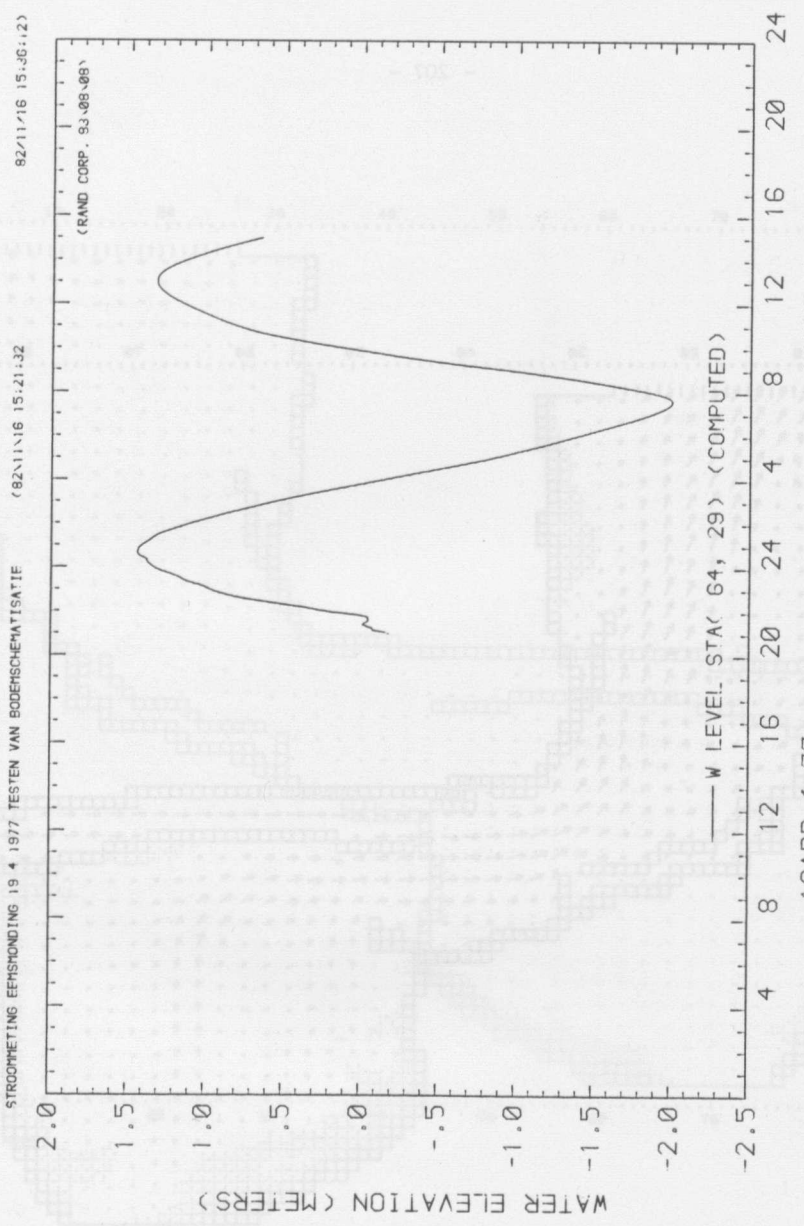


Figure (5-21) Time history at P

The conclusions which can be drawn from this example is that the numerical drying and flooding procedure does not reduce the stability of the FDM of chapter 4. Even wiggles are not being introduced by this procedure.

b. Tidal inlet Eastern Scheldt

In this section we describe an application of the FDM (4.2-4) to the flow in a part of the tidal inlet of the Eastern Scheldt. The Eastern Scheldt is an estuary in the southern part of the Netherlands, see figure (5-16). A large storm surge barrier is under construction at the entrance of this estuary. To support the construction works a sequence of numerical models has been set up by Rijkswaterstaat in collaboration with Dr. J. J. Leendertse of The Rand Corporation, USA. The grid size of this sequence varies from 400 m to 45 m. In this section we describe the application of the FDM of chapter 4 to a model at the tidal inlet with a 90 m grid.
Again we use this example to study one particular aspect that concerns the effect of the size of the timestep τ on the flow patterns. The following numerical values were used:

$\Delta x = \Delta y = 90$ m
$\eta \simeq 0.02$
$\delta_o = 0.025$ m
$\delta_1 = 0.5$ m
$\nu = 15$ m^2/s

The values $\tau = 60$ s and $\tau = 300$ s were chosen for the timestep, which for this estuary implies Courant numbers of approximately 17.5 and 87. According to phase and amplitude errors calculated for purely initial value problems with uniform depth, both timesteps are sufficient to obtain accurate results. Yet for the latter the flow patterns are calculated rather inaccurately as follows from comparison of figures (5-22) and (5-23) relating to maximum velocities at flood and figures (5-24) and (5-25) representing maximum velocities at ebb. The tidal range of the model was about 3 m. Stable results were obtained for Courant numbers of about 100. Accuracy required a much smaller Courant number, which confirms the conclusion of Benque et al. [1].

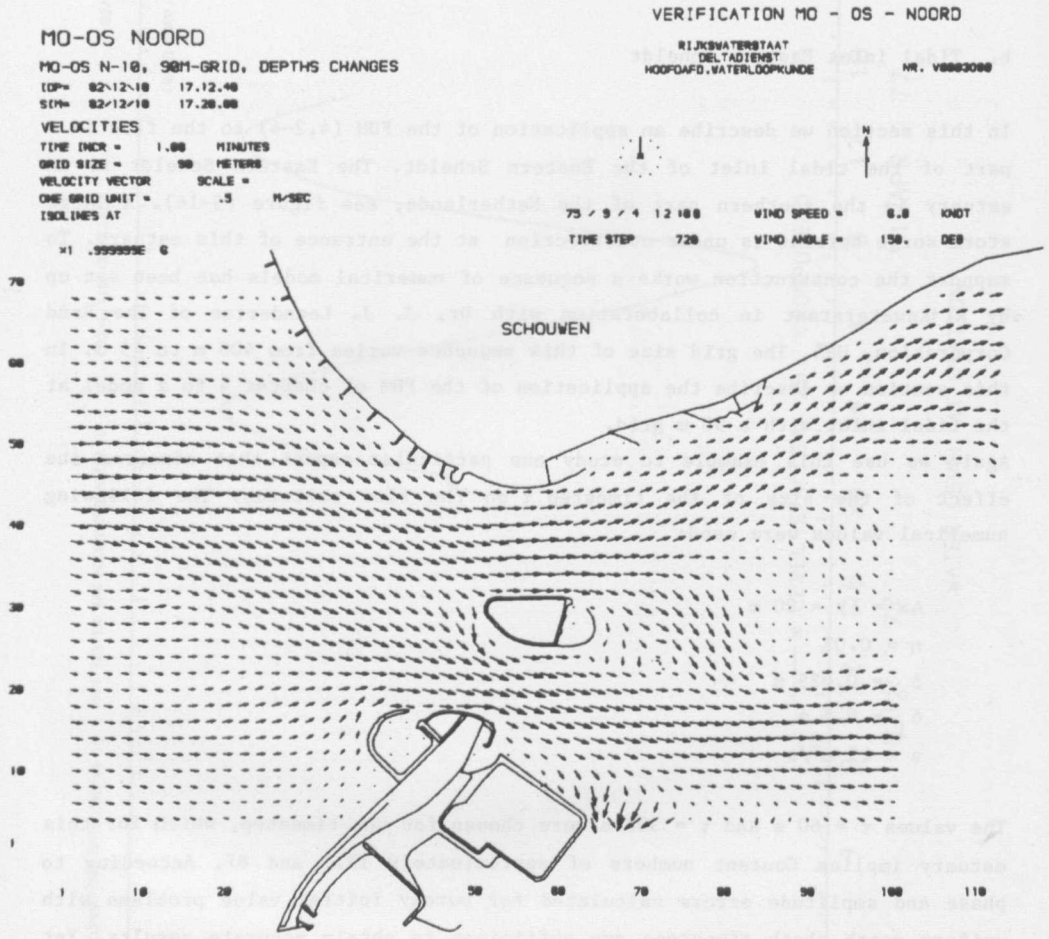


Figure (5-22) Maximum velocities at flood, τ = 60 s

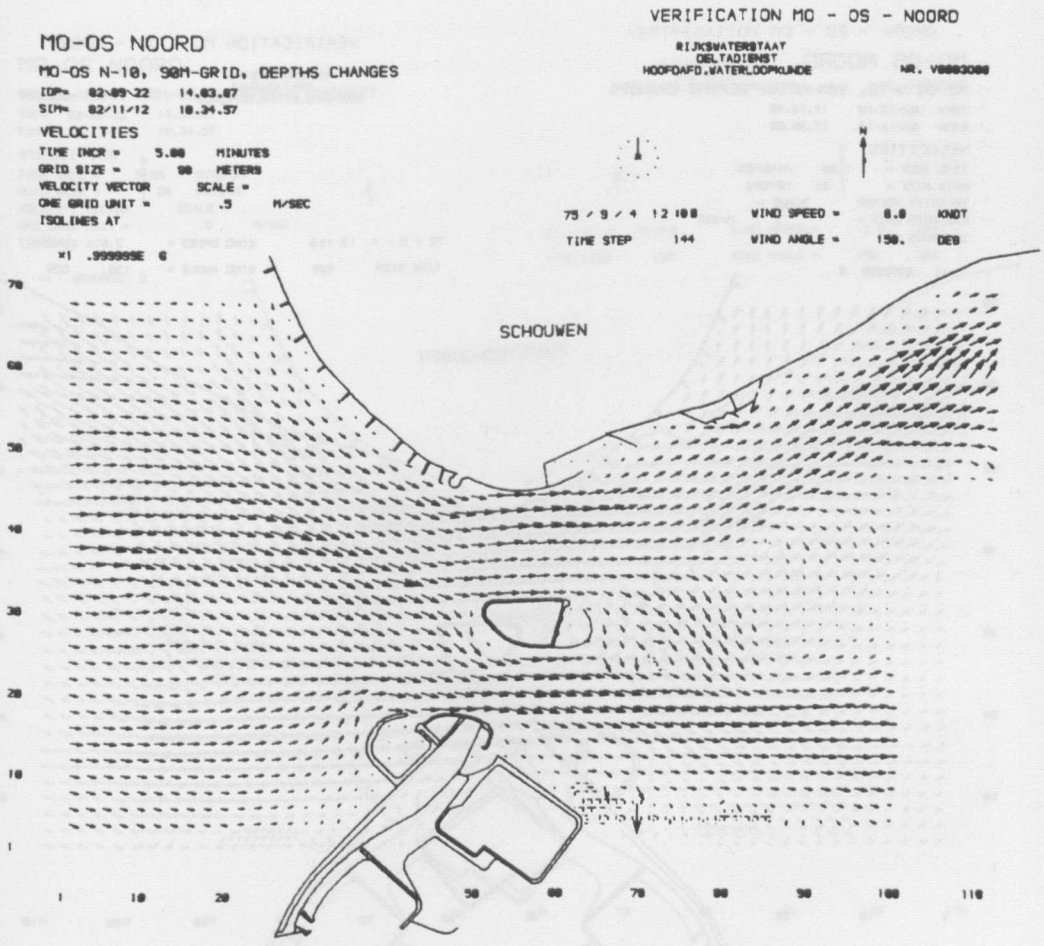


Figure (5-23) Maximum velocity at flood, τ = 300 s

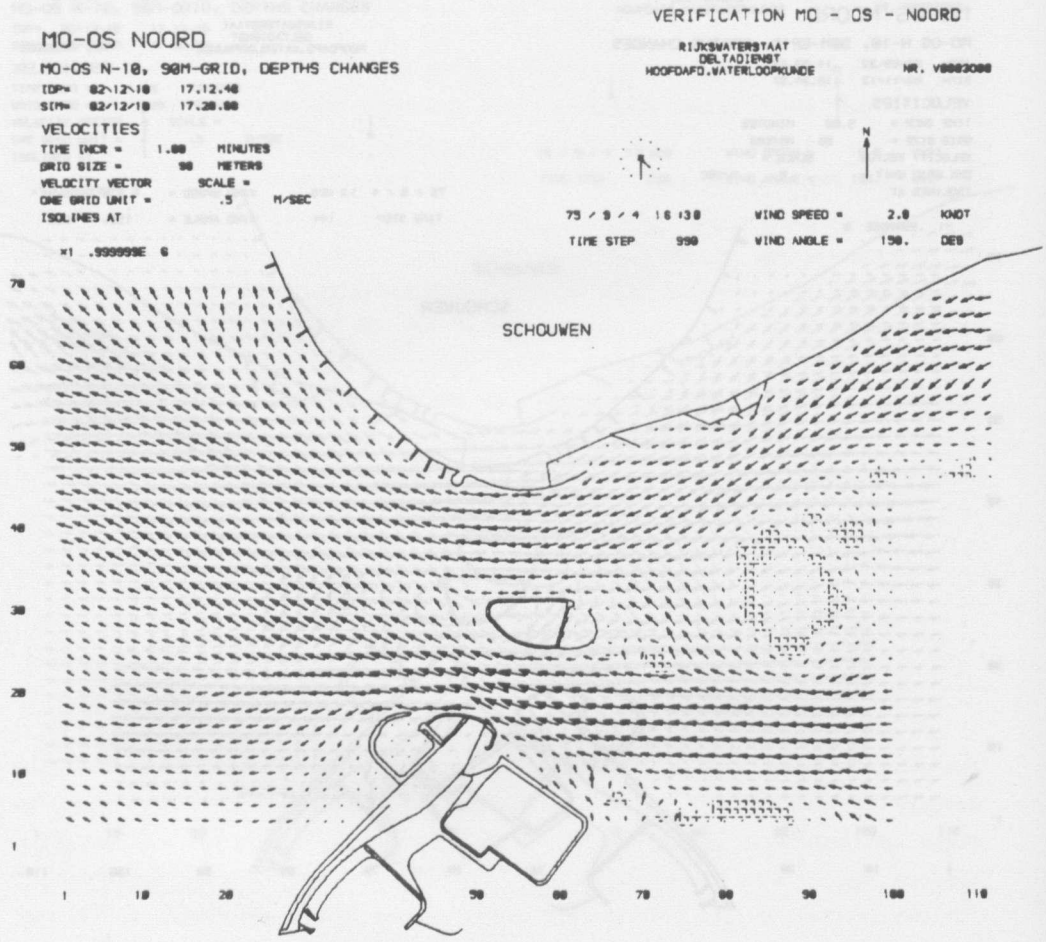


Figure (5-24) Maximum velocity at ebb, τ = 60 s

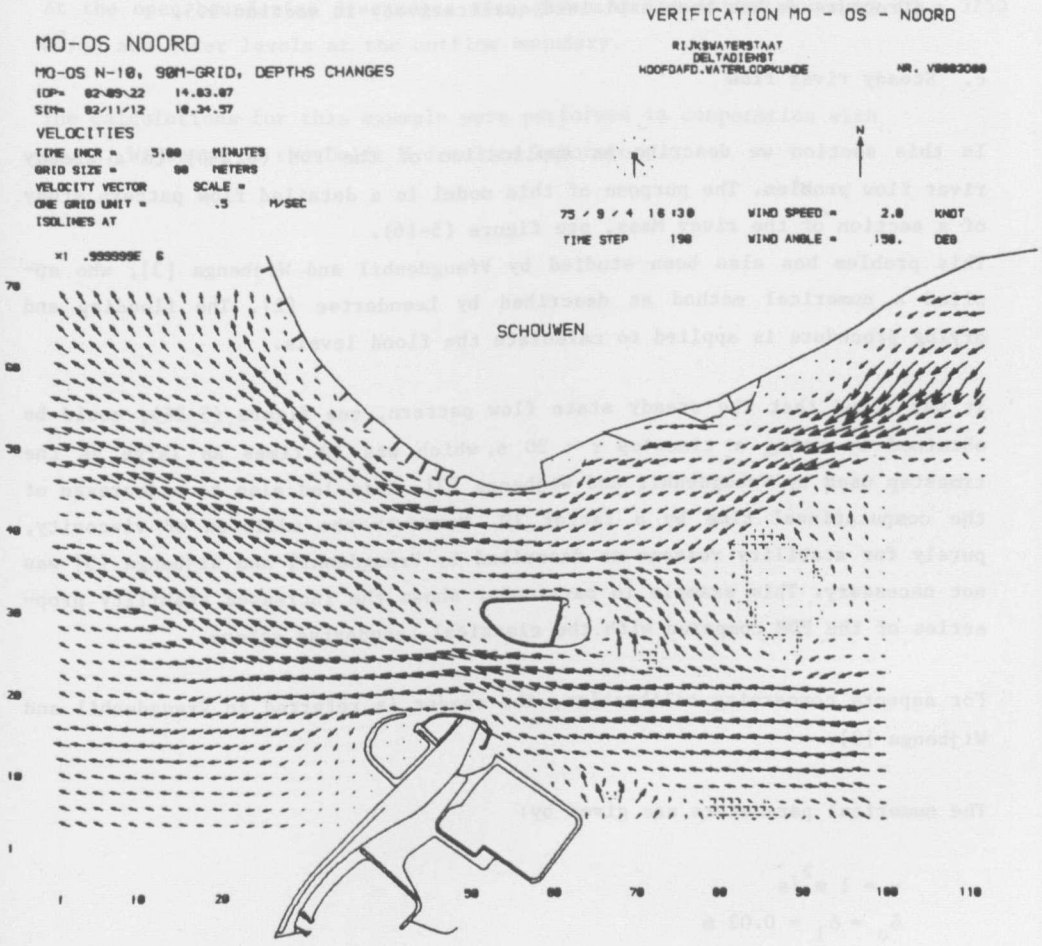


Figure (5-25) Maximum velocity at ebb, τ = 300 s

The conclusions which can be drawn from this example are twofold:

(i) Also for practical applications the stability of the FDM of chapter 4 is practically unrestricted.

(ii) The accuracy limits the maximum timestep. This is possibly due to the ADI structure as has been explained qualitatively in section 3.5.

c. Steady river flow

In this section we describe an application of the FDM (4.2-4) to a steady river flow problem. The purpose of this model is a detailed flow pattern study of a section of the river Maas, see figure (5-16).
This problem has also been studied by Vreugdenhil and Wijbenga [3], who applied a numerical method as described by Leendertse [2]. The flooding and drying procedure is applied to calculate the flood levels.

It was found that the steady state flow pattern, see figure (5-26), could be obtained by using a timestep τ = 20 s, which was 10 times as large as the timestep used by Vreugdenhil and Wijbenga [3]. This led also to a decrease of the computational time by a factor 10. Moreover the addition of viscosity, purely for stability reasons as described by Vreugdenhil and Wijbenga [3] was not necessary. This example in particular shows the increased stability properties of the FDM compared with the classical Leendertse scheme.

For aspects concerning calibration, the reader is referred to Vreugdenhil and Wijbenga [3].

The numerical parameters are given by:

$$\nu = 1 \text{ m}^2/\text{s}$$
$$\delta_o = \delta_1 = 0.02 \text{ m}$$
$$\Delta x = \Delta y = 30 \text{ m}$$
$$\tau = 20 \text{ s}$$

The Chezy coefficients were calculated according to the White-Colebrook formula given by

$$C = 18^{10} \log (12\ H/\mu) \qquad (5.2\text{-}2)$$

where μ denotes the bottom roughness coefficient.

At the open boundaries discharges were given at the inflow boundary ($Q = 3750\ m^3/s$) and water levels at the outflow boundary.

The calculations for this example were performed in cooperation with ir. A. Wijbenga of the Delft Hydraulics Laboratory.

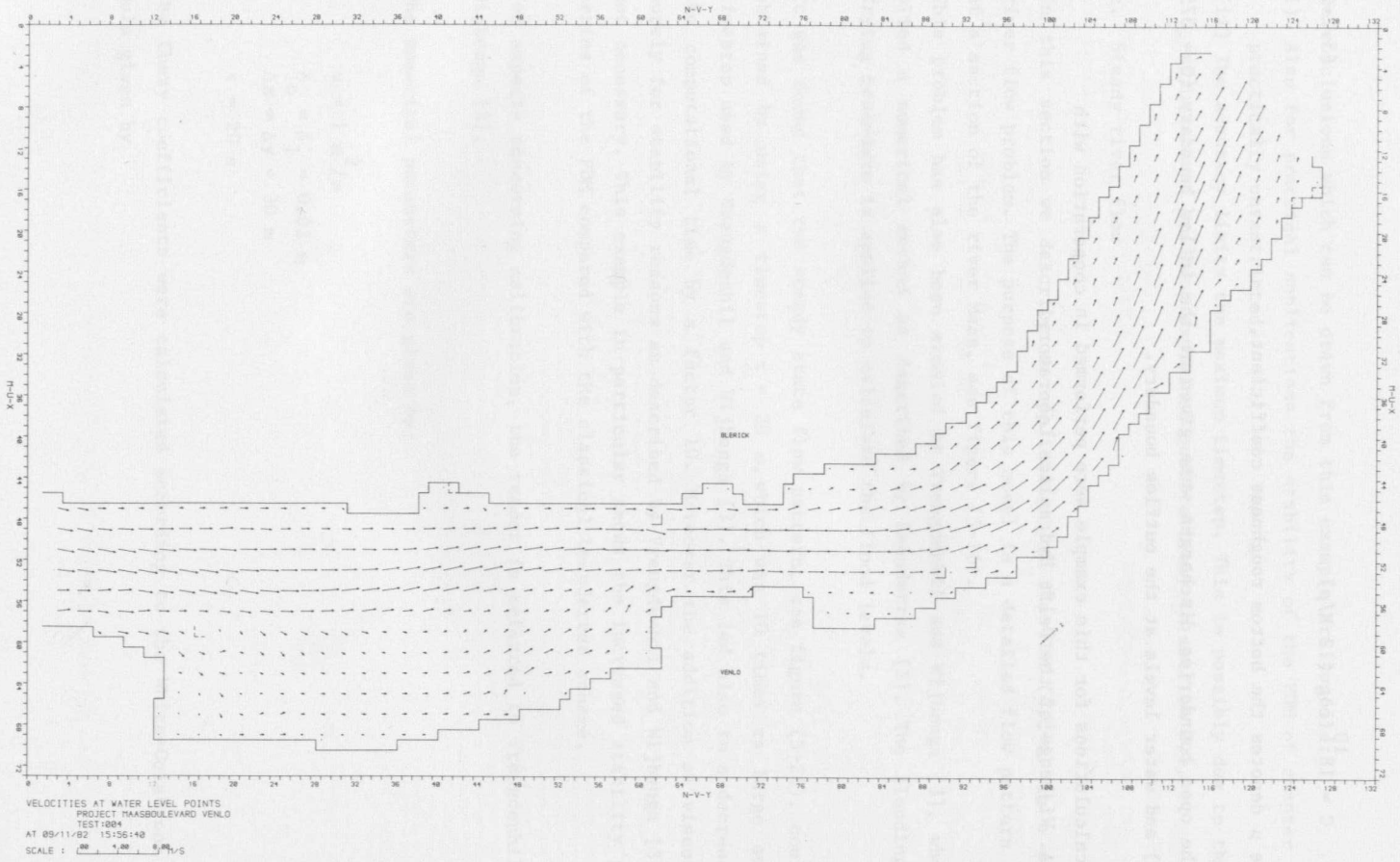


Figure (5-26) Steady riverflow

5.3 Concluding remarks

The FDM of chapter 4 is a robust method that is applicable to a wide range of simulation problems varying in size from shallow seas to tidal flumes in hydraulic laboratories.

The unconditional stability of the FDM has been confirmed by practical experiments just as has the absence of artificial viscosity. The ability of the FDM to represent the real viscosity accurately is very important for the simulation of complicated flow patterns.

ADI structure of the FDM seems to limit the maximum timestep. Because the implicit equations can be solved very efficiently while the storage requirements are minimal, we believe the FDM is competitive to a fully implicit method as described by Benqué et al. [1]. Moreover, for this fully implicit method the maximum timestep is limited by the advection approximation.

REFERENCES TO CHAPTER 5

1. BENQUE, J.P., J.A. CUNGE, J. FEUILLET, A. HAUGUEL and F.M. HOLLY,
New Method of Tidal Current Computation,
Journal of the Waterway, Port, Coastal and Ocean Division, ASCE, 1982, pp 396-417.
2. LEENDERTSE, J.J.,
Aspects of a Computational Model for Long-Period Water Wave Propagation,
Rand Corporation, Memorandum RM-5294-PR, Santa Monica, 1967.
3. VREUGDENHIL, C.B. and J.H.A. WIJBENGA,
Computation of Flow Patterns in Rivers,
Journal of the Hydraulics Division, ASCE, V108, 1982, pp 1296-1310.
4. WANG, L.X.,
Experiments on Unsteady Separating Flow with a Free Surface,
Delft University of technology, Department of Civil Engineering,
Internal Report No. 7-82, 1982.

6 Conclusions

A numerical method has been developed that is applicable to a wide range of practical, shallow water flow problems.

Considering the computational labour per timestep and the storage requirements, this method is very efficient, especially if its robustness is taken into account. Its numerical dissipation is minimal.

Probably more important than the method itself is the description of the step-by-step process by which it has been developed and the number of details that were considered. Only by considering many details, such as tidal flats or boundary treatment for complicated geometries, which from a mathematical point of view are perhaps not very interesting, can a numerical method be developed, that is capable of solving practical problems.

Obviously the method described here is not the only possible numerical method for SWE. If minimal storage requirements and robustness are considered equally important, this method is very efficient. If the storage requirements are less stringent, then the method can easily be modified into a fully implicit method. The approximation of the advection operator is in fact already fully implicit. This would circumvent the disadvantageous effects of the ADI structure with respect to the maximum timestep.

The staggered grid seems a very important aspect to guarantee robustness and simplicity of boundary condition procedures.

Appendix , Notation

C	Chezy coefficient
Cf	Courant number
E_{+x}, E_{-x}, etc.	finite difference operators
f	coriolis parameter
$F^{(x,y)}$	external forcing functions of momentum equations
g	acceleration due to gravity
h	water depth below some plane of reference
H	total water depth
S_{+x}, S_{ox}, etc.	finite difference operators
t	time
u	velocity in x direction
v	velocity in y direction
x,y	spatial coordinates
α	weighting factor for non slip and perfect slip boundary conditions
δ_o	threshold for flooding or drying
δ_1	threshold for application of dissipative numerical approximations
$\Delta x, \Delta y$	spatial grid sizes
ε	weighting factor for non reflective part of open boundary conditions
ζ	water level above some plane of reference
η	Manning coefficient
μ	bottom roughness coefficient for White Colebrook formula
ν	viscosity coefficient
τ	time increment

Appendix , Notation (continued)

$u_x = \frac{\partial u}{\partial x}$,

$u_y = \frac{\partial u}{\partial y}$, .

$u_t = \frac{\partial u}{\partial t}$,

u at m, n, k denotes: $u^k_{m,n}$,

u_{ox} at m, n, k denotes: $(u^k_{m+\frac{1}{2},n} - u^k_{m-\frac{1}{2},n})/\Delta x$,

u_{oy} at m, n, k denotes: $(u^k_{m,n+\frac{1}{2}} - u^k_{m,n-\frac{1}{2}})/\Delta y$,

u_{ot} at m, n, k denotes: $(u^{k+\frac{1}{2}}_{m,n} - u^{k-\frac{1}{2}}_{m,n})/\tau$,

u_{+x} at m, n, k denotes: $(u^k_{m+1,n} - u^k_{m,n})/\Delta x$,

u_{-x} at m, n, k denotes: $(u^k_{m,n} - u^k_{m-1,n})/\Delta x$,

u_{+y} at m, n, k denotes: $(u^k_{m,n+1} - u^k_{m,n})/\Delta y$,

u_{-y} at m, n, k denotes: $(u^k_{m,n} - u^k_{m,n-1})/\Delta y$,

$\overrightarrow{u}^x$ at m, n, k denotes: $(u^k_{m+\frac{1}{2},n} + u^k_{m-\frac{1}{2},n})/2$,

$\overrightarrow{u}^y$ at m, n, k denotes: $(u^k_{m,n+\frac{1}{2}} + u^k_{m,n-\frac{1}{2}})/2$ and

$\overrightarrow{\overrightarrow{u}}$ at m, n, k denotes: $(u^k_{m+\frac{1}{2},n+\frac{1}{2}} + u^k_{m-\frac{1}{2},n+\frac{1}{2}} + u^k_{m+\frac{1}{2},n-\frac{1}{2}} + u^k_{m-\frac{1}{2},n-\frac{1}{2}})/4$.

Summary

Calculations of velocities and water levels in shallow seas, estuaries or rivers are often based on shallow water equations, which have many methods for numerical solution. Often these methods fail when they are applied to situations with a complicated flow pattern. Numerical problems such as spurious "wiggles" or instabilities are often solved with the addition of numerical dissipation.

A description is given of a method that is applicable to many problems varying from the entire North Sea to a tidal flume in a hydraulic laboratory. This method is developed step by step. First, simple examples illustrate important notions of numerical analysis such as convergence and stability. Overspecification of boundary conditions leading to useless numerical solutions is also illustrated by means of simple examples.

For a simple advection equation a number of numerical methods are compared. A few new and efficient methods are introduced. Because of their structure these methods can be implemented for the approximation of the advection operator, within ADI methods for the approximation of shallow water equations.

For simple, linear, shallow water equations the advantages of "staggered" grids are explained. "Staggered" grids are not only efficient but are also useful for eliminating wiggles.

By means of the aforementioned advection method the Leendertse scheme is modified such that some theoretical disadvantages of this scheme are eliminated.

This modified scheme is extended, including details concerning boundary treatment and tidal flats, to nonlinear, shallow water flow problems. Some varying examples illustrate the applicability of the method.

Samenvatting

Ondiep water vergelijkingen vormen dikwijls de basis voor de berekening van snelheden en waterstanden in ondiepe zeeën, estuaria of rivieren. Voor de numerieke oplossing van deze vergelijkingen zijn veel methoden bekend. Dikwijls echter falen deze methoden indien zij worden toegepast voor situaties met een ingewikkeld stromingspatroon. Numerieke problemen die zich dan kunnen voordoen, zoals "2Δx golven" of instabiliteit worden dikwijls opgelost door toevoeging van numerieke dissipatie.

Dit werk bevat de beschrijving van een methode die toepasbaar is op een ruim scala problemen variërend bijvoorbeeld van de gehele Noordzee tot een getijgoot in een waterloopkundig laboratorium, zonder dat numerieke viskositeit een overwegende rol speelt.

De methode wordt stap voor stap ontwikkeld. Eerst worden aan de hand van eenvoudige voorbeelden enkele belangrijke begrippen uit de numerieke wiskunde zoals stabiliteit en convergentie nader toegelicht. Eveneens met eenvoudige voorbeelden wordt geïllustreerd hoe overspecifikatie van randvoorwaarden leidt tot onbruikbare numerieke oplossingen.

Vervolgens wordt voor een eenvoudige advectie vergelijking een aantal numerieke methoden vergeleken. Enkele nieuwe en efficiënte methoden worden geïntroduceerd die vanwege hun struktuur zeer wel in te passen zijn als numerieke advectieve operator in ADI methoden voor de numerieke oplossing van ondiep water vergelijkingen.

Voor vereenvoudigde lineaire ondiep water vergelijkingen worden de voordelen van "gestaggerde" roosters uitgelegd. "Gestaggerde" roosters zijn niet alleen efficiënt maar vormen ook een methode om zogenaamde "2Δx golven" te elimineren.

Door middel van de reeds genoemde advectieve methode wordt het Leendertse schema gemodificeerd zodanig dat enkele theoretische nadelen van deze methode worden geëlimineerd. Dit wordt uitgelegd voor lineaire vergelijkingen. Dit schema wordt tot in details, die betrekking hebben op randvoorwaarden en droogvallende platen, uitgebreid voor de numerieke oplossing van praktische, niet lineaire stromingsproblemen in ondiep water.

Aan de hand van enkele uiteenlopende voorbeelden wordt de toepasbaarheid van de methode geïllustreerd.

Curriculum vitae

De schrijver van dit proefschrift is geboren op 13 juli 1948 te Zaandam.
Na het behalen van het diploma H.B.S.-B begon hij in 1966 met de studie geodesie aan de Technische Hogeschool te Delft.
In 1967 vervolgde hij zijn studie bij de afdeling algemene wetenschappen van de T.H. Delft. In 1973 slaagde hij voor het examen van wiskundig ingenieur. Zijn afstudeerrichting was numerieke wiskunde.
Van 1973 tot 1976 was hij, als wetenschappelijk ambtenaar, verbonden aan het rekencentrum van de Landbouw Hogeschool te Wageningen. Daarna, van 1976 tot 1981, was hij als project-ingenieur werkzaam bij het Waterloopkundig Laboratorium te Delft. Daarbij werd een groot deel van de ervaring opgedaan die aan dit proefschrift ten grondslag ligt.
Sinds 1981 is hij, eerst als projectleider en vervolgens als onderafdelingshoofd, in dienst bij de
Dienst Informatie Verwerking van Rijkswaterstaat. Hier kreeg hij tevens de gelegenheid dit proefschrift af te ronden.

S T E L L I N G E N

behorende bij

On the Construction of Computational
Methods for Shallow Water Flow Problems

door

G.S. Stelling

Stelling I.

Bij toepassing van de matrix methode voor de bestudering van stabiliteit van numerieke methoden voor de benadering van partiële differentiaal vergelijkingen wordt een vorm van stabiliteit onderzocht die niet noodzakelijk convergentie impliceert.

Stelling II.

Door toepassing van gestaggerde roosters wordt, ten opzichte van methoden gebaseerd op niet gestaggerde roosters, niet alleen de efficiëntie van numerieke methoden voor ondiep water vergelijkingen verhoogd maar ook de robuustheid.

Stelling III.

Voor de numerieke benadering van de advectie operator van ondiep water vergelijkingen in de buurt van randen is de orde van nauwkeurigheid niet van invloed. Slechts van belang is dat men een stabiele en niet een tot z.g. "wiggles" aanleiding gevende benadering kiest.

Stelling IV.

Bij ADI schema's voor ondiep water vergelijkingen kan de maximale tijdstap bij willekeurige gebieden worden beperkt doordat het numerieke afhankelijkheidsgebied per tijdstap niet noodzakelijk het exacte afhankelijkheidsgebied volledig bevat.

Stelling V.

Benadering van de advectie operator van ondiep water vergelijkingen door middel van karakteristieke interpolatie methoden in combinatie met "operator splitting" veroorzaakt voor stationaire problemen toevoeging van numerieke diffusie van de eerste orde, dit ongeacht de orde van het gebruikte interpolatie polynoom.

Stelling VI.

Voor getijberekeningen van estuaria dient men de beginvoorwaarden bij voorkeur met hoogwater te laten samenvallen.

Stelling VII.

Bij koppeling van 1- en 2-dimensionale numerieke modellen voor stromingsproblemen in ondiep water, waarbij de modellen gebaseerd zijn op verschillende rooster-strukturen, verdienen koppelingsschema's op basis van karakteristieke vergelijkingen de voorkeur.

G.S. Stelling, Coupling 1-D and 2-D horizontal flow models, Proceedings Int. Conference Numerical Methods for Coupled Problems, Swansea, 1981.

Stelling VIII.

De in dit proefschrift beschreven methode voor de benadering van ondiep water vergelijkingen is te beschouwen als een methode gebaseerd op "operator splitting", die echter geen aanleiding geeft tot speciale problemen met betrekking tot de numerieke weergave van randvoorwaarden op de tussenniveau's.

Stelling IX.

Voor de numerieke approximatie van 1-dimensionale ondiep water vergelijkingen is het box schema bij uitstek geschikt door de afwezigheid van louter numerieke golven en de aanwezigheid van met de analytische oplossing overeenkomende eigenvectoren. Daardoor is de benadering van niet-reflecterende randvoorwaarden met behulp van het box schema vrijwel exact.

G.K. Verboom, G.S. Stelling en M.J. Officier;
Boundary Conditions for the Shallow Water Equations,
Engineering Applications for Computational Hydraulics, Volume I,
(M.B. Abbott en J.A. Cunge ed.) Pitman Publishing, 1982.

Stelling X.

Bij de berekening van langskrachten op schepen in sluizen met vul- en lediggingssysteem in de hoofden is het verschil tussen de resultaten verkregen op basis van de starre schip theorie of de flexibele schip theorie gering. Teneinde stabiliteitsproblemen bij numerieke berekeningen te reduceren verdient daarom de flexibele schip theorie de voorkeur.

J.P.Th. Kalkwijk,
Hydrodynamic forces and ship motions induced by surges in a navigation lock,
Proefschrift, T.H. Delft, 1973.

J. Bosma,
Langskrachten op schepen in sluizen met vul- en ledigingssysteem in de hoofden,
Waterloopkundig Laboratorium Delft, Rapport R1222/M1481, 1978.

G.S. Stelling,
Rekenschema's voor de water- en scheepsbeweging in een schutsluis,
Waterloopkundig Laboratorium Delft, Rapport S105, 1978.

K. den Boer,
Nadere analyse van langskrachten op schepen in sluizen met vul- en ledigingssystemen via hoofden, bodem of wanden,
Waterloopkundig Laboratorium Delft,
Rapport R1222/M1481-II, 1979.

Stelling XI.

Bij de invoering van informatica in het onderwijs verdient het aanbeveling dit vakgebied vooral te doceren als onderdeel van bestaande opleidingen in uiteenlopende richtingen en in mindere mate als zelfstandig specialisme.

Stelling XII.

Het volgens sommigen welhaast vaststaande feit dat computers binnenkort het schaakspel kunnen spelen op het niveau van grootmeesters, terwijl velen dit spel beschouwen als een aangename bezigheid, illustreert dat het een illusie is te menen dat computers uitsluitend gebruikt kunnen worden voor de automatisering van saai en onaangenaam werk.

Stelling XIII.

Een eenvoudige 7/8 tuigage is ongeschikt voor toerzeiljachten.
Dit wordt veroorzaakt door mogelijke instabiliteit van de spanningsverdeling in de hoofdwanten.