Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Fast and robust solution methods for the water quality equations

A thesis submitted to the
Delft Institute of Applied Mathematics
in partial fulfillment of the requirements

for the degree

MASTER OF SCIENCE
in
APPLIED MATHEMATICS

by

A. Morais

Delft, the Netherlands
June 2011

MSc THESIS APPLIED MATHEMATICS

“Fast and robust solution methods for the water quality equations”

A. Morais

Delft University of Technology

Daily supervisor	**Responsible professor**
Ir. L. Postma	Prof.dr.ir. C. Vuik

Other thesis committee members

Dr.ir. M. Genseberger	Dr. M. Borsboom

June 2011	Delft, the Netherlands

Contents

Chapter 1

Introduction

Deltares is an independent institute with high knowledge in the field of water, soil and subsoil. The institute is consulted by public authorities, engineering agencies or other companies for solving problems concerning the safety and environment of the society. One particular expertise of Deltares is doing quantitative water quality research. For these assessments models are used to solve the corresponding cases. In this thesis we consider methods how to solve these water quality models. An important objective is to improve the solution methods in order to obtain an efficient and accurate estimation of the quality of the water.

First we present in the next chapter the water quality model. This model is a mathematical description of the transport and reaction processes of substances in the water. Together with initial and boundary conditions the model for studying the water quality is complete.

In Chapter 3 we explain the problems we encounter when the water quality model is solved by the current numerical methods. This will be explained for both the time-dependent as the stationary case. The main objective is to improve the existing numerical methods, such that robustness and efficiency is obtained for the solution methods.

In Chapter 4 we present the Finite Volume Method for the discretization of the water quality equations. This method maintains the conservation property of the transport process. Next some definitions will be given for the Finite Volume scheme to hold such that the solution is mathematically and physically correct.

In Chapter 5 we present the current solution methods to tackle the problems discussed above for both the time independent and dependent case. In Chapter 6 we present other methods found in the literature to solve the problems defined in Chapter 2.

Finally, in the last two chapters we conclude about the current solution methods and explain how we will continue in the near future to improve these methods. The new solution methods will be applied for the Eems-Dollard case and if possible also to the Hong Kong case.

Chapter 2

The water quality model

In water a lot of different substances can be found. All these species move along each other or interact with each other. Therefore, for the study of the water quality the description is needed of the behaviour of these substances in the water. The behaviour is modelled in terms of transport and water quality processes. For the first kind two important types of transport can be considered: the advective transport and the diffusive transport. The former one is the transport due to the motion of the fluid. The substances are carried in the direction of the stream. The latter is transport due to random movement of the molecules and is also called molecular diffusion. The second type of processes contains physical-, chemical- and biological processes for the substances. For each substance all these processes can be expressed in a single equation, the water quality equation. Also sources and/or sinks are included in the equation.

The water quality equations are defined by the advection-diffusion-reaction equation. For each relevant substance in the water we have this type of PDE for its concentration. This equation describes the change in concentration due to the transport and water quality processes described above. Together with the PDE's of other species this forms our water quality model.

Let I be the total number of substances. For every substance $i \in I$ we have the following partial differential equation

$$\frac{\partial c}{\partial t} - \nabla \cdot (D \nabla c) + \nabla(\underline{u} c) = p, \tag{2.1}$$

where $c(\underline{x}, t)$ is the concentration of substance i in the water, D the diffusion coefficient, $\underline{u}$ the velocity vector and p represents the water quality processes for the substance in the water. The first term of the equation describes the change in time, the second and third term of the left hand side are the diffusion and advection terms respectively. The minus sign originates from the fact that diffusion causes net transport from higher to lower concentrations. Furthermore, all the coefficients in the equation are local and time independent..

A wide range of substances can be included in the water quality model (see also [9]), such as:

- conservative substances(salinity, chloride)
- decayable substances
- suspended sediment

- temperature
- nutrients(ammonia, nitrate, phosphate, silicate)
- organic matter
- dissolved oxygen
- algae
- bacteria
- heavy metals
- organic micro-pollutants

The term p on the right-hand side of Equation (2.1) consists of source terms $S(t)$ and water quality processes $f_R(c,t)$. Changes by sources include the addition of mass by waste loads and the extraction of mass by intakes. Water quality processes convert one substance to another, so there is interaction between several substances. Therefore the function p may depend on the concentration c of substance i and on the concentration c of other substances j, with $j \in I$. A wide range of these water quality processes can be given in the model, see also [9]. A few examples are:

- sedimentation
- reaeration of oxygen
- algae growth and mortality
- mineralisation of organic substances
- (de)nitrification
- adsorption of heavy metals
- volatilisation of organic micro-pollutants

Having explained the right-hand side of the water quality equation (2.1) it can be rewritten as

$$\frac{\partial c}{\partial t} - \nabla \cdot (D\nabla c) + \nabla(\underline{u}c) = f_R(c,t) + S(t). \tag{2.2}$$

An example of defining a water quality process is the following first order decay reaction

$$f_R = -kc, \quad \text{with } k \in \mathbb{R} \setminus \{0\}.$$

To complete the model formulation for the water quality both the initial and the boundary conditions must be specified . The conditions in general formulation are

$$\begin{aligned} c(\underline{x},0) &= c_0(\underline{x}) && (IC), && (2.3)\\ c|_{\underline{x}\in\partial\Omega} &= k_1\frac{\partial c}{\partial n} + k_2 c = g(\underline{x}) && (BC), && (2.4)\end{aligned}$$

with k_1 and k_2 given functions and g a given function defined on the boundary.

Chapter 3

Problem formulation

The water quality model is defined by partial differential equations (PDE). These are advection-diffusion-reaction equations and they describe the change in concentration due to the transport of substances in the water and their interaction with other substances. The formulation of the model is presented in the previous chapter. In this chapter we explain the concept of applying numerical methods to the water quality model. During the solution procedure some problems occur which we will present for both the time-dependent as time-independent case.

3.1 Time-dependent case

To solve the time-dependent water quality equations numerical methods are used. Together with the numerical model a grid is specified. The numerical scheme depends on the time and therefore time steps must be defined. In general a fine grid will lead to more accurate results than a coarser grid. Likewise for smaller time steps higher accuracy can be obtained than for larger time steps. Although high accuracy is desired it has the main disadvantage that computational costs are very high. Below we will discuss the relation between accuracy and costs for the spatial and time discretization separately.

There are different numerical methods for discretizing a partial differential equation (PDE) in space. For the spatial discretization distinction can be made between low order schemes and high order schemes. Low order schemes are less accurate compared with high order schemes. To achieve the same percentage of accuracy as the high order scheme the low order scheme must use a finer grid. The use of a finer grid automatically leads to higher costs and therefore a higher order method is mostly preferred above a low order scheme. But using low order numerical model can still be useful, since it has properties which are not present for high order schemes.

Besides the accuracy of the spacial discretization also the computational cost is an important aspect. The number of grid cells has a large effect on the computation time. So more cells corresponds with higher costs. The size of the cells for the one-dimensional case is denoted by Δx. The size of each grid cell is determined by mainly to parameters and will be presented below. The formula for the grid size is given by

$$\Delta x = \epsilon_L L,$$

where L is the length scale and $\epsilon_L \ll 1$. The parameter ϵ_L depends on the demanded accuracy and the accuracy of the spatial discretization method. Furthermore, the grid size for each cell also depends on the length scale of the corresponding area. For areas in the water with high activity of the transport and water quality processes the length scale is smaller than for areas with low activity. Consider for example a sudden release of waste in the river. In this case the length scale of the region near the release point is small in comparison with the region further from the release point. Furthermore, the length scale also varies in time. The length scale increases when the processes will reach their steady state.

To reduce the computation time the size of the grid cells must be increased by raising the parameter ϵ_L. By the definition of this parameter this can be reached by using a higher order spatial discretization method, e.g. high order upwind. For the discretization of the spatial terms high order schemes will use in general a larger value for ϵ_L than for low order schemes in order to have sufficient accuracy for the numerical solution. This approach has as additional benefit an increase in the accuracy.

For the time discretization a similar analysis can be done. Distinction can be made between the use of explicit and implicit numerical schemes. For explicit schemes a linear system of equations has to be solved whereas for implicit schemes a non-linear system of equations must be solved. But on the other hand, for the latter case larger time steps can be used than for the explicit case. This can be explained by the fact that for the explicit case a CFL condition must be satisfied in order to obtain stability for the numerical solution. For the implicit case no restriction on the time step is given, hence large time steps are taken. For the advection-diffusion equation the CFL condition reads

$$\Delta t \leq \frac{\Delta x^2}{u\Delta x + 2D}.$$

This condition holds for each of the directions in space. One can simply deduce from the CFL condition that when a fine grid is taken to increase the accuracy, then the time step is even much smaller. Although this will improve the numerical solution, it is computationally very expensive and therefore not desired.

Hence the time step Δt has a large effect on the computation time. The time step is also determined by two parameters. The formula for the time step yields

$$\Delta t = \epsilon_T T,$$

where T is the time scale and $\epsilon_T \ll 1$. The parameter ϵ_T depends on the required accuracy. Furthermore, the time step depends on the time scale for certain areas in the water. The time scale is for example smaller for the region near a discharge due to fast changes in the concentration, i.e. steep gradients for the concentration profile. Also it varies in time, since the time scale increases when steady state will be reached. Only for the explicit case we have a combination of the last two formulas to determine the time step. The formula is given by

$$\Delta t \leq \min\{\epsilon_T T, \frac{\Delta x^2}{u\Delta x + 2D}\}.$$

If the size of the CFL value is smaller than $\epsilon_T T$, then the time step is taken smaller than necessary.

Since for explicit schemes the computational costs are low, i.e. solving a linear system, we prefer to use this type of scheme as much as possible even we have to deal with a time restriction. Otherwise a implicit scheme is applied and a large time step can be used.

We may conclude that for solving the water quality model the correct spatial and time discretization method must be applied. By correct we mean to increase ϵ_T and ϵ_L in order to permit larger time steps and larger grid cells while keeping the accuracy high. This will probably lead to an accurate and computationally efficient numerical solution. Our goal is to construct a numerical model that is both robust (stable and accurate) and efficient.

3.2 Stationary case

For the water quality also the time-independent case can be studied. The corresponding equation reads

$$-\nabla \cdot (D\nabla c) + \nabla(\underline{u}c) = f_R(c). \tag{3.1}$$

Solving the equation above results in the steady state solution of the water quality model. Since no time-derivative is included in the equation it corresponds with solving a system of non-linear equations. In case of linear and independent water quality processes f_R, i.e. each f_R is a linear function of a single c_i, already good solvers exists to find the stationary solution. This forms no problem at all.

In reality the water quality processes are non-linear and/or dependent (f_R is a (non-)linear function of concentrations of several substances). A way to tackle this problem is to solve the time-dependent equation (2.1) and let $T \to \infty$ in the numerical model. Furthermore, the water quality processes are taken explicitly in the computation. In practice this method appears to be very time consuming for large water quality problems.

A better approach is to take the 'complex' water quality processes implicitly in the computation. This probably leads to a considerable improvement in the computation of the steady state.

Chapter 4

The numerical model

In the previous chapter the water quality model is described by means of a partial differential equation. Because the coefficients in the water quality model (2.1) are space and time dependent the PDE can not be solved analytically. Therefore in this thesis we will look at numerical methods to deal with this problem. One particular discretization procedure is the Finite Volume Method (FVM), which deals with the integral formulation of equation (2.1). The Finite Volume Method will be discussed below.

4.1 Finite Volume Method

Other important numerical methods for solving PDE's is the Finite Difference Method or the Finite Element Method. But the FVM is preferred since it is a fully mass conservative method in comparison with these other methods. Since for physical problems we deal with conservation of mass a numerical method must be used to preserve this property.

To apply the Finite Volume Method we first divide the domain of interest Ω completely into disjoint volume cells $V_i \subset \Omega$. This is the first step in the Finite Volume Method and is called grid generation.

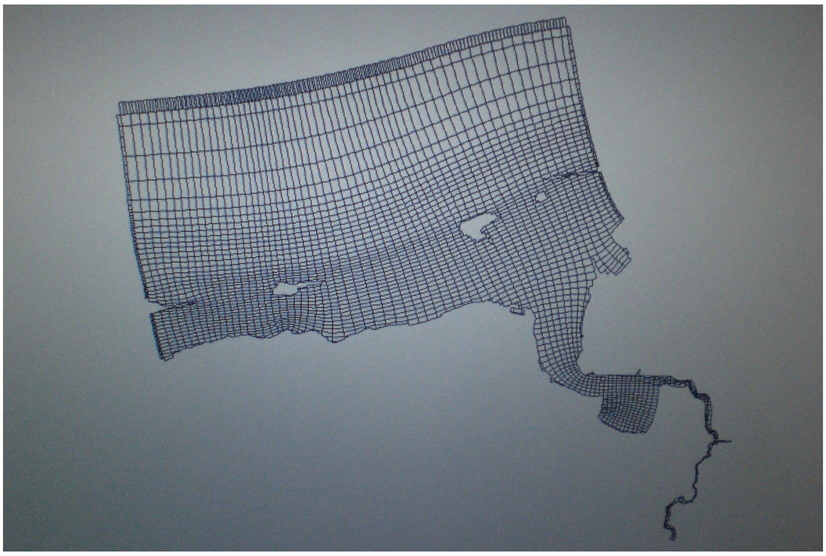

Figure 4.1: Example of a grid generation for the Eems-Dollard region

In the next step we integrate the water quality equation (2.1) piecewise over each volume cell V_i

$$\int_{V_i} [\frac{\partial c}{\partial t} - \nabla \cdot (D \nabla c) + \nabla(\underline{u} c)] \, dV = \int_{V_i} p \, dV, \quad i = 1, \ldots, N,$$

where V_i is the volume of cell i in the domain Ω.

Before proceeding with the final step we rewrite the equation above to a system of (non-linear) equations.

Since the time-derivative and integral can be interchanged the equation can be rewritten as

$$\frac{d}{dt} \int_{V_i} c \, dV - \int_{V_i} [\nabla \cdot (D \nabla c) - \nabla(\underline{u} c)] \, dV = \int_{V_i} p \, dV, \quad i = 1, \ldots, N.$$

Applying the Gauss' divergence theorem to the second term in the previous equation leads to

$$\frac{d}{dt} \int_{V_i} c \, dV - \oint_{\Gamma_i} [D \nabla c \; - \underline{u} c \,] \cdot \underline{n} \, d\Gamma = \int_{V_i} p \, dV, \quad i = 1, \ldots, N, \tag{4.1}$$

where Γ_i represents the total surface area of the cell and $\underline{n}$ is the unit vector normal to the surface pointing outward.

An important property of the FVM is that due to the piecewise integration of the PDE, the equation is expressed in average values. The quantities for the water quality case are defined as

$$\begin{aligned} c_i &= \frac{1}{|V_i|} \int_{V_i} c(\underline{x}, t) \, dV, & (4.2) \\ c_{ij} &= \frac{1}{|\Gamma_{ij}|} \int_{\Gamma_{ij}} c(\underline{x}, t) \, d\Gamma, & (4.3) \\ u_{ij} &= \frac{1}{|\Gamma_{ij}|} \int_{\Gamma_{ij}} \underline{u}(\underline{x}, t) \, d\Gamma, & (4.4) \\ p_i &= \frac{1}{|V_i|} \int_{V_i} p(\underline{x}, t) \, dV. & (4.5) \end{aligned}$$

In the first equation we defined the average concentration over the i^{th} cell at time t, in the second and third equation the average concentration and average velocity respectively over the interface of the i^{th} and j^{th} cell at time t and in the final equation the average water quality process over the i^{th} cell at time t. Here $| \cdot |$ represents the volume/area of the cell/boundary and Γ_{ij} is the joint-boundary/interface of cell i and j.

Furthermore we define the deviations for the concentration and the velocity

$$\begin{aligned} \hat{c}_{ij}(\underline{x}, t) &= c(\underline{x}, t) - c_{ij}, \quad \underline{x} \in \Gamma_{ij}, & (4.6) \\ \underline{\hat{u}}_{ij}(\underline{x}, t) &= \underline{u}(\underline{x}, t) - \underline{u}_{ij}, \quad \underline{x} \in \Gamma_{ij}, & (4.7) \end{aligned}$$

Using the definitions (4.2) and (4.5) Equation (4.1) is rewritten as

$$\frac{d|V_i|c_i}{dt} - \sum_{j\in J_i}[\int_{\Gamma_{ij}} (D\nabla c \cdot \underline{n}_{ij} \; - c\underline{u} \cdot \underline{n}_{ij})\, d\Gamma] \; = |V_i|p_i, \quad i = 1, \ldots, N. \tag{4.8}$$

If we substitute next the definitions (4.6)-(4.7) in the advection part of the summation term and use that the average deviation of the average is zero we get

$$\frac{d|V_i|c_i}{dt} - \sum_{j\in J_i}[\int_{\Gamma_{ij}} (D\nabla c \cdot \underline{n}_{ij} - \hat{c}_{ij}\hat{\underline{u}}_{ij} \cdot \underline{n}_{ij})\, d\Gamma \; - |\Gamma_{ij}|c_{ij}\underline{u}_{ij} \cdot \underline{n}_{ij}] \; = |V_i|p_i, \quad i = 1, \ldots, N. \tag{4.9}$$

The integral term in Equation (4.9) represents the new diffusion term which consist of the molecular diffusion and the turbulent diffusion. In the turbulent diffusion one has the term $\hat{c}_{ij}\hat{\underline{u}}_{ij}$, which can be seen as non-normalized correlation coefficient i.e. a measure of the linear dependence between two random variables. Remember the correlation coefficient for two random variables X and Y which is defined as

$$\rho_{X,Y} = \frac{\sum_i^N (x_i - \overline{x})(y_i - \overline{y})}{\sigma_x \sigma_y},$$

where $\overline{x}$ and $\overline{y}$ are the mean and σ_x and σ_y are standard deviations of X and Y respectively. In our particular case the random variables are the concentration and the velocity (see Equations (4.2) – (4.7)). The turbulent diffusion is a numerical phenomenon. The magnitude of this term depends on the size of the grid, hence the turbulent diffusion term vanishes as the grid is refined.

For the spatial derivatives in (4.9) numerical difference formulas can be used, such as central difference or one-sided difference. Next we use a time integration method to solve equation (4.9). This numerical solution should lead to an accurate solution of the water quality model (2.1). Below we illustrate the FVM described above for a simple one-dimensional example.

Example 4.1

Let's consider the following one-dimensional water quality equation

$$\frac{\partial c}{\partial t} - D(x)\frac{\partial^2 c}{\partial x^2} + u(x)\frac{\partial c}{\partial x} = p(x), \tag{4.10}$$

with x defined on Ω = [a,b] and $c(x, 0) = c_0(x)$.

First we divide our interval Ω in N subintervals with equidistant cell-size. In each volume cell we define a node at the center of the cell.

Integrating the equation above over a volume $V_i = (x_{i-1/2}, x_{i+1/2})$ results in

$$\frac{d}{dt}\int_{x_{i-1/2}}^{x_{i+1/2}} c\, dx - \int_{x_{i-1/2}}^{x_{i+1/2}} [D\frac{\partial^2 c}{\partial x^2} - u\frac{\partial c}{\partial x}]\, dx = \int_{x_{i-1/2}}^{x_{i+1/2}} p\, dx, \quad i = 1, \ldots, N.$$

Next we apply Gauss' divergence theorem to the previous equation. This yields

$$\frac{d}{dt}\int_{x_{i-1/2}}^{x_{i+1/2}} c\,dx - (D_{i+1/2}\frac{\partial c_{i+1/2}}{\partial x} - D_{i-1/2}\frac{\partial c_{i-1/2}}{\partial x}) + (u_{i+1/2}c_{i+1/2} - u_{i-1/2}c_{i-1/2}) = \int_{x_{i-1/2}}^{x_{i+1/2}} p\,dx.$$

Using the average method described above we obtain

$$\frac{dc_i}{dt}h - (D_{i+1/2}\frac{\partial c_{i+1/2}}{\partial x} - D_{i-1/2}\frac{\partial c_{i-1/2}}{\partial x}) + (u_{i+1/2}c_{i+1/2} - u_{i-1/2}c_{i-1/2}) = p_i h,$$

with c_i and p_i given by (4.2) and (4.5) respectively and $D_m = D(x_m, t)$, $u_m = u(x_m, t)$ and $h = x_{i+1/2} - x_{i-1/2}$ for an equidistant grid.

Next we assume a constant diffusion coefficient D and velocity u for each cell i. Using for example central differences for the spatial derivative and central average for the zeroth order derivative we get

$$\frac{dc_i}{dt}h - D_i\frac{c_{i-1} - 2c_i + c_{i+1}}{h} + u_i\frac{c_{i+1} - c_{i-1}}{2} = p_i h. \tag{4.11}$$

A better choice for the difference and average is possible, but at this point this is not relevant since we only illustrate how the PDE can be transformed to a system of equations. After rearrangement equation (4.11) can be written in matrix-vector form as

$$M\frac{d\underline{c}}{dt} + S\underline{c} = \underline{f}, \tag{4.12}$$

where M and S presents the mass and stiffness matrix respectively.

Then one can apply a time integration method to equation (4.12), such as Euler to write the equation in the form

$$(\frac{1}{\Delta t}M - \theta S)\underline{c}^{n+1} = (\frac{1}{\Delta t}M - (1-\theta)S)\underline{c}^{n} + \underline{f}, \tag{4.13}$$

with $\theta = 0$ for Euler Forward and $\theta = 1$ for Euler Backward. For $\theta = 0.5$ we get the method of Crank-Nicholson which has a higher order of accuracy than the Euler methods (second and first order resp.). The letter n denotes the time t_n, i.e. $t_n = n\Delta t$. As starting point we use the initial condition $\underline{c}^0 = c_0(x)$. Hence we are able to solve the discretized water quality equation (4.13) for $\underline{c}$. The value c_i^n is an approximate average value over the i^{th} volume cell at time t_n.

4.2 Convergence of FVM scheme

The FVM leads to a discrete numerical model of a partial differential equation. Several definitions will be introduced below which are important to measure the quality of this model.

An important feature for a FVM scheme (or any other numerical scheme) is that local errors do not grow catastrophically and hence a bound on the global error can be obtained in terms of these local errors. If this description holds for the FVM scheme, then the numerical method is called stable. First we present the definitions for the global and local error.

The global error at a time t_n is given by

$$E^n = q(x, t_n) - c(x, t_n),$$

with q an approximation by the FVM scheme and c the true value. The local truncation error at time t_n is defined as

$$\tau^n = \frac{\mathcal{N}(c(x, t_{n-1})) - c(x, t_n)}{\Delta t},$$

where $\mathcal{N}(\cdot)$ represents the numerical operator.

Stability for a numerical method is given in the following definition.

Definition 4.2
Let $||\cdot||$ be some norm, then a numerical method is stable if

$$||\mathcal{N}(q^n)|| \leq ||\mathcal{N}(q^0)||,$$

where $\mathcal{N}(\cdot)$ represents the numerical operator mapping the approximate solution at one time step to the approximate solution at the next time step, i.e. $q^{n+1} = \mathcal{N}(q^n)$.

If definition 4.2 holds, then the global error is bounded for each time step. Since $q^0 = q(x, t_0) = c(x, 0) + E^0$, we have due to the boundedness that a small perturbation in the initial condition leads to a small change in the solution.

Next we discuss if the FVM scheme is consistent with the PDE. This means that the local truncation error vanishes as the grid is refined.

Definition 4.3
A method is called consistent with the partial differential equation if the local truncation error at time t_n in some norm satisfies

$$\lim_{\Delta t \to 0, \Delta x \to 0} ||\tau^n|| = 0, \quad \text{with x fixed.}$$

The dominant term of the truncation error determines the order of accuracy of the numerical method. We say that a method is accurate of order s_1 in time and accurate of order s_2 in space if

$$||E^N|| = O(\Delta t^{s_1}) + O(\Delta x^{s_2}), \tag{4.14}$$

where N is the total number of time steps, i.e. $N\Delta t = T$.

But what can be said about the numerical solution? Does this solution approximates sufficiently well the real unknown solution? Having stability and consistency for a numerical method we have indeed convergence of the numerical solution to the real solution according to the Fundamental Theorem of numerical methods for PDE's. This theorem can be summarized as

$$\text{consistency } + \text{ stability } \Longrightarrow \text{ convergence.}$$

Definition 4.4
A method is convergent at time t_n in some norm if the global error satisfies

$$\lim_{\Delta t \to 0, \Delta x \to 0} ||E^n|| = 0, \quad \text{with x fixed.}$$

4.3 Positivity of FVM scheme

For the computation of the numerical solution of the water quality model it is important to obtain non-negative values. Negative values are unphysical and may cause instability for its solution. To avoid these negative values we will see in the next chapter that a correct choice for the spatial derivatives is important.

First we discuss the use of basic difference methods for the spatial derivatives. Remember the FVM scheme for Example 4.1 given in the previous section, which reads

$$\frac{dc_i}{dt}h - D_i\frac{\partial^2 c_i}{\partial x^2}h + u_i\frac{\partial c_i}{\partial x}h = p_i h,$$

with c_i and p_i given by (4.2) and (4.5) respectively and $D_i = D(x_i, t)$, $u_i = u(x_i, t)$ and $h = x_{i+1/2} - x_{i-1/2}$ for an equidistant grid.

Central differences for the second order derivative leads to satisfactory results, i.e. no unphysical behaviour, hence a positive solution. For the first order derivative problems arise. In [5] it is shown that for this case central differences leads to an unstable solution. A good alternative is to use one-sided differences, also called first order upwind. Although this method is less accurate the solution will be positive, i.e. non-negative values for its entries. After applying the correct differences for the spatial derivatives we obtain an ordinary differential equation (ODE) system, such as (4.12). In the next chapter we will present an advanced method for discretizing the spatial derivatives, such that positivity is ensured.

In order to find a positive solution it is necessary for the ODE system to be positive. A definition is given below.

Definition 4.5
An ODE system $\frac{dc}{dt} = F(c(t))$ is called positive, or non-negative if $c(0) \geq 0$ (component-wise) implies $c(t) \geq 0$ for all $t > 0$.

The ODE system can be solved numerically by applying a time-integration method. The definition mentioned above have to be translated to time-integration methods, such that the numerical method can be called positive.

Definition 4.6
A time integration method $c^{n+1} = \phi(c^n)$ is called positive if for all n ≥ 0 holds

$$c^n \geq 0 \Longrightarrow c^{n+1} \geq 0.$$

In practice non-negativity of a time-integration method is difficult to guarantee. In [2] it has been shown that positive time integration methods can be first order accurate only. Euler Backward is the only known method being positive. In the next few chapters we at least try to obtain a positive scheme when the spatial discretization is applied. If negative values occur after the time integration then it must be corrected in some way.

Chapter 5

The current solution methods

In this chapter we present the current solution methods for the two problems defined in Chapter 3. Next to the treatment of these numerical methods simple examples will be given to show their results.

5.1 Solving the water quality equations

In this section we will give the method of flux-limiters for solving the water quality model. First we start with a first order method and continue by building up to more accurate methods.

In the previous chapter the water quality equations are discretized by the Finite Volume Method. This finally led to equation (4.1). The accuracy of the solution depends on the choice for the discretization of the spatial derivative terms, especially for the advective term. The current methods based on flux-limiters are constructed such that it also deals with steep gradients. This situation occurs when there is a sudden release of a substance in the water.

We have seen in Section 4.1 that for the spatial derivative terms several difference formulas can be used. For the diffusive term central differences are used as standard since it is second order accurate and no relevant problems occur in the case of steep gradients. Only for the first order derivative in space special care has to be taken. Therefore only the advection equation equation will be studied.

The methods described below will be illustrated for the one-dimensional homogeneous advection equation given by

$$\frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} = 0, \quad x \in [0, 10], \; t \geq 0, \tag{5.1}$$

where $u > 0$ (substances flow from left to right) is constant and with periodic boundary conditions

$$c(0, t) = c(10, t), \quad t \geq 0, \tag{5.2}$$

and with initial condition

$$c(x, 0) = 1_{[2,4]}(x)\frac{1}{2}(1 - cos(\pi x)) + 1_{[6,8]}(x), \quad x \in [0, 10]. \tag{5.3}$$

Since the water quality model is a conservative system the discretization of the equation above is presented in terms of fluxes coming in and going out. Discretizing equation (5.1) by the FVM leads to the following system of equations

$$c_i^{n+1} = c_i^n - \frac{\Delta t}{\Delta x}(F_{i+1/2}^n - F_{i-1/2}^n) \qquad i = 1, \ldots, N \tag{5.4}$$

where the term $F_{i-1/2}^n$ is some approximation of the average flux along the left boundary $x = x_{i-1/2}$ of control volume V_i. Further c_i represents the average concentration of volume cell V_i. The FVM is chosen to be written in this form, since we are dealing with a conservative system, hence Equation (5.4) is a conservative scheme. Therefore the fluxes at the boundaries are important to study.

5.1.1 The upwind method

First of all we use a simple approximation for this average flux before continuing with more accurate methods. The simplest one is the upwind method, which only uses information coming from one side, dependent on the direction of the stream.

For the upwind method for the advection term we have $F_{i-1/2}^n = uc_{i-1}^n$ and $F_{i+1/2}^n = uc_i^n$. The upwind method is first order accurate. By using the Taylor series expansion for the upwind scheme it can be shown that the method introduces diffusive behaviour. The modified equation yields

$$\frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} = \frac{1}{2}u\Delta x(1 - u\frac{\Delta t}{\Delta x})\frac{\partial^2 c}{\partial x^2}, \tag{5.5}$$

where the right-hand side represents a diffusion term.

It is necessary for the upwind method to satisfy the CFL condition $u\frac{\Delta t}{\Delta x} \leq 1$, such that the numerical upwind scheme is stable and converge to the solution of the differential equation as the grid is refined. The main advantage of using the upwind method is that the solution is monotone and negative values do not appear. A disadvantage is that the upwind method leads to severe damping of the numerical solution. This is caused by the artificial numerical diffusion term in Equation (5.5).

Below we present the features mentioned above in a plot.

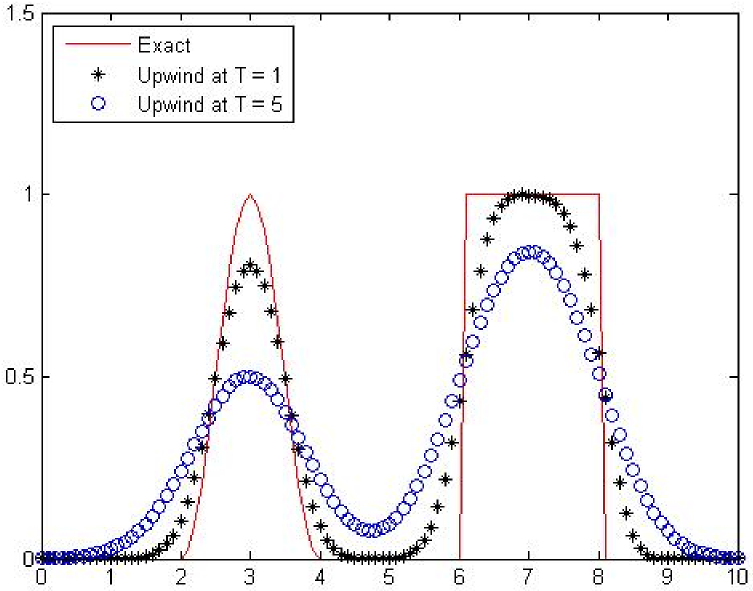

Figure 5.1: Upwind method

5.1.2 Lax-Wendroff method

Also more accurate flux approximations can be used such as the Lax-Wendroff method, which is a second-order accurate method. This method has an extra term to correct for the diffusive upwind part. For the Lax-Wendroff method applied to the one-dimensional advection equation with $u > 0$ we have the following flux approximation

$$F^n_{i-1/2} = \frac{1}{2}u(c^n_{i-1} + c^n_i) - \frac{1}{2}\frac{\Delta t}{\Delta x}u^2(c^n_i - c^n_{i-1}),$$
$$F^n_{i+1/2} = \frac{1}{2}u(c^n_i + c^n_{i+1}) - \frac{1}{2}\frac{\Delta t}{\Delta x}u^2(c^n_{i+1} - c^n_i).$$

The equations above can be rewritten as

$$\begin{aligned} F^n_{i-1/2} &= uc^n_{i-1} + \frac{1}{2}u(c^n_i - c^n_{i-1})(1 - \frac{\Delta t}{\Delta x}u), \\ F^n_{i+1/2} &= uc^n_i + \frac{1}{2}u(c^n_{i+1} - c^n_i)(1 - \frac{\Delta t}{\Delta x}u). \end{aligned}$$

The equations indeed present an anti-diffusion term which is absent for the upwind term. Also for this case the CFL condition $u\frac{\Delta t}{\Delta x} \leq 1$ must be satisfied. The advantage of this method is that the numerical solution is more accurate than the upwind method, especially for smooth parts. It also has no diffusive behaviour, since the numerical diffusion is equal to zero. On the other hand it gives wiggles near steep gradients. As a consequence we may get negative values for our numerical solution which is not desired.

Below we present the features mentioned above in a plot.

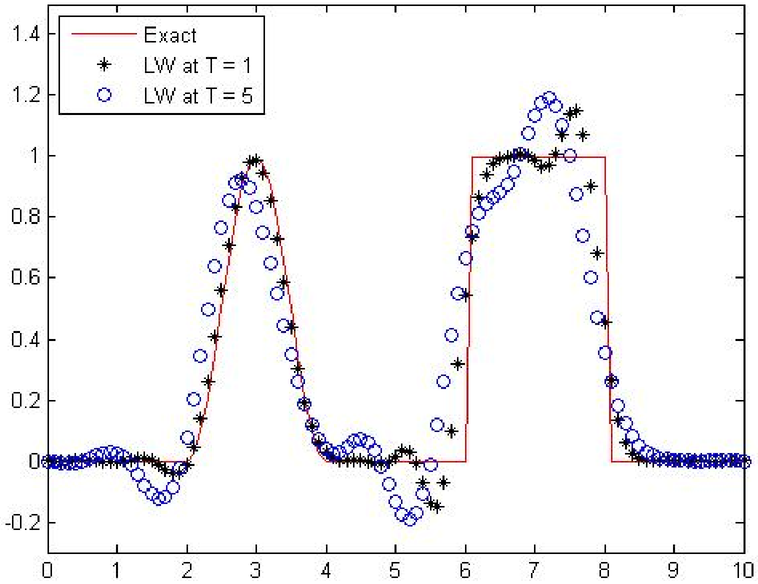

Figure 5.2: Lax-Wendroff method

5.1.3 Flux-limiters

Both the upwind- and Lax-Wendroff method has their limitations for dealing with the advective term. The former introduces extra diffusion of the solution, while the latter one introduces oscillation in the solution. Therefore we introduce flux-limiters, which try to compensate for the errors made by low order methods.

A better approximation for the average flux is to combine the best features of both the first-order method and the second-order method. So generally the flux limiter method is defined as

$$F^n_{i-1/2} = F^n_{i-1,i} = F_L(c_{i-1}, c_i) + l^n_{i-1/2}(F_H(c_{i-1}, c_i) - F_L(c_{i-1}, c_i)), \tag{5.6}$$

which is a convex combination of a low-order flux formula F_L and a high order flux formula F_H. The value for l^n_{ij} determines the weight over the two formulas and is also called the limiter. A limiter of 0 corresponds with a low order method and for a limiter of 1 we obtain a high order method. The term $F_H(c_{i-1}, c_i) - F_L(c_{i-1}, c_i)$ is an anti-diffusion/flux-correction term. This term corrects for the diffusive behaviour made by the first order flux method.

So combining for example the upwind method with any second order method will lead to a monotonous solution with no diffusive behaviour. We present an example in the next section.

In Delft3D-WAQ, a simulation software used at Deltares for solving water quality problems, several flux-limiters can be used. Each of these methods has their own benefits and drawbacks and therefore are applicable to a number of situations. One very important flux-limiter method is the flux corrected transport (FCT) method. In the rest of this section we will consider this type of method.

We will present a flux corrected transport (FCT) method developed by Boris and Book for the one-dimensional advection equation. This explicit flux limiter method involves a step by step method to determine the value of the limiter l_{ij}^n in equation (5.6). The starting point is an explicit discretized equation like (5.4) with initial condition $c^0(x) = c(x, 0)$.

The flux corrected transport(FCT) of Boris and Book.

Below we present the FCT method constructed by Boris and Book [7], but formulated by Zalesak [8], such that it can be extended easily to multi-dimensions.

1. First we compute a first order approximation $\tilde{c}_i^{n+1}$ by means of (5.4) with $l_{ij}^n = 0$ in (5.6), i.e. $F_{ij}^n = F_L(c_i^n, c_j^n)$, where F_L is a low order flux function, i.e. first order upwind. Further $j \in J_i$, where J_i consists of node i and its nearest neighbours. The low order scheme is guaranteed to give monotonic results.

2. Next we define the flux correction by $\Delta F_{ij}^{n+1} = F_H(c_i^n, c_j^n) - F_L(c_i^n, c_j^n)$, where F_H is a high order flux function and $j \in J_i$. It is sometimes referred to as anti-diffusion, since it often corrects the numerical diffusion of the low order flux.

3. Then we apply a prelimiting step. Since the term ΔF_{ij}^{n+1} corrects the diffusive first order flux it should not be diffusive. Hence we set $\Delta F_{ij}^{n+1} = 0$ if $\Delta F_{ij}^{n+1}(\tilde{c}_i^{n+1} - \tilde{c}_j^{n+1}) > 0$, with $j \in J_i$.

4. An important property for the solution is to be monotone, hence no new local maximum or minimum must be created during the iteration steps of the FCT algorithm . Therefore we determine an upper and lower bound for $\tilde{c}_i^{n+1}$ computed by step 1

$$c_i^{\max} = \max_{j \in J_i} \tilde{c}_j^{n+1},$$

$$c_i^{\min} = \min_{j \in J_i} \tilde{c}_j^{n+1},$$

where J_i consists of node i and its nearest neighbours.

These quantities will be used in the next step.

5. Next we define the amount of mass that flows into cell V_i

$$P_i^+ = \sum_{j \in J_i \setminus \{i\}} \max(0, -\Delta F_{ij}^{n+1}).$$

The allowed mass increase is

$$Q_i^+ = |V_i|(c_i^{\max} - \tilde{c}_i^{n+1}),$$

where $|V_i|$ is the volume of cell i.

The fraction of mass that is allowed to flow into the cell is given by

$$R_i^+ = \begin{cases} \min(1, \frac{Q_i^+}{P_i^+}), & P_i^+ > 0, \\ 0, & P_i^+ = 0. \end{cases}$$

For mass decrease we can define analogue the following quantities:

$$\begin{aligned} P_i^- &= \sum_{j \in J_i \setminus \{i\}} \max(0, \Delta F_{ij}^{n+1}), \\ Q_i^- &= |V_i|(\tilde{c}_i^{n+1} - c_i^{\min}), \\ R_i^- &= \begin{cases} \min(1, \frac{Q_i^-}{P_i^-}), & P_i^- > 0, \\ 0, & P_i^- = 0. \end{cases} \end{aligned}$$

The values R_i^+ and R_i^- guarantees no overshoot and undershoot in cell i respectively.

6. In the next step we determine the limiter which is the mass fraction that is allowed by both adjacent cells

$$l_{ij}^{n+1} = \begin{cases} \min(R_j^+, R_i^-), & \Delta F_{ij}^{n+1} \geq 0 \\ \min(R_i^+, R_j^-), & \Delta F_{ij}^{n+1} < 0, \end{cases}$$

where $j \in J_i$.

7. With the previous step we finally update the solution by Equation (5.4) with c_i^n replaced by the first order approximation $\tilde{c}_i^{n+1}$ computed in the first step and with

$$F_{i+1/2}^n = l_{i,i+1}^{n+1} \Delta F_{i,i+1}^{n+1}, \tag{5.7}$$

$$F_{i-1/2}^n = l_{i,i-1}^{n+1} \Delta F_{i,i-1}^{n+1}. \tag{5.8}$$

5.1.4 Local-theta FCT scheme

In this section we describe an efficient and accurate numerical time stepping scheme. This method combines the local-theta method with the FCT-limiter of Boris and Book. The local-theta method uses an optimal local θ rather than a constant value. The method is explained for the one-dimensional advection equation (5.1).

The theta method is a time discretization method and for an ODE system $\frac{dc}{dt} = f(c)$ it is defined as

$$\frac{c_i^{n+1} - c_i^n}{t_{n+1} - t_n} = (1-\theta) f(c_i^n) + \theta f(c_i^{n+1}).$$

In the local theta method we take a local θ per flux in stead of a constant value. So, the local-theta scheme for the 1D homogeneous water quality model (p=0) is given as

$$\Delta x \frac{c_i^{n+1} - c_i^n}{t_{n+1} - t_n} = -[(1 - \theta_{i,i+1}^{n+1}) F_{i+1/2}^n + \theta_{i,i+1}^{n+1} F_{i+1/2}^{n+1} - (1 - \theta_{i,i-1}^{n+1}) F_{i-1/2}^n - \theta_{i,i-1}^{n+1} F_{i-1/2}^{n+1}], \tag{5.9}$$

where $F^n_{i-1/2}$ and $F^n_{i+1/2}$ is some numerical advective flux along the left and right boundary of cell i respectively.

The value for θ^n_{ij} must be chosen as small as possible to minimize the amount of numerical diffusion and large enough to ensure that the scheme is stable, positivity preserving and non-oscillatory. In this way the accuracy of the theta method is improved without loss of robustness. A strategy for obtaining these optimal values is defined for each volume cell and is for the 1D case given by

$$\theta^n_{ij} = \max\{0, \theta^n_i, \theta^n_j\}, \tag{5.10}$$

with

$$\theta^n_i = 1 - \frac{\Delta x_i}{u \Delta t}. \tag{5.11}$$

The FCT-method of Boris and Book presented in Section 5.1.3 combined with Equation (5.9) leads to the local-theta FCT scheme. This method leads again to an increase in the accuracy, since a larger reduction of the numerical diffusion is obtained. This scheme is defined by substitution of the following flux-limiters into equation (5.9)

$$\begin{aligned} F^M_{i-1/2} &= F_L(c^M_{i-1}, c^M_i) + l^{n+1,M}_{i,i-1}(F_H(c^M_{i-1}, c^M_i) - F_L(c^M_{i-1}, c^M_i)), & (5.12)\\ F^M_{i+1/2} &= F_L(c^M_i, c^M_{i+1}) + l^{n+1,M}_{i,i+1}(F_H(c^M_i, c^M_{i+1}) - F_L(c^M_i, c^M_{i+1})), & (5.13) \end{aligned}$$

with $M \in \{n, n+1\}$ and where the limiter l_{ij} is determined according to the algorithm of Boris and Book and F_H and F_L corresponds with a high order and a low order flux function respectively.

The local-theta FCT method makes sure that the solution is computed explicitly as much as possible, i.e. $\theta_{ij} \in [0, \frac{1}{2})$. An explicit scheme is clearly less expensive. According to [1], $\theta_{ij} \in [0, \frac{1}{2})$ introduces unphysical anti-diffusion. This anti-diffusion can be eliminated by raising the local theta coefficients to a minimal value of $\frac{1}{2}$. As side effect we get an increase of the numerical diffusion.

5.1.5 Summary of numerical methods

In the previous sections some important numerical methods were presented for solving the water quality equations. In this section these methods are compared on basis of their properties and accuracy. For the definition of the accuracy we refer to Section 4.2. The results are given in the table below.

Numerical scheme	Order of accuracy	Advantage	Disadvantage
Upwind	$O(\Delta t)$	positive	diffusive
Lax-Wendroff	$O(\Delta t^2)$	zero numerical diffusion	oscillations occur
FCT	$O(\Delta t)$ - $O(\Delta t^2)$	positive and less diffusive	extra anti-diffusion possible
Local-theta FCT	$O(\Delta t)$ - $O(\Delta t^2)$	efficient	unphysical anti-diffusion

5.2 Computing the stationary solution

In this section we present the solution method for computing the stationary solution. In Section 3.1 we already mentioned that linear source terms are no problem when solving the stationary solution. Since the source terms can be non-linear and also depends on the concentration of several substances, these terms are computed explicitly to avoid solving a nonlinear system . This means that the stationary equation is computed in two steps.

Remember the time-dependent water quality equation

$$\frac{\partial c}{\partial t} - \nabla \cdot (D \nabla c) + \nabla(\underline{u} c) = S + f_R, \tag{5.14}$$

where the terms in the left-hand side of the PDE are the diffusion and advection term respectively. The latter two terms are so-called source terms. The term S stands for discharges or 'waste loads' and the term f_R stands for reaction terms or 'processes' f_R.

Suppose that c_i^n is known. The equation of the first step reads

$$\frac{\tilde{c}_i^{n+1} - c_i^n}{\Delta t} = p_i^n,$$

where p_i^n represents the source terms and $\tilde{c}_i^{n+1}$ is an intermediate solution.

Next the second step must be solved. The equation is given by

$$\frac{d|V_i|c_i}{dt} = -(A\underline{c})_i,$$

where A is a matrix representing the discretized transport processes. The right-hand side stands for the i^{th} entry of the vector $A\underline{c}$. In this step the intermediate solution $\tilde{c}_i^n$ is used at the previous time t_n. Note that this equation can be non-linear.

For each time step this will take in total $N+1$ iterations per substance to solve both equations. N iterations for the second step. If we assume a total of M substances with each using the same number of iterations $N+1$, this results in $(N+1) \times M$ iterations per time step. Computationally it can be very inefficient for large values for M and N.

Chapter 6

New numerical methods

In the first section we introduce an implicit flux-limiter method by Kuzmin. In the next section we look at inexact Newton methods for solving non-linear equations.

6.1 Flux-limiter by Kuzmin

In Section 5.1 we described the FCT method of Boris and Book. In this section we present a new FCT procedure by Kuzmin and Turek [4], which is an implicit flux limiter method. Roughly, the main steps to be performed are as follow:

1. Discretize the water quality equations
2. Define a low-order transport operator
3. Define the anti-diffusive fluxes
4. Apply a pre-limiting step
5. Cancel anti-diffusive fluxes for local extrema situations
6. Define the remaining correction factors α
7. Update the solution

The several steps in the algorithm are treaten in more detail below for the one-dimensional homogeneous case. The equation for the one-dimensional water quality equation reads

$$\frac{\partial c}{\partial t} - \frac{\partial}{\partial x}(D\frac{\partial c}{\partial x}) + \frac{\partial}{\partial x}(uc) = 0. \tag{6.1}$$

1. The first step is to perform a spatial discretization to obtain the following semi-discrete problem

$$M_C \frac{dc}{dt} = K^H c, \tag{6.2}$$

where M_C is the mass matrix and K^H is the discrete transport operator which has zero row sum. An operator has zero row sum if the following holds

$$k_{ii}^H = -\sum_{j\neq i} k_{ij}^H. \tag{6.3}$$

So for every entry of the right-hand side of Equation (6.2) we have

$$\begin{aligned}
(K^H c)_i &= \sum_{j=1} k_{ij}^H c_j = k_{ii}^H c_i + \sum_{j\neq i} k_{ij}^H c_j \\
&= -\sum_{j\neq i} k_{ij}^H c_i + \sum_{j\neq i} k_{ij}^H c_j \qquad \text{by (6.3)} \\
&= \sum_{j\neq i} k_{ij}^H (c_j - c_i).
\end{aligned}$$

2. In the second step a low-order transport operator is constructed by

$$K^L = K^H + D, \tag{6.4}$$

where D is designed to eliminate all negative off-diagonal entries of the high-order operator. The operator D is defined as

$$d_{ii} = -\sum_{k\neq i} d_{ik}, \tag{6.5}$$

$$d_{ij} = d_{ji} = \max(0, -k_{ij}^H, -k_{ji}^H), \text{ for all } i < j. \tag{6.6}$$

D is constructed such that it has zero row and column sum, and so have all the properties of generalized diffusion operators including mass conservation. If K^H has nonnegative off-diagonal entries then K^L and K^H are identical.

The semi-discrete low order scheme reads

$$M_L \frac{dc}{dt} = (K^H + D)c = K^L c, \tag{6.7}$$

where M_L is the lumped mass matrix. This is a matrix with only non-zero elements on the diagonal. In the one-dimensional case we have $M_L = M_C$.

For time-schemes other then Euler backward the time step is bounded in order to preserve positivity

$$\Delta t \leq \frac{1}{1-\theta} \min_i (-m_i / k_{ii}^L \mid k_{ii}^L < 0), \tag{6.8}$$

where θ is the degree of implicitness and m_i the i^{th} diagonal element of the lumped mass matrix.

3. In the third step we determine the anti-diffusive fluxes. Therefore we compute first the high order solution c^H by the standard θ-scheme applied on Equation (6.7). Furthermore, a

correction term is included in the computation. The discretized method is given by

$$(M_L - \theta \Delta t K^L)c^H = (M_L + (1-\theta)\Delta t K^L)c^n + F(c^H, c^n), \tag{6.9}$$

where the antidiffusion term F responsible for high spatial accuracy is given by

$$F(c^H, c^n) = -\Delta t(K^L - K^H)[\theta c^H + (1-\theta)c^n]. \tag{6.10}$$

The antidiffusion term is a flux correction similar as in step 2 of the FCT-method of Boris and Book. It is clear that omitting the antidiffusive term leads to a low-order scheme, whereas retaining it leads to the original high-order scheme.

From (6.9) the anti-diffusive fluxes are defined as

$$f_{ij} = -\Delta t d_{ij}[\theta(c_j^H - c_i^H) + (1-\theta)(c_j^n - c_i^n)], \quad f_{ij} = f_{ji}, \;\; i < j.$$

The flux-corrected version of (6.9) can be written in the form

$$m_i c_i^{n+1} - \theta \Delta t \sum_j k_{ij}^L c_j^{n+1} = m_i \tilde{c}_i + \sum_{j \neq i} \alpha_{ij} f_{ij}, \quad \alpha_{ij} = \alpha_{ji}, \tag{6.11}$$

where α_{ij} are the correction factors computed and $\tilde{c}$ represents the positivity-preserving solution to the explicit subproblem

$$m_i \tilde{c}_i = m_i c_i^n + (1-\theta)\Delta t \sum_j k_{ij}^L c_j^n. \tag{6.12}$$

The solution $\tilde{c}$ is an intermediate solution computed at the time instant $t^{n+1-\theta}$ by the explicit low-order scheme.

4. Next we apply a pre-limiting step which is an important component of the FCT limiter. The purpose is to cancel those antidiffusive fluxes that directed down the gradient of $\tilde{c}$. So we prevent the antidiffusive flux to be diffusive. The test to be performed is

$$f_{ij} = 0, \quad \text{if} \quad f_{ij}(\tilde{c}_i - \tilde{c}_j) < 0. \tag{6.13}$$

Without using this step the antidiffusive fluxes that are diffusive would cause wiggles in the numerical solution. Hence no monotonicity-preserving [4].

5. In the following step we first determine the maximum and minimum values for $\tilde{c}_i$

$$c_i^{\max} = \max_{j \in J_i} \tilde{c}_j, \tag{6.14}$$

$$c_i^{\min} = \min_{j \in J_i} \tilde{c}_j, \tag{6.15}$$

where J_i consists of node i and its nearest neighbors. For the one-dimensional case J_i consists of 3 values for interior points x_i, whereas for boundary points x_i J_i consists of two values.

The reason for determining (6.14) and (6.15) is to cancel completely those antidiffusive fluxes which try to accentuate a local maximum or minimum. This is in accordance with the FCT theory and the cancellation step reads

$$\alpha_{ij} = 0 \quad \text{if} \quad \tilde{c}_i = c_i^{\max} \text{ and } f_{ij} > 0, \qquad \text{or} \tag{6.16}$$
$$\alpha_{ij} = 0 \quad \text{if} \quad \tilde{c}_i = c_i^{\min} \text{ and } f_{ij} < 0. \tag{6.17}$$

6. For those fluxes in which the previous step is not applied these fluxes have to be limited such that it maintains positivity. To ensure that the right-hand side of (6.11) remains positive we chose the multiplier Q_i to be

$$Q_i = \begin{cases} Q_i^+ = c_i^{\max} - \tilde{c}_i, & \text{if } \sum_{j\neq i} \alpha_{ij} f_{ij} > 0, \\ Q_i^- = c_i^{\min} - \tilde{c}_i, & \text{if } \sum_{j\neq i} \alpha_{ij} f_{ij} < 0, \\ 1, & \text{if } \sum_{j\neq i} \alpha_{ij} f_{ij} = 0, \end{cases}$$

so that the right-hand side of (6.11) can be written as $m_i\tilde{c}_i + u_i Q_i$ with $u_i = \frac{\sum_{j\neq i}\alpha_{ij}f_{ij}}{Q_i}$. Note that the coefficient u_i is always nonnegative.

Further we define the following quantities according to Zalesak's limiter

$$P_i^{\pm} = \frac{1}{m_i}\sum_{j\neq i} \substack{\max \\ \min}(0, f_{ij}),$$

and

$$R_i^{\pm} = \begin{cases} \min(1, Q_i^{\pm}/P_i^{\pm}), & \text{if } P_i^{\pm} \neq 0, \\ 0, & \text{if } P_i^{\pm} = 0. \end{cases}$$

The values for $R_i^{\pm}$ must lie in the interval [0,1], since the corrected flux must be a fraction of the allowed flux along the boundaries of cell V_i. The values $R_i^{\pm}$ represents the fraction of mass that is allowed to flow into cell V_i.

The exchange of mass through the interface of the cells V_i and V_j is the mass fraction that is allowed by both adjacent cells, so therefore the correction factors (flux limiters) are defined by

$$\alpha_{ij} = \begin{cases} \min(R_i^+, R_j^-), & \text{if } f_{ij} \geq 0, \\ \min(R_j^+, R_i^-), & \text{if } f_{ij} < 0. \end{cases}$$

It is important to note that the computation of the correction factors is in accordance with the cancellation of anti-diffusive fluxes in step 5. Having $Q_i^{\pm} = 0$ implies $\alpha_{ij} = 0$.

7. In the final step we update the solution to c_i^{n+1} according to Equation (6.11).

We once again state that the right-hand side of Equation (6.11) can be written in the following form

$$m_i\tilde{c}_i + \sum_{j\neq i}\alpha_{ij}f_{ij} = m_i\tilde{c}_i + u_i Q_i. \tag{6.18}$$

Using that at node k adjacent to node i the local extremum is attained we rewrite the RHS of the equation above as

$$\begin{aligned} m_i\tilde{c}_i + u_i Q_i &= m_i\tilde{c}_i + u_i(\tilde{c}_k - \tilde{c}_i) \\ &= (m_i - u_i)\tilde{c}_i + u_i\tilde{c}_k, \end{aligned} \tag{6.19}$$

with $u_i \geq 0$.

The FCT scheme (6.11) will remain positive provided that $m_i \geq u_i$. By using the definitions presented in the algorithm above we can show that this condition indeed holds. For a local maximum we have

$$u_i Q_i^+ = \sum_{j\neq i} \alpha_{ij} f_{ij} \leq \sum_{j\neq i} \alpha_{ij} \max(0, f_{ij}) \leq m_i R_i^+ P_i^+ \leq m_i Q_i^+. \tag{6.20}$$

and for a local minimum

$$u_i Q_i^- = \sum_{j\neq i} \alpha_{ij} f_{ij} \geq \sum_{j\neq i} \alpha_{ij} \min(0, f_{ij}) \geq m_i R_i^- P_i^- \geq m_i Q_i^-, \tag{6.21}$$

where Q_i^- is non-positive, hence the condition is also in this case satisfied.

The main difference between the FCT method of Boris and Book and the FCT method by Kuzmin is that the corrected flux of the former is an explicit method and the latter an implicit method, i.e. the corrected flux is computed explicitly and implicitly respectively. In the FCT algorithm of Boris and Book the update c_i^{n+1} is only based on information at the previous time step

$$c_i^{n+1} = c_i^n - \frac{\Delta t}{\Delta x}(F_{i+1/2}^n - F_{i-1/2}^n),$$

with

$$F_{ij}^n = F_L(c_i^n, c_j^n) + l_{ij}(F_H(c_i^n, c_j^n) - F_L(c_i^n, c_j^n)).$$

The FCT method by Kuzmin reads

$$m_i c_i^{n+1} - \theta\Delta t \sum_j k_{ij}^L c_j^{n+1} = m_i \tilde{c}_i^n + \sum_{j\neq i} \alpha_{ij} f_{ij}, \quad \alpha_{ij} = \alpha_{ji},$$

with

$$f_{ij} = -\Delta t d_{ij}[\theta(c_j^H - c_i^H) + (1-\theta)(c_j^n - c_i^n)].$$

Note that for the FCT method by Kuzmin c^H is used instead of c^{n+1} in the anti-diffusive term f_{ij}. This approach require solving two linear systems per time step ((6.9) and (6.11)) instead of solving one non-linear system which is usual for implicit schemes. The implicit numerical scheme of Kuzmin may improve the accuracy of the numerical solution but extra computational costs are introduced. If $\theta = 0$, then the same order of accuracy must be obtained as the explicit scheme of Boris and Book.

6.2 Inexact Newton methods

In this section a new method will be presented for computing the stationary solution of the water quality model. In the previous chapter a "time-step strategy" was used, whereas now an iterative solver will be discussed.

6.2.1 Newton's method

The most familiar method for solving non-linear equations like $F(x) = 0$, is the method of Newton. This method reads as follows:

```
Algorithm 1
Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable vector function and let $x_0$ be given, then
FOR n = 1,2, .... until 'convergence' DO
    1. Solve $F'(x_n)s_n = -F(x_n)$, where $F'$ represents the Jacobian matrix,
    2. Set $x_{n+1} = x_n + s_n$.
END FOR
```

The main advantage of using the Newton method is if x_0 is sufficiently close to the solution x then the convergence of the sequence $\{x_n\}$ is quadratic, in the sense

$$||x_{n+1} - x|| \leq C||x_n - x||^p,$$

where C is some constant, $p = 2$ and $|| \cdot ||$ is some proper norm.

A disadvantage of the Newton Method is that the method is very sensitive to the starting value x_0. If the value is not sufficiently close to x then the method diverges. In practice it can be a rather difficult task to find a proper starting value.

6.2.2 Inexact Newton methods

Another method based on Newton's method is called Inexact Newton Method. This method do not solve the system $F'(x_n)s_n = -F(x_n)$ exactly, but rather give an approximation of it, since computing the exact solution can be very expensive or even infeasible.

The Inexact Newton Method is formulated as follows:

```
Algorithm 2
Let $x_0$ be given then
FOR n = 1,2, ... until 'convergence' DO
    1. Given some $\eta_n \in$ [0,1) a priori, find a $s_n$ that satisfy
           $||F(x_n) + F'(x_n)s_n|| \leq \eta_n||F(x_n)||,$        (6.22)
    2. Set $x_{n+1} = x_n + s_n$.
```

END FOR

Note that the left-hand side of (6.22) is the linearization of F around the point x_n. Some proper norm $||\cdot||$ is used and η_n, which is called a forcing term, forces the left-hand side of Equation (6.22) to be small in a particular way. Actually we have a kind of weak formulation for the Newton Method. Satisfying this condition makes sure that we approach the actual solution in the correct way.

If F and its local linear model disagree at a step and one chooses η_n to be too small, this leads to "oversolving" Equation (6.22). This means that an accurate solution is computed for the inaccurate Newton correction. A less accurate approximation by taking a less smaller η_n may be computationally cheaper and more effective. Therefore the following choices for η_n are defined

$$\begin{aligned}
1.\ \eta_n &= \frac{||F(x_n)|| - ||F(x_{n-1}) - F'(x_{n-1})s_{n-1}||}{||F(x_{n-1})||},\ n = 1, 2, \ldots \text{ and } \eta_0 \text{ given},\\
2.\ \eta_n &= \gamma\frac{||F(x_n)||^2}{||F(x_{n-1})||^2},\ \text{with } \gamma \in [0, 1) \text{ a parameter},\\
3.\ \eta_n &= \min\{\frac{1}{n+2}, ||F(x_n)||\},\\
4.\ \eta_n &= \frac{1}{2^{n+1}}.
\end{aligned}$$

The forcing terms are taken from Eisenstat & Walker (1996), Kelley (2003), Dembo & Steihaug (1983) and Brown & Saad (1990) respectively. For a deeper understanding of these forcing terms we refer to [2], chapter 6.

The speed of the convergence depends on the choice for η_n. Assuming that x_0 is sufficiently close to the actual solution x, the method converges linearly to x. If η_n goes in the limit to zero, then we have superlinear convergence. In case F' satisfies the Lipschitz continuity at x and $\eta_n = O||F(x_n)||$, then the convergence of the method is quadratic.

The definition for linear and quadratic convergence is presented in the section of Newton's method, with $p = 1$ and $p = 2$ respectively. A sequence $\{x_n\}$ converges superlinearly to x if

$$\lim_{n\to\infty} \frac{|x_{n+1} - x|}{|x_n - x|} = 0.$$

6.2.3 Globalized Inexact Newton

In case the Inexact Newton Method diverges it is useful to adapt this method in order to obtain global convergence. Divergence occurs when x_n is not close to x. A possibility is to define a sufficient decrease condition on $||F||$. This condition is formulated as

$$||F(x_n + s_n)|| \leq (1 - t(1 - \eta_n))||F(x_n)||,\ \ 0 < t < 1. \tag{6.23}$$

The Globalized Inexact Newton method reads:

Algorithm 3
Let $x_0, \eta_{\max} \in [0, 1),\ t \in (0, 1)$ and $0 < \lambda_{\min} < \lambda_{\max} < 1$ be given. Then
FOR n = 1,2, ... until 'convergence' DO

```
1. Given some $\eta_n \in [0, \eta_{\max}]$ a priori, find a $s_n$ that satisfy

       $||F(x_n) + F'(x_n)s_n|| \leq \eta_n ||F(x_n)||,$

2. WHILE $||F(x_n + s_n)|| > (1 - t(1 - \eta_n))||F(x_n)||$ DO
    (a) Choose $\lambda \in [\lambda_{\min}, \lambda_{\max}]$
    (b) Set $s_n = \lambda s_n$ and $\eta_n = 1 - \lambda(1 - \eta_n)$
   END WHILE
3. Set $x_{n+1} = x_n + s_n$
END FOR
```

In the first step of the algorithm we try to satisfy the Inexact Newton condition (6.22). Also for this case the forcing terms are defined according to one of the definitions above. For choices 1 and 2 we take $\eta_0 = \frac{1}{2}$.

In the second step we make sure that $||F||$ is decreasing sufficiently in the new direction $x_n + s_n$. If not, then we continue until a correct direction is found. The procedure for finding this new direction is first by halving the Newton step. In case that after two reductions no sufficient decrease is obtained, then a quadratic polynomial $||F(x_n + \lambda s_n)||_2^2$ is build, which is based on the three most recent values of λ. The next λ is the minimizer of the quadratic polynomial, provided that $0.5 \leq \lambda \leq 0.9$. Also η_n must be adapted by this minimizer λ. The adapted Newton step $x_n + \lambda s_n$ is correct if condition (6.23) is satisfied.

6.2.4 Globalized Projected Newton methods

When solving the stationary solution for the water quality model we have to make sure that the values for the concentrations are non-negative. This means that the numerical scheme after the spacial discretization must be positive. The Globalized Inexact Newton method does not guarantee that positivity is preserved when applied to these schemes. To overcome this problem the Globalized Inexact Projected Newton method is defined, which is an adaption of the previous mentioned method. The Globalized Projected Newton method makes sure that after each iteration step each entry of x_n remains non-negative. Therefore it is necessary to assume that a positive solution exists.

For the Globalized Projected Newton method we have the following sufficient decrease condition

$$||F(\mathbb{P}(x_n + s_n))|| > (1 - t(1 - \eta_n))||F(x_n)||, \tag{6.24}$$

where $\mathbb{P}$ is the projection on the positive orthant. Entry j is given as

$$\mathbb{P}_j(x) = \begin{cases} x_j, & \text{if } x_j \geq 0, \\ 0, & \text{if } x_j < 0. \end{cases}$$

The operator projects negative entries to zero and check if this projected value still satisfies the sufficient decrease condition (6.24) on $||F||$.

The method is formulated as:

Algorithm 4

```
Let $x_0, \eta_{\max} \in [0,1), t \in (0,1)$ and $0 < \lambda_{\min} < \lambda_{\max} < 1$ be given. Then
FOR n = 1,2, ... until 'convergence' DO
  1. Given some $\eta_n \in [0, \eta_{\max}]$ a priori, find a $s_n$ that satisfy
         $||F(x_n) + F'(x_n)s_n|| \leq \eta_n ||F(x_n)||,$
  2. WHILE $||F(\mathbb{P}(x_n + s_n))|| > (1 - t(1 - \eta_n))||F(x_n)||$ DO
       (a) Choose $\lambda \in [\lambda_{\min}, \lambda_{\max}]$
       (b) Set $s_n = \lambda s_n$ and $\eta = 1 - \lambda(1 - \eta_n)$
     If such $\lambda$ cannot be found, terminate with failure

     END WHILE
  3. Set $x_{n+1} = \mathbb{P}(x_n + s_n)$
END FOR
```

The algorithm presented above is almost equal to the Globalized Newton Method, except that negative values are not allowed. Since the corresponding decrease condition is a rather strong condition it may occur that the algorithm will break down in step 2. In that case s_n and η_n in step 2 will be adjusted by a line-search procedure. But having a positive numerical scheme and starting sufficient close to the solution x, we may expect that the algorithm converges without any problem to this solution.

Chapter 7

Conclusions

In the previous chapters we presented numerical methods for solving the water quality equations. The current solution methods are sufficiently accurate and efficient, although more improvement is desired for the numerical methods. Solving the water quality problem can be discussed for both the time-independent and stationary case.

For the time-independent case the water quality equations are discretized in space according to the flux-limiter approach. This solution method is capable in dealing with steep gradients in the concentration profile. A special flux-limiter method is the Flux Corrected Transport algorithm of Boris and Book. This procedure reduce the amount of numerical diffusion considerably so that a positive, monotonic-preserving and stable solution is obtained. Combining the FCT method with the local-theta method (an implicit time-discretization method) reduces the numerical diffusion even more so that a more accurate solution is obtained. Furthermore, the efficiency is increased, since an implicit scheme is used. A disadvantage of this FCT method is that the anti-diffusion term can be larger than necessary. By raising the local theta coefficients to a minimal value of $\frac{1}{2}$ this problem can be avoided, but as side effect we get an increase of the numerical diffusion.

For the stationary case on the other hand a raise in the efficiency is desired. The steady state is computed by solving the time-dependent equation with $T \to \infty$. Since the water quality processes are non-linear and dependent on the concentration of several substances these terms are taken explicitly in the computation. This approach is not difficult, since it avoids solving a non-linear system. Though, it is shown that for a large water quality model, i.e. a large number of substances, this strategy can be very time consuming.

For both cases we briefly presented the current solution methods and showed the problems we encounter, which have a negative effect on the accuracy and/or the efficiency. This shows that more improvements can be made in the numerical methods.

Chapter 8

Future work

In the previous chapter we conclude that the current solution methods leads to satisfactory results for the accuracy and efficiency, but more improvement is desired for both cases. In this section we will present our future work in order to reach this goal.

First we want to solve the water quality model for the time-dependent case with a numerical method to obtain more accuracy and higher efficiency. The FCT method by Kuzmin presented in Section 6.1 will be applied. We expect to obtain more improvement, since the FCT method by Kuzmin is an implicit flux-limiter. Furthermore, only linear systems has to be solved. Hence higher efficiency can be realized. Based on the results for the current local theta FCT method we expect also for combining the implicit FCT method with the local theta method more improvement in the accuracy. For the time-dependent situation the following testproblems will be studied

- The 1D homogeneous advection equation with periodic boundary conditions on a non-uniform grid. First the FCT method by Kuzmin will be implemented for this simple problem. Next we will combine the FCT method with the local theta method, so that it can also be applied to this testproblem. The reason for choosing the advection equation is that for this problem the exact solution can be obtained. Hence a comparison between the exact and numerical solution is possible.
- The Molenkamp problem on a structured grid. This is the 2D advection equation with a constant angular velocity which has been chosen such that the exact solution is periodic with a period of 4 hours. The new local theta FCT method will also be implemented for this problem. The results will be compared with the results from the Boris and Book variant of the local theta FCT method.

Next we want to solve the water quality model for the time-independent situation with a numerical method to obtain higher efficiency. The Inexact Newton methods presented in Section 6.2 will be applied. For the 'time-step' strategy the source terms are computed explicitly, but with the new method the source terms are not computed separately. So more improvement can be realized. The following testproblems will be studied for the stationary case

- The 1D stationary water quality model with linear, quadratic and independent water quality processes. The Inexact Newton method will be implemented for this simple problem.

- The previous testproblem, but now with dependent water quality processes. A small number of substances will be used. First we start with two species and continue to up to (at most) ten different species.

- The 2D stationary water quality model, with linear, quadratic and dependent water quality processes. First we start with two different substances and continue to up to (at most) ten different species.

If satisfactory results are obtained for the test problems, then the new methods can be applied to a real 3D case. An interesting and relevant case is the Eems-Dollard region located in the northern Dutch-German area. If there is time available also the Hong-Kong case will be studied for which the current local theta FCT method is already applied. A picture of the Eems-Dollard region is shown below.

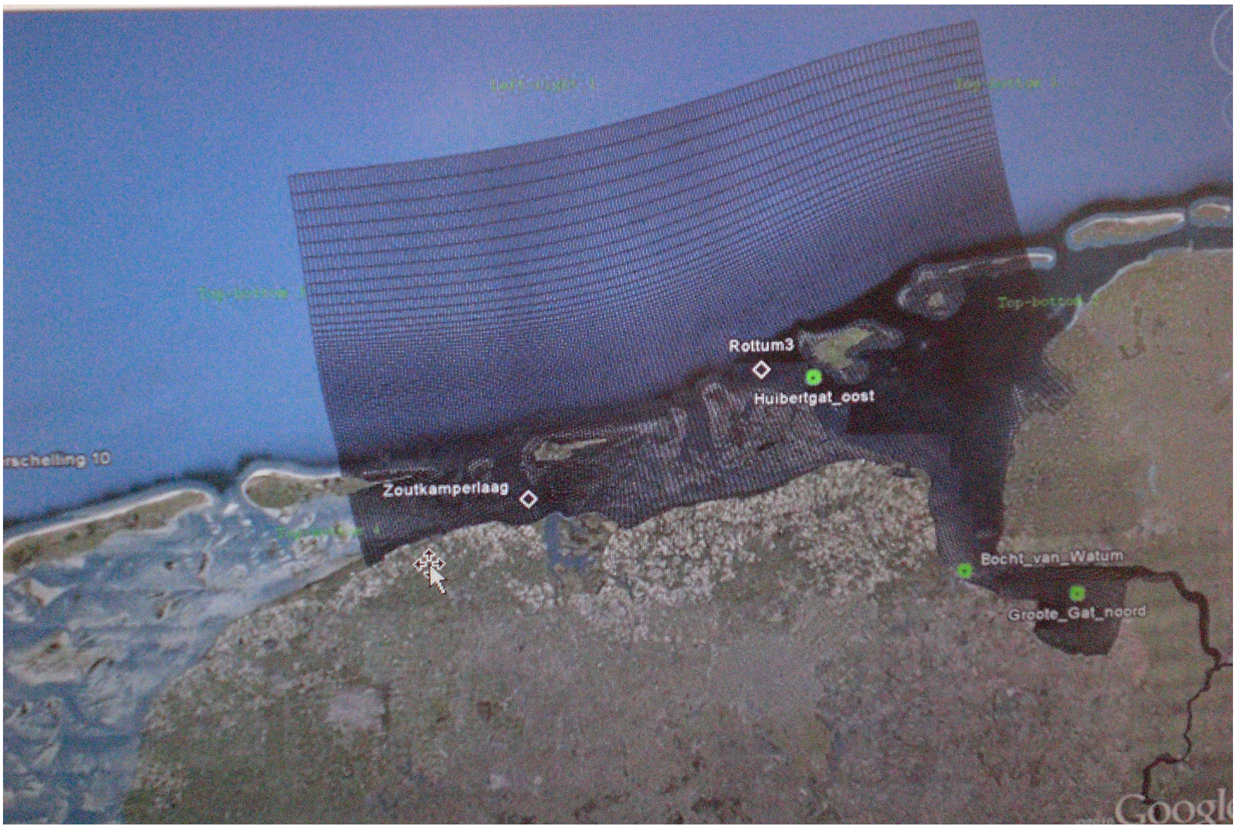

Figure 8.1: The Eems-Dollard region

Appendix A

Current numerical schemes in Delft3D-WAQ

At present, 23 different numerical schemes can be used in Delft3D-WAQ. The most used schemes are briefly presented below.

Scheme 1 The explicit first order upwind scheme.

Scheme 2 Like scheme 1, except that it uses the predictor corrector method for time integration.

Scheme 3 The explicit Lax-Wendroff scheme.

Scheme 4 An Alternation Direction Implicit (ADI) method. It can only be applied in two dimensions on a structured grid. This scheme uses the theta scheme for $\theta = \frac{1}{2}$.

Scheme 5 An explicit FCT scheme a la Boris and Book with Lax-Wendroff flux correction.

Scheme 10 Theta upwind scheme with $\theta = 1$.

Scheme 11 The horizontal and vertical direction are treated separately. In horizontal direction the explicit upwind scheme. In vertical direction the theta scheme with $\theta = \frac{1}{2}$ and central fluxes.

Scheme 12 Like scheme 11, except that it uses an explicit FCT scheme (scheme 5) in the horizontal direction.

Scheme 13 Like scheme 11, except that it uses the theta upwind scheme with $\theta = 1$ in the vertical direction.

Scheme 14 Like scheme 12, except that it uses the theta upwind scheme with $\theta = 1$ in the vertical direction.

Scheme 15 Like scheme 10, except that in horizontal direction the linear systems are solved by means of GMRES with a symmetric GS preconditioner. In the vertical direction a direct method is used.

Scheme 16 Like scheme 15, except that it uses the theta scheme with $\theta = \frac{1}{2}$ and central discretization in the in the vertical direction.

Scheme 19 The horizontal and vertical direction are treated separately. In the horizontal direction an ADI method is used. In the vertical direction central fluxes are used.

Scheme 20 Like scheme 19, except that it uses first order upwind discretization in the vertical direction.

Scheme 21 – 22 The local theta scheme.

Scheme 23 The local theta scheme combined with the FCT scheme a la Boris and Book.

Bibliography

[1] P. van Slingerland, *An accurate and robust finite volume method for the advection diffusion equation*, Delft, June 2007.

[2] S. van Veldhuizen, *Efficient numerical methods for the instationary solution of laminar reacting gas flow problems*, Delft, 2009.

[3] D. Kuzmin, M. Möller, S. Turek, *High − resolution FEM − FCT schemes for multidimensional conservation laws*, University of Dortmund, Dortmund, 2004.

[4] D. Kuzmin, S. Turek, *Flux correction tools for finite elements*, University of Dortmund, Dortmund, 2001.

[5] R.J. Leveque, *Finite Volume Methods for hyperbolic problems*, Cambridge University Press, New York, 2002.

[6] P. Wesseling, *Elements of computational fluid dynamics*, Delft, 2001

[7] J.P. Boris and D.L. Book, *Flux Corrected Transport, I. SHASTA, A fluid transport algorithm that works*, Journal of Computational Physics volume 11, page 38-69, 1973

[8] S.T. Zalesak, *Fully multidimensional flux corrected transport algorithm for fluids*, Journal of Computational Physics volume 31, page 335-362, 1979

[9] Delft3D-WAQ user manual, *Versatile water quality modelling in 1D, 2D or 3D systems including physical, (bio)chemical and biological processes*, Delft, 2009

[10] K. Alhumaizi, *Flux limiting solution techniques for simulation of reaction − diffusion− convection system*, Communications in Nonlinear Science and Numerical Simulation, volume 12, page 953-965, 2007

[11] R. Eymard, T. Gallouet and R. Herbin, *Finite Volume Methods*, page 4-11, 2006