NUMERICAL ANALYSIS OF PHYTOPLANKTON DYNAMICS

Phạm Thị Nguyệt Nga

Printed and bound by Ponsen & Looijen bv.

Cover designed by Phạm Thị Nguyệt Nga.

The phytoplankton picture on the cover is redrawn from the Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam.

CIP-DATA LIBRARY UNIVERSITEIT VAN AMSTERDAM

Phạm Thị Nguyệt Nga

Numerical Analysis of Phytoplankton Dynamics / by Phạm Thị Nguyệt Nga.– Amsterdam: Universiteit van Amsterdam, 2006.– Proefschrift.

ISBN 90-6196-538-1

Subject headings: Integro-advection-diffusion-reaction equations / Time integration / Positivity / Stability / Approximate matrix factorization / Phytoplankton dynamics / Deep chlorophyll maximum / Oscillations.

Numerical Analysis of Phytoplankton Dynamics

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. mr. P.F. van der Heijden
ten overstaan van een door
het college voor promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op dinsdag 21 november 2006, te 12:00 uur

door

Phạm Thị Nguyệt Nga

geboren te Quang Ninh, Vietnam

Promotiecommissie

Promotoren: prof.dr. J.G. Verwer
prof.dr. J. Huisman

Co-promotor: dr. B.P. Sommeijer

Overige leden: prof.dr. P.W. Hemker
prof.dr. A.M. de Roos
dr. J. Brandts
prof.dr. F.J. Weissing
prof.dr. A.E.P. Veldman
prof.dr. R.M.M. Mattheij

Faculteit der Natuurwetenschappen, Wiskunde en Infomatica

NWO
Nederlandse Organisatie voor Wetenschappelijk Onderzoek

THOMAS STIELTJES INSTITUTE
FOR MATHEMATICS

CWI

IBED

Dit onderzoek werd mogelijk gemaakt door het onderzoeksprogamma Computational Science van het Gebiedsbestuur Exacte Wetenschappen van NWO, onder toekenning 635.000.006, en uitgevoerd bij het Centrum voor Wiskunde en Informatica (CWI).

To the memory of my father
To my mother, Nguyet, and Cuong
To the future of Thieu and me

Contents

Preface

Mathematics is an indispensable tool for many different disciplines in science, such as, physics, chemistry, biology, weather forecasting, and finance. It is needed to solve challenging problems arising from those areas. It points out how and why a specific particle would move in space and time; how the reaction between particles would be at a particular time; which species would be the winner in a biological competition; how the weather would be in the next few days; how tomorrow stocks would look like; et cetera. To obtain predictions, first a mathematical model is formulated, based on natural properties of the system. Then, this model is solved by accurate and fast numerical methods. This can be done efficiently by means of computer simulation. When modeling and solving the problems as well as explaining phenomena, analytical manipulations are always involved.

This is exactly what we are doing in the Modeling, Analysis, and Simulation (MAS) department at CWI. This thesis joins in as part of the department, particularly, in the subtheme 'PDEs in the Life Sciences'. The application of this thesis is a problem arising from biology, which is, the dynamics of phytoplankton modelled by an integro-partial differential equation. Solving partial differential equations (PDEs) in general, and the phytoplankton problem in particular, by analytical tools is very difficult (even impossible). Thus, we solve it numerically. The main aim of the thesis is to construct, analyze and test efficient algorithms for the numerical solution of the phytoplankton problem.

The thesis consists of six chapters preceded by an introduction and followed by a summary. The chapters are based on published and submitted papers. Details are listed below:

1. Chapter 1 is based on the paper by N.N. Pham Thi, J. Huisman, and B.P. Sommeijer, entitled *Simulation of three-dimensional phytoplankton dynamics: competition in light-limited environments*, published in Journal of Computational and Applied Mathematics 174, pages 57-77, 2005.

2. Chapter 2 is based on the paper by N.N. Pham Thi, entitled *On positive solutions in a phytoplankton-nutrient model*, published in Journal of Computational and Applied Mathematics 177, pages 467-473, 2005.

3. Chapter 3 is based on the paper, *Positivity for explicit two-step methods in linear multistep and one-leg form*, by N.N. Pham Thi, W. Hundsdorfer, and B.P. Sommeijer, accepted for publication in BIT.

4. Chapter 4 is joint work with B.P. Sommeijer on Implicit-Explicit methods.

5. Chapter 5 is based on the paper by J. Huisman, N.N. Pham Thi, D.M. Karl, and B.P. Sommeijer, entitled *Reduced mixing generates oscillations and chaos in the oceanic deep chlorophyll maximum*, published in Nature 439, pages 322-325, 2006.
6. Chapter 6 is entitled *Analysis of phytoplankton blooming*, a joint work with A. Zagaris, A. Doelman, and B.P. Sommeijer on analysis of the results obtained in Chapter 5, to be submitted.

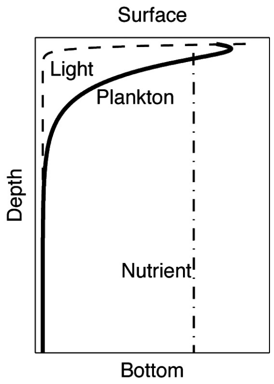
The introductory chapter is meant to help unspecialized readers to understand the motivation as well as the flow of the whole thesis. The summary will summarize the conclusions that we have pointed out in the thesis.

Phạm Thị Nguyệt Nga
Amsterdam, November 2006

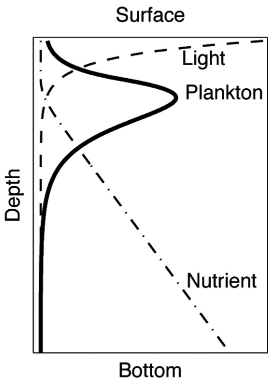
Introduction

Lakes, seas, and oceans are inhabited by large numbers of free-floating microorganisms called phytoplankton. Like grass, trees and other plants, phytoplankton utilize solar energy and carbon dioxide to produce biomass, in a process known as photosynthesis. Phytoplankton photosynthesis forms the basis for nearly all aquatic food webs, and thereby has a major impact on the productivity of aquatic ecosystems. Furthermore, because phytoplankton absorb carbon dioxide during photosynthesis, on a global scale phytoplankton remove nearly as much of the greenhouse gas CO_2 from the atmosphere as all land plants do. As a result, phytoplankton plays a significant role in scenarios on climate change. For these reasons, studies of the growth and population dynamics of phytoplankton is of great interest. These processes are mathematically modelled by integro-partial differential equations (integro-PDEs) of advection-diffusion-reaction type. The major aim of this thesis is to construct, analyze and test efficient algorithms for the numerical solution of these integro-PDEs.

Like all land plants, phytoplankton need both light and nutrients (e.g., nitrogen, phosphorus, iron) to grow. In some regions and seasons, nutrients are abundant and light is the only factor limiting plankton growth (as shown in the left figure above). Owing to shading and absorption, light intensity decreases with depth. Phytoplankton therefore flourish close to the surface of the water column. Below this layer, populations decrease with decreasing light. Phytoplankton species



Concentration profiles when nutrient is abundantly available and light is the only factor limiting phytoplankton growth.



Concentration profiles when both light and nutrient are limiting factors for phytoplankton growth.

that manage to stay in the upper water layer have plenty of light available for photosynthesis and have the additional advantage to shade other species at deeper levels. Although each type of species has a specific light requirement, all species require a similar environment and hence face the competition from the others when light is scarce.

In other regions and seasons, apart from light, nutrient becomes a limiting factor for phytoplankton growth (this situation is shown in the right figure above). Light is supplied from above, and its intensity decreases with depth. Nutrient is supplied from the bottom of the water column, and its concentration increases with depth. Due to these opposing gradients, phytoplankton settle at a depth at which both light and nutrient are sufficient for phytoplankton to grow.

Phytoplankton dynamics in light-limited environments as well as in light-nutrient -limited environments will be studied in this thesis. The thesis consists of six chapters, preceded by this introduction and ends with a summary.

Chapter 1. In the first chapter, we study a model of competition for *light* between phytoplankton species in a *three-dimensional* domain. Here, we have taken into account the influence of (horizontal) water flow, caused by hydrodynamics. The decrease of light intensity with depth appears in the model as an integral over the dynamic phytoplankton concentrations. The resulting phytoplankton model is therefore framed in terms of non-linear integro-PDEs. The spatially discretized system of equations contains a strong coupling of the components. The coupling originates from the three spatial dimensions, the interaction of the various species and the integral term. Due to the stiffness of the discretized system we select an implicit integration method. However, the resulting implicit relations are extremely expensive to solve, caused by the strong coupling of the components. To reduce the amount of work in the linear algebra part, we use an Approximate Matrix Factorization technique. The performance of the complete algorithm is demonstrated on the basis of two test examples. It turns out that unconditional stability (i.e., A-stability) is a very useful property for this application.

Chapter 2. The subject of Chapter 2 is the *coupled phytoplankton-nutrient* model. As it turned out from Chapter 1, the vertical direction is the most important one for the distribution of phytoplankton. Therefore, we study in this chapter the *one-dimensional light-nutrient limited* model. The model is integrated in time by the widely used package VODE [8] (which is also used in Chapter 1). Unfortunately, used in a default setting, this code produces a negative steady state solution. We come up with a remedy to avoid this situation; however, this approach is still far from optimal and needs further study.

Chapter 3. Therefore, in this chapter, we study numerical methods for ordinary differential equations of Linear Multistep (LM) type that do yield a positive solution. Because of accuracy considerations, combined with modest memory demands, we restrict ourselves to second-order two-step methods. We show an improvement of the results obtained in [37] for explicit methods. It turns out that the extrapolated BDF2 method has an optimal positivity property. Moreover, we show that these two-step methods, when formulated in one-leg form, allow a slightly larger time step. It turns out that positivity of the solution can only be guaranteed by imposing a rather severe restriction on the time step.

Chapter 4. In a PDE context, positivity of an integration method also requires a positive spatial discretization (see e.g. the assumption (3.3) in Chapter 3). For the diffusion terms, second-order central discretization is positive. For the advection terms, however, only the first-order upwind method guarantees a positive solution. Unfortunately, this scheme is too inaccurate and produces a large amount of artificial diffusion. A possible remedy is to use the third-order upwind-biased scheme (which has been used in Chapter 1 and Chapter 2) in combination with a limiter. However, this will add additional nonlinearity into the system to be solved. This will be a serious drawback when we integrate the discrete system by a fully implicit method, like VODE. Therefore, implicit-explicit (IMEX) versions of these LM methods are discussed in Chapter 4. In this approach the nonstiff terms in the model (i.e., advection and growth) are treated explicitly whereas the stiff diffusion terms are treated implicitly. Higher efficiency compared to an implicit method is then to be expected. This is because the difficult and nonlinear terms are calculated by straightforward substitutions, whereas the diffusion term, which is simple (and linear for the phytoplankton problem), is treated implicitly for stability reasons.

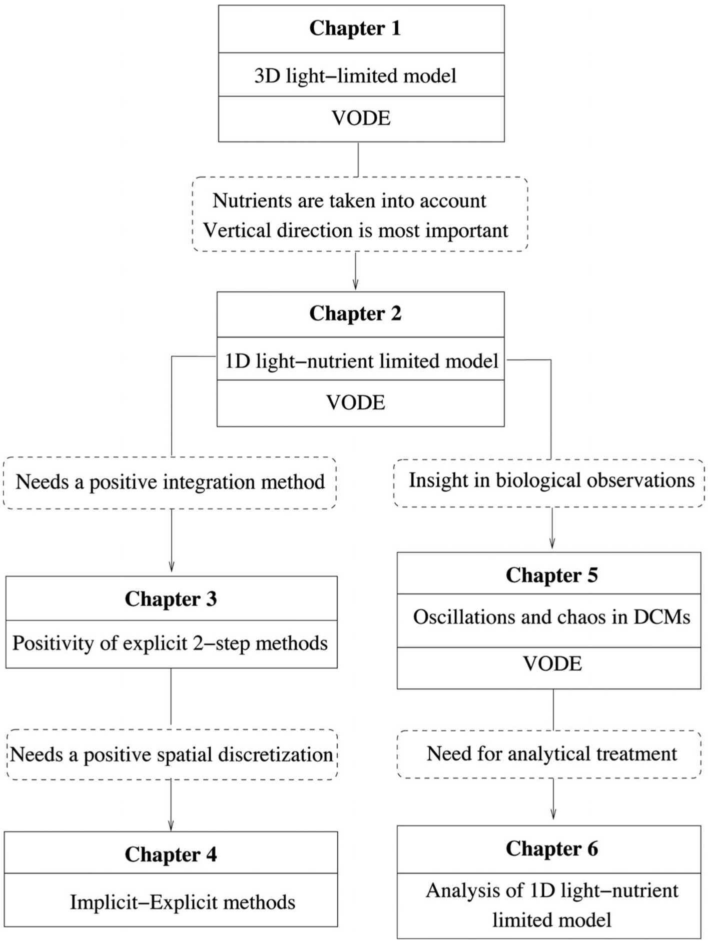
In general, the stability region of an IMEX method is smaller than that of the explicit counterpart [21]. However, the implicitly treated diffusion term gives rise to *real* eigenvalues. Using this property we derive a condition on the parameters in the family of two-step IMEX methods such that the resulting stability is the same as for the fully explicit counterpart.

The advantages of the IMEX approach will be demonstrated by a comparison between IMEX-BDF2 and VODE when applied to the phytoplankton model. The IMEX-BDF2 method is chosen because its stability is as good as the stability of the explicit version of the method, that is, the extrapolated BDF2 method. Moreover, as it has been shown in Chapter 3, this method possesses optimal positivity. The favourable positivity, stability, and damping properties of the extrapolated BDF2 method are demonstrated on the basis of two different advection test examples.

Chapter 5. As already mentioned in Chapter 2, in the light-nutrient limited environment, phytoplankton mainly concentrate at layers with sufficient nutrients welling up from the bottom and sufficient light filtering down from the top. These layers are called deep chlorophyll maxima (DCMs). DCMs are widespread in large parts of the world's oceans. It is often argued that DCMs are *stable* features. Here we show, however, that reduced vertical mixing can generate *oscillations* and even *chaos* in phytoplankton biomass and species composition of DCMs. These fluctuations are caused by a difference in the timescales of two processes: (1) rapid export of sinking plankton, withdrawing nutrients from the euphotic zone and (2) a slow upward flux of nutrients fuelling new phytoplankton production. Climate models predict that global warming will reduce vertical mixing in the oceans. This will suppress the upward flux of nutrients, leading to a decline in oceanic primary production. Our model indicates that reduced mixing will generate more variability in DCMs, with implications for the variability in oceanic primary production and in carbon export into the ocean interior.

Chapter 6. Based on the results found in Chapter 5, it is of great importance to understand for which parameter regimes we may expect no bloom, stationary bloom, or oscillations. An analytical study has been performed in Chapter 6 for a slightly simplified version of the light-nutrient model studied in the preceding chapter. It turns out that it is possible to derive an analytical expression to distinguish between bloom- and nobloom regions in the parameter space. Additional numerical simulations show the validity of this analytical approach.

The coherence of the material presented in the various chapters has been graphically displayed in the following figure.



Chapter 1

Simulation of three-dimensional phytoplankton dynamics: competition in light-limited environments

In this chapter, we develop computational methods for a three-dimensional model of competition for light between phytoplankton species. The competing phytoplankton populations are exposed to both horizontal and vertical mixing. The vertical light-dependence of phytoplankton photosynthesis implies that the three-dimensional model is formulated in terms of integro-partial differential equations that require an efficient numerical solution technique.

Due to the stiffness of the discretized system we select an implicit integration method. However, the resulting implicit relations are extremely expensive to solve, caused by the strong coupling of the components. This coupling originates from the three spatial dimensions, the interaction of the various species and the integral term. To reduce the amount of work in the linear algebra part, we use an Approximate Matrix Factorization technique.

The performance of the complete algorithm is demonstrated on the basis of two test examples. It turns out that unconditional stability (i.e., A-stability) is a very useful property for this application.

1.1 Introduction

Lakes, seas, and oceans are inhabited by large numbers of free-floating microorganisms called phytoplankton. Like grass, trees and other plants, phytoplankton utilize solar energy and carbon dioxide to produce biomass, in a process known as photosynthesis [42]. Phytoplankton photosynthesis forms the basis for nearly all aquatic foodwebs, and thereby has a major impact on the productivity of aquatic ecosystems including fish production. Furthermore, because phytoplankton absorb carbon dioxide during photosynthesis, on a global scale phytoplankton remove nearly as much of the greenhouse gas CO_2 from the atmosphere as all land plants do. As a result, phytoplankton photosynthesis has a major influence on climate change [17, 18]. For these reasons, studies of the growth and population dynamics of phytoplankton is of great interest.

Phytoplankton is the generic name for many different species from a wide variety of taxonomic groups, including cyanobacteria, prochlorophytes, diatoms, coccolithophores, and dinoflagellates [26]. The species composition of the phytoplankton plays an important role. Some species are sinking species, i.e., they have a higher specific weight than water. As a result, they transport their biologically fixed carbon into the deep ocean [3, 29]. Other species are buoyant species since their specific weight is smaller than that of water. These species remain near the surface. All these species essentially require the same environmental resources (light, mineral nutrients, and carbon dioxide) and each phytoplankton species faces competition from the other phytoplankton species when one or more of these essential resources are available in low quantities only. This chapter concentrates on competition for light between phytoplankton species. Competition for light is a major determinant of the species composition of phytoplankton communities [30, 31, 33], as light is the energy source that drives phytoplankton photosynthesis.

In the context of competition for light, physical mixing processes that affect the spatial distributions of the phytoplankton species play a prominent role. In particular, phytoplankton species that manage to stay in the upper water layer have plenty of light available for photosynthesis and have the additional advantage to shade other species at deeper levels. During recent years, several three-dimensional models that combine physical mixing processes and phytoplankton growth have been developed [22, 45, 52, 53]. These biological-physical models have advanced the general understanding of the productivity of marine ecosystems, and play an increasingly important role in oceanographic research. However, in many of the numerical applications of these models the special structure that stems from the vertical light-dependence of phytoplankton photosynthesis has not been fully recognized. Owing to shading, the decrease of light intensity with depth appears in the model as an integral over the dynamic phytoplankton concentrations. The resulting phytoplankton model is therefore framed in terms of integro-partial differential equations. Competition for light results in coupling of the population dynamics of the phytoplankton species through shading, that is, the integro-PDEs are coupled through this integral term. For this complicated model structure, efficient numerical solution techniques that avoid numerical artifacts are indispensable.

We recently outlined an efficient simulation technique for the one-dimensional vertical model formulation of phytoplankton competition in light-limited environments [33]. In this chapter, we extend our approach by incorporation of horizontal water flow. The aim of the chapter is to come up with an efficient numerical technique for the simulation of three-dimensional phytoplankton models that include competition for light between phytoplankton species.

The chapter is organized as follows. In Section 1.2 we formulate our model system which is based on [31, 33]. Section 1.3 deals with the numerical technique to solve this system. Two applications are described in Section 1.4. Section 1.5 is devoted to final remarks and some discussion.

1.2 The model system

We consider a model of competition for light between n species, where we assume that the species interact with one another indirectly, via shading.

Let $\omega_s(x, y, z, t)$, $s = 1, \ldots, n$, denote phytoplankton population densities (cells per unit volume) of n species (the subscripts s indicate the different species) at position (x, y, z) in a three-dimensional domain Ω at time t $(t \geq 0)$. Here z is expressed as the depth of the water column from the surface $(z = 0)$ to the bottom $(z = Z)$, x varies between 0 and X, y runs from 0 to Y. Hence, the domain Ω that we consider has the form of a rectangular basin with vertical boundaries and a flat bottom. This is, of course, far from the actual shape of the lakes that we encounter in nature. However, in this chapter we focus on the construction and analysis of efficient numerical solution techniques for the equations describing the dynamics of phytoplankton (see Eq. (1.6)). Implementing the resulting algorithms paying full attention to all physical details should be a next step and is far beyond the scope of the present chapter.

Continuing with the derivation of our model, we first observe that phytoplankton use energy in sunlight for photosynthesis. In the water column, light intensity L decreases with depth according to

$$L(x, y, z, t) \;=\; L_{in}\, e^{-K_{bg} z} e^{-\int_0^z \left(\sum_{s=1}^{n} r_s \omega_s(x,y,\sigma,t)\right) \mathrm{d}\sigma}. \tag{1.1}$$

At a particular depth, light intensity depends on the incident light intensity L_{in}, the background turbidity K_{bg} due to all non-phytoplankton components in the water and on the total light attenuation of all phytoplankton species above that depth. Here r_s denotes the specific light attenuation coefficient of the s-th species.

The formulation (1.1) explicitly involves light absorption by all phytoplankton species. Thus, the light gradient changes with a change in any species density distribution.

The change in concentration (density distribution) for each species is determined by growth and the local transport process through the partial differential equation (PDE)

$$\frac{\partial \omega_s}{\partial t} \;=\; g_s\,(L)\;\omega_s \;-\left(\frac{\partial \mathcal{I}_s}{\partial x} \;+\; \frac{\partial \mathcal{J}_s}{\partial y} \;+\; \frac{\partial \mathcal{K}_s}{\partial z}\right). \tag{1.2}$$

Here $g_s\,(L(x, y, z, t))$ is the specific growth rate of the s-th species driven by light availability. $\mathcal{I}_s(x, y, z, t)$, $\mathcal{J}_s(x, y, z, t)$ and $\mathcal{K}_s(x, y, z, t)$ are defined below, and are respectively the horizontal and vertical fluxes of the s-th species at position (x, y, z) and time t.

The specific growth rate $g_s\,(L(x, y, z, t))$ in the above equation depends on the balance between the production rate $p_s\,(L(x, y, z, t))$ and the specific loss rate ℓ_s

$$g_s(L) \;=\; p_s(L) \;-\; \ell_s. \tag{1.3}$$

The production rate $p_s(L)$ is modelled by the so-called Monod-function (see e.g. [30])

$$p_s(L) \;=\; \frac{p_{s_{\max}} L}{H_s + L}, \tag{1.4}$$

where $p_{s_{\max}}$ denotes the maximum specific production rate and H_s is the half-saturation constant, both for the s-th species.

The fluxes in Eq. (1.2) depend on the dynamics of the system as determined by the horizontal flow, the vertical velocity, and the transport of phytoplankton by turbulent diffusion

$$\begin{aligned} \mathcal{I}_s(x,y,z,t) \;&=\; a(x,y,z)\,\omega_s(x,y,z,t) \;-\; D_H(x,y,z)\,\frac{\partial \omega_s}{\partial x}(x,y,z,t), \\ \mathcal{J}_s(x,y,z,t) \;&=\; b(x,y,z)\,\omega_s(x,y,z,t) \;-\; D_H(x,y,z)\,\frac{\partial \omega_s}{\partial y}(x,y,z,t), \\ \mathcal{K}_s(x,y,z,t) \;&=\; c_s\,\omega_s(x,y,z,t) \;-\; D_V(x,y,z)\,\frac{\partial \omega_s}{\partial z}(x,y,z,t), \end{aligned} \tag{1.5}$$

where $a(x,y,z)$ and $b(x,y,z)$ are the horizontal velocity components of the water flow, c_s is the vertical velocity of the s-th species (which is positive for sinking phytoplankton and negative for buoyant phytoplankton), and $D_H(x,y,z)$ and $D_V(x,y,z)$ are the horizontal and the vertical turbulent diffusion coefficients. The minus sign in the second terms on the right hand side indicates that turbulent diffusion is in the direction opposite to the concentration gradient. In the above formulation the velocities a and b, as well as the diffusion coefficients D_H and D_V may be space-dependent, whereas the characteristic velocity c_s is taken constant. However, an extension to more general functions (e.g., time-dependent) is straightforward.

Our key system, *the system of integro-partial differential equations*, follows now straightforwardly from substituting (1.1), (1.3), (1.4) and (1.5) into (1.2)

$$\begin{aligned} \frac{\partial \omega_s}{\partial t} = p_s\left(L_{in} e^{-K_{bg} z} e^{-\int_0^z \left(\sum_{j=1}^n r_j \omega_j(x,y,\sigma,t)\right) \mathrm{d}\sigma}\right) \omega_s - \ell_s\,\omega_s - \\ \left[(a\omega_s)_x + (b\omega_s)_y + c_s(\omega_s)_z - (D_H \omega_{s_x})_x - (D_H \omega_{s_y})_y - (D_V \omega_{s_z})_z\right], \end{aligned} \tag{1.6}$$

where $s = 1,\ldots,n$. Here, the subscripts x, y, and z denote the spatial differentiation in the various directions. From this formula, one can see that a change in any of the phytoplankton densities $\omega_j(x,y,\sigma,t)$ $(\sigma < z)$, within the integral term, causes a change in the light intensity, which in turn, influences the population density $\omega_s(x,y,z,t)$ of all species. In other words, the species compete with one another for light.

The boundaries of our rectangular lake are assumed to be 'closed'. By that we mean that phytoplankton cannot enter or leave the domain. In other words, the fluxes $\mathcal{I}_s(x,y,z,t)$, $\mathcal{J}_s(x,y,z,t)$ and $\mathcal{K}_s(x,y,z,t)$ all vanish at the boundaries of the domain Ω, defining the boundary conditions for our integro-PDE system (1.6).

1.3 Numerical approach

In order to find the numerical solution of the system (1.6), we use a technique which is based on the popular Method of Lines (MOL) approach, where space and time discretizations are considered separately [38]. That is, first we derive a large system of ordinary differential equations (ODEs), which is still continuous in time, from the discrete approximations of the spatial differential operators as well as the integral term (Section 1.3.1). Then, that ODE system will be integrated in time numerically (Section 1.3.2).

Using this approach is motivated by the fact that it is easy to combine various discretizations for advection and diffusion with the treatment of the reaction term. Another attractive, practical point is that there exist nowadays many well developed ODE methods and for these methods sophisticated software is freely available.

1.3.1 Spatial discretization

There are many ways to discretize the differential operators on the domain Ω. The purpose is to approximate the solution at a desired accuracy level, with as few grid points as possible. Most simple to use is the equidistant grid

$$\begin{aligned} x_0 &= 0, \quad x_i = (i - \frac{1}{2})\Delta x, \quad i = 1, \ldots, N_1, \quad x_{N_1+1} = X, \\ y_0 &= 0, \quad y_j = (j - \frac{1}{2})\Delta y, \quad j = 1, \ldots, N_2, \quad y_{N_2+1} = Y, \\ z_0 &= 0, \quad z_k = (k - \frac{1}{2})\Delta z, \quad k = 1, \ldots, N_3, \quad z_{N_3+1} = Z, \end{aligned} \tag{1.7}$$

where $\Delta x = X/N_1, \Delta y = Y/N_2, \Delta z = Z/N_3$. Each grid point is imaginarily surrounded by a cell, at the boundaries of which we approximate the derivative of the fluxes (the terms inside the bracket in (1.2)). In the internal intervals, the cell faces lie halfway between the grid points. For the end intervals, the grid points are positioned on the boundary of Ω and coincide with the cell faces. This way of discretizing is based on the so-called finite-volume method [38, 65]. In this way, we obtain *conservation* of the flux quantity since all contributions of the fluxes along the interior cell faces cancel [65].

To be more precise, $\partial \mathcal{I}_s/\partial x$ in the internal points (x_i, y_j, z_k) is approximated by $(\mathcal{I}_{s_{ijk}} - \mathcal{I}_{s_{(i-1)jk}})/\Delta x$, where $\mathcal{I}_{s_{ijk}}$ denotes the flux $\mathcal{I}_s$ at $(x_{i+1/2}, y_j, z_k)$ with $x_{i+1/2} := x_i + \frac{1}{2}\Delta x$, i.e.,

$$\mathcal{I}_{s_{ijk}} = a(x_{i+\frac{1}{2}}, y_j, z_k)\, \omega_s(x_{i+\frac{1}{2}}, y_j, z_k, t) - D_H \frac{\partial \omega_s}{\partial x}(x_{i+\frac{1}{2}}, y_j, z_k, t). \tag{1.8}$$

Here, for simplification, we consider the model with uniform turbulent diffusion coefficients. The approximation to $\mathcal{I}_{s_{ijk}}$ is obtained by using the approach that is nowadays standard in the field of Computational Fluid Dynamics for the numerical solution of advection-diffusion equations [38, 65]. That is, the diffusion term is discretized symmetrically

$$D_H \frac{\partial \omega_s}{\partial x}(x_{i+\frac{1}{2}}, y_j, z_k, t) \approx D_H \frac{w_{s_{(i+1)jk}}(t) - w_{s_{ijk}}(t)}{\Delta x}, \tag{1.9}$$

where $w_{s_{ijk}}(t)$ denotes an approximation to the population density of the s-th species at (x_i, y_j, z_k) and time t. For the advection term, the third-order upwind-biased discretization is used [38]

$$a(x_{i+\frac{1}{2}}, y_j, z_k)\; \omega_s(x_{i+\frac{1}{2}}, y_j, z_k, t) \approx a(x_{i+\frac{1}{2}}, y_j, z_k)\; w_{s_{(i+\frac{1}{2})jk}}(t), \tag{1.10}$$

where[1]

$$w_{s_{(i+\frac{1}{2})jk}} = \begin{cases} \frac{1}{6}\left(-w_{s_{(i-1)jk}} + 5w_{s_{ijk}} + 2w_{s_{(i+1)jk}}\right) & \text{if } a(x_{i+\frac{1}{2}}, y_j, z_k) > 0, \\ \frac{1}{6}\left(2w_{s_{ijk}} + 5w_{s_{(i+1)jk}} - w_{s_{(i+2)jk}}\right) & \text{if } a(x_{i+\frac{1}{2}}, y_j, z_k) < 0. \end{cases} \tag{1.11}$$

This upwind discretization is preferred to the more simple second-order symmetric discretization as the symmetric discretization of the advection term more easily leads to ‘wiggles’ in the numerical solution, which may result in negative solution components. A negative population density is of course not realistic. To reduce this unwanted property one can use the current upwind scheme. Higher order upwind schemes may give still better results. However, for these we need a larger stencil of grid points which makes such methods impractical in simulation with boundary conditions (see e.g. [38] for more details).

We note that, according to the boundary condition, the fluxes $\mathcal{I}_{s_{0jk}}$, $\mathcal{I}_{s_{N_1jk}}$ vanish. Since we lack sufficient upstream information, a symmetric discretization for $\mathcal{I}_{s_{1jk}}$ (or $\mathcal{I}_{s_{(N_1-1)jk}}$) has been used in case of $a(x_{1+\frac{1}{2}}, y_j, z_k) > 0$ (or $a(x_{N_1-\frac{1}{2}}, y_j, z_k < 0)$

$$\begin{aligned} w_{s_{(1+\frac{1}{2})jk}} &= \frac{w_{s_{2jk}} + w_{s_{1jk}}}{2} && \text{if } a(x_{1+\frac{1}{2}}, y_j, z_k) > 0, \\ w_{s_{(N_1-\frac{1}{2})jk}} &= \frac{w_{s_{N_1jk}} + w_{s_{(N_1-1)jk}}}{2} && \text{if } a(x_{N_1-\frac{1}{2}}, y_j, z_k) < 0. \end{aligned} \tag{1.12}$$

A complete approximation to $\mathcal{I}_{s_{ijk}}$ is then obtained by the combination of (1.9)–(1.12). In the same way we obtain approximations for the fluxes $\mathcal{J}_{s_{ijk}}$ and $\mathcal{K}_{s_{ijk}}$.

Using the repeated trapezoidal rule for the integral term within the light function (L in (1.1)) the light intensity at (x_i, y_j, z_k) is approximated by

$$L_{ijk} = L_{in} e^{-K_{bg} z_k} e^{-\sum_{s=1}^{n} r_s \left[\frac{1}{4} w_{s_{ij0}} + \frac{3}{4} w_{s_{ij1}} + w_{s_{ij2}} + \cdots + w_{s_{ij(k-1)}} + \frac{1}{2} w_{s_{ijk}}\right] \Delta z} \tag{1.13}$$

with the solution at the surface, $w_{s_{ij0}}$, extrapolated as $w_{s_{ij0}} = (3w_{s_{ij1}} - w_{s_{ij2}})/2$. The corresponding specific growth rate is then $g_{s_{ijk}} := g_s(L_{ijk}) = p_s(L_{ijk}) - \ell_s$.

---

[1] Here we omit the explicit time-dependence in the notation.

Finally, we arrive at the following set of ODEs

$$\frac{dw_{s_{ijk}}(t)}{dt} = -\frac{\mathcal{I}_{s_{ijk}} - \mathcal{I}_{s_{(i-1)jk}}}{\Delta x} - \frac{\mathcal{J}_{s_{ijk}} - \mathcal{J}_{s_{i(j-1)k}}}{\Delta y} - \frac{\mathcal{K}_{s_{ijk}} - \mathcal{K}_{s_{ij(k-1)}}}{\Delta z} + g_{s_{ijk}}\, w_{s_{ijk}}, \tag{1.14}$$

where s, i, j and k respectively run from 1 to n, N_1, N_2 and N_3.

1.3.2 Time integration

This section deals with the numerical integration of the above derived ODE system, which is still continuous in time and can be written in the form

$$\frac{d\,\boldsymbol{w}(t)}{d\,t} = \boldsymbol{F}\left(\boldsymbol{w}(t)\right), \quad t \geq 0, \tag{1.15}$$

where the vector $\boldsymbol{w}(t) \in \mathbb{R}^N$, $N = n\, N_1\, N_2\, N_3$, contains the components $w_{s_{ijk}}$. This system is *stiff* (has widely spread eigenvalues) due to the fact that the spectral radius of the diffusion part is proportional to the inverse of the square of the grid sizes [38]. To cope with the stiffness of this ODE system, we use an *implicit* BDF integration method, since these methods are known to possess good stability properties. Due to the implicitness we need to solve, in each time step, a system of implicit relations to find the solution at the next point in time, $\boldsymbol{W}_{k+1}$, using previously computed values $\boldsymbol{W}_k, \boldsymbol{W}_{k-1}, \ldots, \boldsymbol{W}_{k+1-m}$. We denote this system by

$$\mathbf{R}(\boldsymbol{W}_{k+1}) := \boldsymbol{W}_{k+1} - \gamma \Delta t \boldsymbol{F}(\boldsymbol{W}_{k+1}) - \sum_{i=1}^{m} \alpha_i \boldsymbol{W}_{k+1-i} = \mathbf{0}, \tag{1.16}$$

where $\boldsymbol{W}_{k+1}$ is an approximation to $\boldsymbol{w}(t)$ at $t = t_{k+1}$, with Δt the current step size and the coefficients α_i and γ are defined by the method in use.

System (1.16) is iteratively solved by the modified Newton method, that is

$$\left[I - \gamma \Delta t \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{w}}\right] \left[\boldsymbol{W}^{j}_{k+1} - \boldsymbol{W}^{j-1}_{k+1}\right] = -\mathbf{R}(\boldsymbol{W}^{j-1}_{k+1}), \quad j = 1, 2, \ldots, \tag{1.17}$$

where I denotes the identity matrix and the Jacobian matrix $\partial \boldsymbol{F}/\partial \boldsymbol{w}$ is occasionally evaluated at certain values $\boldsymbol{W}_k$. The superscript j denotes the iteration index.

The Jacobian matrix has a huge number of entries (Fig. 1.1) since it has the structure of a matrix which is a tensor product of a $N_1 \times N_1$ 5-diagonal band matrix and a $N_2 \times N_2$ 5-diagonal band matrix (both originating from the variable advection and the diffusion parts), a $N_3 \times N_3$ 4-diagonal matrix (due to the constant vertical velocity of each species) plus a lower triangular matrix (due to the integral term), and a full $n \times n$ matrix (due to multi-species competition). This pattern of the Jacobian matrix makes it unfeasible, if not impossible, to solve the linear systems in (1.17) by a direct solver. Such an

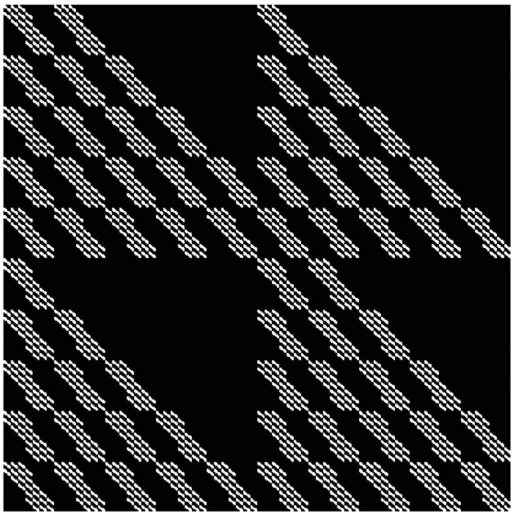


FIGURE 1.1: Structure of the Jacobian matrix in case of 2 species.

approach has a computational complexity which is proportional to the third power of the dimension (i.e., $\mathcal{O}((nN_1N_2N_3)^3)$). An obviously cheaper way to solve the linear systems is given by the following approach: we approximate the first term in the left-hand side of (1.17) as

$$\left[I - \gamma\Delta t\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{w}}\right] \approx \left[I - \gamma\Delta t\frac{\partial \boldsymbol{F}_1}{\partial \boldsymbol{w}}\right]\left[I - \gamma\Delta t\frac{\partial \boldsymbol{F}_2}{\partial \boldsymbol{w}}\right]\left[I - \gamma\Delta t\frac{\partial \boldsymbol{F}_3}{\partial \boldsymbol{w}}\right], \tag{1.18}$$

where $\boldsymbol{F}_1(\boldsymbol{w}(t))$, $\boldsymbol{F}_2(\boldsymbol{w}(t))$ and $\boldsymbol{F}_3(\boldsymbol{w}(t))$ correspond to the three first terms (derivative of fluxes) in the right-hand side of (1.14). Thus, we have removed the derivatives of the specific growth rates $g_{s_{ijk}}$ in the Jacobian matrix and have approximately factorized what remains in $[I - \gamma\Delta t\, \partial\boldsymbol{F}/\partial\boldsymbol{w}]$ in the way as shown in (1.18). The latter simplification is usually termed Approximate Matrix Factorization [4, 14, 28, 38] and can be seen as a form of dimension splitting. The growth rates can be omitted since these terms are non-stiff. Loosely speaking, the growth rates are now handled by simple successive substitution as in Jacobi iteration. All diffusion and advection terms are kept in the Jacobian because these terms contribute to the stiffness. Removing them could cause stability problems.

Recall that the coupling of the various species entered the model through the light function L (cf. (1.6)). Hence, by removing the influence of the growth term in the Jacobian the species have in fact been ‘uncoupled’ as far as the solution of the linear systems is concerned. Of course, the growth term is still present in Eq. (1.16), $\mathbf{R}(\boldsymbol{W}_{k+1}) = \mathbf{0}$, that we have to solve in each step. Moreover, due to the special splitting of $\boldsymbol{F}(\boldsymbol{w}) = \boldsymbol{F}_1(\boldsymbol{w}) + \boldsymbol{F}_2(\boldsymbol{w}) + \boldsymbol{F}_3(\boldsymbol{w})$, where each of the $\boldsymbol{F}_j$ is associated with only one spatial dimension, a further ‘uncoupling’ has been achieved. As a result, the implicit relations can be solved along each grid line separately. Hence, each of the three matrices $[I - \gamma \Delta t\, \partial \boldsymbol{F}_j / \partial \boldsymbol{w}]$, $j = 1, 2, 3$, essentially consists of a large collection of uncoupled band matrices of small dimension (which even could be solved in parallel). The computational complexity in solving such a system is linear in the dimension. Summarizing, the total computational complexity in solving the linear systems in (1.18) by Approximate Matrix Factorization equals $3n(\mathcal{O}(N_1 N_2 N_3))$. Obviously, this is orders of magnitude lower than what we need in a direct approach since there the costs increase with the third power of the total number of unknowns. The only concern that remains is a possibly reduced rate of convergence of the Newton process. However, as we shall see in Section 1.4.3, the averaged number of Newton iterations per time step did not exceed 2 (see Figs. 1.9c and 1.11c), which is of course quite satisfactory.

The time integration technique based on (1.16) has been implemented by Brown et al. [8] in the code VODE. The results presented in this chapter were obtained by adapting VODE w.r.t. the Approximate Matrix Factorization technique defined in (1.18). All strategies in VODE have been left unchanged. This widely used stiff ODE solver is very robust in the sense that it includes all kind of strategies, necessary for automatic integration. VODE is freely available from `http://www.netlib.org/ode/` (both in Fortran and C).

Remark 1.1. The authors of VODE have extended this code by incorporating a Krylov subspace iterative method (GMRES) for solving the linear systems arising in the Newton iteration. The resulting code, termed VODPK, allows the user to define a preconditioner to speed up the convergence of the GMRES-iteration. We refrained from solving our phytoplankton problem with VODPK because already one GMRES iteration is more costly than solving the three band matrices in the Approximate Matrix Factorization approach. Since 5–10 GMRES-iterations is quite common in this kind of applications, it is clear that Approximate Matrix Factorization (in combination with the good convergence of the Newton process) is a very efficient choice. A speed-up by (at least) a factor 10 compared with the VODPK-approach seems to be a realistic estimate.

1.4 Application

1.4.1 Model structure

We consider competition for light between three typical phytoplankton species: a sinking species ($c_s > 0$), a neutrally buoyant species ($c_s = 0$), and a buoyant species ($c_s < 0$). All these three species have similar growth characteristics. However, we assume that the sinking species has a higher maximal specific production rate than the neutrally buoyant species, which in turn has a higher specific production rate than the buoyant species. We note that a high specific growth rate allows proliferation under rather low light conditions. In Table 1.1 we specify all parameters characterizing the three species.

TABLE 1.1: Species parameters (see also [33, Table 2]).

Species	c_s $\left(\frac{\text{cm}}{\text{h}}\right)$	r_s $\left(\frac{\text{cm}^2}{\text{cells}}\right)$	$p_{s_{\max}}$ $\left(\frac{1}{\text{h}}\right)$	H_s $\left(\frac{\mu\text{mol photons}}{\text{cm}^2\,\text{s}}\right)$	ℓ_s $\left(\frac{1}{\text{h}}\right)$	ω_{s_0} $\left(\frac{\text{cells}}{\text{cm}^3}\right)$
Sinking species	$+4.2$	$3.0 \cdot 10^{-7}$	0.04	$1.0 \cdot 10^{-3}$	0.01	50
Neutral species	$+0.0$	$1.5 \cdot 10^{-7}$	0.03	$1.0 \cdot 10^{-3}$	0.01	5000
Buoyant species	-8.3	$1.5 \cdot 10^{-7}$	0.02	$2.0 \cdot 10^{-3}$	0.01	5000

TABLE 1.2: System parameters (see also [33, Table 1]).

X (m)	Y (m)	Z (m)	D_H $\left(\frac{\text{cm}^2}{\text{s}}\right)$	D_V $\left(\frac{\text{cm}^2}{\text{s}}\right)$	K_{bg} $\left(\frac{1}{\text{cm}}\right)$	L_{in} $\left(\frac{\mu\text{molphotons}}{\text{cm}^2\,\text{s}}\right)$
100	100	10	100	10	$2.0 \cdot 10^{-3}$	$3.5 \cdot 10^{-2}$

Our experiments are performed on the domain Ω with $X = Y = 100\,\text{m}$ and $Z = 10\,\text{m}$ using an equidistant grid of $50 \times 50 \times 20$ cells. Hence, the total system consists of 150 000 ODEs. The simulations are carried out with constant turbulences $D_H = 100$, $D_V = 10$, both in cm^2/s. All system parameter values[2] are summarized in Table 1.2.

For the water flow, we will use two different velocity fields. These flow fields are given in analytical form and have been chosen mainly for test purposes. In real-life applications the flow fields have to be computed by a hydrodynamical solver. Then the output of this solver serves as input for the current phytoplankton competition model. To stay as close as possible to a realistic flow, these two fields have been chosen divergence free, reflecting the incompressibility property of water.

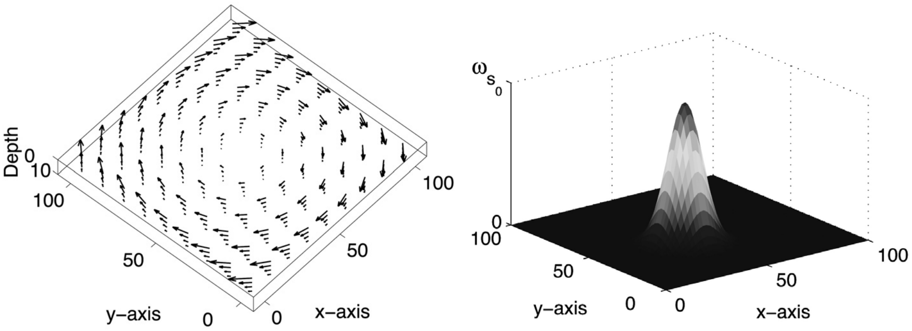
Test example 1.1. We start with the so-called Molenkamp velocity field (see Fig. 1.2a)

$$\begin{aligned} a(x,y,z) &= \frac{2\pi}{5Y}\left(y - \frac{Y}{2}\right) e^{-2z/Z}, \\ b(x,y,z) &= -\frac{2\pi}{5X}\left(x - \frac{X}{2}\right) e^{-2z/Z}. \end{aligned} \tag{1.19}$$

This velocity field has been used by various authors for testing PDE-solvers (see e.g. [38]). It describes a clockwise rotation around the center water column $(X/2, Y/2, z)$, with amplitude decreasing over depth. In the horizontal, the amplitude increases from the center towards the boundary of the domain. It takes $5Xe^{2z/Z}$ (s) for species at depth z to rotate for one cycle.

---

[2] The spatial-intervals and diffusion coefficients are chosen to be in the critical region of



a) Molenkamp velocity field.

b) Initial values of species.

FIGURE 1.2: Input for the Molenkamp test.

For this velocity field, we start with uniform population densities for each species over depth. In the horizontal plane, phytoplankton species are all densely distributed in the location of $x = 40$ m and $y = 40$ m (Fig. 1.2b) according to

$$\omega_s(x, y, z, 0) \;=\; \omega_{s_0} e^{-10^{-6}\left((x-0.4X)^2+(y-0.4Y)^2\right)},$$

where the amplitude ω_{s_0} is given in Table 1.1. Notice that the neutral and the buoyant species are both initialized 100 times more abundant than the sinking species.

Test example 1.2. In the second test example we use the velocity field defined by (see Fig. 1.3a)

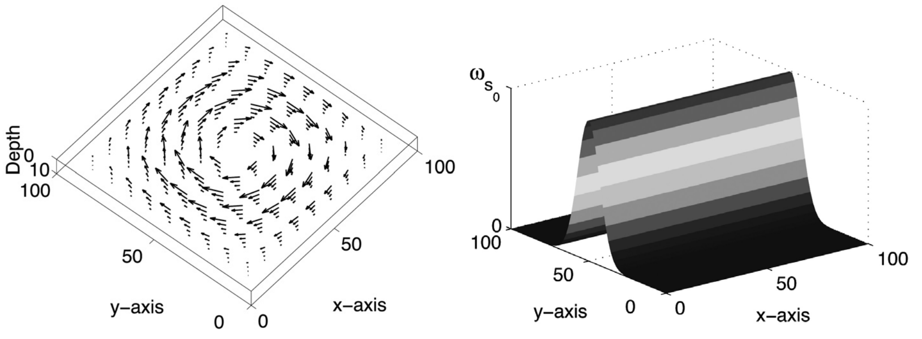
$$\begin{aligned} a(x, y, z) &= -\frac{\pi}{5}\, 10^{-4} X \,\sin^2(\pi \frac{x}{X})\, \sin(2\pi \frac{y}{Y})\, e^{-2z/Z}, \\ b(x, y, z) &= \frac{\pi}{5}\, 10^{-4} Y \,\sin^2(\pi \frac{y}{Y})\, \sin(2\pi \frac{x}{X})\, e^{-2z/Z}. \end{aligned} \tag{1.20}$$

Similar to the Molenkamp velocity field, it is again a clockwise rotation around the center water column $(X/2, Y/2, z)$, with amplitude decreasing over depth. The difference with the Molenkamp field concerns the amplitude of the flow in the horizontal. Here, the flow is minimal in the center and at the boundaries of the domain, whereas in the Molenkamp field the velocities assume their maximal values at the boundaries.

For this second test example, we also start with a vertically uniform population density for each species. In the horizontal plane, phytoplankton species are, however, all densely distributed in the location of $y = 50$ m (Fig. 1.3b) according to

$$\omega_s(x, y, z, 0) \;=\; \omega_{s_0} e^{-10^{-6}(y-0.5Y)^2}.$$

‘bloom development’ found in [33].



a) The velocity field.

b) Initial values of species.

FIGURE 1.3: Input for the second test example.

Remark 1.2. At the end of Section 1.2 we have discussed the boundary conditions for the phytoplankton concentrations. These conditions are of the form $a\omega - D_H \partial\omega/\partial x = 0$, with a being the velocity component in the x-direction (and similar conditions in the other two directions). It should be observed that these zero-flux conditions concern the phytoplankton concentrations and hold for all values of the velocity a. This means that for the Molenkamp test, where $a \neq 0$ at the boundaries, we will have a non-zero slope of the solution at the boundaries because $\partial\omega/\partial x = a\omega/D_H$. In the second test example, where we have zero water velocities at the boundaries, this condition results in a vanishing slope of the phytoplankton concentration (perpendicular to the boundary).

1.4.2 Biological observations

Competition behaviour Since the depth of the water column is not large and the initial concentrations are quite low, the light availability is sufficient in the whole water column. As a result, in the first 10 days the concentrations of all three species increase (see Fig. 1.4). The larger phytoplankton concentrations result in a steeper light gradient. Since the buoyant species has the smallest $p_{s_{\max}}$ and largest H_s (see Table 1.1), its production rate $p_s(L)$, defined in (1.4), will be small, especially on low light conditions. In the competition context, the buoyant species will therefore loose influence and its concentration starts to decrease. Consequently, the population of the neutral species has a chance to rapidly increase until its maximal value (at about 25 days). The light availability at that time is not sufficient for such a large amount of neutral species. Thus, the neutral species no longer grows and starts to decrease towards the steady state (obtained at about 150 days). As the neutral and the buoyant species populations are getting smaller, the sinking species makes use of its high specific growth rate property and the relatively high mixing, to proliferate. Eventually, at the steady state, the sinking species dominates (see also the Figs. 1.5–1.7).

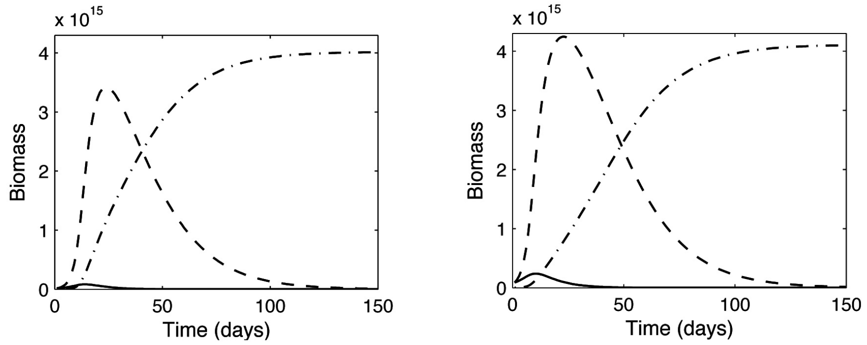
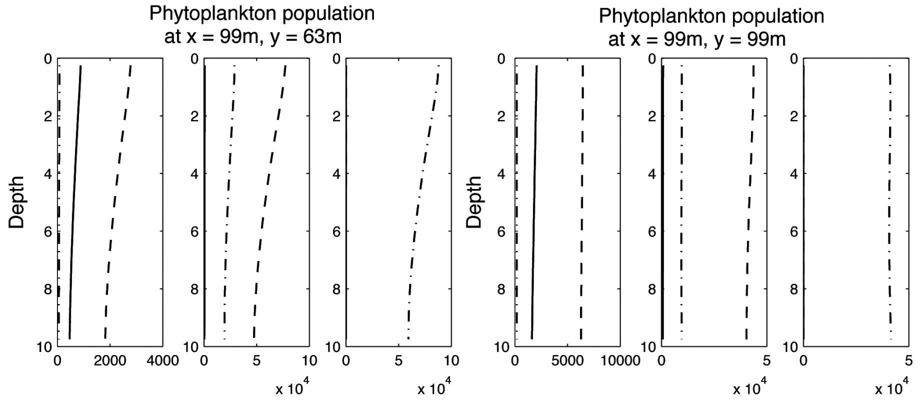


FIGURE 1.4: Biomasses (integral over space) of species in time, shown for the two different velocity fields according to test example 1.1 (left) and test example 1.2 (right). Sinking species: dash-dot line. Neutral species: dash line. Buoyant species: solid line.

We remark that the above competition behaviour is similar to what has been observed in the one-dimensional competition model [33].

Vertical distribution Even though phytoplankton species are distributed uniformly over depth at the onset, all species live close to the surface where ample light is available, while the populations usually decrease towards the bottom, because of darkness (Fig. 1.5). This behaviour was already found in [33].



Test example 1.1.

Test example 1.2.

FIGURE 1.5: Distribution over depth (x and y fixed) of Sinking species (dash-dot line), Neutral species (dash line) and Buoyant species (solid line) at 5 days (left), 25 days (middle) and 150 days (right). The left figures show the results for the Molenkamp test; the right figures for the velocity field of test example 1.2.

It is interesting to observe that the final concentration of the sinking species in case of the second velocity field (i.e., the right-most figure in Fig. 1.5) is almost constant in depth. Although this figure shows the situation in only one particular point in the (x, y)-plane (i.e., at $x = y = 99$ m), a further examination of the full solution reveals that this constant depth-profile is observed for all points in the horizontal plane.

Horizontal distribution In contrast to the aforementioned two aspects (competition behaviour and vertical distribution), the two test examples show a substantial difference with respect to the horizontal distribution.

For the Molenkamp test example we plotted in Fig. 1.6 the horizontal distribution of the three species just below the surface (at $z = 0.25$ m) after 1 day of simulation (left column) and at steady state (right column). We observe that the shapes of the various species are quite similar, whereas the amplitudes largely differ. The change of these amplitudes is in accordance with the time-behaviour of the biomasses (see Fig. 1.4, left panel).

We remark that the particular horizontal shape of the species shown in the right column of Fig. 1.6 was already observed after a few days of simulation and did not change significantly during the remaining part of the integration. Apparently, already after a short period of time all terms involving spatial derivatives are in balance and the only contribution to the right-hand side in (1.6) comes from the growth term.

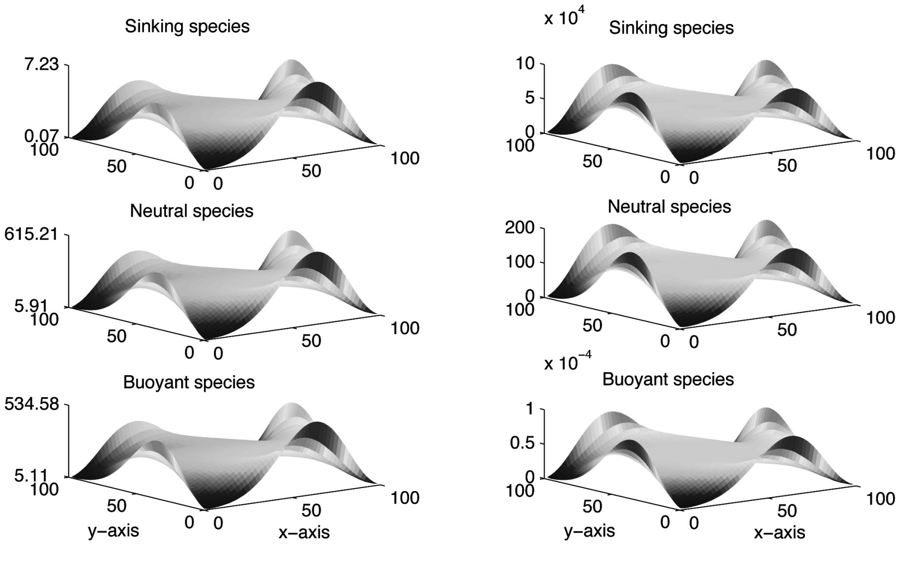


FIGURE 1.6: First test example (Molenkamp test): populations of phytoplankton species at $z = 0.25\,\mathrm{m}$ at $t = 1$ day (left) and $t = 150$ days (right).

Finally, the particular shape of the horizontal distribution can be partly explained by taking into account the boundary conditions that we imposed (see also Remark 1.2 at the end of Section 1.4.1). At all boundaries we use a condition of the form $v\,\omega - D\partial\omega/\partial x = 0$, with v the particular velocity component (see (1.5)). Hence the sign of $\partial\omega/\partial x = v\,\omega/D$ is determined by the sign of v. Using the Molenkamp velocity components as defined in (1.19) the correct slopes at the boundaries can be recognized in the plots.

Moreover, we consider an interior point close to a corner point. Making a Taylor series expansion of the solution in this interior point around the solution in the corner point and using the same reasoning as above for the sign of the derivatives, it can be proved that the solution in the corner point must vanish as shown in Fig. 1.6.

For the second test example the same information is given in Fig. 1.7. Again, the three species show a horizontal distribution which is quite similar and the mutual amplitudes are in accordance with the time-behaviour of the biomasses (see the right panel in Fig. 1.4). The main difference with the first test example is that eventually the horizontal structure has disappeared: the solutions are completely flat in the horizontal (except for the Buoyant species; the dark regions in the plot are in fact 'wiggles', i.e. numerical artifacts which will be discussed in Section 1.5). This horizontal solution 'profile' is in agreement with the observation that all derivatives (in normal direction) of the solution at the boundary vanish for this velocity field (see (1.20) and Remark 1.2 at the end of Section 1.4.1).

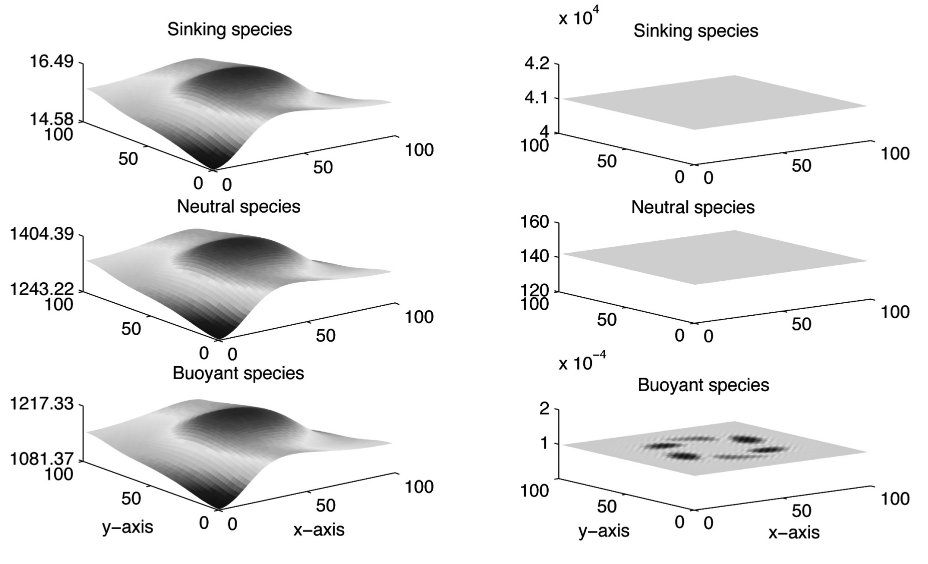


FIGURE 1.7: Second test example: populations of phytoplankton species at $z = 0.25\,\mathrm{m}$ at $t = 1$ day (left) and $t = 150$ days (right).

1.4.3 Numerical observations

To motivate the choice of the numerical techniques described in Sections 1.3.1 and 1.3.2, it is useful to look at certain characteristic numbers, such as the Cell Péclet number, the Courant-Friedrichs-Lewy (CFL) number and the stiffness number of our problem. These numbers, which are discussed in many numerical text books on PDEs (see e.g. [38]), are listed in Table 1.3 (notice that both test examples have the same maximal velocities). Both the CFL number and the stiffness number depend on the time step Δt. In Table 1.3 we have used the value $\Delta t = 200$(s), since this value turns out to be chosen by the BDF2 code in a characteristic integration scenario (see the discussion in the sequel of the section).

TABLE 1.3: $|a|$ and $|b|$ are the maximal velocities in the horizontal direction, $|c_s|$ is the largest velocity of the three species, h: mesh size, D: diffusion coefficient, Δt: time step.

Characteristic number	Horizontal	Vertical
Péclet number	$\lvert a\rvert h/D = 1.2566$	$\lvert c_s\rvert h/D = 0.0115$
CFL number	$(\lvert a\rvert /h + \lvert b\rvert /h)\Delta t = 1.2566$	$\lvert c_s\rvert \Delta t/h = 0.0092$
Stiffness number	$8\,\Delta t\, D/h^2 = 4.0000$	$4\,\Delta t\, D/h^2 = 3.2000$

The diffusion coefficient, which plays an important role in the population dynamics of phytoplankton [31, 33], varies in a wide range from 10^{-1}(cm^2/s) in poorly mixed water, to 10^3(cm^2/s) in well mixed water. This results in a wide range for the Péclet number as well as for the stiffness number. Hence, for small D-values, the Péclet number will be large which motivates the choice for the third-order upwind-biased discretization. On the other hand, for large D-values, the stiffness number enforces to use an implicit method. Since we want to capture the whole spectrum of parameter values in one code, we decided to include upwind discretization as well as an implicit method.

As said in Section 1.3.2, for the time integration we have used VODE (extended with the Approximate Matrix Factorization technique). We observed that the behaviour of VODE is a bit erratic, especially when the code tries to integrate with a high order formula (orders 1 until 5 are available). This erratic behaviour is probably due to the fact that the BDF formulae of order 3 and higher lack a part of the left half of the complex plane in their stability region. The use of these high-order formulae may have led to instabilities, caused by the complex eigenvalues originating from the (discretization of the) advection terms. Therefore, we also applied VODE with the maximal order set to 2, since the BDF formulae of order 1 and order 2 are unconditionally stable (A-stable). In the results described below, this mode will be denoted by BDF2, whereas the application of the full code will be denoted by VODE.

We will now discuss the behaviour of both solvers when applied to our two test examples on the time interval $[0, T]$, with $T = 2.5 \cdot 10^5$(s). For the first test example (the Molenkamp test), this interval corresponds to 5 rotations for each point at the surface. The results for the first test example are listed in Table 1.4.

From this table one can see that BDF2 is more efficient for the large tolerances, both in terms of CPU time and number of steps. Only for the very small tolerances VODE is more efficient, but those tolerance values are not realistic for our application (the spatial discretization is of order two).

TABLE 1.4: Test example 1.1. Output for BDF2 (upper part) and VODE (lower part). RTOL: relative tolerance, NST: number of steps, NNI: number of Newton iterations, NJE: number of Jacobian evaluations, CFN: number of non-linear convergence failures, ETF: number of error test failures, Q: order used in the final step, CPU: CPU time (seconds), GRERR: global relative error in L_2 norm.

RTOL	NST	NNI	NJE	CFN	ETF	Q	CPU	GRERR
10^{-2}	855	1583	26	7	6	2	466	6.23e–3
10^{-3}	1219	1920	24	2	15	2	601	8.79e–4
10^{-4}	1767	2588	32	1	21	2	767	9.55e–5
10^{-5}	2950	3747	53	2	46	2	1233	8.46e–6
10^{-6}	5268	6532	93	2	78	2	2161	2.46e–6
10^{-2}	1547	1989	30	3	4	2	608	8.42e–3
10^{-3}	1789	2487	56	17	14	3	797	5.41e–4
10^{-4}	3471	5302	59	1	10	4	1642	4.58e–5
10^{-5}	2592	4080	48	2	14	3	1239	9.48e–6
10^{-6}	4584	5766	77	0	33	4	1904	1.02e–6

Fig. 1.8 presents an accuracy/cost plot. This figure confirms the better performance of BDF2 in the low accuracy range. The global relative error GRERR is the time integration error for the semi-discrete system (1.15). This error has been obtained by comparing the numerical solution with a reference solution, obtained with a very small tolerance value.

For a better understanding of these results, we will have a closer look at the time integration statistics for one particular RTOL value, i.e. 10^{-3}. Initially, VODE increases both the step size and the order as we can see in Fig. 1.9a and b. Then at steps 125 and 126 two consecutive convergence failures occur. VODE reacts with a reduction of the step size with a factor 4 for each failure, maintaining order 5. Next, several error test failures occur which cause a further reduction of the step size. Then VODE decides to lower the order to 4 and then to 3, due to an error test failure. After a next convergence failure, the order is further reduced to 2.

As can be seen in Fig. 1.9a and b, VODE seems to have trouble in finding an appropriate time step and an appropriate order, especially from step number

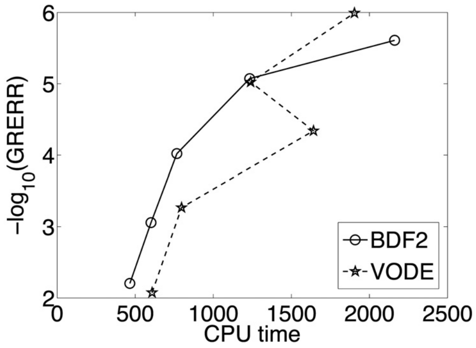
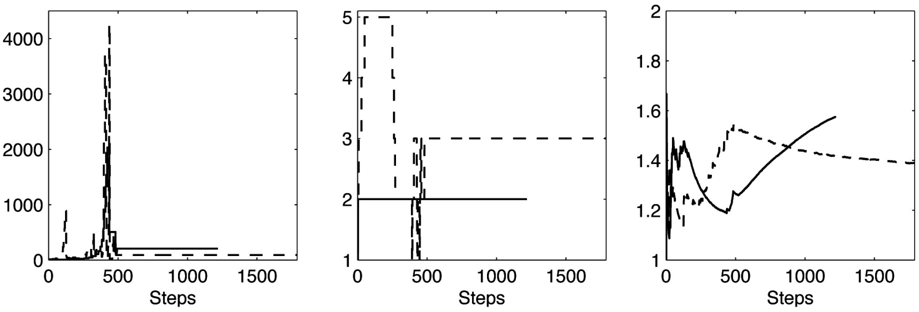


FIGURE 1.8: Test example 1.1. Efficiency plot for BDF2 and VODE.



a) The time step Δt. b) Order used. c) Averaged NNI/NST.

FIGURE 1.9: Test example 1.1. BDF2 (solid line) and VODE (dashed line) for RTOL=10^{-3}.

390 to 500, where other convergence failures occur. Finally, beyond step number 500, the time step and order are fixed at 89(s) and 3, respectively. With this choice, VODE successfully reaches the end of the integration interval without any failure (see Fig. 1.9a).

The behaviour of BDF2 is different. Here we observe a modest increase of the step size for the first 436 steps. Then, after a sudden increase of the step size, also BDF2 encounters two convergence failures at steps 436 and 481. After a reduction by a factor 4 for each failure, the step size is now appropriate to reach the end of the interval without any failure.

Observe that it is remarkable that the lower order BDF2 mode completes the integration with the constant step size $\Delta t = 206$(s) while VODE, using order 3, seems to feel comfortable with the constant step size $\Delta t = 89$(s) to satisfy the same tolerance criterion for the local error.

Finally, we will discuss the convergence behaviour which may have suffered from the fact that we replaced the Newton matrix in (1.17) by the right-hand side of (1.18).

As discussed in Section 1.3.2, the linear system in (1.17) has a huge number of entries. Therefore, we solved this complicated system by removing the growth term contributions and by successively solving three band-structured systems (see (1.18)) within each modified Newton iteration. In spite of this simplification, Newton's process still works very well: on average (taken over the steps), both modes need less than 2 Newton iterations per time step. This is shown in Fig. 1.9c.

For the second test example (defined in (1.20)), the codes show a behaviour which resembles the behaviour that we obtained for the first test example: again, VODE behaves rather erratic, in the sense that the global error is far from a monotone function of the costs. Furthermore, we again found that BDF2 is more efficient unless a very stringent tolerance-value is used. As can be seen in Fig. 1.10, these two properties are even more pronounced than in the first test example. These observations indicate that VODE encounters stability problems when using a high-order formula, due to the advection terms in the model.

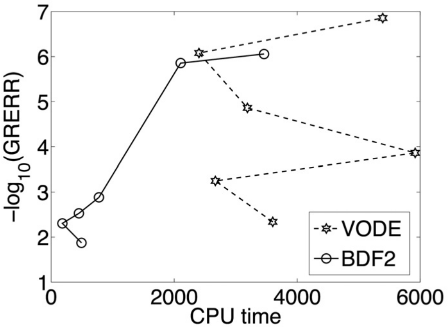


FIGURE 1.10: Test example 1.2. Efficiency plot for BDF2 and VODE.

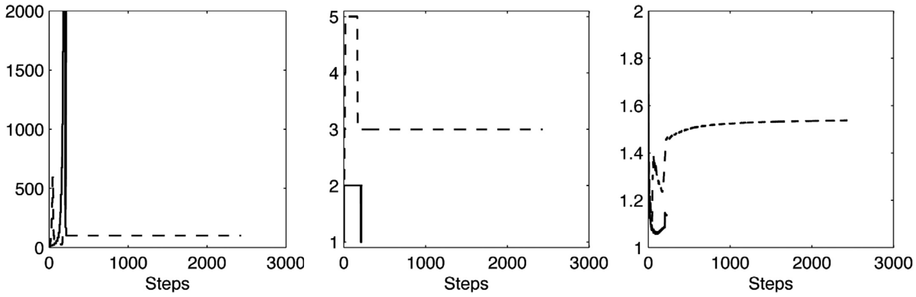
This conclusion is supported by the statistical data collected in Table 1.5: for RTOL=$10^{-2}, 10^{-4}$, and 10^{-5}, VODE integrates the last part of the integration interval using a fourth-order method and the resulting number of steps (and CPU time) are significantly higher than in case of a third-order formula (as VODE did for RTOL=10^{-3}, 10^{-6}). A close inspection of the performance of both solvers for RTOL=10^{-3} is shown in Fig. 1.11.

In Fig. 1.11a we see that both codes try to substantially increase the step size, in particular BDF2. However, then a convergence failure reduces the step size to a more realistic value. VODE settles at a constant step size of 102(s) until the end of the integration interval. After an initial increase of the order to 5, VODE completes the integration with order 3. BDF2 prefers the second-order formula, except for a few steps at the end where Backward Euler has been used (see Fig. 1.11b).

Finally, from Fig. 1.11c we conclude that the use of Approximate Matrix Factorization in the Newton process requires not more than 1.5 iteration (on average) which is a quite satisfactory convergence behaviour.

TABLE 1.5: Test example 1.2. Output for BDF2 (upper part) and VODE (lower part). The entries in this table have the same meaning as in Table 1.4.

RTOL	NST	NNI	NJE	CFN	ETF	Q	CPU	GRERR
10^{-2}	422	815	12	3	0	2	493	1.34e–2
10^{-3}	226	258	7	2	1	2	182	4.96e–3
10^{-4}	557	684	15	3	3	2	450	2.97e–3
10^{-5}	1034	1180	24	5	3	2	776	1.32e–3
10^{-6}	2636	3374	52	5	1	2	2105	1.40e–6
10^{-2}	3248	5128	55	0	11	4	3602	4.58e–3
10^{-3}	2431	3737	42	1	6	3	2669	5.67e–4
10^{-4}	4009	5346	69	1	7	4	5919	1.37e–4
10^{-5}	3100	5199	52	0	9	4	3191	1.38e–5
10^{-6}	2589	3915	47	2	8	3	2403	8.35e–7



a) The time step Δt. b) Order used. c) Averaged NNI/NST.

FIGURE 1.11: Test example 1.2. BDF2 (solid line) and VODE (dashed line) for RTOL=10^{-3}.

1.5 Discussion

In this final section, we will briefly summarize the approach and discuss our findings. Here, we distinguish between results from an application (biological) perspective and from an algorithmic (numerical) point of view.

Biological discussion

Although there exist several buoyant phytoplankton species, particularly among the cyanobacteria, most phytoplankton species have a slightly higher density than water. They sink. This particularly applies to many diatom species that are heavily armoured with silica, which may result in relatively high sinking velocities of diatoms compared to other phytoplankton species. Our simulation results show that such a sinking phytoplankton species wins when in competition with neutrally buoyant and buoyant phytoplankton. Since buoyant phytoplankton species should generally have better access to light, why is it that buoyant species may lose in competition? How does a sinking phytoplankton species manage to become dominant? Recent work has shown that the outcome of competition for light between buoyant and sinking phytoplankton species depends on the intensity of vertical mixing [29, 31, 33]. During weak vertical mixing, characterized by a low value of the vertical turbulent diffusion coefficient, buoyant species will float upwards and are superior competitors for light. In our simulations, however, the water column is intensely mixed, as the value of the vertical turbulent diffusion coefficient is relatively high (10 cm^2/s, Table 1.2). Such mixing conditions are characteristic of winter and early spring in the temperate climate zone, with little or no temperature stratification of the water column [39]. As a result, all phytoplankton species are more or less uniformly mixed over depth (Fig. 1.5). Furthermore, given the high horizontal diffusivities, all species are also distributed in a similar form along the horizontal plane (Figs. 1.6 and 1.7). In this case, when all species are similarly distributed in both the horizontal and vertical, the superior competitor for light is the species with lowest critical light intensity (sensu [30]). That is, the strongest competitor for light is the species that is best adapted to grow under low light conditions. In our simulations, the sinking species grows better under low light conditions than the neutrally buoyant and buoyant species, because it has a higher maximum specific production rate ($p_{s_{\max}}$, in Table 1.1) than the other species, combined with a relatively low half-saturation constant (H_s, in Table 1.1). As a consequence, the sinking species has a higher production rate at low light intensities than the other species. In conclusion, consistent with numerous field observations (see e.g. [57]), our results point out that sinking species that are able to grow well under low-light conditions (like many diatoms) can be strong competitors for light that may dominate the phytoplankton of temperate regions during intense mixing in winter and early spring.

Numerical discussion

The integro-PDE system modelling the three-dimensional phytoplankton dynamics has been solved in two steps.

First, we discretized the spatial differential operators as well as the integral term. The diffusion term has been discretized symmetrically. For the advection term, we used the third-order upwind-biased discretization. The integral term is approximated using the repeated trapezoidal role.

Next, the resulting ODE system has been integrated in time. Due to the stiffness, we selected an implicit approach, viz. the family of BDF methods. The automatic integrator VODE, which has actually been used, is based on this family of implicit methods. Since the structure in the Jacobian gives rise to a laborious linear algebra owing to coupling of the competing species through the integral term of the integro-PDEs, we implemented the Approximate Matrix Factorization technique in VODE. The effect of this approach is that the coupling of the unknowns in the linear systems is drastically reduced. As a matter of fact, the only coupling that remains is in one spatial direction. As a result, the total computational complexity for the Approximate Matrix Factorization is linear in N, the total number of unknowns, whereas directly solving the total linear system would require $\mathcal{O}(N^3)$ operations. Needless to say that such a direct approach is simply unfeasible with N-values as large as $1.5 \cdot 10^5$, as we have used in our experiments. We have also argued (see Remark 1.1 in Section 1.3.2) that the Approximate Matrix Factorization approach will be much more efficient than an iterative (Krylov-based) technique to solve the linear systems within the Newton process. We have made plausible that a speed-up factor of 10 is quite realistic.

Experiments with two test examples indicate that unconditional stability (A-stability) is a very useful property for the efficient and reliable solution of the phytoplankton dynamics model. Therefore, in the range of realistic (i.e., low) accuracies, the version of VODE in which we restricted the order to 2 shows a superior behaviour. However, based on what we experienced with this code, we believe that the numerical approach to solve this particular application can be improved upon. For example, as already mentioned in Section 1.4.2, the second test problem shows ‘wiggles’, i.e., small oscillations superimposed on a smooth solution (see the right panel in Fig. 1.7). The origin of such oscillations can be twofold: (i) spatial discretization and (ii) time integration.

For the spatial discretization of the advection terms we used the third-order upwind-biased scheme. Although rather accurate and better than a symmetric discretization, this choice does not guarantee that ‘wiggles’ will be absent. In fact, this is only achieved with the first-order upwind discretization. This choice, however, has the disadvantage of low accuracy and the introduction of a large amount of artificial diffusion. A possible remedy to avoid the unwanted oscillations and to maintain a high-order is to combine the third-order upwind-biased discretization with limiters (see e.g. [38], p. 215). A disadvantage is, however, that such a technique introduces additional nonlinearity in the scheme which is a drawback when implemented in a fully implicit method such as used in VODE.

The second source of oscillations stems from the time integration method. Also here we encounter an order 1-barrier. Hence, the A-stability of the BDF2 method is not a sufficient condition to suppress ‘wiggles’. Indeed, the only so-called positive method possessing this property is the BDF of order 1, i.e., Backward Euler. In passing, we remark that we also applied the Backward Euler scheme to the second test example and observed that the ‘wiggles’ were still present, although to a much smaller extent. Hence, in our application, the spatial discretization seems to be the main reason for the ‘wiggles’.

Based on the above considerations and the characteristics given in Table 1.3, an implicit-explicit approach might be a promising alternative. With this approach we mean that the non-stiff terms in the model (i.e., growth and advection) are treated by an explicit time integration method and the (stiff) diffusion terms are treated implicitly. In this setting, a third-order upwind-biased discretization combined with limiters only marginally complicates the algorithm and Approximate Matrix Factorization can still be used to solve the three-dimensional diffusion part. In this way we can avoid the Newton process since the diffusion terms are linear. This approach will be subject of Chapter 4.

Chapter 2

On positive solutions in a phytoplankton-nutrient model

On the basis of an application from aquatic ecology, we discuss the behaviour of the widely used time integration package VODE by Brown et al. [8]. When used in a default setting this code smoothly produces a negative steady state solution, which is not realistic in this application.

2.1 Introduction

Phytoplankton, the generic name of microorganisms living in lakes, seas and oceans [26], are at the basis of the aquatic foodweb. Their role for a proper functioning of the aquatic ecosystem has been recognized for a long time and has been widely studied both empirically [57] as well as theoretically [31, 15].

For their primary production of biomass, phytoplankton use photosynthesis [42], a process where solar energy (light) and carbon dioxide are utilized. Due to the sequestration of carbon dioxide, phytoplankton have a significant impact on the reduction of the greenhouse effect on a global scale (see e.g. [17]).

In many regions (and some seasons) light availability is the major factor limiting phytoplankton growth [70]. In other regions, and seasons, phytoplankton growth is largely influenced by the availability of nutrients, such as nitrogen, phosphorous, and iron (see [56, 7, 10]).

In this chapter we consider a model in which both limiting factors, light and nutrient, are taken into account. These two factors give rise to contrasting gradients since light is coming from above, whereas nutrients are supplied at the sediment (see e.g. [43, 72, 13]). As a result, the vertical distribution of the phytoplankton population can be quite heterogeneous in the sense that a large aggregation of phytoplankton is formed at a subsurface depth, where both light and nutrient are just sufficiently available to sustain a population.

2.2 The mathematical model

Here, we describe the phytoplankton-nutrient model for one single species (a multi-species extension of the model can be found in Chapter 5). The mono-species formulation is sufficient for the purpose of this chapter: showing the peculiar behaviour of the time integrator VODE [8].

We consider a water column in which the depth co-ordinate z runs from $z = 0$ (the surface) to $z = z_B$ (the bottom). Furthermore, let $\omega(z,t)$ denote the population density of a phytoplankton species at vertical position z at time $t \geq 0$. The distribution of phytoplankton is determined by the combined effect of growth (the main biological factor) and local transport processes (the main physical factor) through the partial differential equation

$$\frac{\partial \omega}{\partial t} = g\,\omega - \frac{\partial J}{\partial z}, \tag{2.1}$$

where g and J are, respectively, the growth rate and the flux at depth z at time t.

The flux J is determined by the convective transport, due to the settling speed v and the diffusive transport, due to mixing,

$$J(z,t) = v\,\omega(z,t) - D(z)\,\frac{\partial \omega}{\partial z}(z,t), \tag{2.2}$$

where $D(z)$ is the space-dependent mixing rate.

In our model, the growth rate g is assumed to depend on the light intensity $\mathcal{L}$ and the nutrient concentration $\mathcal{N}$. In fact, it depends on the balance between the production rate p and the specific loss rate ℓ as given by

$$g(\mathcal{L},\mathcal{N}) = p(\mathcal{L},\mathcal{N}) - \ell. \tag{2.3}$$

Here, the loss rate is assumed constant and represents grazing by zooplankton, mortality, excretion, etc. The production rate p determines the growth of phytoplankton and is defined by the two limiting environmental resources (i.e., light and nutrient) in the following way (see e.g. [43, 72]),

$$p(\mathcal{L},\mathcal{N}) = \mu \min\left(\frac{\mathcal{L}}{L_H + \mathcal{L}}, \frac{\mathcal{N}}{N_H + \mathcal{N}}\right), \tag{2.4}$$

where μ, L_H and N_H, respectively denote the maximum specific production rate and the half-saturation constants of light and nutrient.

Owing to shading and absorption, light intensity (the energy source) decreases with depth. At a particular depth, light intensity depends on the incident light intensity L_{in}, the light absorption coefficient of the water in the absence of phytoplankton K_{bg} and on the total light attenuation of the phytoplankton species above that particular depth (see e.g. [33]),

$$\mathcal{L}(z,t) = L_{in}\, e^{-K_{bg} z}\, e^{-r \int_0^z \omega(\sigma,t) d\sigma}, \tag{2.5}$$

where r denotes the specific light attenuation coefficient.

The change in the nutrient concentration $\mathcal{N}$ – the second environmental source that drives the growth rate g – is governed by the total amount of nutrient converted by the phytoplankton species and by the turbulent diffusion

$$\frac{\partial \mathcal{N}}{\partial t}(z,t) \;=\; -\,\alpha\, p(\mathcal{L},\mathcal{N})\,\omega(z,t) + \frac{\partial}{\partial z}\left(D(z)\,\frac{\partial \mathcal{N}}{\partial z}(z,t)\right), \tag{2.6}$$

where α denotes the conversion factor.

In summary, the phytoplankton-nutrient dynamics are studied through the following *system of integro-partial differential equations* (integro-PDEs)

$$\begin{aligned} \frac{\partial \omega}{\partial t} &= \Big(p(\mathcal{L},\mathcal{N}) - \ell\Big)\,\omega \;-\; \left[v\,\frac{\partial \omega}{\partial z} \;-\; \frac{\partial}{\partial z}\left(D(z)\,\frac{\partial \omega}{\partial z}\right)\right], \\ \frac{\partial \mathcal{N}}{\partial t} &= -\,\alpha\, p(\mathcal{L},\mathcal{N})\,\omega + \frac{\partial}{\partial z}\left(D(z)\,\frac{\partial \mathcal{N}}{\partial z}\right). \end{aligned} \tag{2.7}$$

To complete the model, we prescribe the following boundary conditions:

- Phytoplankton remains between two system boundaries: $J(z,t) = 0$ at $z = 0$ and $z = z_B$.
- There is no nutrient entering or leaving the surface: $\partial\mathcal{N}/\partial z = 0$ at $z = 0$.
- Nutrient is supplied from the sediment: $\mathcal{N}(z_B,t) = N_B$.

Remark 2.1. In this chapter we are interested in the long term behaviour of the solution. Therefore, we derive an explicit expression for the biomass of phytoplankton (i.e., the depth integral over the concentrations) in steady state. Dividing the second equation in (2.7) by α, adding to the first equation in (2.7) and integrating over space, results in

$$\frac{\partial}{\partial t}\int_0^{z_B}\left(\omega + \frac{1}{\alpha}\mathcal{N}\right)dz = -\ell\int_0^{z_B}\omega dz + \frac{D}{\alpha}\frac{\partial \mathcal{N}}{\partial z}(z_B,t),$$

where we have used the boundary conditions. Hence in steady state we obtain for the biomass of phytoplankton

$$\int_0^{z_B}\omega dz = \frac{D}{\ell\alpha}\frac{\partial \mathcal{N}}{\partial z}(z_B,t). \tag{2.8}$$

2.3 Numerical approach

For the numerical solution of the model (2.7) we follow the so-called Method of Lines approach. That is, we first discretize the spatial differential operators as well as the integral term in the light function. Subsequently, the resulting large system of ordinary differential equations (ODEs), which is still continuous in time, will be integrated numerically.

For the spatial discretization we use the same method as described in [33] and Chapter 1: symmetric second-order discretizations for the diffusion terms (both in the equations for the phytoplankton species and for the nutrient) and third-order upwind-biased formulae for the advection terms. For the integral term the repeated trapezoidal rule has been used. For a detailed discussion on these aspects, the interested reader is referred to [38].

The resulting stiff ODE system has been integrated in time by means of an implicit method, because an explicit time integrator has to obey an extremely severe time step restriction to avoid numerical instabilities. For the concept of stiffness, as well as other technical details about the time integration process, we refer to [24, 38]. For our simulations, we have selected the code VODE of Brown et al. [8]. This widely used code is based on a family of implicit backward differentiation formulas and proved to be efficient in many cases. VODE includes all kinds of strategies, necessary for automatic integration. It is available both in C and Fortran and is freely downloadable from the internet site `http://www.netlib.org/ode/`.

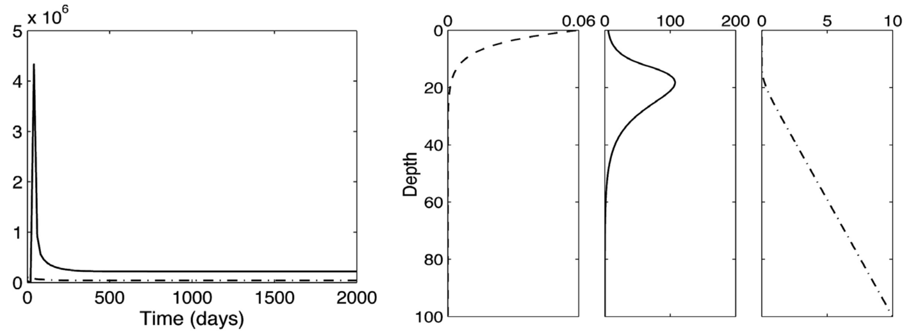
2.4 Application

We will discuss the simulation results of the phytoplankton-nutrient model for a sinking species (i.e., $v > 0$), the parameters of which have been specified in Table 2.1. The environmental parameters are given the values $D = 1\,\mathrm{cm}^2/\mathrm{s}$, $K_{bg} = 2 \cdot 10^{-3}\,\mathrm{cm}^{-1}$, $z_B = 100\,\mathrm{m}$, $N_B = 10\,\mu\mathrm{mol\ nutrient/cm}^3$, and $L_{in} = 6 \cdot 10^{-2}\,\mu\mathrm{mol\,photons/cm}^2\,\mathrm{s}$. For the initial condition we assume that nutrient is uniformly distributed over depth (i.e., equal to $10, \forall z$), whereas phytoplankton is initialized according to a Gaussian profile with a maximum of $50\,\mathrm{cells/cm}^3$ at 50 m (halfway down). Based on the experience that we obtained with the simulations described in Chapter 5, we use an equidistant spatial grid with 500 points. Such a grid is sufficiently fine to adequately capture the spatial variation in the solution.

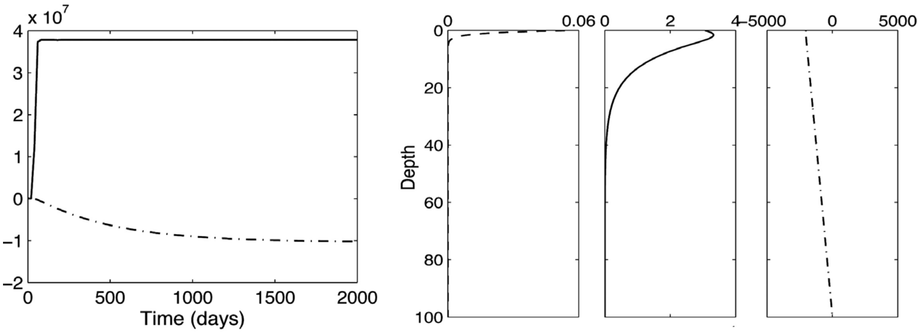
TABLE 2.1: Species parameters.

v $(\frac{\mathrm{cm}}{\mathrm{h}})$	r $(\frac{\mathrm{cm}^2}{\mathrm{cells}})$	μ $(\frac{1}{\mathrm{h}})$	L_H $(\frac{\mu\mathrm{mol\ photons}}{\mathrm{cm}^2\,\mathrm{s}})$	N_H $(\frac{\mu\mathrm{mol\ nutrient}}{\mathrm{cm}^3})$	ℓ $(\frac{1}{\mathrm{h}})$	α (const)
$+4.2$	$3 \cdot 10^{-7}$	0.04	$2 \cdot 10^{-3}$	$2 \cdot 10^{-3}$	0.01	$2 \cdot 10^{-3}$

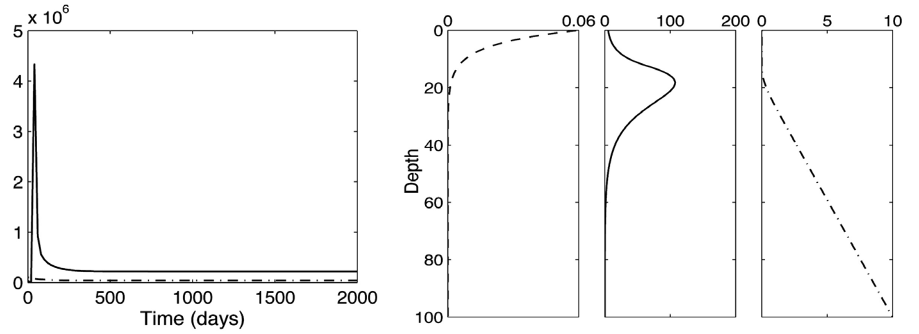
To first obtain insight in the exact ODE solution, VODE was applied with an extremely stringent value for the tolerance parameter. This solution is shown in Fig. 2.1a. We see that initially the phytoplankton biomass strongly increases (the peak near $t = 0$). This is due to a bloom of plankton near the water surface, since light and nutrient are amply available. This bloom causes a substantial consumption of nutrient, which starts to decrease in the upper water level. As a consequence, also the phytoplankton concentration reduces over there (because of a lack of food) and its maximum shifts downward to a position where light and



a) The exact solution.



b) Solution without positivity control.



c) Solution with positivity control.

FIGURE 2.1: Left: phytoplankton biomass (solid line) and total amount of nutrient (dash-dotted line) as functions of time. Right: distributions over depth (vertical axis, in m) of light intensity (dashed line), phytoplankton density (solid line) and nutrient concentration (dash-dotted line) at steady state.

nutrient are just sufficiently available to sustain a phytoplankton population. Already after about 400 days a stable stationary profile has been established, which is shown in the right panel of Fig. 2.1a. In passing we remark that, using a numerical approximation for $\partial \mathcal{N}/\partial z$ at the bottom, it is easily verified that relation (2.8) is indeed satisfied by the stationary solution.

Next we solve the problem with a coarser resolution in time: setting the tolerance parameter to 10^{-4} is more realistic in the present context of PDEs. However, using this value, we observe an unacceptable behaviour of VODE: it produces a *fundamentally wrong solution* (even negative components occur), which is shown in Fig. 2.1b. Although the integration statistics (such as number of time steps, number of rejected steps, number of Newton iterations to solve the nonlinear implicit relations, etc.) did not give rise to any suspicion, VODE returned a steady state with a completely wrong phytoplankton biomass and a negative value for the total amount of nutrient. This is of course not feasible (and hence unacceptable) and we did not observe such a behaviour before with this time integrator. Here we remark that it is known (see e.g., [38, page 187]) that linear multistep methods of order larger than 1 (which are used in VODE) have to obey a time step restriction for positivity reasons. However, in all previous simulations, the local error control in VODE prevented the code to converge to an unrealistic (i.e., negative) steady state solution. As an additional experiment, we applied VODE with the maximum order set to 1 (viz. the Backward Euler method, which has no step size restriction w.r.t. positivity). This test indeed yielded the correct solution, however at a high price, since now the time steps were kept very small to meet the accuracy requirements.

To prevent the unwanted situation corresponding to Fig. 2.1b and to avoid the excessive costs when using an extremely small tolerance (or first-order method), we implemented a provisional remedy by adapting the control strategy in VODE: after each time step, all solution components are checked for positivity. If one or more negative components occur, the step will be rejected and the time step is drastically reduced for a new try. Fig. 2.1c shows the results. Due to this 'brute force' strategy, the solution remains positive and is in fact quite close to the exact solution, shown in Fig. 2.1a. From this example, it is obvious that positivity-control is of utmost importance for this application.

2.5 Discussion

We have shown the behaviour of VODE when applied to a model from microbiology. In a default setting, VODE's behaviour is not satisfactory, since it yields negative solutions.

Hence, the lesson to learn is that using VODE as a 'black box' solver needs precaution in case of an application where the positivity of the solution is a prerequisite.

We have implemented an 'ad hoc' strategy to enforce positivity but this approach is far from optimal. Therefore, in the next chapter, we will concentrate on efficient time integration methods that guarantee positivity of the solution.

Chapter 3

Positivity for explicit two-step methods in linear multistep and one-leg form

Positivity results are derived for explicit two-step methods in linear multistep form and in one-leg form. It turns out that, using the forward Euler starting procedure, the latter form allows a slightly larger step size with respect to positivity.

3.1 Introduction

We consider the initial value problem for a positive system of ordinary differential equations (ODEs) in $\mathbb{R}^m$

$$\begin{aligned} \boldsymbol{w}'(t) &= \boldsymbol{F}(t, \boldsymbol{w}(t)), \\ \boldsymbol{w}(0) &= \boldsymbol{W}_0. \end{aligned}$$

With positivity (actually, preservation of non-negativity) we mean that the solution vector $\boldsymbol{w}(t) \geq 0$, $\forall\, t > 0$ if $\boldsymbol{W}_0 \geq 0$. Here, and in the sequel, such inequalities are to be understood componentwise. For such systems of ODEs we will study whether we can obtain a similar property for the numerical solutions $\boldsymbol{W}_n \approx \boldsymbol{w}(t_n)$, $t_n = n\Delta t$, Δt being the time step, and throughout this chapter it is assumed that $\boldsymbol{W}_0 \geq 0$. Positivity is a natural requirement in applications where the solution represents, for example, densities or concentrations. Moreover, the positivity condition is closely related to avoiding undershoots near steep gradients (see e.g. [38]). In [37], the related concept of monotonicity with semi-norms for linear multistep methods has been studied. Here we focus on positivity and adapt the results obtained in [37].

In Section 3.2 we will present an extension in the case of explicit two-step methods with forward Euler start-up (to compute $\boldsymbol{W}_1$), and we will point out the best method with respect to positivity, i.e., $\boldsymbol{W}_n \geq 0$ for $n \geq 1$, whenever

$\boldsymbol{W}_0 \geq 0$. In Section 3.3 we consider the corresponding one-leg form and show that this allows a slightly larger step size.

3.2 Positivity for linear two-step methods

Consider the following explicit linear two-step scheme

$$\boldsymbol{W}_{n+2} = \sum_{j=0}^{1} \Big[-\alpha_j \boldsymbol{W}_{n+j} + \beta_j \,\Delta t\, \boldsymbol{F}(t_{n+j}, \boldsymbol{W}_{n+j})\Big]. \tag{3.1a}$$

Observe that the freedom in scaling the coefficients has been used to set the coefficient in front of $\boldsymbol{W}_{n+2}$ equal to 1. In the one-leg form we will use a different scaling.

The scheme (3.1a) is of second-order accuracy if

$$\alpha_0 = 1-\xi,\ \alpha_1 = \xi - 2,\ \beta_0 = \frac{\xi}{2} - 1,\ \beta_1 = \frac{\xi}{2} + 1, \tag{3.1b}$$

where ξ is a free parameter. We note that the scheme is zero-stable (stable for the trivial equation $\boldsymbol{w}'(t) = \boldsymbol{0}$, see [38]) if the condition $-1 \leq \alpha_0 < 1$ is satisfied, i.e., if $0 < \xi \leq 2$. In the remainder of this chapter we shall always deal with methods that are second-order accurate and zero-stable. In [37], both implicit and explicit methods have been analyzed. In this section we will extend the results obtained in that paper for the explicit methods. For monotonicity results with higher-order methods, we refer to [23, 36].

Following Shu [62], the step in (3.1a) is written as a linear combination of scaled forward Euler steps yielding

$$\boldsymbol{W}_{n+2} = -\sum_{j=0}^{1} \alpha_j \Big[\boldsymbol{W}_{n+j} + c_j \Delta t\, \boldsymbol{F}(t_{n+j}, \boldsymbol{W}_{n+j})\Big], \quad c_j = -\frac{\beta_j}{\alpha_j}. \tag{3.2}$$

In the sequel we assume that there exists a $\Delta t_{FE} > 0$ such that

$$\boldsymbol{v} + \Delta t \boldsymbol{F}(t, \boldsymbol{v}) \geq 0 \quad \text{for all} \quad \boldsymbol{v} \geq 0,\ \ t \geq 0,\ \ 0 < \Delta t \leq \Delta t_{FE}. \tag{3.3}$$

Many ODEs originating from advection-diffusion problems indeed have a right-hand side function $\boldsymbol{F}$, for which (3.3) is a relevant condition (see e.g. [38]). Then, if

$$\beta_j \geq 0 \ \text{ and } \ \alpha_j \leq 0, \ \text{ for } j = 0, 1, \tag{3.4}$$

and hence $c_j \geq 0$, the terms within the square brackets in (3.2) are non-negative under the step size restriction $0 \leq c_j \Delta t \leq \Delta t_{FE}$, $j = 0, 1$. Therefore, $\boldsymbol{W}_{n+2} \geq 0$ for all $\Delta t \leq \min(\frac{1}{c_0}, \frac{1}{c_1})\Delta t_{FE}$, for *arbitrary* values of $\boldsymbol{W}_0, \boldsymbol{W}_1, \cdots, \boldsymbol{W}_{n+1} \geq 0$.

However, for the class of explicit second-order two-step methods, condition (3.4) for β_0 leads to $\xi \geq 2$. Combining this with the zero-stability requirement

$0 < \xi \le 2$ gives $\xi = 2$ as the only possible value. This, however, results in $c_1 = \infty$ and hence $\Delta t \le 0$. Indeed, for $\xi = 2$ we obtain

$$\boldsymbol{W}_{n+2} = \Big[\boldsymbol{W}_n - \boldsymbol{W}_{n+1}\Big] + \Big[\boldsymbol{W}_{n+1} + 2\Delta t \boldsymbol{F}(t_{n+1}, \boldsymbol{W}_{n+1})\Big].$$

Although the second term gives a positive contribution for $\Delta t \le \frac{1}{2}\Delta t_{FE}$, the first term can be negative for arbitrary positive $\boldsymbol{W}_n$ and $\boldsymbol{W}_{n+1}$ which may result in $\boldsymbol{W}_{n+2} < 0$.

Fortunately, if we consider an appropriate starting procedure, a positive result can be obtained [37, 36]. If $\boldsymbol{W}_1$ is obtained by the forward Euler method, i.e.,

$$\boldsymbol{W}_1 = \boldsymbol{W}_0 + \Delta t \boldsymbol{F}(t_0, \boldsymbol{W}_0), \tag{3.5}$$

we have $\boldsymbol{W}_1 \ge 0$ for all $\Delta t \le \Delta t_{FE}$ (see (3.3)). By introducing a non-negative parameter θ, which will be specified later, and subsequently subtracting and adding $\theta^j \boldsymbol{W}_{n+2-j}, j = 1, 2, \cdots, n+1$, in (3.1a), in which the added terms with $j = 1, 2, \cdots, n$ are again written in the form of (3.1a), we arrive at

$$\begin{aligned} \boldsymbol{W}_{n+2} = \; & (-\alpha_1 - \theta)\boldsymbol{W}_{n+1} + \beta_1 \Delta t\, \boldsymbol{F}_{n+1} \\ & + \sum_{j=0}^{n-1} \theta^j \Big[(-\alpha_0 - \theta\alpha_1 - \theta^2)\boldsymbol{W}_{n-j} + (\beta_0 + \theta\beta_1)\Delta t \boldsymbol{F}_{n-j}\Big] \\ & + \; \theta^{n-1}\Big[\theta^2 \boldsymbol{W}_1 - \theta\alpha_0 \boldsymbol{W}_0 + \theta\beta_0 \Delta t \boldsymbol{F}_0\Big], \qquad n \ge 0, \end{aligned} \tag{3.6}$$

where $\boldsymbol{F}_j$ denotes $\boldsymbol{F}(t_j, \boldsymbol{W}_j)$. Since $\boldsymbol{W}_1$ was calculated by the forward Euler method and $\alpha_1 = -1 - \alpha_0$ (see (3.1b)), this relation can be written as

$$\begin{aligned} \boldsymbol{W}_{n+2} = \; & (-\alpha_1 - \theta)\boldsymbol{W}_{n+1} + \beta_1 \Delta t\, \boldsymbol{F}_{n+1} \\ & + \sum_{j=0}^{n-1} \theta^j \Big[(1-\theta)(\theta - \alpha_0)\boldsymbol{W}_{n-j} + (\beta_0 + \theta\beta_1)\Delta t \boldsymbol{F}_{n-j}\Big] \\ & + \; \theta^{n}\Big[(\theta - \alpha_0)\boldsymbol{W}_0 + (\theta + \beta_0)\Delta t \boldsymbol{F}_0\Big], \qquad n \ge 0. \end{aligned} \tag{3.7}$$

Considering this step as a linear combination of scaled forward Euler steps, we see that if all coefficients are non-negative, i.e.,

$$\begin{aligned} & -\alpha_1 - \theta \ge 0, && \beta_1 \ge 0, \\ & (1-\theta)(\theta - \alpha_0) \ge 0, && \beta_0 + \theta\beta_1 \ge 0, \\ & \theta - \alpha_0 \ge 0, && \theta + \beta_0 \ge 0, \end{aligned} \tag{3.8}$$

then $\boldsymbol{W}_{n+2} \ge 0$ under the step size restriction $\Delta t \le \gamma(\theta)\Delta t_{FE}$, where

$$\begin{aligned} \gamma(\theta) \; & = \min\left(\frac{-\alpha_1 - \theta}{\beta_1}, \; \frac{(1-\theta)(\theta - \alpha_0)}{\beta_0 + \theta\beta_1}, \; \frac{\theta - \alpha_0}{\theta + \beta_0}\right) \\ & =: \min\left(A(\theta), B(\theta), C(\theta)\right). \end{aligned} \tag{3.9}$$

Obviously, the larger $\gamma(\theta)$, the better are the positivity properties of the scheme.

The conditions (3.8) define an eligible θ-interval, viz. $\theta \in [\theta_{\min}, \theta_{\max}]$, where

$$\begin{aligned}\theta_{\min} &= \max(\alpha_0, -\tfrac{\beta_0}{\beta_1}, -\beta_0) = -\beta_0,\\ \theta_{\max} &= \min(-\alpha_1, 1).\end{aligned}$$

Observe that $A(\theta)$, $B(\theta)$ and $C(\theta)$ are monotonic decreasing functions of θ (recall the condition $0 < \xi \leq 2$). Therefore, we obtain the maximal $\gamma(\theta)$-value

$$\begin{aligned}\gamma_{\max} &= \min\left(A(\theta_{\min}),\ B(\theta_{\min}),\ C(\theta_{\min})\right)\\ &= \begin{cases} B(\theta_{\min}) = \frac{\xi}{2-\xi} & \text{if} \quad 0 < \xi \leq \frac{2}{3}, \\ A(\theta_{\min}) = \frac{2-\xi}{2+\xi} & \text{if} \quad \frac{2}{3} \leq \xi \leq 2. \end{cases}\end{aligned}$$

Using the above considerations we can formulate the following theorem on the positivity condition for explicit linear two-step methods.

Theorem 3.1. *If $\boldsymbol{W}_1$ is obtained by the forward Euler method (3.5) then the explicit two-step method (3.1) is positive under the step size restriction $\Delta t \leq \gamma_{\max}\Delta t_{FE}$ where*

$$\gamma_{\max} = \min\left(\frac{\xi}{2-\xi}, \frac{2-\xi}{2+\xi}\right). \tag{3.10}$$

This $\gamma_{\max}$ is plotted in Fig. 3.1. The ascending part of the $\gamma_{\max}$-curve (i.e., for $0 < \xi < \frac{2}{3}$) gives an improvement of the bound in [37]. We note that in that paper only the minimum of $A(\theta)$ and $B(\theta)$ was considered in (3.9), leading to a different value of $\theta_{\min}$. The forward Euler starting procedure (3.5) was introduced afterwards, but this does not lead to a positivity result for $0 < \xi < \frac{2}{3}$.

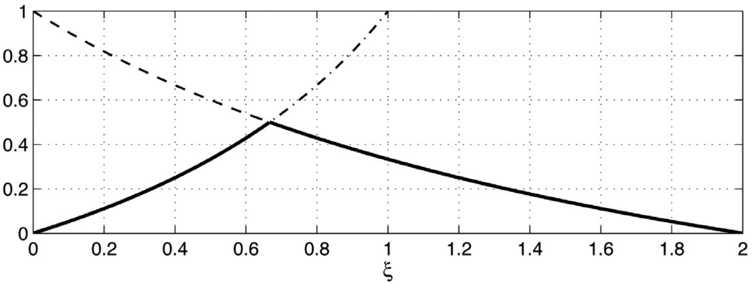


FIGURE 3.1: The values of $\gamma_{\max}$ (solid), $A(\theta_{\min})$ (dashed), and $B(\theta_{\min})$ (dash-dotted) as functions of ξ.

From Fig. 3.1 we see that, within the class of explicit second-order two-step method, the optimal method with respect to positivity is the $\xi = \frac{2}{3}$ method (known as the extrapolated BDF2 method [38]). The resulting value for $\gamma_{\max}$ is $\frac{1}{2}$.

Remark 3.1. In (3.6), the sequence of subtracting and adding $\theta^j \boldsymbol{W}_{n+2-j}$ was performed until $j = n+1$. In [37] these terms were subtracted and added up to $j = n$. It has been proved that the latter choice has no advantages compared with the choice made in (3.6), i.e., does not lead to a more relaxed condition on Δt (see the Appendix at the end of this chapter).

3.3 Positivity for one-leg methods

One-leg schemes were introduced by Dahlquist [12] to facilitate the analysis of linear multistep methods. Therefore, it is of interest to study the positivity properties of methods in the one-leg form. Similar to the preceding section, we will consider explicit methods. We will see that the results are slightly better than those derived for the linear multistep methods.

A natural scaling for one-leg methods is to require $\beta_0 + \beta_1 = 1$. Starting from the linear multistep method (3.1) we multiply the coefficients by a factor $\frac{1}{\xi}$ to obtain

$$\alpha_2 \boldsymbol{W}_{n+2} = \sum_{j=0}^{1} \Big[-\alpha_j \boldsymbol{W}_{n+j} + \beta_j \, \Delta t \, \boldsymbol{F}(t_{n+j}, \boldsymbol{W}_{n+j}) \Big], \tag{3.11a}$$

where

$$\alpha_0 = \frac{1}{\xi} - 1, \ \ \alpha_1 = 1 - \frac{2}{\xi}, \ \ \alpha_2 = \frac{1}{\xi}, \ \ \beta_0 = \frac{1}{2} - \frac{1}{\xi}, \ \ \beta_1 = \frac{1}{2} + \frac{1}{\xi}. \tag{3.11b}$$

Since $\xi > 0$ we have

$$0 < \alpha_2 = -(\alpha_1 + \alpha_0). \tag{3.12}$$

The one-leg form of (3.11a) reads

$$\begin{aligned} \alpha_2 \boldsymbol{W}_{n+2} &= -\alpha_1 \boldsymbol{W}_{n+1} - \alpha_0 \boldsymbol{W}_n + \Delta t \boldsymbol{F}\left(\bar{t}, \overline{\boldsymbol{W}}_{n+2}\right), \\ \overline{\boldsymbol{W}}_{n+2} &= \beta_1 \boldsymbol{W}_{n+1} + \beta_0 \boldsymbol{W}_n, \end{aligned} \tag{3.13}$$

where $\bar{t} = \beta_1 t_{n+1} + \beta_0 t_n = t_n + \beta_1 \Delta t$. This one-leg method is second-order accurate if the coefficients satisfy (3.11b).

In order to obtain positivity of $\boldsymbol{W}_n$, $n \geq 1$, we follow an approach used in [35] for a class of BDF2-type methods, applied to linear problems. Let us define

$$\boldsymbol{V}_n = \boldsymbol{W}_n - \theta \boldsymbol{W}_{n-1}, \quad \theta \in [0, 1), \quad n \geq 1. \tag{3.14}$$

Furthermore, we introduce the coefficients

$$\begin{aligned} &\alpha_1^* = -\alpha_1 - \alpha_2\theta, \quad && \alpha_2^* = -\alpha_0 - \alpha_1\theta - \alpha_2\theta^2 = (1-\theta)(\alpha_2\theta - \alpha_0), \\ &\beta_1^* = \beta_1, && \beta_2^* = \beta_0 + \beta_1\theta. \end{aligned} \tag{3.15}$$

The parameter θ in (3.14) and (3.15) will be chosen such that the coefficients in (3.15) satisfy

$$\alpha_j^* \geq 0, \quad \beta_j^* \geq 0, \quad j = 1, 2. \tag{3.16}$$

Assuming

$$\boldsymbol{V}_1 \geq 0, \quad \boldsymbol{W}_0 \geq 0, \tag{3.17}$$

we have the following theorem.

Theorem 3.2. *Suppose that* $\Delta t \leq \mathcal{C}\Delta t_{FE}$, *with* $\mathcal{C} = \min\left(\frac{\alpha_1^*}{\beta_1^*}, \frac{\alpha_2^*}{\beta_2^*}\right)$, *and* θ *is such that the conditions (3.16) and (3.17) are satisfied. Then* $\boldsymbol{V}_n \geq 0$ *and* $\boldsymbol{W}_n \geq 0$ *for all* $n \geq 1$.

Proof. The formulae (3.13)–(3.14) give

$$\begin{aligned} \alpha_2 \boldsymbol{V}_{n+2} &= \alpha_1^* \boldsymbol{V}_{n+1} + \alpha_2^* \boldsymbol{W}_n + \Delta t \boldsymbol{F}\left(\bar{t}, \overline{\boldsymbol{W}}_{n+2}\right), && (3.18) \\ \overline{\boldsymbol{W}}_{n+2} &= \beta_1^* \boldsymbol{V}_{n+1} + \beta_2^* \boldsymbol{W}_n. && (3.19) \end{aligned}$$

Adding $\mathcal{C}\overline{\boldsymbol{W}}_{n+2}$ to both sides in equation (3.18) we obtain

$$\alpha_2 \boldsymbol{V}_{n+2} = (\alpha_1^* - \mathcal{C}\beta_1^*)\boldsymbol{V}_{n+1} + (\alpha_2^* - \mathcal{C}\beta_2^*)\boldsymbol{W}_n + \mathcal{C}\overline{\boldsymbol{W}}_{n+2} + \Delta t \boldsymbol{F}\left(\bar{t}, \overline{\boldsymbol{W}}_{n+2}\right).$$

The coefficients in this relation are non-negative, due to the definition of $\mathcal{C}$ and (3.12). Therefore, $\boldsymbol{V}_{n+2} \geq 0$ if

$$\boldsymbol{V}_{n+1} \geq 0, \quad \boldsymbol{W}_n \geq 0, \quad \mathcal{C}\overline{\boldsymbol{W}}_{n+2} + \Delta t \boldsymbol{F}\left(\bar{t}, \overline{\boldsymbol{W}}_{n+2}\right) \geq 0. \tag{3.20}$$

The term $\mathcal{C}\overline{\boldsymbol{W}}_{n+2} + \Delta t \boldsymbol{F}\left(\bar{t}, \overline{\boldsymbol{W}}_{n+2}\right)$ can be seen as a scaled forward Euler step. Thus, it is non-negative if $\overline{\boldsymbol{W}}_{n+2} \geq 0$ and $\Delta t \leq \mathcal{C}\Delta t_{FE}$. From (3.19) and (3.16) we see that $\overline{\boldsymbol{W}}_{n+2} \geq 0$ if

$$\boldsymbol{V}_{n+1} \geq 0 \quad \text{and} \quad \boldsymbol{W}_n \geq 0. \tag{3.21}$$

Combining (3.20) and (3.21) we have

$$\boldsymbol{V}_{n+2} \geq 0 \text{ if } \boldsymbol{V}_{n+1} \geq 0 \text{ and } \boldsymbol{W}_n \geq 0. \tag{3.22}$$

By assumption, we know that $\boldsymbol{V}_1 \geq 0$ and $\boldsymbol{W}_0 \geq 0$ (see (3.17)). Thus, relation (3.14) gives $\boldsymbol{W}_1 = \boldsymbol{V}_1 + \theta\boldsymbol{W}_0 \geq 0$. Moreover, (3.22) yields $\boldsymbol{V}_2 \geq 0$. As a result, $\boldsymbol{W}_2 = \boldsymbol{V}_2 + \theta\boldsymbol{W}_1 \geq 0$ (see (3.14)). Again by (3.22) we obtain $\boldsymbol{V}_3 \geq 0$ which results in $\boldsymbol{W}_3 = \boldsymbol{V}_3 + \theta\boldsymbol{W}_2 \geq 0$, etc. for all $n \geq 4$. ☐

Let us now return to assumption (3.17). If $\boldsymbol{W}_1$ is calculated by the forward Euler method then

$$\boldsymbol{V}_1 = \boldsymbol{W}_1 - \theta \boldsymbol{W}_0 = (1-\theta)\boldsymbol{W}_0 + \Delta t \boldsymbol{F}_0 \geq 0$$

under the additional step size restriction $\Delta t \leq (1-\theta)\Delta t_{FE}$.

Using the above considerations we can formulate the following theorem on the positivity condition for the one-leg method.

Theorem 3.3. *If $\boldsymbol{W}_1$ is obtained by the forward Euler method (3.5) and θ is such that condition (3.16) is satisfied, then the one-leg method (3.13) is positive under the step size restriction $\Delta t \leq \gamma^{\mathrm{OL}}(\theta)\Delta t_{FE}$ where*

$$\gamma^{\mathrm{OL}}(\theta) = \min(\mathcal{C}, 1-\theta) = \min\left(\frac{-\alpha_1 - \alpha_2\theta}{\beta_1}, \frac{(1-\theta)(\alpha_2\theta - \alpha_0)}{\beta_0 + \beta_1\theta}, 1-\theta\right). \tag{3.23}$$

It is illustrative to compare this $\gamma^{\mathrm{OL}}(\theta)$ with the $\gamma(\theta)$ derived in (3.9): Condition (3.16) gives $\theta \in [\theta_{\min}, \theta_{\max}]$, where

$$\theta_{\min} = \max(\tfrac{\alpha_0}{\alpha_2}, -\tfrac{\beta_0}{\beta_1}) = -\tfrac{\beta_0}{\beta_1},$$

$$\theta_{\max} = \min(-\tfrac{\alpha_1}{\alpha_2}, 1).$$

Observe that the terms in the minimum function in (3.23) are monotonic decreasing functions of θ. Therefore, the optimal $\gamma^{\mathrm{OL}}(\theta)$-value is obtained at $\theta = \theta_{\min} = \frac{2-\xi}{2+\xi}$ and is given by

$$\gamma^{\mathrm{OL}}_{\max} = \min\left(\frac{2(1+\xi)(2-\xi)}{(2+\xi)^2}, \frac{2\xi}{2+\xi}\right). \tag{3.24}$$

The result is plotted in Fig. 3.2. From this figure we see that the best method with respect to positivity is no longer the method with $\xi = \frac{2}{3}$. The optimal

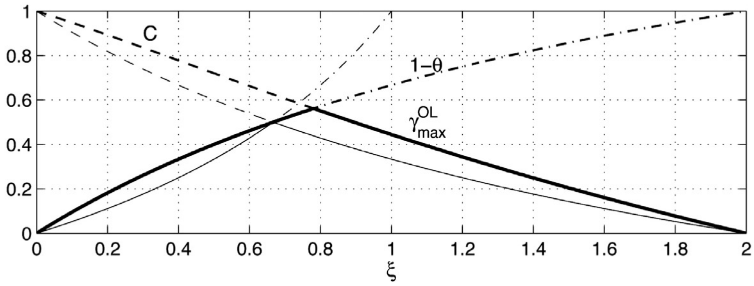


FIGURE 3.2: Step size restriction for positivity of the one-leg methods (thick lines) and of the linear two-step methods (thin lines, obtained from Fig. 3.1).

method with respect to positivity is now the method with $\xi = \frac{1}{4}(\sqrt{17}-1) \approx 0.78$. The corresponding $\gamma_{\max}^{\mathrm{OL}}$ is then $\frac{1}{2}(\sqrt{17}-3) \approx 0.56$. Comparing (3.10) and (3.24) we see that the one-leg method allows a slightly larger time step than the linear two-step method with the forward Euler starting procedure.

Appendix

In this Appendix, we will show that the choice made in [37] has no advantages compared with the choice made in (3.6), i.e., does not lead to a more relaxed condition on Δt.

In (3.6), and hence in (3.7), the sequence of subtracting and adding was performed until $j = n+1$. If we subtract and add $\theta^j \boldsymbol{W}_{n+2-j}$ up to $j = n$ into (3.1a), as it was done in [37], we arrive at

$$\begin{aligned}\boldsymbol{W}_{n+2} = \; & (-\alpha_1 - \theta)\boldsymbol{W}_{n+1} + \beta_1 \Delta t\, \boldsymbol{F}_{n+1} \\ & + \sum_{j=0}^{n-2} \theta^j \Big[(1-\theta)(\theta - \alpha_0)\boldsymbol{W}_{n-j} + (\beta_0 + \theta\beta_1)\Delta t \boldsymbol{F}_{n-j}\Big] \\ & + \theta^{n-1}\Big[(-\alpha_0 - \theta\alpha_1)\boldsymbol{W}_1 + (\beta_0 + \theta\beta_1)\Delta t \boldsymbol{F}_1 - \theta\alpha_0 \boldsymbol{W}_0 + \theta\beta_0 \Delta t \boldsymbol{F}_0\Big],\end{aligned} \tag{3.25}$$

with $n \geq 1$.

The first two lines of the right-hand side in (3.25) is non-negative if

$$\max\left(\alpha_0, -\frac{\beta_0}{\beta_1}\right) = -\frac{\beta_0}{\beta_1} \leq \theta \leq \min(-\alpha_1, 1), \tag{3.26}$$

under the step size restriction

$$\Delta t \leq \min\big(\Gamma_2, A(\theta), B(\theta)\big)\Delta t_{FE}, \tag{3.27}$$

where $\Gamma_2 = \frac{2-\xi}{2+\xi}$ is to guarantee positivity of $\boldsymbol{W}_2$ (see [37]), because $\boldsymbol{W}_2$ acts as the initial value of the sequence in the summation term in (3.25). The definition of $A(\theta)$ and $B(\theta)$ is given in (3.9).

Since we start with the forward Euler method (3.5), the term within the square brackets in the last line of (3.25) is written as

$$(-\alpha_0 - \theta\alpha_1 - \delta)\boldsymbol{W}_1 + (\beta_0 + \theta\beta_1)\Delta t \boldsymbol{F}_1 + (-\theta\alpha_0 + \delta)\boldsymbol{W}_0 + (\theta\beta_0 + \delta)\Delta t \boldsymbol{F}_0, \tag{3.28}$$

where δ is introduced for optimality reasons and will be determined below. This term is positive if

$$-\alpha_0 - \theta\alpha_1 - \delta \geq 0, \quad \beta_0 + \theta\beta_1 \geq 0, \quad -\theta\alpha_0 + \delta \geq 0, \quad \theta\beta_0 + \delta \geq 0, \tag{3.29}$$

under the step size restriction $\Delta t \leq f(\delta)\Delta t_{FE}$, where

$$f(\delta) = \min\left(\frac{-\alpha_0 - \theta\alpha_1 - \delta}{\beta_0 + \theta\beta_1}, \frac{-\theta\alpha_0 + \delta}{\theta\beta_0 + \delta}\right). \tag{3.30}$$

In summary, $\boldsymbol{W}_{n+2} \geq 0$ if the conditions (3.26) and (3.29) are satisfied and

$$\Delta t \;\leq\; \widetilde{\gamma}(\theta)\Delta t_{FE},$$

with

$$\widetilde{\gamma}(\theta) = \min\big(\Gamma_2, A(\theta), B(\theta), D(\theta)\big)$$

and

$$D(\theta) = \max_{\delta} f(\delta).$$

Condition (3.29) gives $-\beta_0\theta \leq \delta \leq -\alpha_0 - \theta\alpha_1$ implying $\theta \geq -\frac{\alpha_0}{\beta_0}$ (note $\alpha_1 = 2\beta_0 \leq 0$). Together with (3.26) we have the condition on θ

$$\min(-\alpha_1, 1) \;\geq\; \theta \;\geq\; \max\left(-\frac{\beta_0}{\beta_1}, -\frac{\alpha_0}{\beta_0}\right) = \begin{cases} -\frac{\beta_0}{\beta_1} & \text{if } \xi \geq \frac{2}{3}, \\ -\frac{\alpha_0}{\beta_0} & \text{if } \xi \leq \frac{2}{3}. \end{cases} \tag{3.31}$$

Observe that both the terms within the minimum function in (3.30) are monotonic decreasing functions of δ. Thus, in the interval $[-\beta_0\theta, -\alpha_0 - \theta\alpha_1]$, the maximal $f(\delta)$-value is obtained at $\delta = -\beta_0\theta$ and is given by

$$D(\theta) = \max_{\delta} f(\delta) = \frac{-\alpha_0 - \theta\beta_0}{\beta_0 + \theta\beta_1}. \tag{3.32}$$

As stated in Section 3.2, $A(\theta)$ and $B(\theta)$ are monotonic decreasing with θ. $D(\theta)$ is a monotonic decreasing function of θ if $\xi > \frac{2}{3}$ and is monotonic increasing with θ if $\xi < \frac{2}{3}$. Thus, for $\xi \geq \frac{2}{3}$ the maximal $\widetilde{\gamma}(\theta)$-value is obtained at $\theta = -\frac{\beta_0}{\beta_1}$ (see (3.31)) and is

$$\widetilde{\gamma}_{\max} = \min\big(\Gamma_2, A\left(-\beta_0/\beta_1\right), \infty, \infty\big) = \Gamma_2 = \frac{2-\xi}{2+\xi},$$

which is exactly the same result as given in Theorem 3.3 in [37].

For $\xi \leq \frac{2}{3}$, the maximal $\widetilde{\gamma}(\theta)$-value is obtained either at the intersection points of $A(\theta)$, $B(\theta)$ and $D(\theta)$ or at the endpoints of the θ-interval. By some straightforward calculations, it turns out that $\theta = -\beta_0$, the intersection point of $B(\theta)$ and $D(\theta)$, gives the optimal $\widetilde{\gamma}(\theta)$-value. Therefore,

$$\widetilde{\gamma}_{\max} = \min\big(\Gamma_2, A(-\beta_0), B(-\beta_0), D(-\beta_0)\big) = B(-\beta_0) = \frac{\xi}{2-\xi}$$

and the optimal δ-value is $\delta = -\beta_0\theta = \theta^2$.

Using $\delta = \theta^2$ in (3.28), we arrive at the formulation of (3.7), showing that this formulation *cannot be improved* by changing the upper bound in the summation term from $n-1$ to $n-2$ ($n \geq 2$), as was done in [37].

Chapter 4

Implicit-explicit methods

In Chapter 1 and Chapter 2 we have used the fully implicit code VODE [8] to integrate the phytoplankton model. However, when used as a 'black box' solver, VODE may easily produce negative solution components. This motivated us to study positivity for linear multistep methods. By accuracy considerations combined with modest memory demands, we restrict ourselves to second-order two-step methods. A positivity analysis for these methods has been given in [37] and an improvement for explicit schemes is shown in Chapter 3.

In a PDE context, positivity of an integration method also requires a positive spatial discretization (see the assumption (3.3)). Central discretization of second order of diffusion terms is positive. For advection terms, the third-order upwind-biased scheme, which has been used in Chapters 1 and 2, is not positive, unless a limiting technique is used. However, the use of a limiter will add extra nonlinearity to the system to be solved. This is a serious drawback when integrating with a fully implicit integrator, like VODE. Therefore, we study in this chapter a class of implicit-explicit (IMEX) methods, in which the advection and (non-stiff) growth term are treated explicitly, whereas the stiff diffusion term is treated implicitly. The IMEX methods considered are based on a family of second-order two-step methods of linear multistep type.

We observe that the implicitly treated diffusion term gives rise to *real* eigenvalues. This property has been used in deriving a condition on the parameters in the family of two-step IMEX methods such that the resulting stability region is the same as for the fully explicit counterpart. It turns out that the IMEX-BDF2 method, resulting from the extrapolated BDF2 method, possesses good stability properties when the advection term is discretized by upwind schemes. Combined with the optimal positivity (Chapter 3) the extrapolated BDF2 method is superior within the class of explicit two-step methods. This is illustrated by two different advection test examples.

Finally, the resulting IMEX-BDF2 method shows advantages over VODE when both are applied to the three-dimensional phytoplankton problem studied in Chapter 1.

4.1 Introduction

In many applications, positivity of numerical solutions, such as biological species densities and chemical concentrations, is of utmost importance. A numerical method lacking the positivity property can give rise to small oscillations, i.e., “wiggles” superimposed on a smooth solution. When solving a space-time dependent problem, for instance an advection-diffusion-reaction problem, such oscillations can originate from spatial discretization as well as from time integration.

For the spatial discretization, the second-order central scheme can be used to discretize diffusion terms, while only the first-order upwind scheme guarantees the absence of “wiggles” coming from discretization of advection terms. However, this choice has the disadvantage of low accuracy and the introduction of a large amount of artificial diffusion. A possible remedy to avoid unwanted oscillations and maintain high accuracy of the solution is to use the third-order upwind-biased discretization scheme in combination with limiters. However, this technique introduces additional nonlinearity in the scheme, which is a serious drawback when implemented in a fully implicit time integration method.

For the time integration, it is known that only first-order methods, e.g. the Backward Euler method, can possess unconditional positivity, while higher order methods lead to a restriction on the step size. However, the accuracy of first-order methods is usually too low in practical applications. This disadvantage can only be overcome by reducing the time step which leads to an expensive time integration process. On the other hand, a high order scheme is more complicated to implement. Therefore, as a compromise, we restrict ourselves to a second-order linear two-step scheme. The second-order in time is usually sufficiently accurate in a PDE context, whereas the overhead in implementing a two-step method is still rather modest. Positivity analysis for two-step methods has been given in [37] and an improvement for explicit schemes is shown in Chapter 3.

Unlike implicit schemes which are expensive because implicit relations have to be solved in each integration step, explicit schemes are cheap since the solution is found by explicit substitutions. However, explicit time integration methods are only suitable for non-stiff terms, such as advection and (non-stiff) reaction terms, because of their limited stability region. For stiff terms, e.g. diffusion terms, which have a large range of eigenvalues, implicit schemes are preferred because of their usually large stability regions. Therefore, a suitable mixture of implicit (IM) and explicit (EX) methods called IMEX methods, might result in a more efficient algorithm. With this approach, the combination of the third-order upwind-biased scheme and limiters applied to advection terms and implemented explicitly no longer complicates the algorithm. In addition, in case of a linear diffusion term, one Newton iteration suffices to solve the relation in the implicit part of the IMEX method.

It is often seen that the stability region of the IMEX method is smaller than that of the explicit counter part (see [21] and [38]). The eigenvalues of the implicitly treated diffusion terms are real. This property can be exploited to construct a class of IMEX methods whose stability region coincides with that of the explicit counter part (see also [21]). Within this class, we will choose a method with a good stability constant.

4.2 Mathematical model and numerical approach

Consider the advection-diffusion-reaction problem

$$\begin{aligned} \frac{\partial \omega}{\partial t}(x,t) + \nabla \cdot (a\,\omega(x,t)) &= \nabla \cdot (D \nabla \omega(x,t)) + g(\omega(x,t)),\ x \in \Omega, \\ a\,\omega(x,t) - D\nabla\omega(x,t) &= 0,\ x \in \partial\Omega, \\ \omega(x,0) &= \omega_0(x) \geq 0,\ \forall x \in \Omega, \end{aligned} \tag{4.1}$$

where $t \geq 0$, $\Omega \subset \mathbb{R}^n$, the velocity field a is a vector of dimension $n \geq 1$, D is a diagonal matrix of positive diffusion coefficients $D_i, i = 1, \cdots, n$, and $g(\omega)$ denotes the reaction term. Since the application that we have in mind is a phytoplankton model (see Section 4.4), we have chosen in problem (4.1) for homogeneous Neumann conditions, i.e., all fluxes vanish at the boundary $\partial\Omega$ of the domain Ω.

In order to find the numerical solution of problem (4.1), we follow the popular Method of Lines (MOL) approach. By this we mean that the spatial discretization and time integration are considered separately. We will combine various spatial discretization schemes for different terms, i.e., advection, diffusion and reaction terms. The resulting ordinary differential equations (ODEs), which are still continuous in time, will be integrated numerically by a suitable ODE method.

4.2.1 Spatial discretization

For simplicity of notation, this section considers the spatial discretization of advection and diffusion terms on a *one*-dimensional equidistant grid. An extension to the *multi*-dimensional form is straightforward. The approximation of the reaction term depends on the particular form of $g(\omega)$ (see Section 4.4 on phytoplankton dynamics, where we have a rather complicated g-function)[1].

The discretization scheme used here is based on the so-called finite-volume method, in which we surround each grid point x_i by a cell, at the boundaries of which we approximate the flux $\mathcal{F}(x) = a\omega - D\frac{\partial \omega}{\partial x}$. Then, conservation is achieved if $\frac{\partial \mathcal{F}}{\partial x}(x_i)$ is approximated by $(\mathcal{F}(x_{i+1/2}) - \mathcal{F}(x_{i-1/2}))/\Delta x$, where $x_{i\pm 1/2} := x_i \pm \frac{1}{2}\Delta x$ and Δx denotes the grid size. Here, the velocity a and the diffusion coefficient D are supposed to be independent of ω and given exactly at the grid points $x_{i\pm1/2}$, at which ω as well as its derivative $\frac{\partial \omega}{\partial x}$ need to be approximated.

---

[1] In this section, we omit the presence of t.

For the diffusion term, we use the second-order central scheme. That is,

$$\frac{\partial \omega}{\partial x}(x_{i+\frac{1}{2}}) \approx \frac{w_{i+1} - w_i}{\Delta x},$$

where w_i stands for the approximation of ω at grid point x_i. In spite of its simplicity, this discretization scheme is often successfully used in actual applications.

For the advection term, one can use the first-order upwind scheme. Although it is positive, this scheme is in most cases too inaccurate and produces a large amount of artificial diffusion. On the other hand, common higher-order discretizations may produce oscillating solutions (because of negative off-diagonal elements in the Jacobian matrix, see [38]), which might result in negative solution components. That is of course unphysical in many applications, such as those involving densities and concentrations. In order to have better accuracy than first-order, but still positivity, we can change to a high-order discretization by a technique called limiting (see e.g. [38]). That is,

$$\omega(x_{i+\frac{1}{2}}) \approx \begin{cases} w_i + \psi(\Theta_i)\,(w_{i+1} - w_i), & \text{if } a(x_{i+\frac{1}{2}}) \geq 0, \\ w_{i+1} + \psi(\frac{1}{\Theta_{i+1}})\,(w_i - w_{i+1}), & \text{if } a(x_{i+\frac{1}{2}}) < 0, \end{cases}$$

where

$$\Theta_i = \frac{w_i - w_{i-1}}{w_{i+1} - w_i} \tag{4.2}$$

and the limiter function ψ satisfies the condition

$$0 \leq \psi(\Theta) \leq 1, \quad 0 \leq \frac{1}{\Theta}\psi(\Theta) \leq \mu, \quad \text{for all } \Theta \in \mathbb{R},$$

and contains information on the discretization scheme. Associated to the third-order upwind-biased scheme, the limiter function ψ can be chosen as [38, 44]

$$\psi(\Theta) = \max\big(0, \min(1, \frac{1}{3} + \frac{1}{6}\Theta, \Theta)\big), \tag{4.3}$$

which aims to maintain the high-accuracy property of the third-order scheme in smooth regions (where $\Theta \approx 1$), whereas close to the extremum (where $\Theta \leq 0$), the limiter switches to the first-order upwind flux preventing the solution from oscillations.

Different from the discretization of the diffusion term whose contribution to the spectral radius is proportional to $(\Delta x)^{-2}$ (called stiff), the impact of the advection discretization scheme is smaller because its contribution to the spectral radius is only proportional to $(\Delta x)^{-1}$ (called non-stiff). Moreover, from relation (4.2) we see that the limiter adds nonlinearity to the numerical scheme, which is a drawback when using a fully implicit time integration. Therefore, the advection part will be treated explicitly, whereas the diffusion part is integrated implicitly. This mixture is known as an IMEX method. The computational advantage of an IMEX method is that the difficult, nonlinear advection part is treated explicitly, which is simple, whereas the diffusion part, which needs to be treated implicitly for stability reasons, is also quite cheap since it results in the solution of a linear system with a band structure in the matrix.

4.2.2 Time integration

After discretizing the above spatial operators, we obtain an ODE system which is still continuous in time and needs to be integrated. This system of ODEs will be written in the form

$$\frac{d\boldsymbol{w}(t)}{dt} = \boldsymbol{F}_E(t, \boldsymbol{w}(t)) + \boldsymbol{F}_I(t, \boldsymbol{w}(t)), \tag{4.4}$$

where the vector $\boldsymbol{w}$ contains all discrete solution components, $\boldsymbol{F}_E$ originates from the discretization of the non-stiff terms (advection and reaction terms), and $\boldsymbol{F}_I$ comes from the discretization of the stiff terms (diffusion terms).

In order to integrate (4.4) in time, we use an IMEX two-step method (see Section 4.3), i.e.,

$$\begin{aligned}\mathbf{R}(\boldsymbol{W}_{n+2}) := \boldsymbol{W}_{n+2} \ + \sum_{j=0}^{1} \alpha_j \boldsymbol{W}_{n+j} \ - \\ \Delta t_{n+1}\Big(\sum_{j=0}^{1} \beta_j^* \, \boldsymbol{F}_E(t_{n+j}, \boldsymbol{W}_{n+j}) + \sum_{j=0}^{2} \beta_j \, \boldsymbol{F}_I(t_{n+j}, \boldsymbol{W}_{n+j})\Big) \\ = \mathbf{0},\end{aligned} \tag{4.5}$$

where $\boldsymbol{W}_{n+j}$ indicates the approximate solution of (4.4) at time level t_{n+j} and $\Delta t_{n+1} = t_{n+2} - t_{n+1}$ is the time step. Provided that we start with a positive spatial discretization, our purpose is to find the coefficients α_j, β_j^*, and β_j such that the method (4.5) produces positive solutions and, in addition, has good stability properties.

In order to solve the above implicit relation, which is implicit in $\boldsymbol{F}_I$ only, one can use the modified Newton method. That is,

$$\left[I - \beta_2 \Delta t_{n+1} \frac{\partial \boldsymbol{F}_I}{\partial \boldsymbol{w}}\right] \left[\boldsymbol{W}_{n+2}^{k} - \boldsymbol{W}_{n+2}^{k-1}\right] = -\mathbf{R}(\boldsymbol{W}_{n+2}^{k-1}), \quad k = 1, 2, \cdots, \tag{4.6}$$

where the superscript k denotes the iteration index, I is the identity matrix, and $\partial \boldsymbol{F}_I / \partial \boldsymbol{w}$ is the Jacobian matrix that is occasionally evaluated along the numerical solution.

4.3 Analysis of the IMEX method

As a stepping-stone of the IMEX method presented in this section, a general second-order linear two-step method with variable step sizes is considered first.

4.3.1 General linear two-step method for $\boldsymbol{w}'(t) = \boldsymbol{F}(t, \boldsymbol{w}(t))$

Consider the following linear two-step scheme

$$\boldsymbol{W}_{n+2} + \sum_{j=0}^{1} \alpha_j \boldsymbol{W}_{n+j} = \Delta t_{n+1} \sum_{j=0}^{2} \beta_j \, \boldsymbol{F}(t_{n+j}, \boldsymbol{W}_{n+j}) \tag{4.7a}$$

with the current time step $\Delta t_{n+1} = t_{n+2} - t_{n+1} > 0$. Let $r = \Delta t_{n+1}/\Delta t_n$. By inserting the exact solution values into the scheme (4.7a) and using the Taylor expansion around t_{n+1} the scheme is of second-order accuracy if

$$\begin{aligned} &1 + \alpha_0 + \alpha_1 = 0, \\ &1 - \alpha_0/r - \beta_0 - \beta_1 - \beta_2 = 0, \\ &1 + \alpha_0/r^2 + 2\beta_0/r - 2\beta_2 = 0, \end{aligned}$$

yielding

$$\begin{aligned} &\alpha_0 = r - \xi, && \alpha_1 = -1 - r + \xi, \\ &\beta_0 = -\frac{r+1}{2} + \frac{\xi}{2r} + r\eta, && \beta_1 = \frac{r+1}{2} + \frac{\xi}{2r} - (r+1)\eta, \quad \beta_2 = \eta, \end{aligned} \tag{4.7b}$$

where ξ and η are free parameters. We note that the scheme is zero-stable (stable for the trivial equation $\boldsymbol{w}'(t) = \boldsymbol{0}$, see [38]) if the condition $-1 \leq \alpha_0 < 1$ is satisfied.

In modern practice, ODE problems are usually integrated with variable step sizes. This approach leads to small (large) step sizes in regions of rapid (slow) variation of the solution, resulting in efficient computations in term of CPU time versus accuracy. However, for the sake of the analysis of stability and positivity, a constant step size Δt will be used. Hence, in the sequel of this subsection we will set $r = 1$.

Stability

Stability of linear multistep methods is studied mainly for *linear* ODEs and in particular for the familiar scalar test equation $w' = \lambda w(t)$ with $\lambda \in \mathbb{C}^-$ representing eigenvalues of the semi-discrete system. For this scalar test equation, the scheme (4.7a) leads to the recursion

$$(1 - z\beta_2)W_{n+2} + (\alpha_1 - z\beta_1)W_{n+1} + (\alpha_0 - z\beta_0)W_n = 0, \tag{4.8}$$

where $z = \lambda \Delta t$. This recursion has the characteristic polynomial

$$(1 - z\beta_2)\zeta^2 + (\alpha_1 - z\beta_1)\zeta + (\alpha_0 - z\beta_0). \tag{4.9}$$

The *stability region* $\mathcal{S} \subset \mathbb{C}$ of the linear multistep method is defined as the set of all z for which the sequence $\{W_n\}$ in (4.8) is bounded. It is well known that this requirement is equivalent to the *root condition* of the characteristic polynomial (4.9), that is, the set of all z for which the roots $\zeta_i, i = 1, 2$, of (4.9) satisfy

$$|\zeta_i| \leq 1 \text{ and } |\zeta_i| < 1 \text{ if } \zeta_i \text{ is not simple.} \tag{4.10}$$

To determine $\mathcal{S}$, we use that on the boundary $\partial\mathcal{S}$, at least one of the roots of (4.9) has modulus 1. Therefore, on $\partial\mathcal{S}$ we have

$$(1 - z\beta_2)e^{2i\phi} + (\alpha_1 - z\beta_1)e^{i\phi} + (\alpha_0 - z\beta_0) = 0,$$

i.e.,

$$z = \frac{e^{2i\phi} + \alpha_1 e^{i\phi} + \alpha_0}{\beta_2 e^{2i\phi} + \beta_1 e^{i\phi} + \beta_0}, \quad \phi \in [0, 2\pi].$$

The method is called *A-stable* if

$$\mathcal{S} \supset \{z \in \mathbb{C} : \text{Re } z \leq 0\}, \tag{4.11a}$$

i.e., λ is supposed to lie in the left half-plane, and is called *A(α)-stable* if

$$\mathcal{S} \supset \mathcal{W}_\alpha = \{z \in \mathbb{C} : |\arg(-z)| \leq \alpha, z \neq 0\}, \tag{4.11b}$$

i.e., λ is supposed to lie in the left half-plane within an infinite wedge with angle α with respect to the negative real axis. The scheme (4.7) with $r = 1$ is A-stable iff $\beta_2 \geq \frac{1}{2}$ (see [37]).

4.3.2 The IMEX two-step method

Applying the scheme (4.7a) to the ODE system (4.4)

$$\frac{d\boldsymbol{w}}{dt} = \boldsymbol{F}(t, \boldsymbol{w}) = \boldsymbol{F}_E(t, \boldsymbol{w}) + \boldsymbol{F}_I(t, \boldsymbol{w})$$

and extrapolating the non-stiff leading term $\boldsymbol{F}_E(t_{n+2}, \boldsymbol{W}_{n+2})$ as

$$\boldsymbol{F}_E(t_{n+2}, \boldsymbol{W}_{n+2}) \simeq (r + 1)\, \boldsymbol{F}_E(t_{n+1}, \boldsymbol{W}_{n+1}) - r\, \boldsymbol{F}_E(t_n, \boldsymbol{W}_n)$$

we obtain the IMEX-scheme (4.5), with the coefficients $\alpha_0, \alpha_1, \beta_0, \beta_1$ and β_2 as given in (4.7b) and

$$\begin{aligned} \beta_0^* &= \beta_0 - r\beta_2 &&= -\frac{r+1}{2} + \frac{\xi}{2r}, \\ \beta_1^* &= \beta_1 + (r+1)\beta_2 &&= \frac{r+1}{2} + \frac{\xi}{2r}, \end{aligned} \tag{4.12}$$

which are the same as setting η in (4.7b) to zero. The scheme is of second-order accuracy, because it originates from the second-order two-step scheme and second-order extrapolation (see [38], p. 387).

Stability

In this subsection, we study the stability of the IMEX method (4.5) applied to the scalar test equation $w' = \lambda w + \kappa w$, where λ and κ denote the eigenvalues of the non-stiff part and the stiff part, respectively. Again, we will perform the analysis for the case of a constant step size Δt. Stability of (4.5) is determined by the location of the roots of the characteristic equation

$$\zeta^2 + \sum_{j=0}^{1} \alpha_j \zeta^j - \lambda \Delta t \sum_{j=0}^{1} \beta_j^* \zeta^j - \kappa \Delta t \sum_{j=0}^{2} \beta_j \zeta^j = 0. \tag{4.13}$$

That is, a root ζ has to obey the requirement $|\zeta| \leq 1$, with strict inequality for multiple roots. Dividing this equation by ζ^2, replacing $\lambda\Delta t$ and $\kappa\Delta t$ by λ and κ, respectively, and substituting

$$z = 1/\zeta, \quad A(z) = 1+\alpha_1 z+\alpha_0 z^2, \quad B(z) = \beta_1^* z+\beta_0^* z^2, \quad C(z) = \beta_2+\beta_1 z+\beta_0 z^2,$$

the characteristic equation (4.13) is transformed into

$$A(z) - \lambda B(z) - \kappa C(z) = 0, \tag{4.14}$$

where $|z| \geq 1$ is the requirement for stability, with strict inequality for multiple roots. Thus, the necessary condition for stability is

$$\kappa \neq \varphi_\lambda(z) := \Big[A(z) - \lambda B(z)\Big]/C(z), \text{ for all } |z| < 1. \tag{4.15}$$

This condition implies that φ_λ must map a unit circle into the *complement* of the region containing κ. Note that, if the derivation of this criterion was used for the equation (4.13) instead of (4.14), the exterior of the unit circle (i.e., $|\zeta| > 1$) would had been considered. This is of course not a good choice, emphasizing the beauty of the transformation $z = 1/\zeta$ that has been used. Apart from the possibility of multiple roots, criterion (4.15) is also a sufficient condition for stability.

Under the criterion (4.15), Frank et al. [21] determined what condition on the location of λ ensures stability of the method for all κ in the left half-plane. However, this usually leads to a less favorable stability region, i.e., smaller than the stability region denoted by $\mathcal{S}$ of the explicit version (obtained by setting $\boldsymbol{F}_I \equiv 0$). This is due to the demand of A-stability with respect to κ. In our application however, κ lies on the negative real axis, i.e., $\kappa \in \mathcal{W}_0$, because it originates from central discretization of the diffusion term (Section 4.2). Therefore, the A-stability requirement is too demanding for this application. The question is whether or not the IMEX method is stable for arbitrary $\lambda \in \mathcal{S}$ if the knowledge of $\kappa \in \mathcal{W}_0$ is taken into account.

Because $\kappa \in \mathcal{W}_0$, $\varphi_\lambda(z)$ therefore must lie in the complement of the wedge $\mathcal{W}_0$, i.e., $|\arg(\varphi_\lambda(z))| < \pi$, for all $|z| < 1, \lambda \in \mathcal{S}$. We know that

$$\begin{aligned} |\arg(\varphi_\lambda(z))| &= |\arg(A(z) - \lambda B(z)) - \arg(C(z))| \\ &\leq |\arg(A(z) - \lambda B(z))| + |\arg(C(z))|. \end{aligned}$$

Thus, if we have

$$|\arg(A(z) - \lambda B(z))| \leq \frac{1}{2}\pi + \nu \quad \text{and} \quad |\arg(C(z))| < \frac{1}{2}\pi - \nu, \tag{4.16}$$

for all $|z| < 1, \lambda \in \mathcal{S}$, the IMEX method is stable for arbitrary $\lambda \in \mathcal{S}$ (see [21] for a more general result).

To determine ν, we note that a polynomial $c + bz + az^2$ with roots z_1 and z_2 can be written as $c + bz + az^2 = c(1 - \frac{z}{z_1})(1 - \frac{z}{z_2})$. Thus, $A(z) - \lambda B(z) = (1 - \zeta_1 z)(1 - \zeta_2 z)$, where ζ_1 and ζ_2 are the characteristic roots of the explicit method, i.e., the roots of (4.13) with $\kappa = 0$. As mentioned in Section 4.3.1, for any choice of λ in $\mathcal{S}$, these roots satisfy $|\zeta_2| \leq |\zeta_1| \leq 1$ (see (4.10)). Thus, $1 - \zeta_1 z$ and $1 - \zeta_2 z$ map the unit circle $((0,0),1)$ into two circles $((1,0),|\zeta_1|)$ and $((1,0),|\zeta_2|)$ in the *right* half-plane. By geometrical considerations we have

$$\begin{aligned} |\arg(A(z) - \lambda B(z))| &\leq |\arg(1 - \zeta_1 z)| + |\arg(1 - \zeta_2 z)| \\ &\leq \arcsin|\zeta_1| + \arcsin|\zeta_2| \\ &\leq \tfrac{\pi}{2} + \arcsin(\max\nolimits_{\lambda \in \mathcal{S}} |\zeta_2|), \qquad \forall |z| \leq 1, \lambda \in \mathcal{S}. \end{aligned}$$

The second inequality changes to an equality when $|z| = 1$. The last inequality changes to an equality when $\lambda \in \partial\mathcal{S}$ (see [21]). Therefore,

$$\nu = \arcsin(\max_{\lambda \in \partial\mathcal{S}} |\zeta_2|). \tag{4.17}$$

Remark 4.1. Being solutions of (4.13) with $\kappa = 0$ and λ replacing $\lambda\Delta t$, ζ_1 and ζ_2 satisfy

$$\begin{aligned} \zeta_1 + \zeta_2 &= -\alpha_1 + \lambda\beta_1^*, \\ \zeta_1\,\zeta_2 &= \alpha_0 - \lambda\beta_0^*. \end{aligned}$$

Multiplying these two equations by β_0^* and β_1^*, respectively, we obtain

$$\beta_0^*\,(\zeta_1 + \zeta_2) + \beta_1^*\,\zeta_1\,\zeta_2 = -\alpha_1\,\beta_0^* + \alpha_0\,\beta_1^*.$$

Thus, for $\lambda \in \partial\mathcal{S}$, we get $\zeta_1 = e^{i\phi}$ and

$$\zeta_2 = \frac{\alpha_0\beta_1^* - \alpha_1\beta_0^* - \beta_0^* e^{i\phi}}{\beta_0^* + \beta_1^* e^{i\phi}}, \quad \phi \in [-\pi, \pi]. \tag{4.18}$$

Remark 4.2. We know that $C(z) = \beta_2 + \beta_1 z + \beta_0 z^2$. Its argument is easily found if $\beta_1 = 0$. Because, in that case $C(z) = \beta_2 + \beta_0 z^2 = \beta_2(1 + \frac{\beta_0}{\beta_2} z^2)$. Hence, $\arg(C(z)) = \arg\beta_2 + \arg\left(1 + \frac{\beta_0}{\beta_2}z^2\right) = \arg\left(1 + \frac{\beta_0}{\beta_2}z^2\right)$ if $\beta_2 \geq 0$. The assumption $\beta_2 \geq 0$ is usually imposed for important properties of implicit methods, such as monotonicity, positivity and boundedness (see Section 3.3). We note that $1 + \frac{\beta_0}{\beta_2}z^2$ maps the unit circle $((0,0),1)$ into the circle $\left((1,0), \left|\frac{\beta_0}{\beta_2}\right|\right)$, which is completely in the right half-plane if $\left|\frac{\beta_0}{\beta_2}\right| \leq 1$. If this condition is satisfied, we obtain

$$|\arg(C(z))| = \left|\arg\left(1 + \frac{\beta_0}{\beta_2} z^2\right)\right| \leq \arcsin\left|\frac{\beta_0}{\beta_2}\right|.$$

Therefore, we consider the IMEX method with $\beta_1 = 0$ leading to $\eta = \frac{2+\xi}{4}$ (cf. (4.7b)), i.e,

$$\begin{aligned} \boldsymbol{W}_{n+2} + (\xi - 2)\boldsymbol{W}_{n+1} + (1-\xi)\boldsymbol{W}_n = & \\ (\tfrac{\xi}{2}+1)\Delta t\,\boldsymbol{F}_E(t_{n+1}, \boldsymbol{W}_{n+1}) + (\tfrac{\xi}{2}-1)\Delta t\,\boldsymbol{F}_E(t_n, \boldsymbol{W}_n) + & \\ \tfrac{2+\xi}{4}\,\Delta t\,\boldsymbol{F}_I(t_{n+2}, \boldsymbol{W}_{n+2}) + \tfrac{3\xi-2}{4}\,\Delta t\,\boldsymbol{F}_I(t_n, \boldsymbol{W}_n). & \end{aligned} \tag{4.19}$$

With this η-value, the implicit version (obtained by setting $\boldsymbol{F}_E \equiv 0$) is A-stable, because $\eta > \frac{1}{2}$. Moreover, $\left|\frac{\beta_0}{\beta_2}\right| = \frac{|3\xi-2|}{2+\xi} \leq 1$. Hence,

$$|\arg(C(z))| \leq \arcsin\frac{|3\xi - 2|}{2+\xi}. \tag{4.20}$$

Formula (4.18) gives

$$|\zeta_2| = \left|\frac{(2-\xi)e^{i\phi} - (2\xi^2 - 3\xi + 2)}{(2+\xi)e^{i\phi} - (2-\xi)}\right|,$$

from which we derive

$$|\zeta_2|^2 = f(X) := \frac{\mathcal{A} - \mathcal{B}X}{\mathcal{C} - \mathcal{D}X},$$

where $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D} \geq 0$ and

$$\begin{aligned} \mathcal{A} &= [\,(2-\xi)^2 + (2\xi^2 - 3\xi + 2)^2\,]/\,2, \\ \mathcal{B} &= (2-\xi)(2\xi^2 - 3\xi + 2), \\ \mathcal{C} &= \xi^2 + 4, \\ \mathcal{D} &= 4 - \xi^2, \\ X &= \cos\phi. \end{aligned}$$

Since $f'(X) = \frac{-2\xi^4(\xi-2)^2}{(\mathcal{C}-\mathcal{D}X)^2} \leq 0$, we have

$$\max_{-1 \leq X \leq 1} f(X) = f(-1).$$

Therefore,

$$|\zeta_2|^2 \leq \frac{\mathcal{A} + \mathcal{B}}{\mathcal{C} + \mathcal{D}}.$$

Using (4.17) we then have

$$\nu = \arcsin\sqrt{\frac{\mathcal{A}+\mathcal{B}}{\mathcal{C}+\mathcal{D}}} = \arcsin\frac{(\xi-1)^2+1}{2}. \tag{4.21}$$

Combining (4.16) and (4.20)–(4.21) we formulate the following theorem.

Theorem 4.1. *The IMEX method (4.19) is stable for all $\lambda \in \mathcal{S}$ and $\kappa \in \mathcal{W}_0$ if ξ satisfies*

$$\arcsin\frac{|3\xi-2|}{2+\xi} < \left(\frac{\pi}{2} - \arcsin\frac{(\xi-1)^2+1}{2}\right). \tag{4.22}$$

The corresponding ξ-interval is shown in Fig. 4.1.

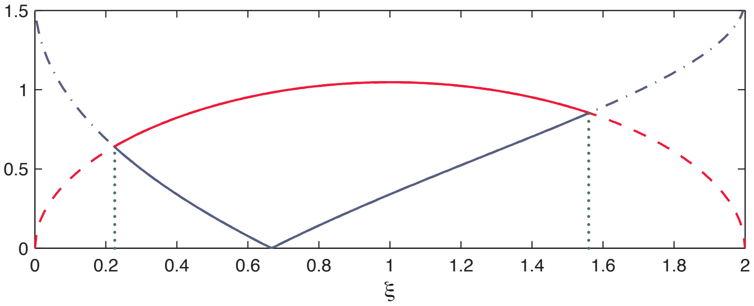


FIGURE 4.1: Dash-dotted line: $\arcsin\frac{|3\xi-2|}{2+\xi}$. Dashed line: $\frac{\pi}{2} - \arcsin\frac{(\xi-1)^2+1}{2}$. Solid lines show the interval of ξ-value in which the IMEX method (4.19) is stable for all $\lambda \in \mathcal{S}$. The boundaries of this interval are indicated by dotted lines.

Example 4.1. Consider the IMEX-BDF2 method ($\xi = \frac{2}{3}, \eta = \frac{2}{3}$)

$$\begin{aligned} \boldsymbol{W}_{n+2} - \tfrac{4}{3}\boldsymbol{W}_{n+1} + \tfrac{1}{3}\boldsymbol{W}_n = \tfrac{4}{3}\Delta t\boldsymbol{F}_E(t_{n+1}, \boldsymbol{W}_{n+1}) - \tfrac{2}{3}\Delta t\boldsymbol{F}_E(t_n, \boldsymbol{W}_n) + \\ \tfrac{2}{3}\Delta t\boldsymbol{F}_I(t_{n+2}, \boldsymbol{W}_{n+2}). \end{aligned} \tag{4.23}$$

From Fig. 4.1 we see that the method (4.23) satisfies condition (4.22) and hence is stable for arbitrary $\lambda \in \mathcal{S}$.

Example 4.2. The inequality (4.22) applied to the IMEX-ADAMS method ($\xi = 1, \eta = \frac{3}{4}$), i.e.,

$$\begin{aligned} \boldsymbol{W}_{n+2} - \boldsymbol{W}_{n+1} = \; & \tfrac{3}{2}\Delta t\boldsymbol{F}_E(t_{n+1}, \boldsymbol{W}_{n+1}) - \tfrac{1}{2}\Delta t\boldsymbol{F}_E(t_n, \boldsymbol{W}_n) + \\ & \tfrac{3}{4}\Delta t\boldsymbol{F}_I(t_{n+2}, \boldsymbol{W}_{n+2}) + \tfrac{1}{4}\Delta t\boldsymbol{F}_I(t_n, \boldsymbol{W}_n), \end{aligned} \tag{4.24}$$

gives stability of the method (4.24) for all $\lambda \in \mathcal{S}$.

Example 4.3. The IMEX method (4.19) with $(\xi = \frac{18}{13}, \eta = \frac{11}{13})$, i.e.,

$$\begin{aligned}\boldsymbol{W}_{n+2} - \tfrac{8}{13}\boldsymbol{W}_{n+1} - \tfrac{5}{13}\boldsymbol{W}_n = \\ \tfrac{22}{13}\Delta t\boldsymbol{F}_E(t_{n+1}, \boldsymbol{W}_{n+1}) - \tfrac{4}{13}\Delta t\boldsymbol{F}_E(t_n, \boldsymbol{W}_n) + \\ \tfrac{11}{13}\Delta t\boldsymbol{F}_I(t_{n+2}, \boldsymbol{W}_{n+2}) + \tfrac{7}{13}\Delta t\boldsymbol{F}_I(t_n, \boldsymbol{W}_n),\end{aligned} \tag{4.25}$$

satisfies (4.22), see also Fig. 4.1. Hence, the method is stable for all $\lambda \in \mathcal{S}$.

This IMEX method has been constructed with a special reason. The ξ-value $\frac{18}{13}$ has been chosen to give its explicit version (obtained by setting $\boldsymbol{F}_I \equiv 0$) the same CFL number[2] as the explicit version of the IMEX-BDF2 method (4.23). Here, this CFL number is based on a linear advection model, discretized by the third-order upwind-biased scheme and turns out to be 0.46 (see Table 4.1 and [38, p. 389]). Hence, with respect to stability, the step sizes of both methods have to satisfy the same condition.

As we have seen, the stability of the IMEX method with $\boldsymbol{F}_I$ possessing a negative spectrum is largely determined by the stability region $\mathcal{S}$ of the explicit version of the method. Therefore we plot in Fig. 4.2 the stability regions $\mathcal{S}$ of the above example methods.

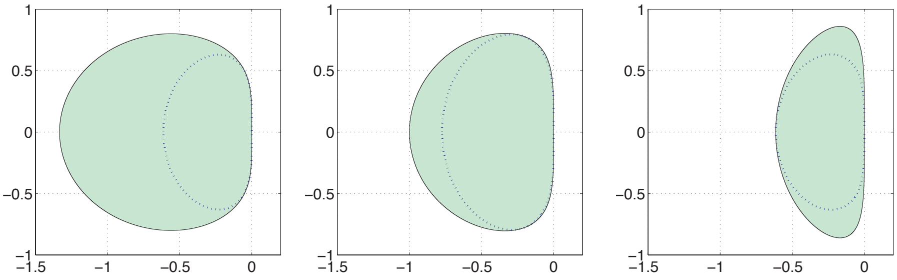


FIGURE 4.2: Filled regions: Stability regions $\mathcal{S}$ of explicit methods with $\xi = \frac{2}{3}$ (left), $\xi = 1$ (middle) and $\xi = \frac{18}{13}$ (right). Dotted lines: Scaled eigenvalues of the third-order upwind-biased discretization scheme.

4.4 Numerical illustrations

We will start this section by considering a simple test model, i.e., a one-dimensional, scalar, constant-coefficient advection model, integrated with constant step sizes. The ultimate test example is a realistic application from marine ecology, i.e., a three-dimensional model of phytoplankton dynamics, which is presented in terms of a nonlinear integro-PDEs of advection-diffusion-reaction

---

[2] That is, the maximal value of $\frac{|a|\Delta t}{\Delta x}$, where a denotes the advection coefficient.

type. However, before treating the full phytoplankton model we will consider - as an intermediate study - a simplified version of this test model: (i) only the advection terms with a space-dependent velocity field are taken into account, and (ii) we restrict ourselves to two space dimensions. For the one- and two-dimensional *advection* problems, the explicit versions of the IMEX methods will be used. The absence of stiff diffusion terms justifies this choice. The major aim of these tests is to select the best ξ-value from the ones that we have discussed so far. To be more specific, we will compare the methods with $\xi = \frac{2}{3}$, 1, and $\frac{18}{13}$, i.e., the methods (4.23), (4.24), and (4.25), all with $\boldsymbol{F}_I \equiv 0$. As we will see, the method with $\xi = \frac{2}{3}$ shows the best performance. Therefore in the final section discussing the full plankton model, only this ξ-value will be used. Furthermore, the corresponding IMEX method will be compared with VODE, a widely-used time integrator based on the fully implicit BDF-methods. Here we remark that VODE has also been used in the preceding chapters to solve the phytoplankton model.

4.4.1 One-dimensional advection test

Consider the one-dimensional advection problem

$$\frac{\partial \omega}{\partial t} + \frac{\partial \omega}{\partial x} = 0, \quad t > 0, \quad 0 \leq x \leq 1, \tag{4.26}$$

with periodic boundary condition and initial value $\omega(x,0) = (\sin(\pi x))^{100}$. When positivity of the numerical solution is a prerequisite, it is essential to use a positive spatial discretization scheme. For instance, the non-limited third-order upwind-biased scheme gives rise to a negative solution (Fig. 4.3a), whereas we obtain a positive solution if the limiter (4.3) is used (Fig. 4.3b). Therefore, in this test model, as well as in all subsequent applications, we will only consider the limited third-order upwind-biased scheme.

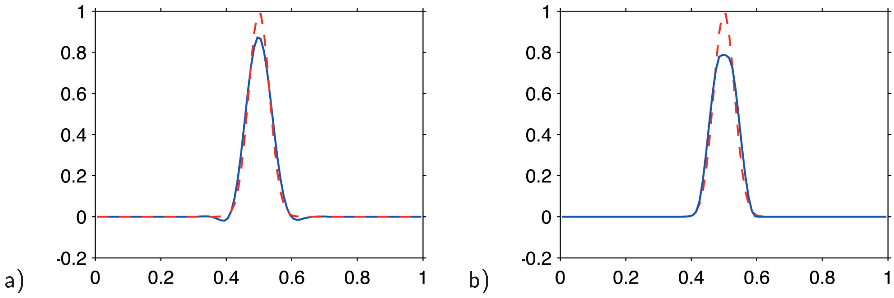


FIGURE 4.3: ODE solutions (solid lines), at $t = 1$, obtained by the third-order upwind-biased scheme without (left) and with (right) limiter. The dashed lines indicate the exact PDE solution. $\Delta x = 10^{-2}$.

Test on stability

At points with large spatial variation in the solution, the limiter switches to the first-order upwind scheme, to prevent a negative solution; in smooth regions, the third-order upwind-biased scheme is used. Let ν_1 and ν_3 denote the CFL numbers, i.e., maximal value of $\Delta t/\Delta x$, of the first-order upwind scheme and the third-order upwind-biased scheme, respectively. The ν_1- and ν_3-values of the explicit versions of the methods (4.23), (4.24), and (4.25) are listed in Table 4.1.

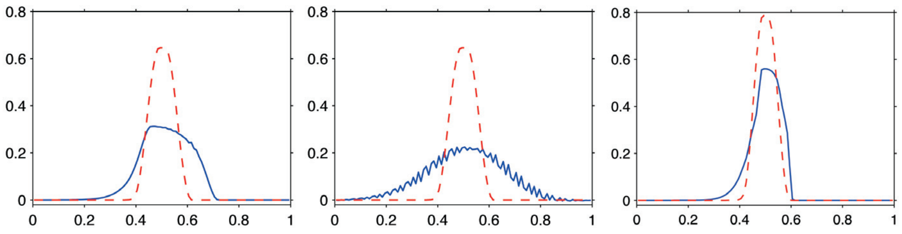
TABLE 4.1: CFL numbers of the $\xi = \frac{18}{13}$, $\xi = 1$, and $\xi = \frac{2}{3}$ methods.

	$\xi = \frac{18}{13}$ (Special method)	$\xi = 1$ (ADAMS method)	$\xi = \frac{2}{3}$ (EBDF2 method)
ν_1	0.3076	0.5000	0.6666
ν_3	0.4615	0.5801	0.4617

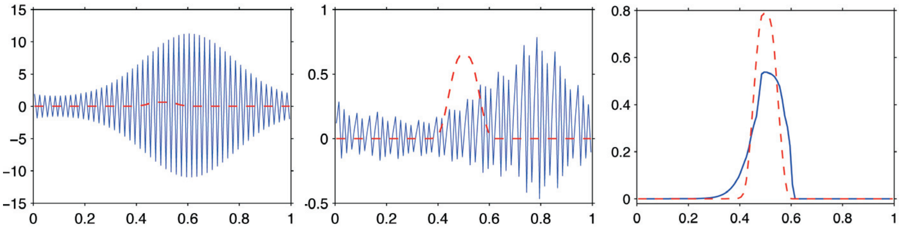
Note, however, that CFL numbers are the result of a stability analysis applied to the *linear* advection model problem, whereas the use of a limiter gives rise to a *nonlinear* model. Nevertheless, it is an interesting experimental observation that stability is critically determined by the value of $\min(\nu_1, \nu_3)$, as we will demonstrate now.

Let us first consider the $\xi = \frac{18}{13}$ and $\xi = 1$ methods. Taking $\Delta t/\Delta x$ slightly smaller than the critical value $\min(\nu_1, \nu_3) = \nu_1$, both methods are stable (Figs. 4.4a,b). However, for $\Delta t/\Delta x$ slightly larger than $\min(\nu_1, \nu_3)$ these two methods give unstable solutions (Figs. 4.4d,e). This behaviour can be explained, somewhat heuristically, as follows: due to a lack of positivity and/or the development of local instabilities, it might happen that some oscillations show up in the solution (see Figs. 4.4a,b). In that case, the limiter will switch to the first-order upwind scheme with, as an effect, smoothing of the solution. In fact, the more oscillatory the solution is, the more often this scheme is used. Consequently, stability of the methods is mainly determined by ν_1, which is smaller than ν_3 for the $\xi = \frac{18}{13}$ and $\xi = 1$ methods (see Table 4.1). Therefore, when $\Delta t/\Delta x$ exceeds $\nu_1 = \min(\nu_1, \nu_3)$ the CFL condition is frequently violated and these methods show an unstable behaviour.

Next, we consider the EBDF2 method, i.e., $\xi = \frac{2}{3}$. As shown in Fig. 4.4c, this method behaves stable for $\Delta t/\Delta x \leq \min(\nu_1, \nu_3) = \nu_3$. When $\Delta t/\Delta x$ exceeds the ν_3-value some oscillations occur due to a lack of positivity and/or the development of local instabilities (see Fig. 4.5a). Then, the limiter reacts with a switch to the first-order upwind discretization and, again, the stability is mainly determined by ν_1. Since $\nu_1 > \nu_3$ for the EBDF2 method such a switch to first order will favour the stability of the method. Different from the two above methods, this method is still stable for $\Delta t/\Delta x > \min(\nu_1, \nu_3)$ (see Figs. 4.4f and 4.5a). However, if $\Delta t/\Delta x$ exceeds $\nu_1 = \max(\nu_1, \nu_3)$, instabilities quickly develop (see Fig. 4.5b).

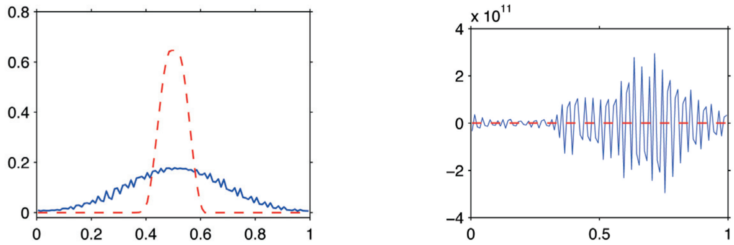


a) $\xi = \frac{18}{13}$, $\frac{\Delta t}{\Delta x} = 0.30$. b) $\xi = 1, \frac{\Delta t}{\Delta x} = 0.49$. c) $\xi = \frac{2}{3}$, $\frac{\Delta t}{\Delta x} = 0.46$.



d) $\xi = \frac{18}{13}$, $\frac{\Delta t}{\Delta x} = 0.31$. e) $\xi = 1, \frac{\Delta t}{\Delta x} = 0.51$. f) $\xi = \frac{2}{3}$, $\frac{\Delta t}{\Delta x} = 0.48$.

FIGURE 4.4: Test on stability of the $\xi = \frac{18}{13}$, $\xi = 1$, and $\xi = \frac{2}{3}$ methods. The time interval is sufficiently large, i.e., $t = 4$, to allow development of instability. The ODE solution is indicated by dashed lines. $\Delta x = 10^{-2}$.



a) $\frac{\Delta t}{\Delta x} = 0.65$. b) $\frac{\Delta t}{\Delta x} = 0.67$.

FIGURE 4.5: The EBDF2 method with $\Delta x = 10^{-2}$ and $t = 4$.

In summary, the experimentally observed critical stability condition is largely determined by the $\min(\nu_1, \nu_3)$-value. Moreover, for methods with $\nu_1 < \nu_3$, like the $\xi = \frac{18}{13}$ and $\xi = 1$ methods, the use of a limiter is a disadvantage with respect to stability. For methods having $\nu_1 > \nu_3$, such as the EBDF2 method, the use of a limiter allows a larger time step. An overview of the ν_1- and ν_3-values of the class of explicit second-order two-step methods, parametrized by ξ, is shown in Fig. 4.6. This figure suggests to choose $\xi \approx 0.9$ since that ξ-value yields $\nu_1 = \nu_3 \approx 0.57$ resulting in the largest possible $\min(\nu_1, \nu_3)$-value. However, we refrain from considering this ξ-value since it yields a positivity constant $\gamma_{\max} \approx 0.4$ (see Fig. 3.1). Because the positivity constant is usually more restrictive than the CFL-number, we prefer a method with an optimal

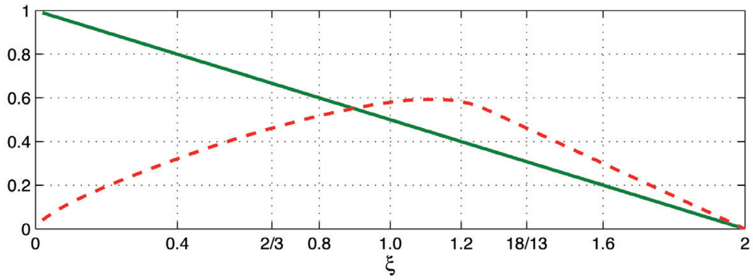


FIGURE 4.6: The CFL numbers ν_1 (solid) and ν_3 (dashed) of explicit second-order two-step methods.

$\gamma_{\max}$. This criterion is in favour of the EBDF2 method.

Test on positivity

In Figs. 4.4a,b $\Delta t/\Delta x$ satisfied the stability requirement for both individual schemes, i.e., the first-order upwind scheme and the third-order upwind-biased scheme. Nevertheless, the solutions show oscillations which are probably due to neglecting the positivity requirement. We recall from Section 3.2 the positivity condition

$$\Delta t \leq \gamma_{\max}\, \Delta t_{FE}. \tag{4.27a}$$

Using limiter (4.3), we obtain (see [38, p. 221])

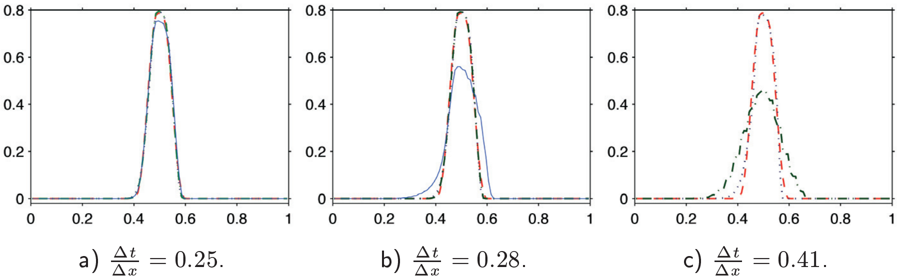
$$\Delta t_{FE} = \frac{1}{2}\Delta x. \tag{4.27b}$$

The values of $\gamma_{\max}$ for the three above methods are listed in Table 4.2 (see also (3.10)). If we take Δt such that the positivity condition (4.27) is satisfied, the solutions do not oscillate and closely resemble the ODE solution as shown in Fig. 4.3b.

TABLE 4.2: Positivity constants $\gamma_{\max}$ of the $\xi = \frac{18}{13}$, $\xi = 1$, and $\xi = \frac{2}{3}$ methods.

	$\xi = \frac{18}{13}$ (Special method)	$\xi = 1$ (ADAMS method)	$\xi = \frac{2}{3}$ (EBDF2 method)
$\gamma_{\max}$	$\frac{2}{11}$	$\frac{1}{3}$	$\frac{1}{2}$

We remark that the value of Δt_{FE} in (4.27b) was found by the sufficient condition for positivity of the forward Euler method. In practice, however, this turns out to be too restrictive. In theory, the $\xi = \frac{18}{13}$, $\xi = 1$ and $\xi = \frac{2}{3}$ methods are positive for $\frac{\Delta t}{\Delta x} = \frac{1}{11}$, $\frac{1}{6}$, and $\frac{1}{4}$, respectively. Nevertheless, these three



a) $\frac{\Delta t}{\Delta x} = 0.25$. b) $\frac{\Delta t}{\Delta x} = 0.28$. c) $\frac{\Delta t}{\Delta x} = 0.41$.

FIGURE 4.7: Test on positivity of the $\xi = \frac{18}{13}$ (solid), $\xi = 1$ (dash-dotted), and $\xi = \frac{2}{3}$ (dotted) methods. Solutions are plotted at $t = 1$. The ODE solution is indicated by dashed lines. $\Delta x = 10^{-2}$.

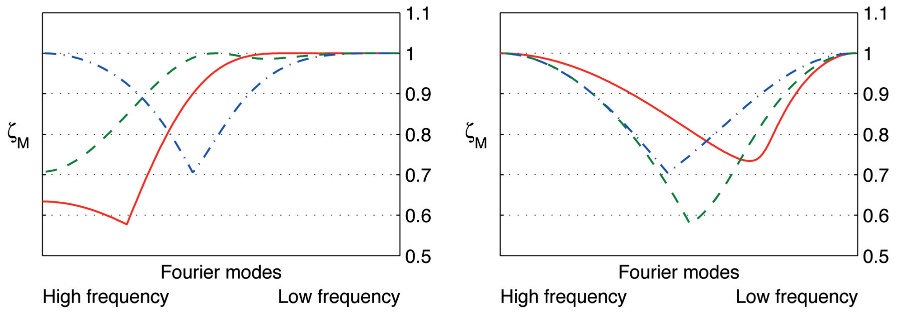
methods still give positive and non-oscillatory solutions for larger values of $\frac{\Delta t}{\Delta x}$ (see Fig. 4.7). In the first panel, the three solutions are positive, although the $\xi = \frac{18}{13}$ method shows a less accurate solution. If we increase $\frac{\Delta t}{\Delta x}$, this method gives rise to an oscillatory solution (second panel). The two other methods still give positive and accurate solutions. If we further increase the $\frac{\Delta t}{\Delta x}$-value, the ADAMS solution starts to oscillate (see Fig. 4.7c). The first-order upwind scheme is then often used, resulting in extra damping of the solution. The EBDF2 method still behaves satisfactory. The $\xi = \frac{18}{13}$ method is not included in this panel, simply because it is unstable. From Fig. 4.7, we conclude that the larger the $\gamma_{\max}$-value, the better the method performs with respect to positivity. In this respect, the EBDF2 method shows the best performance.

Damping property

Apart from better stability and positivity properties, there is a third reason why the EBDF2 method is superior to the other methods: a stronger damping of the high frequency modes. We shall say that a method has good damping properties if the moduli of the characteristic roots are significantly smaller than 1, especially for the high frequencies which often are a cause of oscillations. The characteristic roots ζ_1 and ζ_2 depend on the coefficients of the method, on Δt, and on the eigenvalues of the discretized system.

Let us first consider the situation where Δt has been chosen as large as possible w.r.t. stability (i.e., using the maximal Courant number). In Fig. 4.8 we plot the values of $\zeta_M := \max(|\zeta_1|, |\zeta_2|)$ of the EBDF2 method, the ADAMS method, and the $\xi = \frac{18}{13}$ method as a function of the spectrum of the ODE system. Here we distinguish between the third-order upwind-biased discretization scheme (left panel) and the first-order upwind discretization scheme (right panel). We show both spatial discretization schemes, since the limiter (4.3) that we use may cause a change from one scheme to the other.

For the high frequency modes, we obtain the value $\zeta_M \approx 0.63$ for the EBDF2 method and $\zeta_M \approx 0.71$ for the ADAMS method (Fig. 4.8a). Both values are substantially below one, showing strong damping of these methods when the



a) Third-order upwind-biased scheme

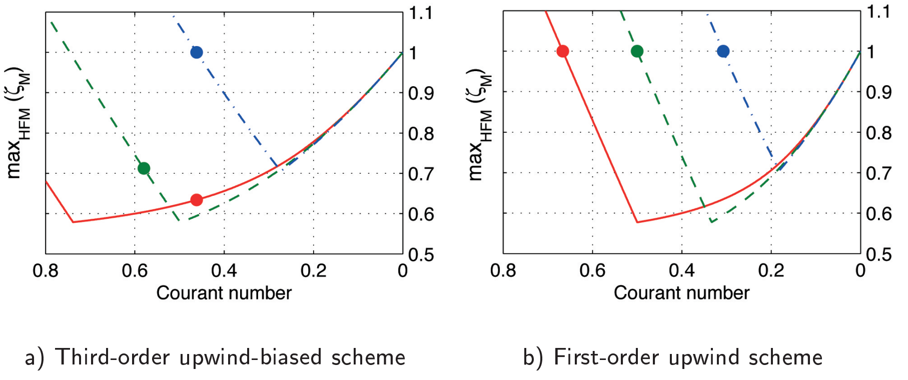
b) First-order upwind scheme

FIGURE 4.8: The maximum of moduli of the characteristic roots of the EBDF2 method (solid lines), the ADAMS method (dashed lines), and the $\xi = \frac{18}{13}$ method (dash-dotted lines) when their maximal Courant numbers are used.

third-order upwind-biased discretization scheme is used. Hence, the damping properties of the EBDF2 method are slightly better. The $\xi = \frac{18}{13}$ method does not damp the high frequency modes, because it has $\zeta_M \approx 1$. This observation is confirmed by the stability regions shown in Fig. 4.2: observing that the high frequency modes correspond to the left part of the (scaled) eigenvalue curve, the damping property of a method can be roughly measured by the distance between this part of the curve and the left boundary of the stability region. The EBDF2 method indeed shows the largest distance, whereas the $\xi = \frac{18}{13}$ method has a zero-distance. For the first-order upwind scheme, we see from Fig. 4.8b that $\zeta_M \approx 1$ for all three methods, showing that the high frequency modes are hardly damped in this case.

In the above considerations, the damping property was studied for maximal Courant numbers of the methods (see Table 4.1). It is interesting to see how the damping property of these methods depends on the Courant number. As stated above, we are particularly interested in the damping of the high frequency modes. Therefore, we plot in Fig. 4.9 the values $\max_{\mathrm{HFM}}(\zeta_M)$, where HFM indicates that we only consider the upper part of the spectrum. Again, the two panels correspond to the two discretization schemes.

Fig. 4.9 shows some interesting results: in the left panel (i.e., the third-order upwind-biased scheme) we see that the EBDF2 and ADAMS methods have $\max_{\mathrm{HFM}}(\zeta_M)$ less than 1 when the Courant number reaches the associated critical value (indicated by dots in the figures). This indicates that the stability is determined by the low-frequency modes for these methods. Again, this can be understood if we consider the stability region in Fig. 4.2: the eigenvalues corresponding to low-frequency modes are very close to the boundary of the stability region. The situation is different for the method with $\xi = \frac{18}{13}$, where the high-frequencies determine the CFL-number.



a) Third-order upwind-biased scheme b) First-order upwind scheme

FIGURE 4.9: Damping property of EBDF2 method (solid lines), ADAMS method (dashed lines), and the $\xi = \frac{18}{13}$ method (dash-dotted lines) for different values of the Courant number. The dots correspond to the maximal Courant numbers of these methods.

For the first-order upwind scheme (Fig. 4.9b) the maximal time step is completely determined by the high frequencies since the points for which $\max_{\mathrm{HFM}}(\zeta_M)$ equals 1 coincide with the critical value of the CFL-numbers of all methods.

As a conclusion of this subsection, we can say that the EBDF2 method is superior to the other integration methods in terms of stability, positivity, and damping properties.

4.4.2 Two-dimensional advection problem

In this section we will consider a simplified version of the phytoplankton model studied in Section 4.4.3. To that end we restrict ourselves to two spatial dimensions and concentrate on the most important term, i.e., the advection term. Therefore, we will still use the explicit version of the IMEX methods, similarly as in the preceding section. An important difference with that section is that we will now use variable step sizes based on local error control. Furthermore, the parameters that we need in this test model have been given realistic values, in correspondence with the full plankton model.

Again we will compare the various ξ-values characterizing the methods and we will see that the conclusion is the same as the one we have drawn for the model problem: the EBDF2 method (based on $\xi = \frac{2}{3}$) is to be preferred.

The 2D advection model that we consider in this subsection is given by

$$\frac{\partial \omega}{\partial t} + \frac{\partial (a_1 \omega)}{\partial x} + \frac{\partial (a_2 \omega)}{\partial y} = 0. \qquad (4.28)$$

The space-time domain is given by $(x, y) \in [0, X] \times [0, Y]$, $t \in [0, T]$ with $X = Y = 10^4$ cm, $T = 5$ days. The divergence-free velocity field $a = (a_1, a_2)$

describes the horizontal water flow and is given by (see Fig. 4.10a)

$$\begin{aligned} a_1(x,y) &= -\frac{\pi}{5}\,10^{-4}X\,\sin^2(\pi\frac{x}{X})\,\sin(2\pi\frac{y}{Y}), \\ a_2(x,y) &= +\frac{\pi}{5}\,10^{-4}Y\,\sin^2(\pi\frac{y}{Y})\,\sin(2\pi\frac{x}{X}), \end{aligned} \tag{4.29}$$

both components are in cm/s. The solution ω represents the concentration of phytoplankton and is initialized as (see Fig. 4.10b)

$$\omega_0(x,y) = C\,e^{-10^{-6}(y-0.5Y)^2}, \quad C = 50\,\text{cells/cm}^3. \tag{4.30}$$

The problem has been discretized on an uniform 50×50 spatial grid. For some particular points in time, Fig. 4.11 shows the solution.

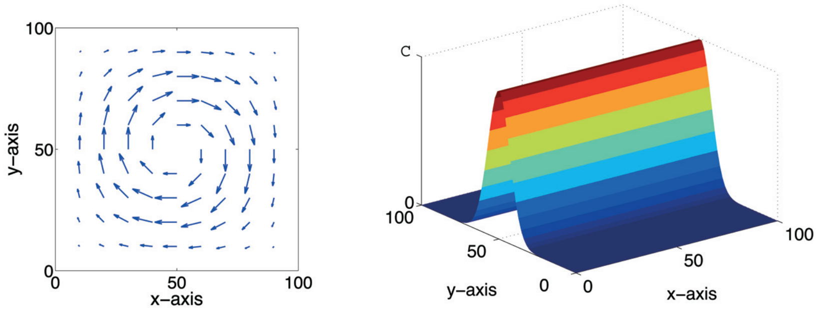
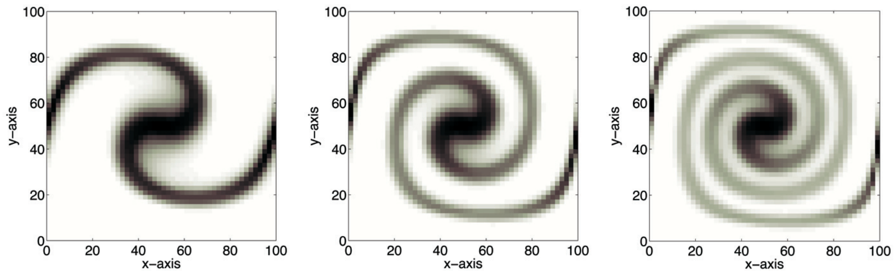


FIGURE 4.10: Left panel: the velocity field $a = (a_1, a_2)$. Right panel: the initial value. Along the axes, the units are presented in m and this will be the default unit throughout similar plots.



$t = 0.1$ day. $t = 0.2$ day. $t = 0.3$ day.

FIGURE 4.11: Snapshots of the solution at particular points in time. Dark coloured regions correspond to larger solutions.

We solve the resulting ODE system by the method (4.7) with $\eta = 0$. The variable step size that we use is subject to two restrictions: (i) the value that is suggested by the local error control mechanism, and (ii) by the CFL condition. The actual time step that we use is the minimum of these two. The local error control is based on the first-order estimator given by (5.9) in [38, p. 204]. That is,

$$\frac{r}{1+r}||\boldsymbol{W}_{n+2} - (1+r)\boldsymbol{W}_{n+1} + r\boldsymbol{W}_n||_2,$$

where $r = \Delta t_{n+1}/\Delta t_n$ is the step size ratio. For the first step, the formula (5.11) in [38, p. 204] is used, i.e.,

$$\frac{1}{2}||\boldsymbol{W}_1 - \boldsymbol{W}_0 - \Delta t_0 \boldsymbol{F}(t_0, \boldsymbol{W}_0)||_2.$$

The step size is chosen such that the estimated local error is close to the specified tolerance. For crude tolerance values it might happen that the CFL condition is more restrictive. Therefore, the new step size Δt_{n+1} is always restricted by

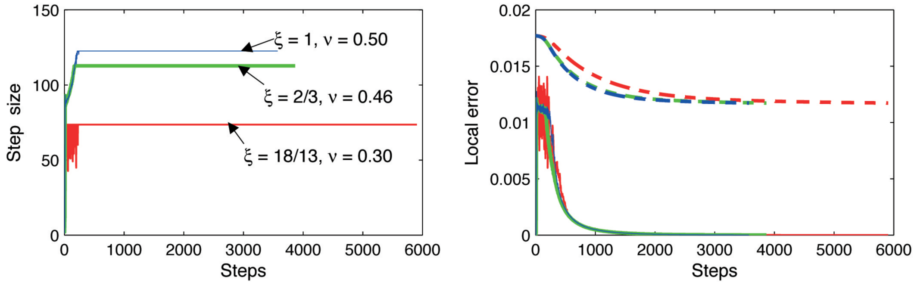
$$\Delta t_{n+1} \max_{(x,y)} \left(\frac{|a_1(x,y)|}{\Delta x} + \frac{|a_2(x,y)|}{\Delta y} \right) \leq \nu. \tag{4.31}$$

Hence, for the velocity (4.29), and using $\Delta x = \Delta y$, the maximal value for the step size is set to $\nu \frac{20\Delta x}{3\sqrt{3}\pi}$.

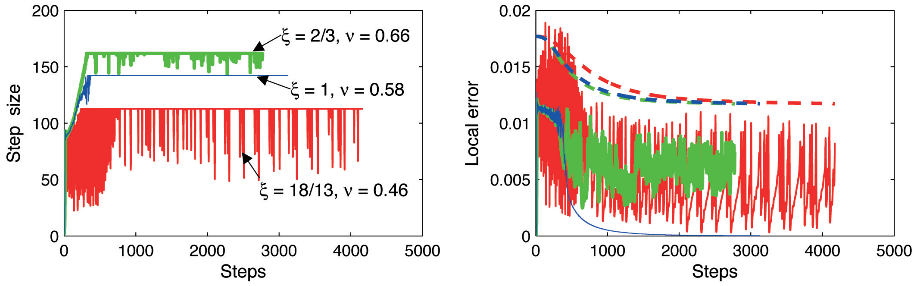
For several values of ν, Table 4.3 contains integration statistics (the total number of steps and number of rejected steps) for the various methods with $\xi = \frac{18}{13}$, $\xi = \frac{2}{3}$, and $\xi = 1$. The actual ν-values are the critical CFL-numbers ν_1 and ν_3 as listed in Table 4.1. Detailed performance of these methods with $\nu = \min(\nu_1, \nu_3)$ and $\nu = \max(\nu_1, \nu_3)$ is shown in Figs. 4.12a and b, respectively. From the table and the figure, the $\xi = \frac{2}{3}$ method is seen to be the most robust integrator. This will now be discussed in some detail:

TABLE 4.3: Two-dimensional advection problem. NST (= number of accepted steps + number of rejected steps) and NETF (= number of error test failures, i.e., number of rejected steps) for the $\xi = \frac{18}{13}$, $\xi = \frac{2}{3}$, and $\xi = 1$ methods. The tolerance is set to 10^{-3}. $T = 5$ days.

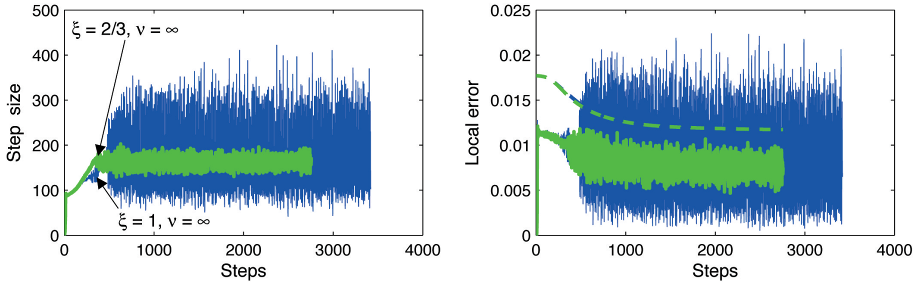
ν	$\xi = \frac{18}{13}$		$\xi = \frac{2}{3}$		$\xi = 1$	
	NST	NETF	NST	NETF	NST	NETF
0.30	5907	0	5887	0	5887	0
0.46	4171	18	3866	0	3867	0
0.50	4177	68	3571	0	3574	0
0.58	4458	381	3111	0	3124	0
0.66	Unstable		2775	0	2795	0
∞	Unstable		2756	0	3420	612



a) The step size restriction determined by $\nu = \min(\nu_1, \nu_3)$.



b) The step size restriction determined by $\nu = \max(\nu_1, \nu_3)$.



c) No step size restriction. The $\xi = \frac{18}{13}$ method is simply unstable and hence is not included here.

FIGURE 4.12: Two-dimensional advection problem: Performance of the $\xi = \frac{18}{13}$ (red lines), $\xi = \frac{2}{3}$ (green lines) and $\xi = 1$ (blue lines) methods. Left: the step size. Right: the L_2 local truncation error; the dashed lines indicate the corresponding (same colour, dash style) reference tolerance, i.e., RTOL $||\boldsymbol{W}_{n+2}||_2$, RTOL$=10^{-3}$.

Setting ν equal to $\min(\nu_1, \nu_3)$ (Fig. 4.12a), both the individual spatial discretization schemes, i.e., the first-order upwind scheme and the third-order upwind-biased scheme, are stable. All three integration methods perform well. The local errors (the right panel) are always smaller than the reference tolerances, especially in the last part of the integration, and error test failures did not occur. After the transient phase all methods have increased the step size to the maximal, CFL based value and this step size has been used until the end of the integration interval. Consequently, the $\xi = \frac{18}{13}$ method needs the largest number of steps because it has the smallest $\min(\nu_1, \nu_3)$-value.

Next, we set ν equal to $\max(\nu_1, \nu_3)$ (Fig. 4.12b), and hence one of the individual discretization schemes is locally unstable. Except for the $\xi = 1$ method, the other two methods give an oscillatory behaviour of the Δt-curve.

With $\nu = \max(\nu_1, \nu_3) = 0.46$ the $\xi = \frac{18}{13}$ method integrates with a wild-oscillatory Δt-curve (left panel, Fig. 4.12b). The local error (right panel) is quite close to the reference tolerance. In the first 600 steps, the local error often exceeds the tolerance. As a result, a lot of failures occur (NETF=18, Table 4.3). Although the $\xi = \frac{2}{3}$ and $\xi = \frac{18}{13}$ methods have the same ν_3-value, their ν_1- and $\gamma_{\max}$-values differ substantially (see Tables 4.1 and 4.2). With the same $\nu = \nu_3 = 0.46$ value, the EBDF2 method (in Fig. 4.12a) performs superior to the $\xi = \frac{18}{13}$ method (in Fig. 4.12b) in the sense that it integrates smoothly, without any failures.

With $\nu = \max(\nu_1, \nu_3) = 0.66$, the EBDF2 method also integrates with non-constant step sizes (left panel, Fig. 4.12b), although no error test failure occurs. This is because the code detects that the local error is close to the tolerance.

Compared with the $\xi = 1$ method, the EBDF2 method has a larger error constant. Moreover, since its $\max(\nu_1, \nu_3)$-value is larger than that of the $\xi = 1$ method, the chosen step size is also larger. Hence, the local error of the EBDF2 method turns out to be much larger than that of the $\xi = 1$ method. As a result of the small error, the $\xi = 1$ method always integrates with a constant step size in the last part of the integration (left panel, Fig. 4.12b).

In some applications, however, it might happen that we do not have prior knowledge of the CFL condition. Therefore, it is of interest to see how the codes behave if the step size is only determined by the local error control (i.e., we set $\nu = \infty$). To answer this question, we have integrated the system by the $\xi = \frac{2}{3}$ and $\xi = 1$ methods with $\nu = \infty$. The result is plotted in Fig. 4.12c. The $\xi = \frac{18}{13}$ method is not included here because it already shows unsatisfactory behaviour for $\nu = 0.46$. From the figure we see that both methods choose non-constant step sizes. However, the Δt-curve of the $\xi = 1$ method oscillates much more prominent. The local error of the EBDF2 method is always below the tolerance curve. In contrast, the local error of the $\xi = 1$ method frequently exceeds the reference tolerance. As a consequence, an enormous number of error test failures occur (see the last row in Table 4.3). This is due to the fact that the $\xi = 1$ method has less favourable positivity and damping properties (see Section 4.4.1). We therefore conclude that the EBDF2 method give the most robust integration.

4.4.3 Model of phytoplankton dynamics

In this final subsection we study the full phytoplankton model in a three dimensional domain $\Omega = [0, X] \times [0, Y] \times [0, Z]$, where $X = Y = 10^4$ cm and $Z = 10^3$ cm, discretized on a $50 \times 50 \times 20$ grid. Here, we briefly summarize the mathematical model. For more details of the model, interested readers are referred to Chapter 1.

The PDE for the phytoplankton concentration ω reads

$$\begin{aligned} \frac{\partial \omega}{\partial t} + \nabla \cdot (a\,\omega) &= \nabla \cdot (D \nabla \omega) + (p(L) - \ell)\,\omega \quad \text{in } \Omega, && (4.32a) \\ a\,\omega - D\nabla\omega &= 0 \quad \text{at } \partial\Omega. && (4.32b) \end{aligned}$$

Here, $D = diag(D_1, D_2, D_3)$ represents the diffusion coefficients and is supposed to be constant and positive. The growth function p is defined by (see Chapter 1) $p(L) = p_{\max}\frac{L}{H+L}$, where H is the so-called half-value, $p_{\max}$ is the maximal growth rate, and ℓ denotes the death rate which is assumed constant. The solution dependent, nonlinear function L

$$L = L_{in} e^{-K_{bg} z} e^{-k \int_0^z \omega(x,y,\sigma,t)\, d\sigma} \tag{4.33}$$

denotes the light which penetrates into a water column. Here, L_{in} denotes the incident light intensity (i.e., the light at the water surface, corresponding to $z = 0$) and K_{bg} is the so-called background attenuation due to all non-phytoplankton components in the water. Hence, K_{bg} defines the turbidity of the water. Finally k is the light attenuation coefficient of the phytoplankton and is species-specific. The derivation of (4.33) is based on the assumption that the light gradient is described by Lambert-Beer's law (see [33] for details). Due to the integral in the light function we see that the unknown concentration ω at a certain depth z depends on all concentrations above that depth (due to absorption and shading). This integral term is discretized by the repeated trapezoidal rule. The parameter values needed in the reaction term are listed in Table 4.4.

TABLE 4.4: Parameter values (see also Tables 1.1 and 1.2).

ℓ $\left(\frac{1}{\mathrm{h}}\right)$	$p_{\max}$ $\left(\frac{1}{\mathrm{h}}\right)$	H $\left(\frac{\mu\mathrm{molphotons}}{\mathrm{cm}^2\,\mathrm{s}}\right)$	L_{in} $\left(\frac{\mu\mathrm{molphotons}}{\mathrm{cm}^2\,\mathrm{s}}\right)$	K_{bg} $\left(\frac{1}{\mathrm{cm}}\right)$	k $\left(\frac{\mathrm{cm}^2}{\mathrm{cells}}\right)$
0.01	0.04	10^{-3}	$3.5 \cdot 10^{-2}$	$2.0 \cdot 10^{-3}$	$3 \cdot 10^{-7}$

The divergence-free velocity field $a = (a_1, a_2, a_3)$ is chosen as (see Sections 4.4.2

and 1.4.1)

$$\begin{aligned} a_1(x,y,z) &= -\frac{\pi}{5}\,10^{-4}X\,\sin^2(\pi\frac{x}{X})\,\sin(2\pi\frac{y}{Y})\,e^{-2z/Z}, \\ a_2(x,y,z) &= +\frac{\pi}{5}\,10^{-4}Y\,\sin^2(\pi\frac{y}{Y})\,\sin(2\pi\frac{x}{X})\,e^{-2z/Z}, \\ a_3 &= c. \end{aligned}$$

Here, a_1 and a_2 are components of the horizontal water flow and are the z-dependent extension of the velocity field (4.29); c is the vertical velocity of phytoplankton species and is in cm/s. We will consider a neutrally-buoyant species, that is a species with a vertical velocity $c = 0$. The phytoplankton concentration is again initialized by (4.30).

The semi-discretized problem will be integrated in time by an IMEX method, where the advection and reaction terms are treated explicitly and the diffusion term implicitly. The actual method that we use is a variable step size version of the IMEX scheme (4.19). Setting η in (4.7b) equal to $\frac{1}{2}(1+\frac{\xi}{r(r+1)})$ so that we have $\beta_1 = 0$, the variable time step version of (4.19) reads

$$\begin{aligned} &\boldsymbol{W}_{n+2} + (-1-r+\xi)\boldsymbol{W}_{n+1} + (r-\xi)\boldsymbol{W}_n = \\ &\tfrac{1}{2}\left(r+1+\tfrac{\xi}{r}\right)\Delta t_{n+1}\boldsymbol{F}_E(t_{n+1},\boldsymbol{W}_{n+1}) + \tfrac{1}{2}\left(-(r+1)+\tfrac{\xi}{r}\right)\Delta t_{n+1}\boldsymbol{F}_E(t_n,\boldsymbol{W}_n) + \\ &\tfrac{1}{2}\left(1+\tfrac{\xi}{r(r+1)}\right)\Delta t_{n+1}\boldsymbol{F}_I(t_{n+2},\boldsymbol{W}_{n+2}) + \tfrac{1}{2}\left(-1+\tfrac{\xi(2r+1)}{r(r+1)}\right)\Delta t_{n+1}\boldsymbol{F}_I(t_n,\boldsymbol{W}_n), \end{aligned} \tag{4.34}$$

where $r = \Delta t_{n+1}/\Delta t_n$ is the step size ratio. As shown in Sections 4.4.1 and 4.4.2, the $\xi = \frac{2}{3}$ value gives the most robust integration. Therefore, from now on we will integrate the ODE system by the method (4.34) with $\xi = \frac{2}{3}$, i.e., the IMEX-BDF2 method.

Using this ξ-value, we will show that the IMEX method is more efficient than the well-known integrator VODE [8], based on the fully implicit BDF methods.

A comparison of IMEX-BDF2 and VODE

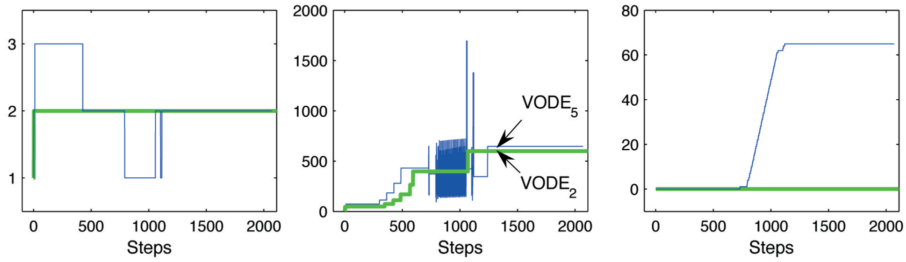
Now we compare the IMEX-BDF2 method with the well-known integrator VODE [8], both applied to the three-dimensional phytoplankton model (4.32). We remark that this model also has been studied in Chapter 1, where VODE was used for the time integration. However, in Chapter 1 the spatial discretization was based on the third-order upwind-biased scheme *without* a limiter. Here we will always apply the limiter (4.3) because in the comparison of IMEX-BDF2 and VODE we will confine ourselves to time integration aspects. Hence, both solvers should start from the same positive system of ODEs. Moreover, in this comparison we do not want to be hampered by negative solution values induced by a non-positive spatial discretization. We will start with mentioning some general properties of VODE.

VODE is based on the backward differentiation formulae (BDF) using a variable order running from 1 to 5, whereas IMEX-BDF2 method is of fixed second-order accuracy. The high-order formulae in VODE need information from the past. Consequently, more storage is required. Furthermore, the higher the order, the smaller the wedge of the stability region in the negative half plane (see [24, p. 251] and [38, p. 179]). VODE offers the possibility to fix the maximal order to be used during the integration process. If we define VODE_q to be VODE with maximum order q, then VODE_2 is A-stable, whereas $\text{VODE}_{q>2}$ is only A(α)-stable (see (4.11)). The angle α decreases as q increases (e.g. $\alpha = 90^{\circ}$ if $q = 2$ and $\alpha = 51.84^{\circ}$ if $q = 5$). When A-stability is an important property, the higher-order methods will be less efficient. This will be illustrated below.

Test with several versions of VODE Consider the advection-diffusion-reaction problem (4.32) with $D_1 = D_2 = 10\,\text{cm}^2/\text{s}$ and $D_3 = 1\,\text{cm}^2/\text{s}$. To capture interesting aspects of the integrators, we consider the problem on a longer time interval, i.e., $T = 10$ days. Setting the tolerance to 10^{-4}, we integrated the ODE system by VODE_2 and VODE_5, and compare the performances of these two versions of the code. In Fig. 4.13 we plot the order as well as the step size that these versions choose for the integration. Initially, VODE_5 increases the order up to 3 (see Fig. 4.13a). Then, it decides to lower the order to 2 and further down to 1. However, this low order is too inaccurate and a large number of convergence failures in the Newton process occur (see Fig. 4.13c). From Fig. 4.13b we see that the code has problems in choosing a correct time step. It decides to increase the order to 2, allowing a larger time step. The code settles at this order as well as a constant step size until the end of the integration interval.

Different from VODE_5, VODE_2 keeps using the order 2 and gradually increases the step size. Then it integrates the last part with a constant step size, without any convergence failure. This step size turns out to be slightly smaller than the one chosen by VODE_5 in the last part of the integration interval. Consequently, the total number of steps taken by VODE_2 is slightly larger than that of VODE_5 (Fig. 4.13). However, the Newton process converges faster. The total number of Newton iterations of VODE_2 and VODE_5 are 2743 and 3272, respectively. In other words, the two codes are of similar efficiency.

Next, to increase the stiffness of the problem, we consider the case where $D_1 = D_2 = 20\,\text{cm}^2/\text{s}$ and $D_3 = 10\,\text{cm}^2/\text{s}$, while all other parameter values are left unchanged. The results obtained by the two codes are plotted in Fig. 4.14 (see also the first block of results in Table 4.5). From this figure we see that the majority of the steps taken by VODE_5 are based on the third-order formula. Nevertheless, the chosen time step is much smaller than that of VODE_2. As a result, VODE_5 needs a much larger number of steps to reach the end point of the integration. This indicates that increasing the order to 3 is not a good choice. This is counterintuitively since we expected that a higher order method would allow larger time steps to satisfy the criterion on the local error. The results observed for this test indicate that VODE_5 has encountered stability problems.

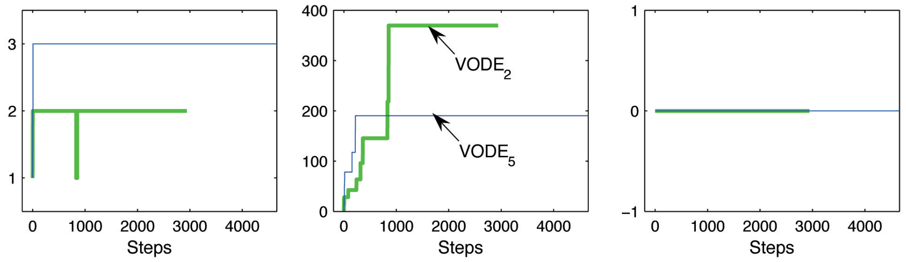


a) Order used.

b) Step size.

c) Convergence failures.

FIGURE 4.13: Comparison between performances of VODE$_2$ (green lines) and VODE$_5$ (blue lines). $D_1 = D_2 = 10\,\mathrm{cm}^2/\mathrm{s}$, $D_3 = 1\,\mathrm{cm}^2/\mathrm{s}$. $T = 10$ days.



a) Order used.

b) Step size.

c) Convergence failures.

FIGURE 4.14: Comparison between performances of VODE$_2$ (green lines) and VODE$_5$ (blue lines). $D_1 = D_2 = 20\,\mathrm{cm}^2/\mathrm{s}$, $D_3 = 10\,\mathrm{cm}^2/\mathrm{s}$. $T = 10$ days.

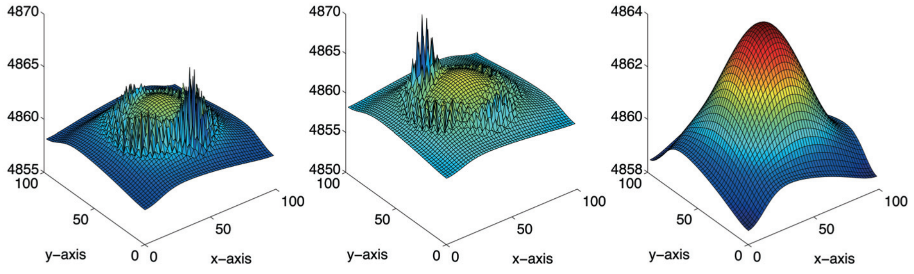
Global comparison To reduce the amount of linear algebra in solving the systems, both VODE and IMEX-BDF2 have been equipped with the Approximate Matrix Factorization (AMF) technique [28]. In this way, solving one system with a rather nasty structure is converted into the successive solution of three systems with much simpler algebra involved. Each system associates with only one spatial dimension. As a result, the implicit relations can be solved along each grid line separately (see Chapter 1). For the IMEX approach, the systems to be solved are even more simple, i.e., systems with tridiagonal matrices.

A disadvantage of AMF is that it has a less rigorous damping of high frequency modes [16, 28]. Moreover, using AMF we need more iterations in the Newton process. For the IMEX method, we loose the nice property of the linear system: one Newton iteration is no longer sufficient to obtain the new approximation $\boldsymbol{W}_{n+2}$, in spite of the fact that the implicitly treated diffusion terms are *linear*[3]. In addition, AMF will slow down the convergence of the Newton process[4]. In VODE, the systems are *non-linear*, due to the limiting advection discretization. Hence, we need more than one Newton iteration per step for solving the systems, with and without the AMF technique.

Due to the disadvantages of AMF, we should be careful in stopping the Newton iteration process. It is common practice in ODE software to require that

$$||\text{iteration error}|| \;\leq\; \kappa \;||\text{local error}||,$$

with κ a small number to avoid that the iteration error severely interferes with the local error. The default value used in VODE is $\kappa = 0.1$. This value has been used in the two aforementioned tests. For the stiff test (with $D_1 = D_2 = 20\,\text{cm}^2/\text{s}$, $D_3 = 10\,\text{cm}^2/\text{s}$) we plot in Figs. 4.15a,b the solutions obtained by VODE$_5$ and VODE$_2$, respectively, using $\kappa = 0.1$. Both solutions show the unsatisfactory property of high-frequency oscillations.



a) b) c)

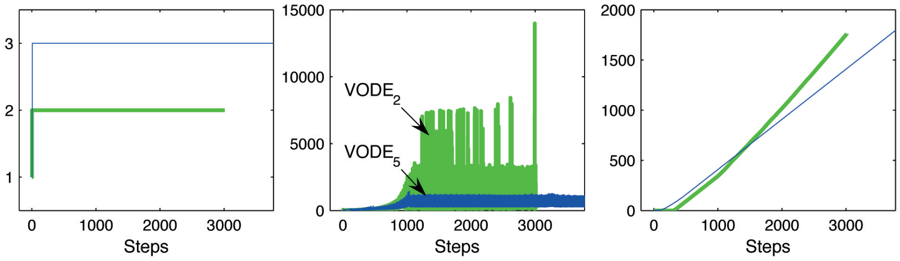
FIGURE 4.15: Solutions at $z = 0.25\,\text{m}$ obtained by VODE$_5$ (left), VODE$_2$ (middle) and VODE$_5$/VODE$_2$ with the stringent Newton solver (right). $D_1 = D_2 = 20\,\text{cm}^2/\text{s}$, $D_3 = 10\,\text{cm}^2/\text{s}$. $T = 10$ days.

Therefore, as a next test, we choose κ as small as 10^{-4} to be sure that the iteration error is negligible with respect to the local error. Hence, using such a stringent stopping criterion we might say that we have 'solved' the implicit relations and that VODE really uses the underlying BDF formulae. As a result of this small κ-value we indeed obtain a smooth solution, which is shown in Fig. 4.15c. Some integration statistics are listed in Table 4.5 (second block of results) and plotted in Fig. 4.16. Comparing the results shown in Fig. 4.14

---

[3] For linear system: take one Newton step to obtain $\boldsymbol{W}^{k+1}_{n+2}$ from an arbitrary $\boldsymbol{W}^{k}_{n+2}$. Then, $\boldsymbol{W}^{k+1}_{n+2}$ always satisfies the implicit relation. However, when AMF is used, this no longer holds, because $((I - \beta_2 \Delta t J_1)(I - \beta_2 \Delta t J_2)(I - \beta_2 \Delta t J_3))^{-1}\,(I - \beta_2 \Delta t J) \neq I$, where $\partial \boldsymbol{F}_I / \partial \boldsymbol{w} \equiv J = J_1 + J_2 + J_3$, the Jacobian matrices associated with different spatial dimensions.

[4] Proof for a scalar problem: the iteration error ϵ^k satisfies the recursion $\epsilon^{k+1} = (1 - M)\,\epsilon^k$, where $M = ((1 - \beta_2 \Delta t \lambda_1)(1 - \beta_2 \Delta t \lambda_2)(1 - \beta_2 \Delta t \lambda_3))^{-1}\,(1 - \beta_2 \Delta t \lambda)$ and $\lambda = \lambda_1 + \lambda_2 + \lambda_3$, eigenvalues of systems in different spatial dimensions. For increasing stiffness, we have $\lambda_i \to \infty$, $i = 1, 2, 3$, and hence $M \to 0$.



a) Order used. b) Step size. c) Convergence failures.

FIGURE 4.16: VODE$_2$ (green lines) and VODE$_5$ (blue lines) with the stringent Newton solver. $D_1 = D_2 = 20\,\mathrm{cm}^2/\mathrm{s}$, $D_3 = 10\,\mathrm{cm}^2/\mathrm{s}$. $T = 10$ days.

($\kappa = 10^{-1}$) with those in Fig. 4.16 ($\kappa = 10^{-4}$), we observe some remarkable differences: the small κ-value results in larger step sizes, particularly for the VODE$_2$ version (compare Δt in Fig. 4.14b and Fig. 4.16b). However, large time steps in combination with a stringent stopping criterion results in a dramatic increase of the number of convergence failures. Obviously, this small κ-value leads to many Newton iterations and hence many $\boldsymbol{F}$-evaluations (NFEE and NFIE in Table 4.5). Also, as a consequence of the large number of convergence failures (NCF in Table 4.5), the Jacobian matrix is frequently updated (NJE in Table 4.5). After each failure, the step size is reduced. Since the implicit relation has been 'solved' due to the small κ-value, the local error control is easily satisfied (the step is accepted). As a consequence, the local error estimator often suggests to largely increase the step size (for the next step). With such a large step size, however, a convergence failure is likely to occur. This explains the wildly-oscillating Δt-curves in Fig. 4.16b.

As a conclusion from this test we may say that VODE, when equipped with the AMF technique, needs a stringent stopping criterion in the Newton process to guarantee a solution which is free of wiggles, however at a high price. Since a smooth numerical solution is important (especially in long term integrations) we think that a stringent κ-value is unavoidable.

Performance of IMEX-BDF2 This version of VODE will be compared to the IMEX-BDF2 method that we also applied to this problem. Since IMEX-BDF2 also uses the AMF technique, we choose the same small $\kappa = 10^{-4}$ in the stopping criterion for Newton. It should be remarked that a small value for κ is less critical, since the implicit systems in the IMEX method only involve the linear diffusion terms. Moreover, an accurate Jacobian matrix is available which needs to be calculated only once. The results for the IMEX-BDF2 method are given in the last block in Table 4.5. The solution obtained by this method is smooth and very similar to the VODE solution. Comparing the performance of both solves we see that the IMEX-BDF2 method is more efficient. It results in a

Table 4.5: Integration statistics of IMEX-BDF2 and VODE. NST: number of steps, NETF: number of error test failures, NNI: number of Newton iterations, NCF: number of convergence failures, NFEE: number of $\boldsymbol{F}_E$ evaluations, NFIE: number of $\boldsymbol{F}_I$ evaluations, NJE: number of Jacobian evaluations, q_{end}: last order used, RERR: relative L_2 integration error. The superscript * indicates that the stringent Newton solver is used. The subscripts $_{0.46}$ and $_{0.66}$ denote that the IMEX-BDF2 method has used the stringent Newton process and integrated with $\Delta t_{\max} = 0.46\,\frac{20\Delta x}{3\sqrt{3}\pi}$ and $\Delta t_{\max} = 0.66\,\frac{20\Delta x}{3\sqrt{3}\pi}$, respectively. $D_1 = D_2 = 20\,\mathrm{cm}^2/\mathrm{s}$, $D_3 = 10\,\mathrm{cm}^2/\mathrm{s}$. $T = 10$ days.

	VODE$_5$	VODE$_2$	VODE$^*_5$	VODE$^*_2$	IMEX$_{0.46}$	IMEX$_{0.66}$
NST	4658	2943	3772	3012	7824	5549
NETF	1	1	1	1	0	0
NNI	6214	5170	18582	15544	11643	8589
NCF	0	0	1793	1759	0	0
NFEE	6215	5171	18583	15545	7825	5550
NFIE	6215	5171	18583	15545	11644	8590
NJE	78	49	3587	3127	1	1
q_{end}	3	2	3	2	2	2
Last Δt	190	370	274	516	113	162
NNI/NST	1.33	1.76	4.92	5.16	1.49	1.55
RERR	$1.1{\cdot}10^{-4}$	$8.5{\cdot}10^{-5}$	$1.1{\cdot}10^{-7}$	$1.1{\cdot}10^{-4}$	$4.4{\cdot}10^{-6}$	$1.9{\cdot}10^{-4}$

much smaller number of Newton iterations as well as number of $\boldsymbol{F}$-evaluations (NFEE and NFIE). The integration is performed without any failures (both NETF=0 and NCF=0).

Comparison per step Finally, we will discuss the advantages and disadvantages of the two integrators within one step. Since VODE is a fully implicit solver, the advection, diffusion, and reaction terms are all treated implicitly. This is more expensive than the IMEX-approach. Firstly, due to the variable velocity field and the nonlinear reaction term, the Jacobian is non-symmetric and varies over the steps. As a result, many Jacobian evaluations are needed (see row 7 in Table 4.5). For the IMEX method, however, the implicitly treated operator consists of linear diffusion terms only, resulting in a symmetric and constant Jacobian matrix (one Jacobian evaluation is needed). Secondly, because VODE is fully implicit, NFEE always equals NFIE, whereas, for the IMEX-method NFEE = NFIE − (NNI − NST), i.e., NFEE is substantially smaller than NFIE. Finally, since the limiting technique is used, extra nonlinearity is introduced into the ODE system, resulting in additional complexity in solving the systems in VODE (more Newton-iterations per step, see row 10 in Table 4.5).

TABLE 4.6: Comparison between VODE and IMEX-BDF2 integrators.

VODE	IMEX-BDF2 method
Global comparison	
– Varying orders (1 to 5)	– Fixed order 2
– More storage is needed	– Less storage is needed
– May allow a large Δt, with respect to stability	– Δt is restricted by the CFL condition
Comparison per step	
– All terms are treated implicitly	– Advection and reaction terms are treated explicitly; only diffusion terms are treated implicitly
– Jacobian matrix is non-symmetric, approximated, and varies in time	– Jacobian matrix is symmetric, exact, and constant in time
– Many Jacobian evaluations are needed	– Only one Jacobian evaluation is needed
– Number of $\boldsymbol{F}_E$-evaluations equals the number of Newton-iterations	– Only one $\boldsymbol{F}_E$-evaluation per step
– Limiter adds extra complexity in solving the system	– Limiter has no influence on the system to be solved

In Table 4.6 we have summarized the advantages and disadvantages of the two integrators, i.e., VODE and IMEX-BDF2.

4.5 Summary and conclusions

We have analyzed implicit-explicit integration methods to solve the advection-diffusion-reaction equations, in which we treated the non-stiff advection and reaction parts explicitly and the stiff diffusion part implicitly. In this way, an efficient approach is obtained, because the implicitly treated operator consists of the diffusion term only (which is linear for the phytoplankton model studied in this thesis), while the difficult and nonlinear advection and reaction terms are treated explicitly.

The linear multistep methods underlying the IMEX scheme forms a family of second-order two-step formulae. This choice is motivated by accuracy considerations combined with modest memory demands. The IMEX method was constructed by extrapolating the non-stiff leading term (see Section 4.3.2). Then, the stability of the resulting method was analyzed. Often, the stability region of the method is smaller than that of the explicit version. However, using the fact that the eigenvalues of the implicitly treated operator are negative and real, we derived a class of methods whose stability regions coincide with those of the explicit parts (Section 4.3.2).

Next, we illustrated the above theory by three applications:

First of all, the one-dimensional model advection test was studied (Section 4.4.1). For the spatial discretization, we always used the third-order upwind-biased scheme combined with the limiter (4.3). In regions with large spatial variation, this limiter switches to the first-order upwind scheme. For this test example the IMEX method reduced to the explicit integration method because of the absence of the diffusion term. The step size was taken constant.

With this application, we found that stability of the method is critically determined by the minimum of the CFL numbers ν_1 and ν_3 of the two individual discretization schemes, i.e., the first-order upwind scheme and the third-order upwind-biased scheme, respectively. This is interesting since CFL numbers are the result of stability analysis applied to a *linear* advection model problem, whereas the use of a limiter gives rise to a *nonlinear* model.

Furthermore, we also showed that for the methods having $\nu_1 > \nu_3$, e.g. the EBDF2 method, the use of a limiter is advantageous with respect to stability. However, for the methods having $\nu_1 < \nu_3$, e.g. the methods with $\xi = 1$ and $\xi = \frac{18}{13}$, the use of a limiter is a disadvantage. The $\xi = 1$ method was considered because it corresponds to the popular ADAMS-type method. The $\xi = \frac{18}{13}$ method was studied since it has the same ν_3-value as the EBDF2 method. Comparing the performances of these methods, the EBDF2 method showed the most robust integration with respect to positivity (recall that the EBDF2 method possesses the optimal positivity constant, see Chapter 3) and with respect to stability (because it has $\nu_1 > \nu_3$). Moreover, it has the best damping of high frequency modes (see Section 4.4.1).

Next, we studied the two-dimensional advection problem, with space dependent velocities. Here we used variable step sizes in the time integration. Although the problem in this application is more complex than the preceding model problem, we again observed that the EBDF2 method behaves most satisfactory (Section 4.4.2).

Finally, as suggested by the above applications, we selected the integration method with $\xi = \frac{2}{3}$ and applied it to the full phytoplankton problem (Section 4.4.3). Due to the presence of the diffusion term, we integrated the ODE system by the IMEX method. This method has been designed with the aim to have computational advantages over a fully implicit integration method. Therefore, we compared the performances of this method and the fully implicit integrator VODE [8]. The IMEX method turned out to be more efficient in terms of number of Newton iterations (NNI), number of $\boldsymbol{F}$-evaluations (NFEE and NFIE), number of convergence failures (NCFN), and number of Jacobian evaluations (NJE), see Table 4.5, as well as in terms of complexity, e.g. the symmetry and variability of the Jacobian and the influence of the limiter on the system to be solved (see Table 4.6).

Chapter 5

Reduced mixing generates oscillations and chaos in the oceanic deep chlorophyll maximum

Deep chlorophyll maxima (DCMs) are widespread in large parts of the world's oceans [68, 11, 51, 49, 47, 66, 27]. These deep layers of high chlorophyll concentration reflect a compromise of phytoplankton growth exposed to two opposing resource gradients: light supplied from above and nutrients supplied from below. It is often argued that DCMs are stable features. Here we show, however, that reduced vertical mixing can generate oscillations and chaos in phytoplankton biomass and species composition of DCMs. These fluctuations are caused by a difference between the time scales of two processes: (1) rapid export of sinking plankton, withdrawing nutrients from the euphotic zone and (2) a slow upward flux of nutrients fuelling new phytoplankton production. Climate models predict that global warming will reduce vertical mixing in the oceans [59, 6, 60, 61]. Our model indicates that reduced mixing will generate more variability in DCMs, thereby enhancing variability in oceanic primary production and in carbon export into the ocean interior.

5.1 Introduction

In oligotrophic waters, where the surface mixed layer is depleted of nutrients, subsurface maxima in chlorophyll concentration and phytoplankton biomass are often found (Fig. 5.1). Such deep chlorophyll maxima are permanent features in large parts of the tropical and subtropical oceans [68, 11, 51, 49, 47]. Furthermore, seasonal DCMs commonly develop in temperate regions [49, 66], and even in the polar oceans [27] when nutrients are depleted in the surface layer with the onset of the summer season. It is generally believed that DCMs are stable features tracking seasonal changes in light and nutrient conditions. However, here we extend recent phytoplankton models [19, 25, 43, 29, 32] to show that the phytoplankton populations of DCMs can exhibit sustained fluctuations.

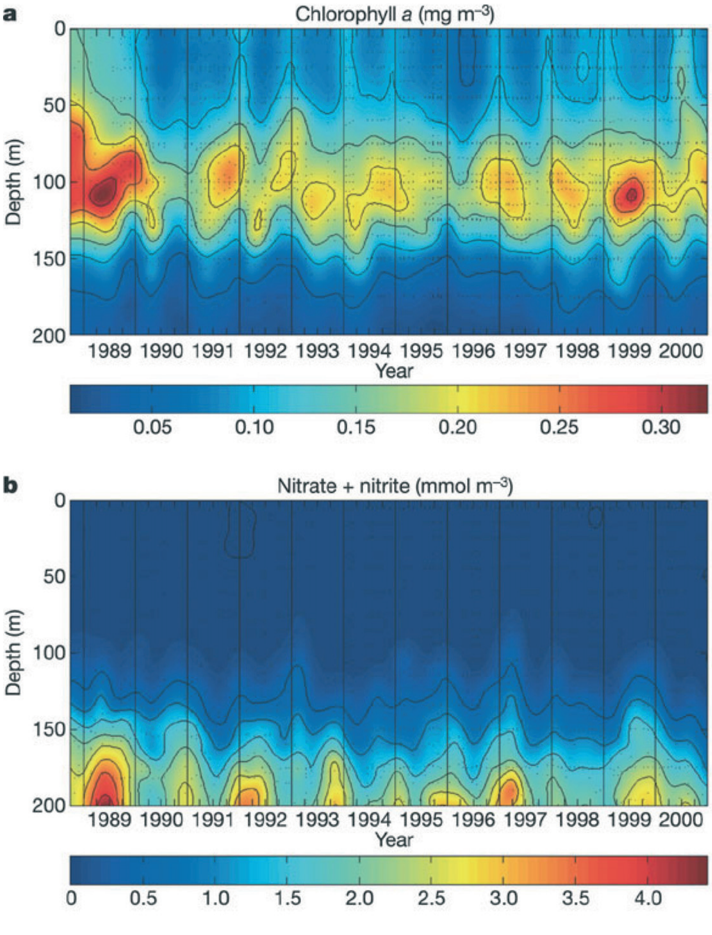


FIGURE 5.1: Time course of the DCM at Station ALOHA, in the subtropical Pacific Ocean, North of Hawaii. **a**, Chlorophyll *a*. **b**, Nitrate and nitrite. Data were obtained from the Hawaii Ocean Time-series (HOT) program.

5.2 Mathematical model

Consider a vertical water column. Let z indicate the depth in the water column. Let P denote the phytoplankton population density (number of cells per m^3). The population dynamics of the phytoplankton can be described by a reaction-advection-diffusion equation [19, 25, 43, 29, 32, 55]:

$$\begin{aligned}\frac{\partial P}{\partial t} &= \text{growth} - \text{loss} - \text{sinking} + \text{mixing} \\ &= \mu(N,I)P - mP - v\frac{\partial P}{\partial z} + \kappa\frac{\partial^2 P}{\partial z^2},\end{aligned} \tag{5.1}$$

where $\mu(N,I)$ is the specific growth rate of the phytoplankton as an increasing saturating function of nutrient availability N and light intensity I, m is the specific loss rate of the phytoplankton, v is the phytoplankton sinking velocity, and κ is the vertical turbulent diffusivity. The nutrient dynamics in the water column can be described as [19, 25, 43]:

$$\begin{aligned}\frac{\partial N}{\partial t} &= -\,\text{uptake} + \text{recycling} + \text{mixing} \\ &= -\,\alpha\,\mu(N,I)P + \epsilon\,\alpha\,mP + \kappa\frac{\partial^2 N}{\partial z^2},\end{aligned} \tag{5.2}$$

where α is the nutrient content of the phytoplankton, and ϵ is the proportion of nutrient in dead phytoplankton that is recycled. We assume that light intensity, I, decreases exponentially with depth according to Lambert-Beer's law, owing to light absorption by the phytoplankton population, by water, and by dissolved substances [29, 32]. To complete the model, we use zero-flux boundary conditions for the phytoplankton. Furthermore, we assume a zero-flux boundary condition for nutrient at the surface, while nutrient is replenished from below with a fixed concentration N_B at the bottom of the water column. The model formulation and simulation methods are described in further detail in Section 5.5, Supplementary Information. The model is parameterized for clear ocean water, reflecting the North Pacific subtropical gyre [47, 41] (Fig. 5.1).

5.3 Simulations

5.3.1 Results for a constant environment

In a first model simulation, with a turbulent diffusivity of $0.5\,\text{cm}^2\,\text{s}^{-1}$, nutrients in the top layer are gradually depleted by the phytoplankton. The nutricline slowly moves downwards, tracked by the phytoplankton population, until the population settles at a stable equilibrium at which the downward flux of consumed nutrients equals the upward flux of new nutrients (Fig. 5.2a). Thus, a stable DCM develops. For lower values of turbulent diffusivity, however, the model predicts that the phytoplankton population in the DCM will oscillate. Depending on the parameter settings the fluctuations in the DCM may range from mild oscillations (Fig. 5.2b) to pronounced chlorophyll peaks (Fig. 5.2c). To investigate this phenomenon further, we ran numerous simulations using a wide range of turbulent diffusivities. For comparison, vertical turbulent diffusivities in the ocean interior are typically on the order of $0.1\,\text{cm}^2\,\text{s}^{-1}$ to $1\,\text{cm}^2\,\text{s}^{-1}$ (refs. [48, 63, 20]). The model simulations predict that the DCM becomes unstable when turbulent diffusivity is in the lower end of the realistic range (Fig. 5.3a). By a cascade of period doublings, reduced turbulent mixing can even generate chaos in the DCM (Fig. 5.3b).

The mechanism underlying these fluctuations is a difference in time scale between the sinking flux of phytoplankton and the upward diffusive flux of nutrients. This might be called an 'advection-diffusion instability'. At low

Constant environment

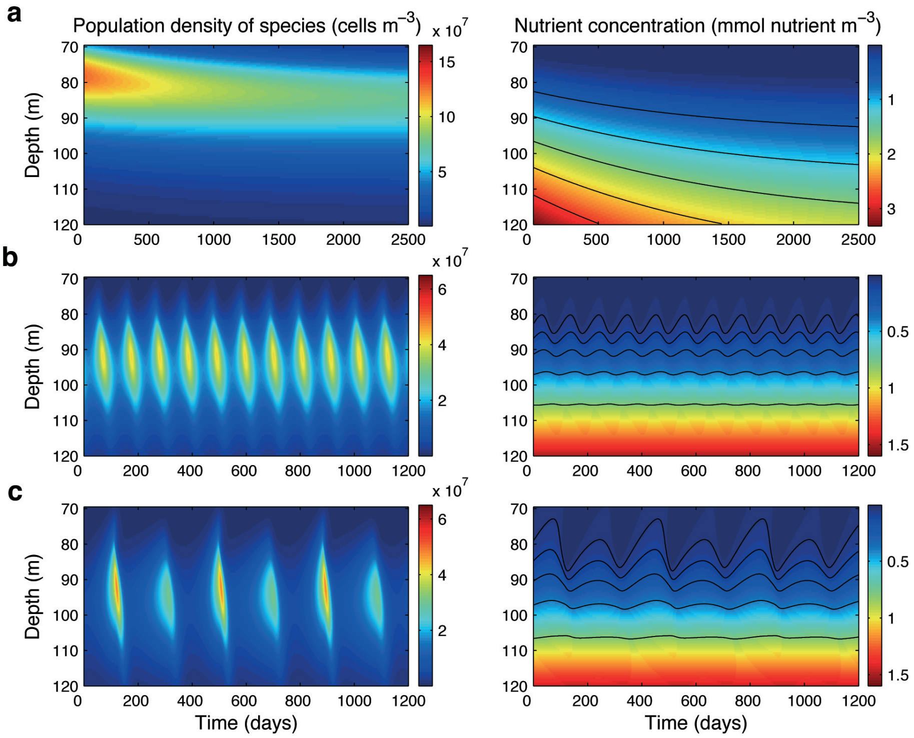


FIGURE 5.2: Model simulations at different intensities of vertical mixing. For **a**-**f**, the left panel shows phytoplankton dynamics (P) and the right panel shows nutrient dynamics (N). See Section 5.5, Supplementary Information, for parameter values. **a**-**c**, Constant environment. **a**, Stable DCM ($\kappa = 0.50\,\mathrm{cm}^2\,\mathrm{s}^{-1}$). **b**, Mild oscillations in the DCM ($\kappa = 0.20\,\mathrm{cm}^2\,\mathrm{s}^{-1}$). **c**, Large-amplitude oscillations in the DCM, with double periodicity ($\kappa = 0.12\,\mathrm{cm}^2\,\mathrm{s}^{-1}$).

diffusivity, the phytoplankton sink fast compared to the slow upward flux of nutrients. Thereby, the light conditions of the sinking phytoplankton deteriorate and the phytoplankton population declines. The declining phytoplankton population loses control over the upward nutrient flux, allowing new nutrients to diffuse further upwards. The upward flux of nutrients reaches a depth at which light conditions are suitable for growth. This fuels the next peak in the DCM. Indeed, model simulations indicate that the sinking flux has an important role in these oscillations, as oscillations were not observed with neutrally buoyant phytoplankton (results not shown). The period and amplitude of the DCM oscillations increase with increasing phytoplankton sinking velocity (Fig. 5.3c). The period and amplitude decrease with increasing vertical diffusivity (Fig. 5.3d).

Seasonal environment

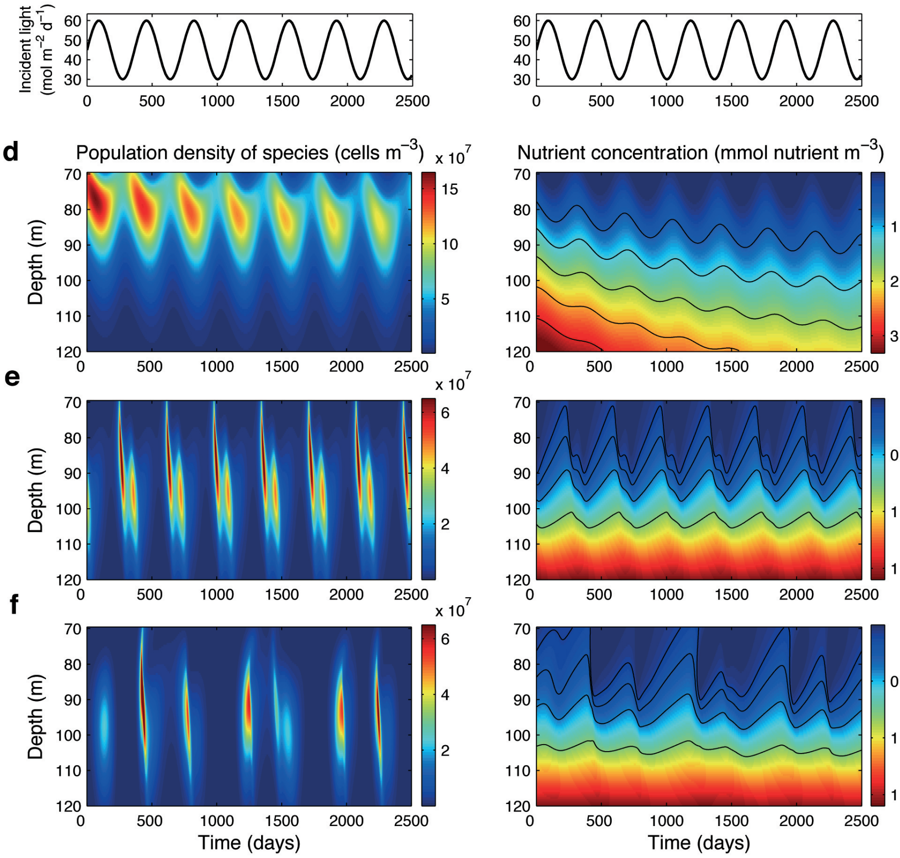


FIGURE 5.2: **d**-**f**, Seasonal environment, in which the model is forced by seasonal changes in incident light intensity [47]. **d**, DCM tracks seasonal variability ($\kappa = 0.50\,\mathrm{cm}^2\,\mathrm{s}^{-1}$). **e**, Double periodicity of DCM locked in a seasonal environment ($\kappa = 0.14\,\mathrm{cm}^2\,\mathrm{s}^{-1}$). **f**, Chaotic DCM in a seasonal environment ($\kappa = 0.08\,\mathrm{cm}^2\,\mathrm{s}^{-1}$).

Thus, the oscillations become more pronounced if the time scale of sinking is fast compared to the time scale of the upward flux of nutrients.

5.3.2 Results for a seasonal environment

Detailed ocean time series indicate that seasonal changes in light conditions have a large effect on the dynamics of DCMs [47] (see also Fig. 5.1). To add more realism to the model, we therefore forced the model by seasonal changes in incident light intensity typical for the North Pacific subtropical gyre [47],

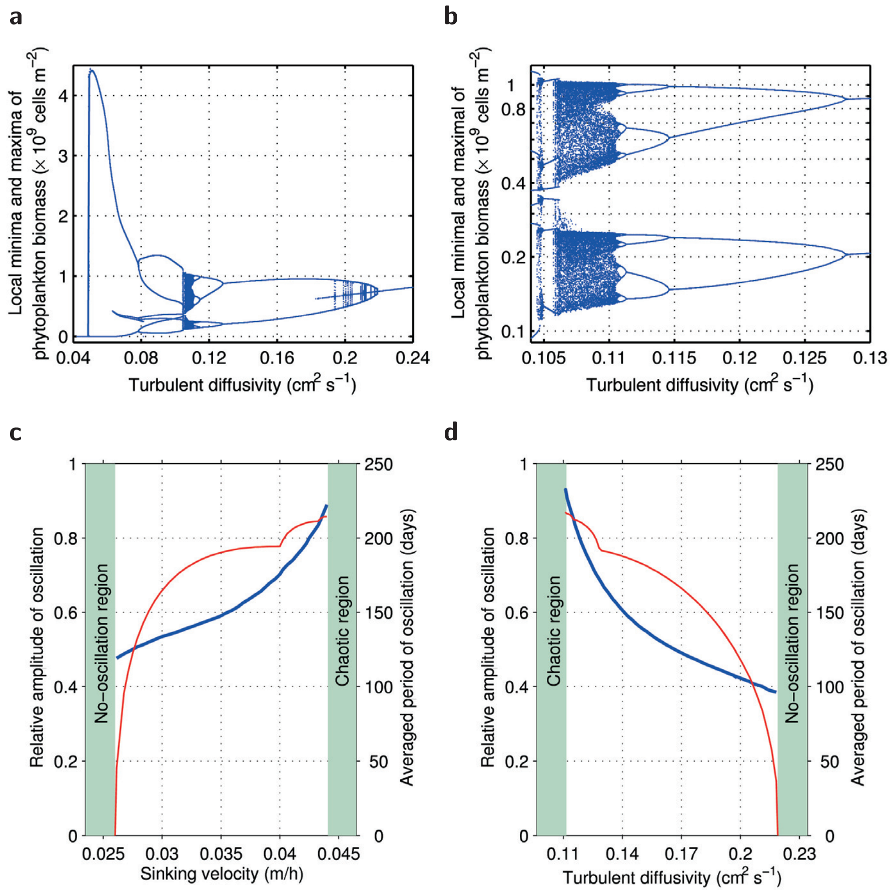


FIGURE 5.3: Bifurcation patterns generated in a constant environment. **a**, Bifurcation diagram showing the local minima and maxima of the phytoplankton population as a function of turbulent diffusivity. **b**, Detail of the chaotic region in the bifurcation diagram. **c**, The period (blue line) and relative amplitude (red line) of the oscillations increase with phytoplankton sinking velocity. **d**, The period (blue line) and relative amplitude (red line) of the oscillations decrease with vertical turbulent diffusivity. In **a** and **b** phytoplankton population density is integrated over the upper 300 m of the water column. See Section 5.5, Supplementary Information, for parameter values.

with a winter minimum of 30 mol photons $\mathrm{m}^{-2}\,\mathrm{d}^{-1}$ and a summer maximum of 60 mol photons $\mathrm{m}^{-2}\,\mathrm{d}^{-1}$. At high turbulent diffusivity, the DCM tracks the seasonal changes in light conditions (Fig. 5.2d). When turbulent diffusivity is reduced, the DCM exhibits a phenomenon known as phase locking, in which oscillations are squeezed within the seasonal cycle (Fig. 5.2e). For even lower turbulent diffusivities, seasonal forcing generates irregular phytoplankton blooms with chaotic multi-annual variability (Fig. 5.2f). Thus, similar to findings for other nonlinear oscillators [58, 64], fluctuating DCMs show even more complex dynamics in a seasonal environment than in a constant environment.

5.3.3 Multispecies in competition context

In reality, DCMs consist of multiple phytoplankton species with different growth rates, nutrient and light requirements, and sinking velocities. How would such a diverse assemblage respond to fluctuations in the DCM? To address this issue, we developed a multi-species version of our DCM model, analogous to earlier phytoplankton competition models [32, 33]. The model is again forced by seasonal changes in incident light intensity. An example is shown in Fig. 5.4, where we assume that the blue species has a lower sinking velocity ($0.1\,\mathrm{m\,d^{-1}}$; resembling pico- and nanoplankton) than the red and green species ($1\,\mathrm{m\,d^{-1}}$; resembling sinking diatoms). Furthermore, the blue species is a better nutrient competitor, whereas the red and green species are better competitors for light. Simulations show that all three species persist in this non-equilibrium environment, which confirms earlier notions that oscillations and chaos promote phytoplankton biodiversity [34]. Periods with co-dominance of the three species are alternated with periods in which either the blue species or the red and green species dominate (Fig. 5.4e). Furthermore, there is a subtle but consistent vertical zonation, with the blue species (better nutrient competitor) inhabiting the nutrient-depleted upper zone of the DCM, while the red and green species (superior light competitors) peak several meters deeper in the light-deprived part of the DCM. The model predicts that phytoplankton species with relatively high sinking velocities (red and green species) show larger fluctuations than small phytoplankton species with low sinking velocities (blue species; Fig. 5.4c-e).

5.4 Discussion

5.4.1 Conclusions

Although simple models can offer only abstractions of real-world phenomena, our model adequately reproduces many features of real-world DCMs. First, the model predicts that DCMs form at a similar depth of $\sim 100\,\mathrm{m}$ and span a similar depth range as observed in clear oceans waters [43] (Figs. 5.1, 5.2). Second, consistent with observations, the model predicts that nutrients are depleted to near-zero levels above the DCM while the nutrient concentration increases linearly with depth below the DCM [43] (Fig. 5.4e). Third, detailed ocean time-series measurements from the subtropical North Pacific confirm the prediction of a vertical zonation of species, with different species assemblages dominating at different depths [67] (see Fig. 5.5, in Section 5.5, Supplementary Information). Fourth, these ocean time series confirm the prediction that the seasonal light cycle gives rise to seasonal patterns in chlorophyll and nutrient concentrations in the DCM [47] (Fig. 5.1). Fifth, the time series support the idea that plankton populations in the DCM show additional fluctuations superimposed upon

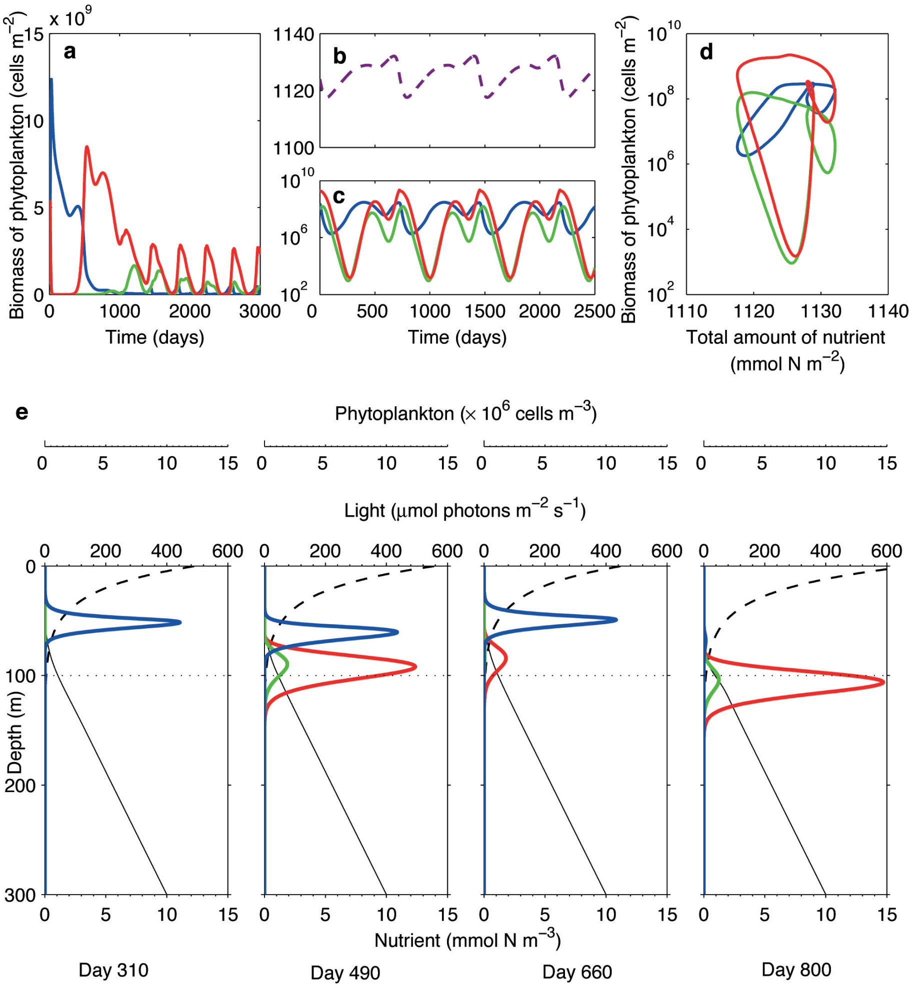


FIGURE 5.4: Competition between 3 phytoplankton species in an oscillating DCM. The model (with $\kappa = 0.12\,\mathrm{cm}^2\,\mathrm{s}^{-1}$) is forced by the same seasonal changes in incident light intensity as in Figs. 5.2d-f. **a**, Initial time course of the phytoplankton species. **b**-**c**, On the long run, the nutrient concentration (**b**) and the phytoplankton species (**c**) settle at a periodic attractor. **d**, Phase plane illustrating the periodic attractor of the phytoplankton species. **e**, Time series of consecutive depth profiles within a single period. Coloured lines: depth profiles of the 3 phytoplankton species; dashed line: light intensity; black line: nutrient concentration. In **a**-**d** phytoplankton population density and nutrient concentration are integrated over the upper $300\,\mathrm{m}$ of the water column. See Section 5.5, Supplementary Information, for parameter values.

the seasonal cycle, often with multi-annual variability in phytoplankton biomass and species composition [47, 41, 67] (see Section 5.5, Supplementary Information). Sixth, as predicted by the model, the time series tentatively suggest that phytoplankton species with relatively high sinking velocities show larger variability than phytoplankton species with low sinking velocities (see Section 5.5, Supplementary Information). In total, time-series data support the theoretical prediction that deep chlorophyll maxima can exhibit sustained non-equilibrium dynamics, driven by a combination of external forces and the complex internal dynamics of DCMs.

5.4.2 Epilogue

Climate models predict that global warming will increase the stability of the vertical stratification in large parts of the oceans [59, 6]. This will reduce vertical mixing and suppress the upward flux of nutrients, leading to a decline in oceanic primary production [6, 60, 61]. Our model predicts that the same process of reduced vertical mixing may induce oscillations and chaos in the phytoplankton of the DCM, generated by the difference in time scale between the sinking flux of phytoplankton and the upward flux of nutrients. Thus, counter-intuitively, increased stability of the water column due to global warming may destabilize the phytoplankton dynamics in the DCM, with implications for oceanic primary production, species composition and carbon export.

5.5 Supplementary Information

5.5.1 Introduction

In the Supplementary Information, we provide detailed information on (1) the model simulations, (2) the competition model, and (3) fluctuations in phytoplankton species composition in a long-term time series of the deep chlorophyll maximum in the North Pacific subtropical gyre.

5.5.2 Model simulations

We consider a vertical water column of one unit surface area. Let z indicate the depth in the water column, where z runs from 0 at the surface to a maximum depth z_B at the bottom. Let P denote the phytoplankton population density (number of cells per m^3), and let N denote the nutrient concentration in the water column. The dynamics of the phytoplankton population and the nutrient concentration in the water column is described by a system of two partial differential equations [19, 25, 43, 29, 32, 55]:

$$\frac{\partial P}{\partial t} = \mu(N,I)P - mP - v\frac{\partial P}{\partial z} + \kappa\frac{\partial^2 P}{\partial z^2}, \tag{5.3}$$

$$\frac{\partial N}{\partial t} = -\alpha\,\mu(N,I)P + \epsilon\,\alpha\,mP + \kappa\frac{\partial^2 N}{\partial z^2}, \tag{5.4}$$

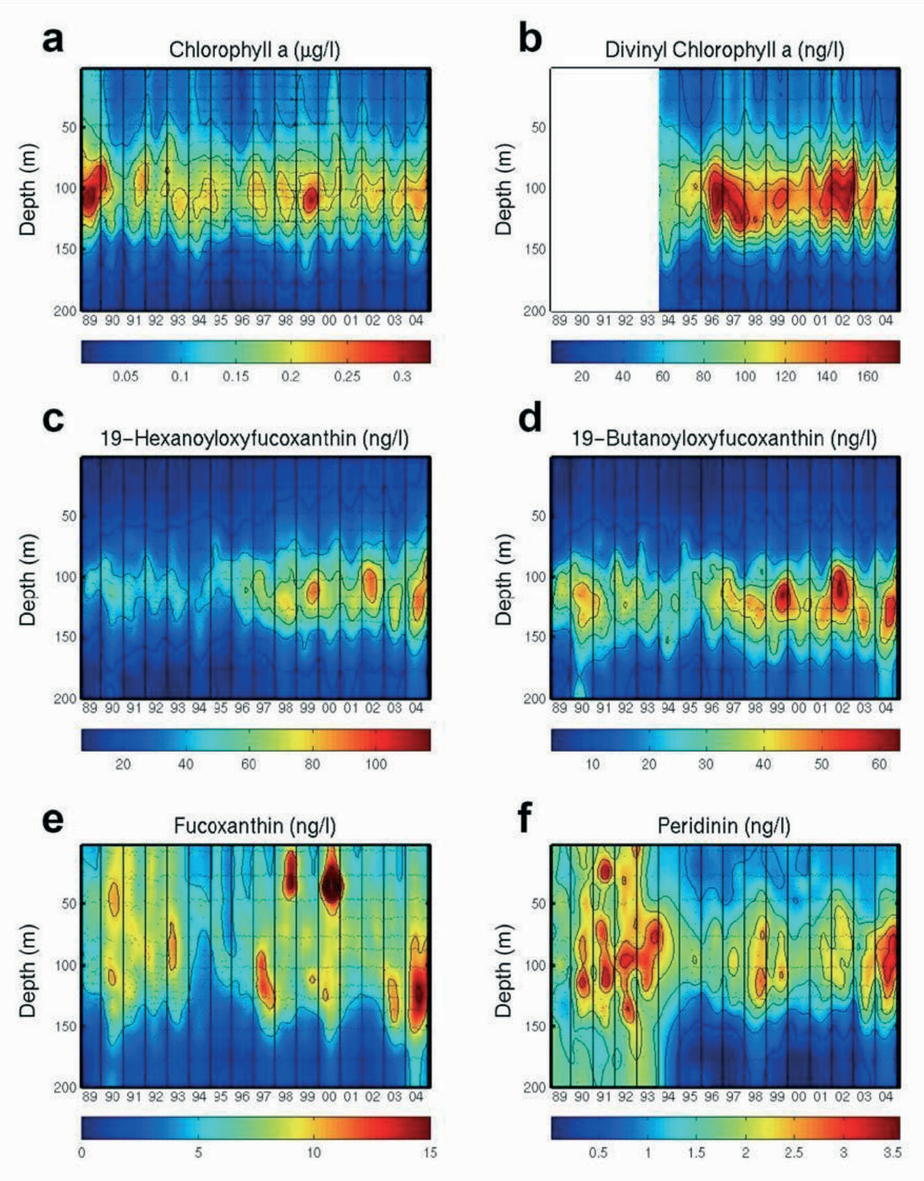


FIGURE 5.5: Time series of different algal pigment biomarkers at Station ALOHA, in the subtropical Pacific Ocean north of Hawaii. The pigment biomarkers are diagnostic for different phytoplankton groups [46, 50, 2]: **a**, total chlorophyll *a* (all phytoplankton). **b**, divinyl chlorophyll *a* (*Prochlorococcus*), which has been measured from 1994 onwards. **c**, 19'-hexanoyloxyfucoxanthin (prymnesiophytes). **d**, 19'-butanoyloxyfucoxanthin (pelagophytes). **e**, fucoxanthin (diatoms). **f**, peridinin (dinoflagellates). Data were obtained from the Hawaii Ocean Time-series (HOT) program, and are publicly available at `http://hahana.soest.hawaii.edu/hot/hot-dogs`.

where $\mu(N, I)$ is the specific growth rate of the phytoplankton as an increasing saturating function of nutrient availability N and light intensity I, m is the specific loss rate of the phytoplankton, v is the phytoplankton sinking velocity, κ is the vertical turbulent diffusivity, α is the nutrient content of the phytoplankton, and ϵ is the proportion of nutrient in dead phytoplankton that is recycled.

We assume that the specific growth rate of the phytoplankton follows the Monod equation [30], and is determined by the resource that is most limiting according to Von Liebig's 'law of the minimum' [69]:

$$\mu(N, I) = \mu_{\max} \min\Big(\frac{N}{H_N + N}, \frac{I}{H_I + I}\Big), \tag{5.5}$$

where $\mu_{\max}$ is the maximum specific growth rate, H_N and H_I are the half-saturation constants for nutrient-limited and light-limited growth, respectively, and min denotes the minimum function. We note that our findings are robust. We found similar results for other formulations of the specific growth rate (e.g., multiplicative functions).

Light intensity, I, is supplied from above and decreases exponentially with depth according to Lambert-Beer's law [29, 32]:

$$I = I_{in} \exp\Big(-K_{bg} z - k \int_0^z P(t, \sigma) d\sigma\Big), \tag{5.6}$$

where I_{in} is the incident light intensity, K_{bg} is the background turbidity of the water column, k is the specific light absorption coefficient of the phytoplankton, and σ is an integration variable accounting for the non-uniform phytoplankton population density distribution. To complete the model, we assume zero-flux boundary conditions for the phytoplankton. Furthermore, we assume a zero-flux boundary condition for nutrient at the surface, while nutrient is replenished from below with a fixed concentration N_B at the bottom of the water column.

Numerical approach. The integral in (5.6) introduces a nonlocal term in the model. As a result, the model is a system of integro-partial differential equations (integro-PDEs), which is computationally quite demanding. Numerical simulation of the model was based on a finite volume method, with spatial discretisation of the differential operators as well as the integral term. The advection terms were discretised by a third-order upwind biased formula, the diffusion terms by a symmetric second-order formula, and the integral term by the repeated trapezoidal rule [38]. The resulting system of stiff ordinary differential equations was integrated over time using an implicit integration method [8] implemented in the computer code VODE (`http://www.netlib.org/ode/`). A detailed presentation of our simulation techniques, with tests of the accuracy and numerical stability of the simulations, is presented elsewhere (see Chapter 1 and [33]).

The model was parameterized for clear ocean water [42, 47], with realistic turbulent diffusivities [48, 63, 20], and growth kinetics typical for nutrient-limited and light-limited phytoplankton [19, 25, 43, 29, 32]. An overview of the parameter values used in the simulations, together with their units and interpretation, is given in Table 5.1.

5.5.3 Competition model

The multi-species version of our DCM model, used in Fig. 5.4, is a straightforward extension of the single-species model outlined above [32, 33]. Suppose that a total number of n phytoplankton species compete for nutrients and light in the DCM. We assume that each species has its own growth and loss characteristics, its own sinking velocity, and its own nutrient content. The species interact only indirectly, via their shared resources. That is, all phytoplankton species consume nutrient and absorb light. Hence, the population dynamics of the different species and the nutrient dynamics in the water column are described as:

$$\begin{aligned} \frac{\partial P_i}{\partial t} &= \mu_i(N,I)P_i - m_i P_i - v_i\frac{\partial P_i}{\partial z} + \kappa\frac{\partial^2 P_i}{\partial z^2}, \quad i = 1,\cdots,n, && (5.7)\\ \frac{\partial N}{\partial t} &= -\sum_{j=1}^{n}\alpha_j\,\mu_j(N,I)P_j + \sum_{j=1}^{n}\epsilon_j\,\alpha_j\,m_j P_j + \kappa\frac{\partial^2 N}{\partial z^2}, && (5.8)\end{aligned}$$

Furthermore, the vertical light gradient is now described as:

$$I = I_{in}\exp\Big(-K_{bg}z - \sum_{j=1}^{n}k_j\int_0^z P_j(t,\sigma)d\sigma\Big). \qquad (5.9)$$

All other aspects of the multi-species model are identical to the single-species model described above.

5.5.4 Fluctuations in phytoplankton species composition

It is interesting to compare the theoretical predictions with long-term time series of deep chlorophyll maxima. For nearly two decades, scientists in the Hawaii Ocean Time-series (HOT) program have collected physical, chemical, and biological measurements of the DCM in the North Pacific subtropical gyre [47, 46, 40, 41]. Samples were taken at approximately monthly intervals at Station ALOHA, North of Hawaii (22° 45'N, 158° 00'W). The sampling and measurement protocols are described in full detail at `http://hahana.soest.hawaii.edu`, and the time-series data are publicly available at `http://hahana.soest.hawaii.edu/hot/hot-dogs`.

Among the key parameters collected at Station ALOHA, phytoplankton community composition is routinely estimated using a combination of flow cytometry [54, 9] and algal pigment biomarkers [46, 40]. The latter relies on the fact that various algal groups have diagnostic pigment markers. These diagnostic pigments, measured with high-performance liquid chromatography (HPLC)

TABLE 5.1: Parameter values and their interpretation

Symbol	Interpretation	Units	Value
Independent variables			
t	Time	h	–
z	Depth	m	–
Dependent variables			
P	Population density	$\mathrm{cells\,m^{-3}}$	–
I	Light intensity	$\mu\mathrm{mol\,photons\,m^{-2}\,s^{-1}}$	–
N	Nutrient concentration	$\mathrm{mmol\,nutrient\,m^{-3}}$	–
Parameters			
I_{in}	Incident light intensity	$\mu\mathrm{mol\,photons\,m^{-2}\,s^{-1}}$	600
K_{bg}	Background turbidity	$\mathrm{m^{-1}}$	0.045
k	Absorption coefficient of phytoplankton	$\mathrm{m^2\,cell^{-1}}$	6×10^{-10}
z_B	Depth of the water column	m	300
κ	Vertical turbulent diffusivity	$\mathrm{cm^2\,s^{-1}}$	0.12
$\mu_{\max}$	Maximum specific growth rate	$\mathrm{h^{-1}}$	0.04
H_I	Half-saturation constant of light-limited growth*	$\mu\mathrm{mol\,photons\,m^{-2}\,s^{-1}}$	20
H_N	Half-saturation constant of nutrient-limited growth*	$\mathrm{mmol\,nutrient\,m^{-3}}$	0.025
m	Specific loss rate	$\mathrm{h^{-1}}$	0.01
α	Nutrient content of phytoplankton	$\mathrm{mmol\,nutrient\,cell^{-1}}$	1×10^{-9}
ϵ	Nutrient recycling coefficient	dimensionless	0.5
v	Sinking velocity*	$\mathrm{m\,h^{-1}}$	0.042
N_B	Nutrient concentration at z_B	$\mathrm{mmol\,nutrient\,m^{-3}}$	10

*Note: Figure 5.4 uses different sinking velocities and half-saturation constants for different species. Red species: $v = 0.042\,\mathrm{m\,h^{-1}}$, $H_I = 20\,\mu\mathrm{mol\,photons\,m^{-2}\,s^{-1}}$, $H_N = 0.0425\,\mathrm{mmol\,nutrient\,m^{-3}}$; green species: $v = 0.042\,\mathrm{m\,h^{-1}}$, $H_I = 25\,\mu\mathrm{mol\,photons\,m^{-2}\,s^{-1}}$, $H_N = 0.0165\,\mathrm{mmol\,nutrient\,m^{-3}}$; blue species: $v = 0.0042\,\mathrm{m\,h^{-1}}$, $H_I = 98\,\mu\mathrm{mol\,photons\,m^{-2}s^{-1}}$, $H_N = 0.0150\,\mathrm{mmol\,nutrient\,m^{-3}}$.

[71], can be used to develop an algorithm attributing total chlorophyll into specific taxonomic groups [46, 50]. The predictions of the algorithm have been compared to direct electron microscopic analyses with generally good correspondence [2]. We note that a novel algal group (pelagophytes [2]) known to contain 19'-butanoyloxyfucoxanthin was misidentified as "chrysophytes" in earlier publications [46] from Station ALOHA. The DCM at Station ALOHA is dominated by *Prochlorococcus*, which averaged over the years made up nearly 40% of the total chlorophyll, followed by other cyanobacteria (24%), prymnesiophytes (22%), and pelagophytes (13%) [46]. Diatoms and dinoflagellates also occur, but are relatively less abundant.

Variations in specific pigment biomarkers have been used to track taxonomic changes in the phytoplankton assemblage at Station ALOHA. Chlorophyll *a*, which is present in all phytoplankton groups, has the highest concentrations in a depth range of 80 to 120 m. It displays both seasonal and inter-annual variation (Fig. 5.5a). Divinyl chlorophyll *a* has been measured since 1994, and is characteristic of *Prochlorococcus*, the most abundant phytoplankton group at Station ALOHA. *Prochlorococcus* was highly abundant in the DCM during the years 1996-2002, but less abundant in 1994-1995 and 2004 (Fig. 5.5b). The pigments 19'-hexanoyloxyfucoxanthin and 19'butanoyloxyfucoxanthin reach the highest concentrations in the 100 to 140 m depth range, and show seasonal, inter-annual, as well as inter-decadal variation (Fig. 5.5c,d). These two pigments typify the prymnesiophytes and pelagophytes, respectively. They were particularly abundant in the years 1999, 2002 and 2004. Diatoms are characterized by the pigment fucoxanthin. Diatoms were abundant in the DCM in the years 1997 and 2004, and near the surface in the years 1998 and 2000, whereas diatoms were quite rare in the years 1994-1996 (Fig. 5.5e). Dinoflagellates, characterized by peridinin, displayed a broad vertical distribution in the years 1990-1992, but a more narrow distribution over the depth range of 80 to 120 m in 1993, 1998, and 2004 (Fig. 5.5f). They were less abundant in 1995-1996 and in 2000.

These time series confirm the presence of vertical zonation in key phytoplankton groups (prediction 3 in Section 5.4.1). Furthermore, the time series also clearly demonstrate both seasonal and multi-annual variability in phytoplankton abundance and phytoplankton species composition (predictions 4 and 5 in Section 5.4.1). Finally, the time series tentatively suggest that phytoplankton species with relatively high sinking velocities, like the diatoms and dinoflagellates (Fig. 5.5e,f), display larger variability than small phytoplankton species with relatively low sinking velocities (Fig. 5.5b-d; prediction 6 in Section 5.4.1), although longer time series will be required to investigate this model prediction in full detail. Thus, at least qualitatively, several of the model predictions are supported by detailed observations of the dominant phytoplankton species in the subtropical North Pacific.

Chapter 6

Analysis of phytoplankton blooming

6.1 Introduction

Based on the interesting results found in the previous chapter, it is of great importance to understand for which parameter regimes we may expect no bloom, stationary bloom, or oscillations. An analytical study will be performed in this chapter for a slightly simplified version of the light-nutrient model studied in the preceding chapter. It turns out that it is possible to derive an analytical expression to distinguish between bloom- and nobloom regions in the parameter space. Additional numerical simulations show the validity of this analytical approach.

6.2 Statement of the problem

We consider the PDE problem

$$W_t = D\,W_{zz} - v\,W_z + [\,\mu\,P(L,N) - l\,]\,W, \tag{6.1a}$$

$$N_t = D\,N_{zz} - \alpha\,\mu\,P(L,N)\,W, \tag{6.1b}$$

for all $(t,z) \in \mathbb{R}_+ \times [0, z_B]$, where $z_B > 0$ is a constant, equipped with the boundary conditions

$$D\,W_z - v\,W|_{z=0,z_B} = 0, \tag{6.2a}$$

$$N_z|_{z=0} = 0, \tag{6.2b}$$

$$N|_{z=z_B} = N_B. \tag{6.2c}$$

Here, $W(z,t)$ and $N(z,t)$ are the concentrations, at depth z and time t, of the phytoplankton population and the nutrient, respectively, D is the diffusion coefficient (taken to be identical for both phytoplankton and nutrient because we work in the turbulent mixing regime), v and l are the sinking speed and specific loss rate, respectively, of phytoplankton (both of which are assumed to be positive), $\alpha > 0$ is a constant, z_B is the depth of the ocean bed, N_B is the constant concentration of nutrient at the bottom of the ocean, and μ denotes the maximal growth rate. Also, $L(z,t)$ is the light intensity at depth z, given by the formula

$$L(z,t) = L_I e^{-K_{bg} z - k \int_0^z W(\zeta,t)d\zeta}, \tag{6.3}$$

where L_I is the intensity of the incident light at the water surface, K_{bg} is the light absorption coefficient due to the non-phytoplankton components in the water column, and k is the light absorption coefficient due to the phytoplankton. Finally, the function $P(L,N)$ is taken to be

$$P(L,N) = \frac{LN}{(L+L_H)(N+N_H)}, \tag{6.4}$$

where L_H and N_H are the half-saturation constants of light and nutrient, respectively. Here, we remark that the function P differs from the corresponding 'minimum' function that we used in Chapters 2 and 5. The choice (6.4) is dictated by our intention to perform analytic calculations on the light-nutrient model. In particular, the two functions have the same qualitative behavior, but the function given in (6.4) has the advantage that it is a smooth function of its arguments.

6.3 Nondimensionalization

It is convenient to recast our model in nondimensional variables. First, we let

$$\begin{aligned} \omega &= W/[W], \\ \tau &= t/[t], \\ x &= z/[z], \\ j &= L/[L], \\ \eta &= N/[N], \\ \ell &= l/\mu, \end{aligned}$$

where $[W]$, $[t]$, $[z]$, $[L]$, and $[N]$ are constants to be chosen appropriately. Next, we introduce the notation

$$p(j,\eta) = P([L]j,[N]\eta) = \frac{j\eta}{(j+j_H)(\eta+\eta_H)},$$

where

$$j_H = \frac{L_H}{[L]} \quad \text{and} \quad \eta_H = \frac{N_H}{[N]}.$$

In terms of the new variables the PDEs (6.1) now read

$$\omega_\tau = \frac{D[t]}{[z]^2}\omega_{xx} - \frac{v[t]}{[z]}\omega_x + \mu\,[t]\,\left(p(j,\eta) - \ell\right)\omega, \tag{6.5a}$$

$$\eta_\tau = \frac{D[t]}{[z]^2}\eta_{xx} - \frac{\alpha\mu[t][W]}{[N]}p(j,\eta)\omega. \tag{6.5b}$$

The boundary conditions (6.2) become

$$\begin{aligned}\left(\omega_x - \frac{v[z]}{D}\omega\right)(0) = \left(\omega_x - \frac{v[z]}{D}\omega\right)\left(\frac{z_B}{[z]}\right) &= 0, && (6.6a)\\ \eta_x(0) &= 0, && (6.6b)\\ \eta\left(\frac{z_B}{[z]}\right) &= \frac{N_B}{[N]}. && (6.6c)\end{aligned}$$

We make the following choice of parameters:

$$\begin{aligned}[W] &= \frac{DN_B}{l\alpha z_B^2},\\ [t] &= 1/\mu,\\ [z] &= z_B,\\ [L] &= L_I,\\ [N] &= N_B,\end{aligned}$$

and we identify the small parameter $\varepsilon = D/(\mu z_B^2)$. Then, Eqs. (6.5) become

$$\begin{aligned}\omega_\tau &= \varepsilon\omega_{xx} - a\,\omega_x + (p(j,\eta) - \ell)\omega, && (6.7a)\\ \eta_\tau &= \varepsilon\left(\eta_{xx} - \frac{1}{\ell}p(j,\eta)\omega\right), && (6.7b)\end{aligned}$$

where $a = v/(\mu z_B)$ is a constant the asymptotic magnitude of which will be tuned in the next sections. The boundary conditions read

$$\begin{aligned}(\varepsilon\omega_x - a\,\omega)(0) = (\varepsilon\omega_x - a\,\omega)(1) &= 0, && (6.8a)\\ \eta_x(0) &= 0, && (6.8b)\\ \eta(1) &= 1. && (6.8c)\end{aligned}$$

Finally, we remark that

$$j(x,\tau) = e^{-\kappa x - r\int_0^x \omega(\chi,\tau)d\chi}, \quad \text{where} \quad \kappa = K_{bg}z_B \quad \text{and} \quad r = \frac{kDN_B}{l\alpha z_B}. \qquad (6.9)$$

6.4 Eigenvalues of the local problem

The PDE model (6.7), equipped with the boundary conditions (6.8), has the steady state

$$\bar{\omega}(x,\tau) = 0, \quad \bar{\eta}(x,\tau) = 1, \quad \text{for all} \quad (\tau, x) \in \mathbb{R}_+ \times [0,1].$$

Bloom occurs for those parameter values for which this steady state becomes *unstable*. In the rest of this chapter, we identify this regime by looking at the eigenvalues associated with this steady state and determining when they lead to an unstable situation.

The linear eigenvalue problem around the steady state (with eigenvalue λ) reads

$$\begin{aligned} \lambda\omega &= \varepsilon\omega_{xx} - a\,\omega_x + (f(x) - \ell)\omega, & (6.10a) \\ \lambda\eta &= \varepsilon\left(\eta_{xx} - \frac{1}{\ell}f(x)\omega\right), & (6.10b) \end{aligned}$$

where the decreasing function

$$f(x) = \Phi\frac{e^{-\kappa x}}{e^{-\kappa x} + j_H} = \frac{\Phi}{1 + j_H e^{\kappa x}}, \quad \text{with} \quad \Phi = \frac{1}{1 + \eta_H}, \tag{6.11}$$

is the linearization of the function $p(j, \eta)$ around the steady state. The corresponding boundary conditions remain unchanged, see (6.8).

Noticing that Eq. (6.10a) for ω does not depend on η, we obtain the associated eigenvalue problem for the rescaled phytoplankton density,

$$\varepsilon\,\omega'' - a\,\omega' + (f(x) - \ell - \lambda)\omega = 0, \tag{6.12}$$

subject to the boundary conditions

$$(\varepsilon\omega' - a\,\omega)\,(0) = (\varepsilon\omega' - a\,\omega)\,(1) = 0. \tag{6.13}$$

This is a singularly perturbed problem[1]. We now proceed with determining the solutions of the problem for different scalings of the coefficient a.

6.5 Case $a = \mathcal{O}(1)$

First, we treat the case $a = \mathcal{O}(1)$ and show that, in this case, the trivial steady state is *stable* for all $\mathcal{O}(1)$ values of the parameters, except for a bifurcation which gives rise to bottom-blooming, rather than a DCM (= Deep Chlorophyll Maximum, see Chapter 5) that we are looking for.

6.5.1 The scaling $\lambda = \mathcal{O}(1)$

First, we assume that the eigenvalues of the problem are $\mathcal{O}(1)$ and we determine these eigenvalues. Based on physical considerations, we postulate that the boundary layer is located at $x = 1$ (a quick calculation verifies the validity of this assumption). In order to find a solution of this type, we use the boundary-layer theory which requires to match the slowly varying outer solution (i.e., the one far away from the boundary-layer) and the rapidly varying inner solution (i.e., the one close to the boundary-layer).

---

[1] That is, a problem where the highest derivative (i.e., of highest order) is multiplied by a small constant ε.

Outer solution

To leading order, the outer solution $\Omega(x) = \sum_{n=0}^{\infty} \varepsilon^n \Omega_n(x)$ satisfies the ODE

$$-a\,\Omega_0' + (f(x) - \ell - \lambda)\Omega_0 = 0, \tag{6.14}$$

subject to the boundary condition

$$\Omega_0(0) = 0. \tag{6.15}$$

Eq. (6.14) is exactly solvable (see the Appendix at the end of this chapter),

$$\Omega_0(x) = C_0 \frac{e^{-\frac{\ell+\lambda}{a}x}}{\left(j_H + e^{-\kappa x}\right)^{\frac{1}{a\kappa(\eta_H+1)}}}, \tag{6.16}$$

where C_0 is a free parameter. The only solution that satisfies the boundary condition (6.15) is the trivial one, $\Omega_0 \equiv 0$.

In fact, it is easy to show inductively that the solution at $\mathcal{O}(\varepsilon^n)$, Ω_n, is identically zero for all $n \in \mathbb{N}$. Indeed, Ω_n satisfies the ODE

$$\Omega_{n-1}'' - a\,\Omega_n' + (f(x) - \ell - \lambda_0)\Omega_n - \sum_{m=1}^{n} \lambda_m \Omega_{n-m} = 0 \tag{6.17}$$

and is subject to the boundary condition

$$\Omega_{n-1}(0) - a\,\Omega_n(0) = 0. \tag{6.18}$$

Since $\Omega_m \equiv 0$, for all $0 \le m \le n-1$, by the induction hypothesis, Eqs. (6.17) and (6.18) reduce to

$$-a\,\Omega_n' + (f(x) - \ell - \lambda_0)\Omega_n = 0$$

and

$$\Omega_n(0) = 0.$$

Thus, Ω_n satisfies the exact same equation and boundary condition as Ω_0, and our above analysis shows that $\Omega_n \equiv 0$. Hence, $\Omega(x) = 0$ for all $x \in [0,1]$.

Inner Solution

Next, we determine the inner solution $\omega(x) = \sum_{n=0}^{\infty} \varepsilon^n \omega_n(x)$ to problem (6.12)–(6.13). First, we change the independent variable to $s = (1-x)/\varepsilon$, to deal with the rapidly varying solution close to $x = 1$. Then, Eq. (6.12) becomes

$$\ddot{\omega} + a\,\dot{\omega} + \varepsilon(\bar{f}(s) - \ell - \lambda)\omega = 0, \tag{6.19}$$

where $\dot{(\cdot)} = d(\cdot)/ds$, subject to the boundary condition

$$(\dot{\omega} + a\,\omega)(0) = 0 \tag{6.20}$$

and where

$$\bar{f}(s) = f(1 - \varepsilon s). \tag{6.21}$$

Expanding $f(1 - \varepsilon s)$ in powers of ε, we obtain

$$\bar{f}(s) = f(1 - \varepsilon s) = f(1) - \varepsilon f'(1)s + \mathcal{O}(\varepsilon^2). \tag{6.22}$$

First, we deal with the leading order problem. The last term on the left hand side of Eq. (6.19) is of higher order with respect to ε. Hence, we obtain, at $\mathcal{O}(1)$, the equation

$$\ddot{\omega}_0 + a\,\dot{\omega}_0 = 0 \tag{6.23}$$

together with the boundary condition

$$(\dot{\omega}_0 + a\,\omega_0)\,(0) = 0. \tag{6.24}$$

Equations (6.23)–(6.24) imply that

$$\dot{\omega}_0 + a\,\omega_0 = 0,$$

whence

$$\omega_0(s) = C_0 e^{-as}, \tag{6.25}$$

where C_0 is a free parameter. Since $a > 0$, all solutions of this form satisfy the matching condition

$$\lim_{s\to\infty} \omega_0(s) = 0 = \lim_{x\to 1} \Omega_0(x), \tag{6.26}$$

and thus no eigenvalues arise at this point.

Next, we look at the $\mathcal{O}(\varepsilon)$ terms of Eqs. (6.19)–(6.20). We obtain

$$\ddot{\omega}_1 + a\,\dot{\omega}_1 + (f(1) - \ell - \lambda_0)\omega_0 = 0 \tag{6.27}$$

and the boundary condition

$$(\dot{\omega}_1 + a\,\omega_1)\,(0) = 0. \tag{6.28}$$

The general solution to this problem is (see the Appendix)

$$\omega_1(s) = C_0 \frac{f(1) - \ell - \lambda_0}{a} \left(s e^{-as} - \frac{1}{a} \right) + C_1 e^{-as}. \tag{6.29}$$

The matching condition

$$\lim_{s\to\infty} \omega_1(s) = 0 = \lim_{x\to 1} \Omega_1(x)$$

implies, then, that

$$C_0 \frac{f(1) - \ell - \lambda_0}{a^2} = 0,$$

whence to leading order, we obtain the eigenvalue

$$\lambda_0 = f(1) - \ell. \tag{6.30}$$

To obtain the $\mathcal{O}(\varepsilon)$ correction to λ, we look at the $\mathcal{O}(\varepsilon^2)$ terms of Eq. (6.19),

$$\ddot{\omega}_2 + a\,\dot{\omega}_2 - (f'(1)s + \lambda_1)\omega_0 = 0, \tag{6.31}$$

where we have used Eq. (6.30) to eliminate the term $(f(1) - \ell - \lambda_0)\omega_1$. The associated boundary condition is

$$(\dot{\omega}_2 + a\,\omega_2)\,(0) = 0. \tag{6.32}$$

The general solution to problem (6.31)–(6.32) is (see the Appendix)

$$\begin{aligned}\omega_2(s) = C_0 &\left[-\frac{f'(1)}{2a}s^2 e^{-as} - \frac{f'(1) + a\lambda_1}{a^2} s e^{-as} + \frac{f'(1) + a\lambda_1}{a^3}\right] \\ &+ C_2 e^{-as}.\end{aligned} \tag{6.33}$$

The matching condition

$$\lim_{s\to\infty} \omega_2(s) = 0 = \lim_{x\to 1} \Omega_2(x)$$

implies, then, that

$$C_0 \frac{f'(1) + a\,\lambda_1}{a^3} = 0,$$

whence we obtain

$$\lambda_1 = -\frac{f'(1)}{a}. \tag{6.34}$$

Thus, the eigenvalue is

$$\lambda = f(1) - \ell - \varepsilon\frac{f'(1)}{a} + \mathcal{O}(\varepsilon^2). \tag{6.35}$$

The equation above shows that we have determined only one eigenvalue, $\lambda = f(1) - \ell - \varepsilon\frac{f'(1)}{a} + \mathcal{O}(\varepsilon^2)$, instead of an infinite set that is guaranteed to exist by Sturm-Liouville theory (see [5]). Thus, our working hypothesis $\lambda = \mathcal{O}(1)$ on the asymptotic magnitude of λ is insufficient as far as the determination of the entire spectrum is concerned.

6.5.2 The scaling $\lambda = \mathcal{O}(1/\varepsilon)$

After several trials with different scalings of λ, that is, $\lambda = \varepsilon^\gamma \Lambda$ with $\gamma \in \mathbb{R}$, it turns out that no eigenvalues arise, except for $\gamma = -1$. Therefore, we consider this only possible scaling $\lambda = \mathcal{O}(1/\varepsilon)$ for the eigenvalues. Writing

$$\lambda = \frac{1}{\varepsilon}\Lambda, \qquad \text{where} \quad \Lambda = \mathcal{O}(1),$$

we obtain from Eqs. (6.12)–(6.13) the eigenvalue problem

$$\begin{aligned} \varepsilon^2\,\omega'' - \varepsilon\, a\,\omega' - (\Lambda - \varepsilon(f(x)-\ell))\,\omega &= 0, \\ (\varepsilon\omega' - a\,\omega)\,(0) \;=\; (\varepsilon\omega' - a\,\omega)\,(1) &= 0. \end{aligned}$$

Performing the change of independent variable $\sigma = x/\varepsilon$ and writing $\dot{(\cdot)}$ for $d(\cdot)/d\sigma$, we obtain the eigenvalue ODE

$$\ddot{\omega} - a\,\dot{\omega} - (\Lambda - \varepsilon(f(\varepsilon\sigma)-\ell))\,\omega = 0 \tag{6.36}$$

subject to the boundary conditions

$$(\dot{\omega} - a\,\omega)\,(0) \;=\; (\dot{\omega} - a\,\omega)\,(1/\varepsilon) = 0. \tag{6.37}$$

Leading order solution to the eigenvalue problem

To leading order, Eqs. (6.36)–(6.37) yield

$$\begin{aligned} \ddot{\omega}_0 - a\,\dot{\omega}_0 - \Lambda_0\,\omega_0 &= 0, \\ (\dot{\omega}_0 - a\,\omega_0)\,(0) \;=\; (\dot{\omega}_0 - a\,\omega_0)\,(1/\varepsilon) &= 0. \end{aligned}$$

The general solution to the ODE above is

$$\omega_0(\sigma) = C_+ \mathrm{e}^{\mu_+\sigma} + C_- \mathrm{e}^{\mu_-\sigma}, \quad \text{where} \quad \mu_\pm = \frac{a \pm \sqrt{a^2 + 4\Lambda_0}}{2}. \tag{6.38}$$

Thus, also, we obtain the expression

$$(\dot{\omega}_0 - a\,\omega_0)\,(\sigma) = C_+(\mu_+ - a)\mathrm{e}^{\mu_+\sigma} + C_-(\mu_- - a)\mathrm{e}^{\mu_-\sigma} \tag{6.39}$$

for the quantity involved in the boundary conditions. Requiring that the boundary conditions are satisfied, we obtain the linear system

$$\begin{pmatrix} \mu_+ - a & \mu_- - a \\ (\,\mu_+ - a\,)\mathrm{e}^{\mu_+/\varepsilon} & (\,\mu_- - a\,)\mathrm{e}^{\mu_-/\varepsilon} \end{pmatrix} \begin{pmatrix} C_+ \\ C_- \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The necessary and sufficient condition for this system to have nonzero solutions is that

$$\det \begin{pmatrix} \mu_+ - a & \mu_- - a \\ (\,\mu_+ - a\,)\mathrm{e}^{\mu_+/\varepsilon} & (\,\mu_- - a\,)\mathrm{e}^{\mu_-/\varepsilon} \end{pmatrix} = -(\mu_+ - a)(\mu_- - a)\left(\mathrm{e}^{\mu_+/\varepsilon} - \mathrm{e}^{\mu_-/\varepsilon}\right) = 0.$$

Since $\Lambda_0 \neq 0$ (otherwise $\Lambda \ll 1/\varepsilon$, which contradicts our assumption on the asymptotic magnitude of Λ), this equation yields

$$\mathrm{e}^{(\mu_+ - \mu_-)/\varepsilon} = \mathrm{e}^{\sqrt{a^2+4\Lambda_0}/\varepsilon} = 1,$$

with solutions

$$\Lambda_0 = -\frac{a^2}{4} - \varepsilon^2 n^2 \pi^2, \qquad \text{for all} \quad n \in \mathbb{N}.$$

Thus, we obtain that

$$\lambda = -\frac{1}{\varepsilon}\frac{a^2}{4} + \mathcal{O}(\varepsilon), \tag{6.40}$$

and thus the trivial steady state is “super-stable” for all $\mathcal{O}(1)$ values of a.

6.5.3 Conclusions for the case $a = \mathcal{O}(1)$

In summary, the full set of eigenvalues is

$$\begin{aligned} \lambda_0 &= f(1) - \ell - \varepsilon \frac{f'(1)}{a} + \mathcal{O}(\varepsilon^2), \\ \lambda_n &= -\frac{1}{\varepsilon}\frac{a^2}{4} - \varepsilon n^2\pi^2, \quad n = 1, 2, \cdots . \end{aligned}$$

We see that λ_n, $n = 1, 2, \cdots$, are super-stable and hence no blooms associated with these eigenvalues may occur. Nevertheless, λ_0 may become unstable if $f(1) > \ell$. When $f(1) > \ell$, the profile arising from the instability is that of bottom-blooming, which is also not what we are looking for.

6.6 Case $a = \mathcal{O}(\sqrt{\varepsilon})$

The above calculations on the stability of the steady state suggest that we should take a closer look into the case $a = \mathcal{O}(\sqrt{\varepsilon})$, since a zero crossing of the eigenvalue of the first mode λ_1, and thus instability, is only possible in that parameter regime. To facilitate calculations, we let

$$a = \sqrt{\varepsilon}\, A, \qquad \text{where} \quad A = \mathcal{O}(1).$$

The eigenvalue problem (6.12)–(6.13) becomes, then,

$$\varepsilon\, \omega'' - \sqrt{\varepsilon}\, A\, \omega' + (f(x) - \ell - \lambda)\omega = 0, \tag{6.41}$$

subject to the boundary conditions

$$\left(\sqrt{\varepsilon}\, \omega' - A\, \omega\right)(0) = \left(\sqrt{\varepsilon}\, \omega' - A\, \omega\right)(1) = 0. \tag{6.42}$$

6.6.1 Transformation into Schrödinger form

First, we transform the dependent variable through

$$w(x) = \mathrm{e}^{-Ax/2\sqrt{\varepsilon}}\omega(x), \tag{6.43}$$

so that Eq. (6.41) can be written in Schrödinger form [5] where the first derivative term is absent. Using the transformation (6.43), we find

$$\begin{aligned} \omega'(x) &= \left(w'(x) + \frac{A}{2\sqrt{\varepsilon}} w(x)\right) \mathrm{e}^{Ax/2\sqrt{\varepsilon}}, \\ \omega''(x) &= \left(w''(x) + \frac{A}{\sqrt{\varepsilon}} w'(x) + \frac{A^2}{4\varepsilon} w(x)\right) \mathrm{e}^{Ax/2\sqrt{\varepsilon}}. \end{aligned}$$

Substitution into Eqs. (6.41)–(6.42) yields

$$\varepsilon\, w'' = Q(x)\, w, \qquad \text{where} \quad Q(x) = \lambda + \ell + \frac{A^2}{4} - f(x), \tag{6.44}$$

subject to the boundary conditions

$$\left(\sqrt{\varepsilon}\, w' - \frac{A}{2}\, w\right)(0) = \left(\sqrt{\varepsilon}\, w' - \frac{A}{2}\, w\right)(1) = 0. \tag{6.45}$$

Defining

$$\begin{aligned} \lambda_* &= f(0) - \ell - \frac{A^2}{4}, && (6.46a)\\ F(x) &= f(0) - f(x) && (6.46b) \end{aligned}$$

and writing

$$\lambda - \lambda_* = -\varepsilon^{\beta}\Lambda, \tag{6.47}$$

where the scaling parameter β is to be determined by our analysis, we recast $Q(x)$ in its final form,

$$Q(x) = F(x) - \varepsilon^{\beta}\Lambda, \quad \text{where} \quad F(x)\,,\, F'(x) \geq 0 \quad \text{for all} \quad x \in [0,1]. \tag{6.48}$$

6.6.2 WKB approximation

Since Q is an increasing function, we have

$$\min_{x\in[0,1]} Q(x) = -\varepsilon^{\beta}\,\Lambda < 0 \quad \text{and} \quad \max_{x\in[0,1]} Q(x) = F(1) - \varepsilon^{\beta}\,\Lambda > 0,$$

as long as $\Lambda > 0$.[2] Hence, the continuous function Q has a zero $x_0 \in [0,1]$, which corresponds to a *turning point* for our Schrödinger equation. A formula for this zero can be obtained readily,

$$\begin{aligned} x_0 &= \frac{1}{\kappa}\log\left(\frac{\Phi j_H + \varepsilon^{\beta}(1+j_H)\Lambda}{\Phi j_H - \varepsilon^{\beta} j_H(1+j_H)\Lambda}\right)\\ &= \varepsilon^{\beta}\frac{(1+j_H)^2\Lambda}{\kappa j_H \Phi} + \mathcal{O}(\varepsilon^{2\beta})\\ &= \varepsilon^{\beta}\frac{\Lambda}{F'(0)} + \mathcal{O}(\varepsilon^{2\beta}). \end{aligned} \tag{6.49}$$

Next, we construct the leading order WKB approximation to the ODE (6.44). Using standard formulas (see [5, Ch. 10, Sect. 1]), we obtain, for the region $[x_0, 1]$ where $Q(x) \geq 0$,

$$w(x) \sim \frac{C_+}{[Q(x)]^{1/4}}\mathrm{e}^{\frac{1}{\sqrt{\varepsilon}}\int_{x_0}^{x}\sqrt{Q(s)}ds} + \frac{C_-}{[Q(x)]^{1/4}}\mathrm{e}^{-\frac{1}{\sqrt{\varepsilon}}\int_{x_0}^{x}\sqrt{Q(s)}ds}. \tag{6.50}$$

---

[2] A WKB analysis of the case $\Lambda < 0$ shows that no eigenvalues $\lambda \geq \lambda_*$ exist (except for the eigenvalue $\lambda \sim f(1) - \ell$ in (6.35), which persists in this case). Thus, the working hypothesis $\Lambda > 0$ is justified.

This solution must satisfy the boundary condition at $x = 1$, since this point lies in its region of validity. Using Eq. (6.50), we calculate, to leading order in ε,

$$\sqrt{\varepsilon}\, w'(x) \sim C_+ \,[Q(x)]^{1/4} \mathrm{e}^{\frac{1}{\sqrt{\varepsilon}}\int_{x_0}^{x} \sqrt{Q(s)}ds} - C_- \,[Q(x)]^{1/4} \mathrm{e}^{-\frac{1}{\sqrt{\varepsilon}}\int_{x_0}^{x} \sqrt{Q(s)}ds}.$$

Using the last two equations to substitute w' and w into Eq. (6.45), we obtain to leading order the condition

$$\begin{aligned} & C_+ \left(\sqrt{Q(1)} - \frac{A}{2}\right) \mathrm{e}^{\frac{1}{\sqrt{\varepsilon}}\int_{x_0}^{1} \sqrt{Q(s)}ds} \\ & - C_- \left(\sqrt{Q(1)} + \frac{A}{2}\right) \mathrm{e}^{-\frac{1}{\sqrt{\varepsilon}}\int_{x_0}^{1} \sqrt{Q(s)}ds} \;=\; 0. \end{aligned} \tag{6.51}$$

6.6.3 Behavior in the region $[0, x_0]$

The solution given in Eq. (6.50) becomes unbounded at x_0 because of the denominator $[Q(x)]^{1/4}$ which vanishes at this point. In the region $[0, x_0]$, then, the WKB approximation (6.50) does not describe the actual solution.

To obtain the solution in that region (of length $\mathcal{O}(\varepsilon^{\beta})$, see (6.49)), we approximate $Q(x)$ by its first-order Taylor expansion around x_0,

$$Q(x) = Q_0'(x - x_0) + \mathcal{O}((x - x_0)^2), \tag{6.52}$$

where

$$Q_0' = Q'(x_0) = F'(0) + \mathcal{O}(\varepsilon^{\beta}). \tag{6.53}$$

Performing also the change of coordinates

$$x = x_0 + \theta\varepsilon^{\beta}\sigma, \tag{6.54}$$

where the $\mathcal{O}(1)$ scaling parameter θ is left unspecified for the time being, we obtain

$$Q(x_0 + \theta\varepsilon^{\beta}\sigma) = \varepsilon^{\beta}\theta Q_0'\sigma + \mathcal{O}(\varepsilon^{2\beta}). \tag{6.55}$$

Next, we calculate

$$w' = \theta^{-1}\varepsilon^{-\beta}\dot{w} \quad \text{and} \quad w'' = \theta^{-2}\varepsilon^{-2\beta}\ddot{w},$$

where $\dot{(\cdot)} = d \cdot /d\sigma$. Letting also $\sigma_0 = \theta^{-1}\varepsilon^{-\beta}x_0$, we recast Eqs. (6.44)–(6.45) in the form

$$\begin{aligned} \varepsilon^{1-2\beta}\,\theta^{-2}\,\ddot{w} \;&=\; \varepsilon^{\beta}\theta Q_0'\sigma\, w, \\ \left(\varepsilon^{1/2-\beta}\,\theta^{-1}\,\dot{w} - \frac{A}{2}\,w\right)(-\theta^{-1}\varepsilon^{-\beta}x_0) \;&=\; 0. \end{aligned}$$

Dominant balance for the first equation dictates that $\beta = 1/3$. Choosing also $\theta = [Q_0']^{-1/3}$, the boundary–value problem above becomes, to leading order,

$$\ddot{w} - \sigma\, w \;=\; 0, \tag{6.56a}$$

$$w(-\sigma_0) \;=\; 0. \tag{6.56b}$$

The ODE (6.56a) is an Airy equation [1, 5], with general solution

$$w(\sigma) = D_+ \mathrm{Bi}(\sigma) + D_- \mathrm{Ai}(\sigma). \tag{6.57}$$

Thus, the boundary condition (6.56b) yields

$$D_+ \,\mathrm{Bi}(-\sigma_0) + D_- \,\mathrm{Ai}(-\sigma_0) = 0. \tag{6.58}$$

6.6.4 Asymptotic matching

The expressions for w given in Eqs. (6.50) and (6.57) should match in a neighborhood of x_0 or, equivalently, of $\sigma = 0$. Since there is only one turning point in our problem, located at x_0, we use the technique presented in [5, Ch. 10, Sect. 4] to perform the matching. First, we calculate

$$w(\sigma) \sim \frac{D_+}{\sqrt{\pi}}\, \sigma^{-1/4}\, \mathrm{e}^{\frac{2}{3}\sigma^{3/2}} + \frac{D_-}{2\sqrt{\pi}}\, \sigma^{-1/4}\, \mathrm{e}^{-\frac{2}{3}\sigma^{3/2}}.$$

Recalling the definition (6.54) and that $\beta = 1/3$, $\theta = [Q_0']^{-1/3}$, we obtain

$$\sigma = \varepsilon^{-1/3}\, [Q_0']^{1/3}\, (x - x_0)$$

and, using this, we recast the asymptotic formula for $w(\sigma)$ in the form

$$\begin{aligned} w\left(\frac{x - x_0}{\varepsilon^{1/3}[Q_0']^{-1/3}}\right) &\sim \varepsilon^{1/12}\, \frac{D_+}{\sqrt{\pi}}\, [Q_0']^{-1/12}\, (x - x_0)^{-1/4}\, \mathrm{e}^{\frac{2}{3\sqrt{\varepsilon}}\sqrt{Q_0'}\,(x-x_0)^{3/2}} \\ &\quad + \varepsilon^{1/12}\, \frac{D_-}{2\sqrt{\pi}}\, [Q_0']^{-1/12}\, (x - x_0)^{-1/4}\, \mathrm{e}^{-\frac{2}{3\sqrt{\varepsilon}}\sqrt{Q_0'}\,(x-x_0)^{3/2}}. \end{aligned}$$

The WKB approximation (6.50) can be also recast in a similar form in an appropriate region around x_0. Using the Taylor approximation (6.52), we obtain

$$[Q(x)]^{-1/4} \sim [Q_0']^{-1/4}\, (x - x_0)^{-1/4} \quad \text{and} \quad \int_{x_0}^{x} \sqrt{Q(s)}ds \sim \frac{2}{3}\, \sqrt{Q_0'}\, (x - x_0)^{3/2}.$$

Thus, Eq. (6.50) yields

$$\begin{aligned} w(x) &\sim C_+\, [Q_0']^{-1/4}\, (x - x_0)^{-1/4}\, \mathrm{e}^{\frac{2}{3\sqrt{\varepsilon}}\sqrt{Q_0'}\,(x-x_0)^{3/2}} \\ &\quad + C_-\, [Q_0']^{-1/4}\, (x - x_0)^{-1/4}\, \mathrm{e}^{-\frac{2}{3\sqrt{\varepsilon}}\sqrt{Q_0'}\,(x-x_0)^{3/2}}. \end{aligned} \tag{6.59}$$

The matching requirement (the limiting Airy and WKB solutions should match in a region around x_0) yields, then, the two conditions

$$C_+\, [Q_0']^{-1/4} = \varepsilon^{1/12}\, \frac{D_+}{\sqrt{\pi}}\, [Q_0']^{-1/12}, \tag{6.60a}$$

$$C_-\, [Q_0']^{-1/4} = \varepsilon^{1/12}\, \frac{D_-}{2\sqrt{\pi}}\, [Q_0']^{-1/12}. \tag{6.60b}$$

6.6.5 Obtaining the eigenvalues

For the system of conditions (6.51), (6.58), and (6.60), to have a nontrivial solution, we need that

$$\det \begin{pmatrix} \mathfrak{A} & \mathfrak{B} & 0 & 0 \\ 0 & 0 & \mathrm{Bi}(-\sigma_0) & \mathrm{Ai}(-\sigma_0) \\ \sqrt{\pi}\,[Q_0']^{-1/6} & 0 & -\varepsilon^{1/12} & 0 \\ 0 & 2\sqrt{\pi}\,[Q_0']^{-1/6} & 0 & -\varepsilon^{1/12} \end{pmatrix} = 0,$$

where

$$\begin{aligned} \mathfrak{A} &= \left(\sqrt{Q(1)} - \frac{A}{2}\right) \mathrm{e}^{\frac{1}{\sqrt{\varepsilon}} \int_{x_0}^{1} \sqrt{Q(s)} ds}, \\ \mathfrak{B} &= -\left(\sqrt{Q(1)} + \frac{A}{2}\right) \mathrm{e}^{-\frac{1}{\sqrt{\varepsilon}} \int_{x_0}^{1} \sqrt{Q(s)} ds}. \end{aligned}$$

A direct calculation of the determinant yields, then, the equation

$$\begin{aligned} 0 \;=\; & 2\,\varepsilon^{1/12}\sqrt{\pi}\,[Q_0']^{-1/6}\left(\sqrt{Q(1)} - \frac{A}{2}\right) \mathrm{e}^{\frac{1}{\sqrt{\varepsilon}} \int_{x_0}^{1} \sqrt{Q(s)} ds}\,\mathrm{Ai}(-\sigma_0) \\ & + \varepsilon^{1/12}\sqrt{\pi}\,[Q_0']^{-1/6}\left(\sqrt{Q(1)} + \frac{A}{2}\right) \mathrm{e}^{-\frac{1}{\sqrt{\varepsilon}} \int_{x_0}^{1} \sqrt{Q(s)} ds}\,\mathrm{Bi}(-\sigma_0). \end{aligned} \tag{6.61}$$

Since $Q(x_0) = 0$ and $Q'(x_0) > 0$, the exponentiated integral appearing in this formula is $\mathcal{O}(1)$, and thus

$$\mathrm{e}^{-\frac{1}{\sqrt{\varepsilon}} \int_{x_0}^{1} \sqrt{Q(s)} ds} \ll \mathrm{e}^{\frac{1}{\sqrt{\varepsilon}} \int_{x_0}^{1} \sqrt{Q(s)} ds}.$$

As a result, the determinant condition (6.61) reads, down to exponentially small terms,

$$\mathrm{Ai}(-\sigma_0) = 0. \tag{6.62}$$

Recalling that $\sigma_0 = \theta^{-1}\varepsilon^{-\beta}x_0$, with x_0 given by Eq. (6.49), and Eq.(6.53), we rewrite this condition in the form

$$\mathrm{Ai}\left(-\frac{\Lambda}{[F'(0)]^{2/3}}\right) = 0, \tag{6.63}$$

whence

$$\Lambda_n = -[F'(0)]^{2/3} c_n = |f'(0)|^{\frac{2}{3}}\,|c_n|\,, \quad n = 1, 2, \cdots.$$

Here, $c_n < 0$ is the $n-$th root of the function $\mathrm{Ai}(x)$, and we have used definition (6.46b) to calculate that $F'(0) = -f'(0) > 0$. Thus, recalling Eq. (6.47), that $\beta = 1/3$ (see Section 6.6.3), we obtain the first few eigenvalues,

$$\lambda_n \sim \lambda_* - \varepsilon^{\frac{1}{3}}\,|f'(0)|^{\frac{2}{3}}\,|c_n|\,, \quad \text{where} \quad \lambda_* = f(0) - \ell - \frac{A^2}{4}. \tag{6.64}$$

As a result, the surface

$$\left[f(0) - \ell - \frac{A^2}{4}\right] - \varepsilon^{\frac{1}{3}}\,|f'(0)|^{\frac{2}{3}}\,|c_1| = 0 \tag{6.65}$$

corresponds to the bifurcation regime in the parameter space[3].

6.7 Numerical simulations

The bifurcation surface is described by Eq. (6.65). First, we recast that expression in terms of the original parameters. Using Eq. (6.11), we calculate

$$\begin{aligned} f(0) &= \frac{\Phi}{1 + j_H} = \frac{1}{(1 + j_H)\,(1 + \eta_H)}\,, \\ f'(0) &= -\frac{\Phi\,\kappa\,j_H}{(1 + j_H)^2} = -\frac{\kappa\,j_H}{(1 + j_H)^2\,(1 + \eta_H)}\,. \end{aligned}$$

Substituting into Eq. (6.65), we obtain

$$\frac{A^2}{4} = \frac{1}{(1 + j_H)\,(1 + \eta_H)} - \ell - \varepsilon^{\frac{1}{3}}\,|c_1|\left(\frac{\kappa\,j_H}{(1 + j_H)^2\,(1 + \eta_H)}\right)^{\frac{2}{3}}. \tag{6.66}$$

Finally, we define the new variable

$$\Delta = A^2 = \frac{a^2}{\varepsilon} = \frac{v^2}{\mu\,D} \tag{6.67}$$

and recall that

$$\Phi = \frac{1}{1 + \eta_H} = \frac{1}{1 + \frac{N_H}{N_B}}.$$

We rewrite Eq. (6.66) in the form

$$\Delta = 4\left[\frac{1}{1 + j_H}\,\Phi - \varepsilon^{\frac{1}{3}}\,|c_1|\left(\frac{\kappa}{2 + j_H + j_H^{-1}}\right)^{\frac{2}{3}}\Phi^{\frac{2}{3}} - \ell\right], \tag{6.68}$$

where $|c_1| \approx 2.338$ (see `http://mathworld.wolfram.com/AiryFunctions.html`).

It remains to suggest a numerical procedure in terms of the *old* parameters which ensures that the two new parameters Δ and Φ vary while the remaining new parameters j_H, ε, κ, and ℓ appearing in (6.68) do not.

---

[3] $\lambda_1 < 0$ leads to stable steady state and hence no blooms occur, whereas $\lambda_1 > 0$ corresponds to instability and therefore a bloom exists.

First, we recall the definitions

$$\begin{aligned}
\Delta &= \frac{v^2}{\mu\, D}, \\
\Phi &= \frac{1}{1+\frac{N_H}{N_B}}, \\
\varepsilon &= \frac{D}{\mu z_B^2}, \\
\kappa &= K_{bg}\, z_B, \\
j_H &= \frac{L_H}{L_I}, \\
\ell &= \frac{l}{\mu}.
\end{aligned}$$

Then, we observe that, in every run, one may *keep* v, μ, N_H, L_H, L_I, and l *fixed*, whereas *vary* D and N_B freely, *vary* z_B in such a way that ε *does not change* (the change in z_B counteracts that of D in the definition of ε), and *vary* K_{bg} so that κ *does not change* (the change in K_{bg} counteracts that of z_B in the definition of κ).

Next, we determine a rectangular region in the parameter plane $(\Delta\,,\,\Phi)$ which contains the bifurcation curve given in Eq. (6.68). First, we observe that the definition of Φ implies that $\Phi \in (0,1)$. Then, the definition of Δ and Eq. (6.68) also imply that $\Delta \in (0\,,\,\Delta_{\max})$, where

$$\Delta_{\max} = 4\left[\frac{1}{1+j_H} - \ell\right].$$

Thus, the rectangle $[0\,,\,1] \times [0\,,\,\Delta_{\max}]$ contains the entire bifurcation curve.

Next, we use the definitions of Δ and Φ to write the original parameters that we vary (D and N_B) in terms of the new parameters we vary (Δ and Φ). We find

$$N_B = \frac{N_H}{\Phi^{-1}-1} \quad \text{and} \quad D = \frac{v^2}{\mu\,\Delta}.$$

To carry out the numerical simulations, we introduce a regular grid on the rectangle $[0,1] \times [0,\Delta_{\max}]$. If each node of the mesh is indexed by the pair $(m\,,\,n)$, then the numerical simulation can run along this blueprint:

```
% Assign fixed values to those parameters that do not change throughout the
% simulation. Values for the rest of the parameters will be assigned at each grid point.
  Set values for $v$, $\mu$, $N_H$, $L_H$, $L_I$, $l$, $\alpha$, $k$.
% Calculate $\ell$ and $j_H$.
  $\ell = \frac{l}{\mu}$,   $j_H = \frac{L_H}{L_I}$.
% Assign fixed values to the new parameters $\varepsilon$ and $\kappa$, which have to remain unchanged
```

```
% throughout the simulation.
  Set values for $\varepsilon$ and $\kappa$.
% Calculate the upper bound for $\Delta$.
  $\Delta_{\max} = 4\left(\frac{1}{1+j_H} - \ell\right)$.
% This is the loop that builds the bifurcation diagram in the $(\Phi\,,\,\Delta)$-plane. Here,
% $\delta\Phi$ and $\delta\Delta$ correspond to the grid sizes in the $\Phi$- and $\Delta$-directions, respectively.
  for $m = 1 : 1/\delta\Phi$.
    % Assign the current $\Phi$-value.
      $\Phi_m = m\,\delta\Phi$.
    % Calculate the corresponding $N_B$-value from the $\Phi$-value.
      $(N_B)_m = N_H/(\Phi_m^{-1} - 1)$.
      for $n = 1 : \Delta_{\max}/\delta\Delta$.
        % Assign the current $\Delta$-value.
          $\Delta_n = n\,\delta\Delta$.
        % Calculate the corresponding $D$-value from the $\Delta$-value.
          $D_n = v^2/(\mu\,\Delta_n)$.
        % Change $z_B$ so as to counteract the change in $D$ and thus keep $\varepsilon$ constant.
          $z_B = \sqrt{\frac{D}{\mu\varepsilon}}$.
        % Change $K_{bg}$ so as to counteract the change in $z_B$ and thus keep $\kappa$ constant.
          $K_{bg} = \frac{\kappa}{z_B}$.
        % Determine the character of the solution at the current grid point.
          integrate PDE $\rightarrow$ decide: 'bloom', 'no bloom', 'oscillation'.
      end.
  end.
```

With the above blueprint, we ran numerous simulations for the full model (6.1), the results of which are plotted in Fig. 6.1. The associated parameter values are listed in Table 6.1. From Fig. 6.1, we see that the analytical curve describing the boundary between no-blooming and blooming (i.e., (6.68)) is indeed asymptotically close to the border found by simulation (between A and B&C). The region B indicates that phytoplankton population (as well as nutrient concentration) tends to a steady state, whereas in region C phytoplankton population oscillates. Although the simulation results have been given, our analysis only shows the boundary between the no-blooming and blooming regions, rather than distinguishing between oscillating and non-oscillating solutions. This topic will be subject of future research.

We will now discuss some interesting observations from Fig. 6.1. We note that Δ and Φ, along the two axes, are proportional to D^{-1} and N_B, respectively.

TABLE 6.1: Values for the parameters that are fixed throughout the simulation. The values of ε and κ have been motivated by the fact that we observed an oscillatory solution for $D = 0.1\,\mathrm{cm}^2/\mathrm{s}$, $z_B = 10^4\,\mathrm{cm}$, and $K_{bg} = 2 \cdot 10^{-3}\,\mathrm{cm}^{-1}$.

v $(\frac{\mathrm{cm}}{\mathrm{s}})$	μ $(\frac{1}{\mathrm{s}})$	N_H $(\frac{\mu\mathrm{mol\ nutrient}}{\mathrm{cm}^3})$	L_H $(\frac{\mu\mathrm{mol\ photons}}{\mathrm{cm}^2\,\mathrm{s}})$	l $(\frac{1}{\mathrm{s}})$
$1.17 \cdot 10^{-3}$	$1.11 \cdot 10^{-5}$	$2.0 \cdot 10^{-3}$	$2.0 \cdot 10^{-3}$	$2.78 \cdot 10^{-6}$

ε (const)	κ (const)	α (const)	L_I $(\frac{\mu\mathrm{mol\ photons}}{\mathrm{cm}^2\,\mathrm{s}})$	k $(\frac{\mathrm{cm}^2}{\mathrm{cells}})$
$9.0 \cdot 10^{-5}$	20	$2.0 \cdot 10^{-3}$	$6.0 \cdot 10^{-2}$	$3.0 \cdot 10^{-7}$

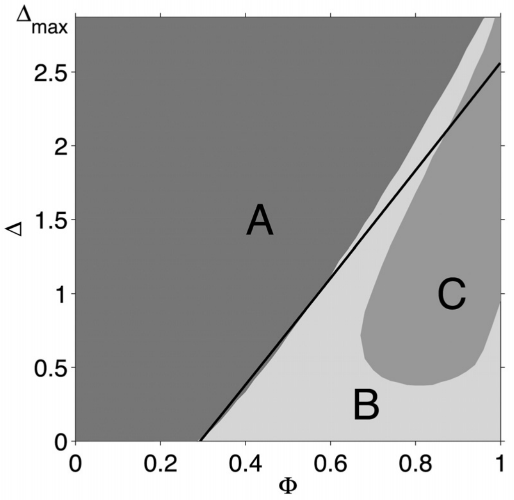


FIGURE 6.1: Simulating regions of no-blooming (A) and blooming (B and C). B: no-oscillations; C: oscillations. The solid line is the analytical curve described by (6.68).

Therefore, the lower boundary of region A corresponds to a large D and small N_B. In other words, this corresponds to an environment with high mixing, but a lack of nutrient. As a consequence, phytoplankton species are not able to survive. For the upper part of region A, no matter how abundant the amount of the nutrient is, sinking phytoplankton has no chance to move upward to the euphotic layer, because the water is poorly mixed. Again, phytoplankton will vanish. When the water is sufficiently mixed and nutrient is sufficiently available, there is always a bloom (regions B and C). For the lower part of region

B, the water is well mixed (D is large) and phytoplankton population tends to a steady state. This is consistent with the observation given in Chapter 5. In region C, the water is relatively-low mixed, and the phytoplankton population will oscillate. Again, this is what has been observed in Chapter 5. Finally, Fig. 6.1 shows that there is no direct transition from region A to region C, i.e., these regions are always separated by region B. Moreover, the distance between these two regions increases as Δ decreases.

Appendix

Derivation of Eq. (6.16)

Equation (6.14), together with Eq. (6.11), yield

$$\int^{\Omega_0(x)} \frac{d\Omega_0}{\Omega_0} = \int^x \frac{f(\sigma) - \ell - \lambda}{a} d\sigma = \frac{1}{a(\eta_H + 1)} \int^x \frac{e^{-\kappa\sigma}}{(j_H + e^{-\kappa\sigma})} d\sigma - \frac{\ell + \lambda}{a} x.$$

Recalling that

$$\frac{e^{-\kappa\sigma}}{j_H + e^{-\kappa\sigma}} = -\frac{(j_H + e^{-\kappa\sigma})'}{\kappa(j_H + e^{-\kappa\sigma})},$$

we obtain

$$\log \Omega_0(x) = -\frac{1}{a\kappa(\eta_H + 1)} \log(j_H + e^{-\kappa x}) - \frac{\ell + \lambda}{a} x + C,$$

where C is an integration constant. Equation (6.16) follows by exponentiation of both members of this equation.

Derivation of Eq. (6.29)

Combining Eqs. (6.25) and (6.27), we find

$$\ddot{\omega}_1 + a\,\dot{\omega}_1 = -C_0(f(1) - \ell - \lambda_0)e^{-as}.$$

Integrating both members once, and recalling the boundary condition (6.28), we find

$$\dot{\omega}_1 + a\,\omega_1 = C_0 \frac{f(1) - \ell - \lambda_0}{a} (e^{-as} - 1),$$

whence we obtain

$$\frac{d}{ds}(e^{as}\omega_1) = C_0 \frac{f(1) - \ell - \lambda_0}{a} (1 - e^{as}).$$

Integrating once again, we find

$$e^{as}\omega_1 = C_0 \frac{f(1) - \ell - \lambda_0}{a} \left(s - \frac{e^{as}}{a}\right) + C_1,$$

where C_1 is a constant of integration. Equation (6.29) now follows.

Derivation of Eq. (6.33)

Combining Eqs. (6.25) and (6.31), we find

$$\ddot{\omega}_2 + a\,\dot{\omega}_2 = C_0(f'(1)s + \lambda_1)e^{-as}.$$

Integrating both members once and recalling the boundary condition (6.32), we find

$$\begin{aligned}\dot{\omega}_2 + a\,\omega_2 &= C_0\left[f'(1)\int_0^s \sigma e^{-a\sigma}d\sigma + \lambda_1\int_0^s e^{-a\sigma}d\sigma\right]\\ &= C_0\left[-\frac{f'(1)}{a}se^{-as} - \frac{f'(1)+a\lambda_1}{a^2}e^{-as} + \frac{f'(1)+a\lambda_1}{a^2}\right],\end{aligned}$$

where we have used that

$$\int_0^s \sigma e^{a\sigma}d\sigma = -\frac{se^{-as}}{a} - \frac{e^{-as}}{a^2} + \frac{1}{a^2}.$$

This equation yields

$$\frac{d}{ds}\left(e^{as}\omega_2\right) = C_0\left[-\frac{f'(1)}{a}s - \frac{f'(1)+a\lambda_1}{a^2} + \frac{f'(1)+a\lambda_1}{a^2}e^{as}\right],$$

which becomes, upon integration,

$$e^{as}\omega_2 = C_0\left[-\frac{f'(1)}{2a}s^2 - \frac{f'(1)+a\lambda_1}{a^2}s + \frac{f'(1)+a\lambda_1}{a^3}e^{as}\right] + C_2.$$

Here, C_2 is a constant of integration. Equation (6.33) now follows.

Summary

In this thesis, we study an application from biology, that is, the dynamics of phytoplankton. The phytoplankton dynamics has been considered in light-limited environments as well as in light-nutrient-limited environments. Mathematically, it is modelled in terms of a system of coupled, nonlinear integro-partial differential equations of advection-diffusion-reaction type. This thesis aims to develop efficient numerical methods for solving this complex system of equations.

The numerical method used in the thesis is based on the simple and flexible Method of Lines approach, in which the spatial discretization and the time integration are considered separately.

The diffusion term has been discretized by the simple and popular second-order central scheme. For the advection term, we use the third-order upwind-biased scheme, which improves the artificial dissipation and artificial dispersion of the lower-order methods. The repeated trapezoidal rule has been used to approximate the integral within the reaction term.

The resulting semi-discretized system of ODEs is stiff, due to the discretization of the diffusion term. Therefore, an implicit integration method has been used. The implicit relation has been solved by the modified Newton method. Due to the strong coupling of the components, mainly caused by the integral term, the Jacobian matrix has a huge number of non-zero entries. To reduce the amount of work, we have 'decoupled' the components by neglecting the influence of the non-stiff, nonlinear reaction term in the Jacobian matrix. For the three-dimensional study (Chapters 1 and 4), we have further 'decoupled' the components by the Approximate Matrix Factorization technique. The resulting convergence behaviour of the Newton process is still satisfactory.

The above 'decoupling' strategies have been adopted in VODE, a popular ODE solver. Using VODE with various integration orders, we concluded that unconditional stability is an important property for integrating the phytoplankton problem (Chapter 1). Moreover, using VODE as a 'black box' solver needs precaution in case of an application, such as ours, where positivity of the solution is a prerequisite (Chapter 2). We came up with a remedy to avoid negative solutions. However, this approach is still far from optimal.

Therefore, as an indispensable step, we focused on integration methods that do yield a positive solution (Chapter 3). Here, we fixed the integration order to two. The results that we found for the explicit two-step methods showed an improvement of the results given in [37]. Moreover, it turned out that the optimal positive method is the well-known extrapolated BDF2 method. Nevertheless, the size of the time step has to obey a severe restriction.

Since a positive integration assumes a positive spatial discretization, we improved the discretization of the advection term by using a limiting technique. The resulting nonlinear system of ODEs, however, is difficult to solve implicitly. Therefore, we further improved the integration aspect by using the implicit-explicit (IMEX) version of the methods analyzed. For advection-diffusion equations the non-stiff, but strongly nonlinear advection part is treated explicitly whereas the stiff diffusion part is treated implicitly. It turns out that this IMEX method possesses the same stability region as the explicit version of the method when applied to the non-stiff part. This valuable result is due to the fact that the implicitly treated operator gives rise to a negative spectrum (Chapter 4).

Apart from the interesting numerical observations, we have learned several amazing properties of the phytoplankton dynamics. Although the model offers only abstractions of real-world phenomena, it adequately reproduces many features of the real-world. Surprisingly, in light-nutrient-limited environments, we observed that plankton populations show strong oscillations and even chaos when vertical mixing of nutrients is reduced (Chapter 5). This may have a negative impact on the food chain of the oceans and on the uptake of the greenhouse gas carbon dioxide into the oceans. On the other hand, it has a positive effect on the number of plankton species and thus on the biodiversity in the oceans.

This model prediction was rather unexpected, because it contradicts conventional wisdom that deep plankton in the oceans would represent a stable system. Consequently, it is of great importance to analyze the phytoplankton oscillations. We have shown numerically in which parameter regime we may expect no bloom, stationary bloom, or oscillations (Chapter 6). Additionally, an analytical analysis has been performed to show the boundary between nobloom- and bloom regions. The results obtained by these two approaches indeed asymptotically match. A further analysis to distinguish between stationary blooms and oscillatory blooms will be subject of future research.

Samenvatting

In dit proefschrift bestuderen we een toepassing vanuit de biologie, namelijk de dynamica van fytoplankton. Zowel een omgeving waarin licht de enige beperkende factor is, alsook een omgeving waarin licht *en* nutriënten beide beperkende factoren zijn, worden beschouwd. Wiskundig wordt de dynamica van fytoplankton gemodelleerd als een stelsel gekoppelde, niet-lineaire integro-partiële differentiaalvergelijkingen van het advectie-diffusie-reactie type. Het doel van dit proefschrift is het ontwikkelen van efficiënte numerieke methoden om dit complexe stelsel vergelijkingen op te lossen.

De in dit proefschrift gebruikte numerieke methode is gebaseerd op de eenvoudige en flexibele Methode-der-Lijnen aanpak, waarin de discretisaties in ruimte en tijd gescheiden worden uitgevoerd.

De diffusieterm is gediscretiseerd door middel van het eenvoudige en veelgebruikte tweede-orde centrale schema. Voor de advectieterm gebruiken we het zogeheten 'upwind-biased' schema van orde drie; dit schema verkleint de artificiële dissipatie en dispersie vergeleken met schema's van lagere orde. De integraal in de reactieterm is benaderd met de geregen trapeziumregel.

Het resulterende semi-discrete stelsel gewone differentiaalvergelijkingen is stijf vanwege de discretisatie van de diffusieterm. Om die reden is een impliciete tijdsintegratiemethode gebruikt. We gebruiken de gemodificeerde methode van Newton om de impliciete relaties op te lossen. De Jacobiaanmatrix bevat een groot aantal elementen ongelijk aan nul; dit wordt veroorzaakt door de sterke koppeling van de componenten, vooral vanwege de gediscretiseerde integraalterm. Om de hoeveelheid rekenwerk te reduceren, hebben we de componenten ontkoppeld door de invloed van de niet-stijve, niet-lineaire reactieterm in de Jacobiaan te verwaarlozen. In de driedimensionale modellen (zie Hoofdstuk 1 en 4) hebben we een verdere ontkoppeling aangebracht door middel van de zogeheten 'Approximate Matrix Factorization' techniek. Het resulterende convergentiegedrag van het Newtonproces blijft bevredigend.

Voornoemde ontkoppelingsstrategieën zijn ingebouwd in VODE, een populaire code voor het oplossen van beginwaardeproblemen. VODE is toegepast met verschillende waarden voor de orde van nauwkeurigheid. Uit de resultaten beschreven in Hoofdstuk 1 concluderen we dat onvoorwaardelijke stabiliteit een belangrijke eigenschap is voor het integreren van het fytoplanktonprobleem. Bovendien is in toepassingen zoals de onze een positieve oplossing een vereiste. Wanneer VODE als 'black box solver' gebruikt wordt, zijn er voorzorgsmaatregelen nodig om dit te bereiken (zie Hoofdstuk 2). We hebben een remedie bedacht om negatieve oplossingen te vermijden. Deze aanpak is echter nog verre van optimaal.

Als een noodzakelijke stap concentreren we ons daarom in Hoofdstuk 3 op methoden die een positieve oplossing genereren. Hierbij fixeren we de integratieorde op twee. Voor expliciete tweestaps methoden vonden we een resultaat dat een verbetering is van de resultaten beschreven in [37]. Bovendien bleek de geëxtrapoleerde BDF2 methode optimaal te zijn wat betreft positiviteit. Niettemin geldt er een strenge restrictie voor de grootte van de tijdstap.

Omdat een positief integratieproces een positieve semi-discretisatie veronderstelt, hebben we de discretisatie van de advectieterm verbeterd door een zogenaamde 'limiter' toe te passen. Het resulterende niet-lineaire stelsel gewone differentiaalvergelijkingen wordt echter lastiger op te lossen met een impliciete methode. Een verdere verbetering van het tijdsintegratieproces wordt daarom verkregen door de impliciet-expliciete (IMEX) variant van de geanalyseerde methoden te gebruiken. Voor advectie-diffusie vergelijkingen wordt het niet-stijve, maar sterk niet-lineaire advectiedeel expliciet en het stijve diffusiedeel impliciet behandeld. Het blijkt dat deze IMEX methode hetzelfde stabiliteitsgebied heeft als de expliciete versie van de methode toegepast op het niet-stijve deel. Dit waardevolle resultaat is te danken aan het feit dat de impliciet behandelde operator aanleiding geeft tot een negatief spectrum (zie Hoofdstuk 4).

Los van de interessante numerieke observaties, zijn we verbazingwekkende eigenschappen van de dynamica van fytoplankton te weten gekomen. Hoewel het model slechts een abstractie van de werkelijkheid is, worden veel kenmerken uit de 'real world' adequaat weergegeven. Verrassend was de observatie dat, in een omgeving waarin licht en nutriënten beide beperkende factoren zijn, fytoplanktonpopulaties sterke oscillaties en zelfs chaotisch gedrag vertonen indien de verticale menging van de nutriënten verkleind wordt (zie Hoofdstuk 5). Dit kan een negatieve uitwerking hebben op de voedselketen en op de opname van het broeikasgas koolstofdioxide in de oceanen. Anderzijds heeft het een positief effect op het aantal soorten fytoplankton en daarmee op de biodiversiteit in de oceanen.

Deze voorspelling van het model was tamelijk onverwacht omdat het indruist tegen de traditionele opvatting dat plankton diep in de oceaan een stabiel systeem zou vormen. Het analyseren van de oscillaties in het fytoplankton is daarom van groot belang. We hebben langs numerieke weg bepaald voor welke parameterwaarden we mogen verwachten dat de populatie zal uitsterven dan wel zal groeien. In het laatste geval maken we onderscheid tussen groei naar een stationaire oplossing en het vertonen van oscillerend gedrag (zie Hoofdstuk 6). Bovendien is, in de parameterruimte, met een analytische aanpak de grens tussen uitsterven en groei bepaald. De resultaten van beide aanpakken komen asymptotisch goed overeen. Een verdere analyse om onderscheid te maken tussen groei naar een stationaire oplossing en oscillaties is onderwerp van toekomstig onderzoek.

Acknowledgements

This thesis is a result of my four year of research at the Center for Mathematics and Computer Science (CWI). First of all, I would like to thank the board of CWI for giving me the opportunity to perform my research.

Next, I would like to express my deep gratitude to my daily supervisor Dr. Ben Sommeijer and my promoters Prof. dr. Jan Verwer and Prof. dr. Jef Huisman, for extending their hands to greet my toddling steps and to make sure that I go firmly and towards the right direction. Under their supervision, I have become mature and confident in research. In addition, I thank them very much for their care of my feelings and personal life.

I am also grateful to my colleagues: Tsogtgerel Gantumur for giving me a nice article and discussion on the project when I arrived; Jan Kok for his perfect C code; Barry Koren for his analysis and idea of a nice flow field; David Karl and Willem Hundsdorfer for joining me in the beautiful articles; Jeff Williams, Antonios Zagaris, and Arjen Doelman for their meaningful contribution on the analysis of phytoplankton oscillations; Carolynne Montijn for her nice style of the thesis; Jan Schipper and Jos van der Werf for the valuable photographs and their help on printing the thesis; Susanne van Dam, Nada Mitrovic, the personal- and financial departments for arranging all kind of documents. I thank all the members of CWI for being nice to me.

I specially thank Margreet Nool for her exciting trip; and Valeriu Savcenco, Svetlana Dubinkina, Yousaf Habib, Ismail Rafatov, and Lubor Frastia for sharing a very nice working atmosphere with me. Valeriu, Svetlana, your partners, and Yousaf, thanks for your interesting discussions in mathematics and life experience and for joining me in many remarkable social activities.

I would like to thank my Vietnamese friends in The Netherlands; each time I meet them I have a snug feeling and a relaxed mind: Thang, Duong, Mai Hoa, Chi, Giang, Hung, Hieu, Kim Hoa, Trung, Hong, Huy, Hoang, Tri, Thanh, Duc, Vinh, Hanh, Kien, Thu, Quoc Anh, Viet, and my little friends Khoi and Minh Thi.

I would like to take this chance to thank my teachers and supervisors at the institutions at which my study was carried out: the Mathematical Research Institute (The Netherlands), the Institute of Mechanics (Hanoi, Vietnam), the University of Science (Hanoi, Vietnam), and the schools I studied in Quang Ninh (Vietnam).

I greatly appreciate the kindness of my friends who supported me to overcome the needy life and study well, i.e., the $9\mathrm{A}_1$-class, Mr. Liem – the manager of the Me Tri hostel, the rooms $10\mathrm{B}_1$ and 304, and my closest friends: Oanh, Huong, Son, Tuyen, Hung, Man, and Ha.

It is now the moment for me to express my gratefulness to my cherished mother, brother and sister for their constant support, eternal love and trust.

It is also a pleasure to send a warm appreciation to my dear husband for his invaluable love and for sharing weal and woe. Thieu, thanks a lot for creating the proper surroundings and the support in preparing my thesis.

For the success of this thesis, I am thankful to my parents-in-law for sympathizing and facilitating my desired study.

Last but not least, I would like to show my heartfelt gratitude to my respectable father for creating the motivations, ambitions, power, and belief in my life and study. Father, you are always my power, I am proud of you.

Bibliography

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Fifth edition, Dover Publications, New York, 1968.

[2] R.A. Andersen, R.R. Bidigare, M.D. Keller, and M. Latasa, *A comparison of HPLC pigment signatures and electron microscopic observations for oligotrophic waters of the North Atlantic and Pacific Oceans*, Deep-Sea Res. II 43, 517-537, 1996.

[3] K.R. Arrigo, D.H. Robinson, D.L. Worthen, R.B. Dunbar, G.R. DiTullio, M. van Woert and M.P. Lizotte, *Phytoplankton community structure and the drawdown of nutrients and* CO_2 *in the Southern Ocean*, Science 283, 365-367, 1999.

[4] R.M. Beam and R.F. Warming, *An implicit finite-difference algorithm for hyperbolic systems in conservation-law form*, J. Comput. Phys. 22, 87-110, 1976.

[5] C.M. Bender and S.A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, Applied Mathematical Sciences 35, Springer-Verlag, Berlin, 1999.

[6] L. Bopp, P. Monfray, O. Aumont, J-L. Dufresne, H.L. Treut, G. Madec, L. Terray, and J.C. Orr, *Potential impact of climate change on marine export production*, Glob. Biogeochem. Cycles 15, 81-99, 2001.

[7] P.W. Boyd, A.J. Watson, C.S. Law, E.R. Abraham, T. Trull, R. Murdoch, D.C.E. Bakker, A.R. Bowie, K.O. Buesseler, H. Chang, M. Charette, P. Croot, K. Downing, R. Frew, M. Gall, M. Hadfield, J. Hall, M. Harvey, G. Jameson, J. LaRoche, M. Liddicoat, R. Ling, M.T. Maldonado, R.M. McKay, S. Nodder, S. Pickmere, R. Pridmore, S. Rintoul, K. Safi, P. Sutton, R. Strzepek, K. Tanneberger, S. Turner, A. Waite, and J. Zeldis, *A mesoscale phytoplankton bloom in the polar Southern Ocean stimulated by iron fertilization*, Nature 407, 695-702, 2000.

[8] P.N. Brown, G.D. Byrne and A.C. Hindmarsh, *VODE: a variable-coefficient ODE solver*, SIAM J. Sci. Statist. Comput. 10, 1038-1051, 1989.

[9] L. Campbell, H. Nolla, and D. Vaulot, *The importance of* Prochlorococcus *to community structure in the central North Pacific Ocean*, Limnol. Oceanogr. 39, 954-961, 1994.

[10] F. Colijn and G.C. Cadée, *Is growth of Wadden Sea phytoplankton light or nutrient limited?*, J. Sea Res. 49, 83-93, 2003.

[11] J.J. Cullen, *The deep chlorophyll maximum: comparing vertical profiles of chlorophyll a*, Can. J. Fish. Aquat. Sci. 39, 791-803, 1982.

[12] G. Dahlquist, *Error analysis for a class of methods for stiff non-linear initial value problems*, Numerical Analysis, Lecture Notes in Mathematics 506, 60-72, 1976.

[13] S. Diehl, S. Berger, R. Ptacnik, and A. Wild, *Phytoplankton, light, and nutrients in a gradient of mixing depths: field experiments*, Ecology 83, 399-411, 2002.

[14] E.G. D'Yakonov, *Difference systems of second order accuracy with a divided operator for parabolic equations without mixed derivatives*, USSR Comput. Math. Phys. 4(5), 206-216, 1964.

[15] U. Ebert, M. Arrayás, N.M. Temme, B.P. Sommeijer, and J. Huisman, *Critical conditions for phytoplankton blooms*, Bull. Math. Biol. 63, 1095-1124, 2001.

[16] C. Eichler-Liebenow, P.J. van der Houwen, and B.P. Sommeijer, *Analysis of approximate factorization in iteration methods*, Appl. Numer. Math. 28, 245-258, 1998.

[17] P.G. Falkowski, *The ocean's invisible forest: marine phytoplankton play a critical role in regulating the earth's climate. Could they also help stop global warming?*, Scient. Amer. 287(2), 38-45, 2002.

[18] P.G. Falkowski, R.T. Barber and V. Smetacek, *Biogeochemical controls and feedbacks on ocean primary production*, Science 281, 200-206, 1998.

[19] K. Fennel and E. Boss, *Subsurface maxima of phytoplankton and chlorophyll: steady-state solutions from a simple model*, Limnol. Oceanogr. 48, 1521-1534, 2003.

[20] T.D. Finnigan, D.S. Luther, and R. Lukas, *Observations of enhanced diapycnal mixing near the Hawaiian ridge*, J. Phys. Oceanogr. 32, 2988-3002, 2002.

[21] J. Frank, W. Hundsdorfer and J.G. Verwer, *On the stability of implicit-explicit linear multistep methods*, Appl. Numer. Math. 25, 193-205, 1997.

[22] P.J.S. Franks and C.S. Chen, *A 3-D prognostic numerical model study of the Georges Bank ecosystem (II). Biological-physical model*, Deep-Sea Res. II 48, 457-482, 2001.

[23] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Review 43, 89-112, 2001.

[24] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems*, Second edition, Springer Series in Computational Mathematics 14, Springer, Berlin, 1996.

[25] B.A. Hodges and D.L. Rudnick, *Simple models of steady deep maxima in chlorophyll and biomass*, Deep-Sea Res. I 51, 999-1015, 2004.

[26] C. van den Hoek, D.G. Mann, and H.M. Jahns, *Algae: an Introduction to Phycology*, Cambridge University Press, Cambridge, 1995.

[27] O. Holm-Hansen and C. D. Hewes, *Deep chlorophyll-a maxima (DCMs) in Antarctic waters: I. Relationships between DCMs and the physical, chemical, and optical conditions in the upper water column*, Polar Biol. 27, 699-710, 2004.

[28] P.J. van der Houwen and B.P. Sommeijer, *Approximate factorization for time-dependent partial differential equations*, J. Comput. Appl. Math. 128, 447-466, 2001.

[29] J. Huisman, M. Arrayás, U. Ebert and B. Sommeijer, *How do sinking phytoplankton species manage to persist?*, Amer. Naturalist 159, 245-254, 2002.

[30] J. Huisman, R.R. Jonker, C. Zonneveld and F.J. Weissing, *Competition for light between phytoplankton species, experimental tests of mechanistic theory*, Ecology 80, 211-222, 1999.

[31] J. Huisman, P. van Oostveen and F.J. Weissing, *Species dynamics in phytoplankton blooms: incomplete mixing and competition for light*, Amer. Naturalist 154, 46-68, 1999.

[32] J. Huisman, J. Sharples, J.M. Stroom, P.M. Visser, W.E.A. Kardinaal, J.M.H. Verspagen, and B. Sommeijer, *Changes in turbulent mixing shift competition for light between phytoplankton species*, Ecology 85, 2960-2970, 2004.

[33] J. Huisman and B.P. Sommeijer, *Population dynamics of sinking phytoplankton in light-limited environments: simulation techniques and critical parameters*, J. Sea Res. 48, 83-96, 2002.

[34] J. Huisman and F.J. Weissing, *Biodiversity of plankton by species oscillations and chaos*, Nature 402, 407-410, 1999.

[35] W. Hundsdorfer, *Partially implicit BDF2 blends for convection dominated flows*, SIAM J. Num. Anal. 38, 1763-1783, 2001.

[36] W. Hundsdorfer and S.J. Ruuth, *On monotonicity and boundedness properties of linear multistep methods*, Math. Comp. 75, 655-672, 2006.

[37] W. Hundsdorfer, S.J. Ruuth and R.J. Spiteri, *Monotonicity-preserving linear multistep methods*, SIAM J. Numer. Anal. 41, 605-623, 2003.

[38] W. Hundsdorfer and J.G. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer Series in Computational Mathematics 33, Springer, Berlin, 2003.

[39] K. Hutter and K. Jöhnk, *Continuum Methods of Physical Modeling: Continuum Mechanics, Dimensional Analysis, Turbulence*, Springer, Berlin, 2004.

[40] D.M. Karl, R.R. Bidigare, and R.M. Letelier, in *Phytoplankton Productivity: Carbon Assimilation in Marine and Freshwater Ecosystems* (eds P.J.LeB. Williams, D.N. Thomas, and C.S. Reynolds), Blackwell Science, Oxford, 222-264, 2002.

[41] D.M. Karl, J.R. Christian, J.E. Dore, D.V. Hebel, R.M. Letelier, L.M. Tupas and C.D. Winn, *Seasonal and interannual variability in primary production and particle flux at Station ALOHA*, Deep-Sea Res. II 43, 539-568, 1996.

[42] J.T.O. Kirk, *Light and Photosynthesis in Aquatic Ecosystems*, Second edition, Cambridge University Press, Cambridge, 1994.

[43] C.A. Klausmeier and E. Litchman, *Algal games: the vertical distribution of phytoplankton in poorly mixed water columns*, Limnol. Oceanogr. 46, 1998-2007, 2001.

[44] B. Koren, *A robust upwind discretization for advection, diffusion and source terms*, in *Numerical Methods for Advection-Diffusion Problems* (eds. C.B. Vreugdenhil and B. Koren), Notes on Numerical Fluid Mechanics 45, Vieweg, Braunschweig, 117-138, 1993.

[45] P. Korpinen, M. Kiirikki, P. Rantanen, A. Inkala and J. Sarkkula, *High resolution 3D-ecosystem model for the Neva Bay and estuary: model validation and future scenarios*, Oceanologia 45, 67-80, 2003.

[46] R.M. Letelier, R.R. Bidigare, D.V. Hebel, M. Ondrusek, C.D. Winn, and D.M. Karl, *Temporal variability of phytoplankton community structure based on pigment analysis.* Limnol. Oceanogr. 38, 1420-1437, 1993.

[47] R.M. Letelier, D.M. Karl, M.R. Abbott, and R.R. Bidigare, *Light driven seasonal patterns of chlorophyll and nitrate in the lower euphotic zone of the North Pacific Subtropical Gyre*, Limnol. Oceanogr. 49, 508-519, 2004.

[48] M.R. Lewis, W.G. Harrison, N.S. Oakey, D. Hebert, and T. Platt, *Vertical nitrate fluxes in the oligotrophic ocean*, Science 234, 870-873, 1986.

[49] A.R. Longhurst, *Ecological Geography of the Sea*, Academic, San Diego, 1998.

[50] M.D. Mackey, D.J. Mackey, H.W. Higgins, and S.W. Wright, *CHEMTAX – A program for estimating class abundances from chemical markers: application to HPLC measurements of phytoplankton.* Mar. Ecol. Prog. Ser. 144, 265-283, 1996.

[51] K.H. Mann and J.R.N. Lazier, *Dynamics of Marine Ecosystems*, Blackwell Science, Oxford, 1996.

[52] D.J. McGillicuddy, A.R. Robinson and J.J. McCarthy, *Coupled physical and biological modeling of the spring bloom in the North Atlantic (II). 3-Dimensional bloom and post-bloom processes*, Deep-Sea Res. I 42, 1359-1398, 1995.

[53] J.R. Moisan, E.E. Hofmann and D.B. Haidvogel, *Modeling nutrient and plankton processes in the California Coastal Transition Zone (II). A three-dimensional physical-bio-optical model*, J. Geophys. Res. 101, 22677-22692, 1996.

[54] B.C. Monger and M.R. Landry, *Flow cytometric analysis of marine bacteria with Hoechst 33342*. Appl. Environ. Microbiol. 59, 905-911, 1993.

[55] A. Okubo and S.A. Levin, D*iffusion and Ecological Problems: Modern Perspectives*, Second edition, Springer, Berlin, 2001.

[56] T. Platt, S. Sathyendranath, O. Ulloa, W.G. Harrison, N. Hoepffner, and J. Goes, *Nutrient control of phytoplankton photosynthesis in the Western North Atlantic*, Nature 356, 229-231, 1992.

[57] C.S. Reynolds, *The Ecology of Freshwater Phytoplankton*, Cambridge University Press, Cambridge, 1984.

[58] S. Rinaldi, S. Muratori, and Y. Kuznetsov, *Multiple attractors, catastrophes and chaos in seasonally perturbed predator-prey communities*, Bull. Math. Biol. 55, 15-35, 1993.

[59] J.L. Sarmiento, T.M.C. Hughes, R.J. Stouffer, and S. Manabe, *Simulated response of the ocean carbon cycle to anthropogenic climate warming*, Nature 393, 245-249, 1998.

[60] J.L. Sarmiento, R. Slater, R. Barber, L. Bopp, S.C. Doney, A.C. Hirst, J. Kleypas, R. Matear, U. Mikolajewicz, P. Monfray, V. Soldatov, S.A. Spall, and R. Stouffer, *Response of ocean ecosystems to climate warming*, Glob. Biogeochem. Cycles 18, doi:10.1029/2003GB002134, 2004.

[61] A. Schmittner, *Decline of the marine ecosystem caused by a reduction in the Atlantic overturning circulation*, Nature 434, 628-633, 2005.

[62] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Statist. Comput. 9, 1073-1084, 1988.

[63] W.D. Smyth, J.N. Moum, and D.R. Caldwell, *The efficiency of mixing in turbulent patches: inferences from direct simulations and microstructure observations*, J. Phys. Oceanogr. 31, 1969-1992, 2001.

[64] J. Vandermeer, L. Stone, and B. Blasius, *Categories of chaos and fractal basin boundaries in forced predator-prey models*, Chaos, Solitons and Fractals 12, 265-276, 2001.

[65] A.E.P. Veldman, *Computational Fluid Dynamics*, Lecture Notes, University of Groningen, The Netherlands, 2001.

[66] E.L. Venrick, *Phytoplankton seasonality in the central North Pacific: the endless summer reconsidered*, Limnol. Oceanogr. 38, 1135-1149, 1993.

[67] E.L. Venrick, *Phytoplankton species structure in the central North Pacific, 1973-1996: variability and persistence*, J. Plankton Res. 21, 1029-1042, 1999.

[68] E.L. Venrick, J.A. McGowan, and A.W. Mantyla, *Deep maxima of photosynthetic chlorophyll in the Pacific Ocean*, Fish. Bull. 71, 41-52, 1973.

[69] J. Von Liebig, *Die Organische Chemie in Ihrer Anwendung auf Physiologie und Pathologie*, Friedrich Vieweg, Braunschweig, 1840.

[70] F.J. Weissing and J. Huisman, *Growth and competition in a light gradient*, J. Theoret. Biol. 168, 323-336, 1994.

[71] S.W. Wright, S.W. Jeffrey, R.F.C. Mantoura, C.A. Llewellyn, T. Bjornland, D. Repeta and N.A. Welschmeyer, *An improved HPLC method for the analysis of chlorophylls and carotenoids from marine phytoplankton*, Mar. Ecol. Prog. Ser. 77, 183-196, 1991.

[72] K. Yoshiyama and H. Nakajima, *Catastrophic transition in vertical distributions of phytoplankton: Alternative equilibria in a water column*, J. Theoret. Biol. 216, 397-408, 2002.