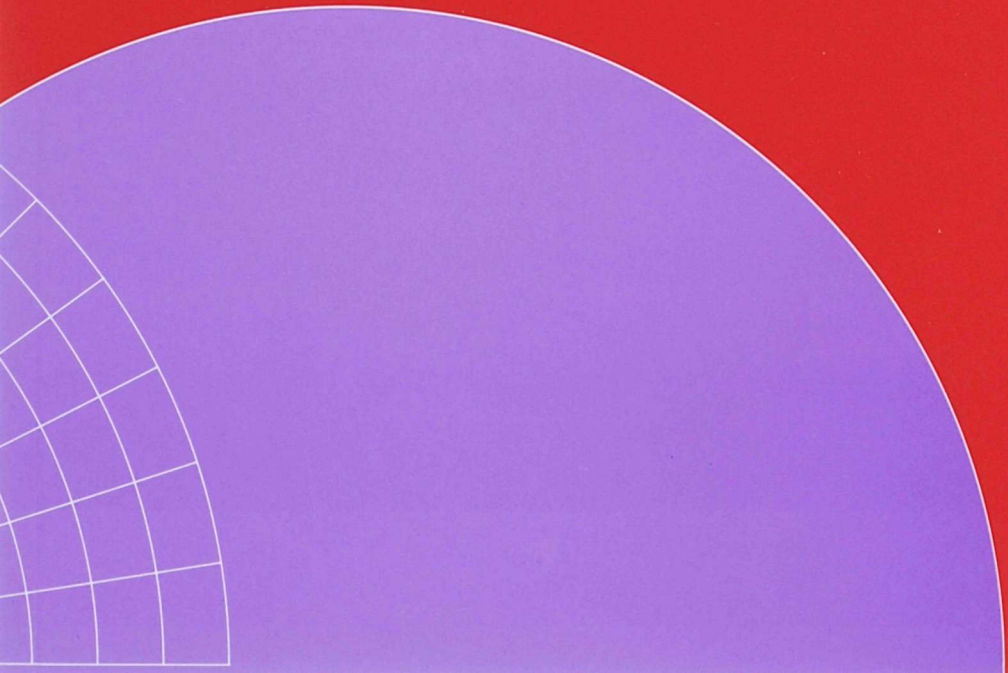
Numerical Methods for Atmospheric Flow and Transport Problems

Debby Lanser



1. Application of a three-term symmetrical Strang splitting to pure, autonomous, initial-value problems of the ADR-type leads to no splitting error between advection, diffusion and chemistry, when, with exact integration of the intermediate steps in the Strang splitting, the chemistry $R(c)$ is linear in c, and the wind field $\mathbf{u}$, the diffusion coefficient matrix K and the chemistry R are independent of the spatial variable $\mathbf{x}$. (Chapter 2)

2. The Shallow Water Equations (SWEs) can be written in conservation form and can therefore be conveniently discretized in space with a finite volume method exploiting the hyperbolic character of the equations. (Chapter 3)

3. The stepsize restriction for an explicit time integration method for solving a semi-discrete system of the global SWEs is significantly alleviated when this system is derived on a combined reduced latitudinal-longitudinal (lat-lon) and stereographic grid instead of on a common uniform or reduced lat-lon grid of similar resolution. (Chapter 3)

4. The third-order A-stable Rosenbrock method maintains its A-stability when applied with approximate matrix factorization. (Chapter 4)

5. As integration method for the semi-discrete SWEs on a global uniform lat-lon grid, Ros3 with AMF is far more efficient than the explicit time integration method RK3. Its superiority is unaffected even when the latter is applied to the semi-discrete SWEs on a reduced lat-lon and stereographic grid of similar resolution. (Chapter 4)

6. Een cabaretier met een gedegen wiskundeopleiding zou verder komen dan grappen over Máxima en minima.

7. Lopend onderzoek is net als een rijdende trein,
   je weet nooit wanneer het stagneert.

8. Indien de kwaliteit van de bolletjes net zo slecht zou zijn als het arrestatiebeleid voor hun slikkers, zou het cellentekort op Schiphol binnen no-time zijn opgelost.

9. Vrouwen kunnen maar één ding tegelijk, mannen doen alles half.

10. De meest persoonlijke stellingen zijn degenen die ontbreken.

Stellingen

bij het proefschrift

"Numerical Methods for Atmospheric Flow and Transport Problems"

door

Debby Lanser

Numerical Methods for Atmospheric Flow and Transport Problems

Debby Lanser

Maart 2002

Numerical Methods for Atmospheric Flow and Transport Problems

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 donderdag 7 maart 2002, te 12:00 uur

door

Debby Lanser

geboren te Rijsenhout

Promotor: prof. dr. J.G. Verwer

Faculteit: Natuurwetenschappen, Wiskunde en Informatica

Perhaps some day in the dim future it will be possible to advance the computations faster than the weather advances and at a cost less than the saving to mankind due to the information gained. But that is a dream.

Lewis Fry Richardson, 1921

Preface

Before you, you find my PhD-thesis. It is the completion of four years of hard work at CWI, the national research institute for mathematics and computer science in the Netherlands. This thesis is based on most of the research I conducted in the field of numerical methods for global atmospheric modelling; A field, that was particularly appealing to me, for it meant doing research in computational fluid dynamics with an application important to everyday life. This research resulted in four articles, which cover the main part of this thesis. These articles are self-contained and the corresponding chapters can be read separately.

- D. Lanser and J.G. Verwer.
  Analysis of operator splitting for advection-diffusion-reaction problems from air pollution modelling.
  J. Comp. Appl. Math., **111**, pp. 201–216, 1999.

- D. Lanser, J.G. Blom and J.G. Verwer.
  Spatial discretization of the shallow water equations in spherical geometry.
  J. Comput. Phys., **165**, pp. 542–565, 2000.

- D. Lanser, J.G. Blom and J.G. Verwer.
  Time integration of the shallow water equations in spherical geometry.
  J. Comput. Phys., **171**, pp. 373–393, 2001.

- D. Lanser.
  A comparison of operator splitting and approximate matrix factorization for the shallow water equations in spherical geometry.
  Technical Report MAS-R0115, CWI, Amsterdam, 2001.
  Submitted for publication.

Dankwoord

The acknowledgments will be written in Dutch. In this part, I would like to thank my colleagues and friends who made these last four years to what they were, sometimes difficult but most of all special. One exception will be made for Dave, who will not understand Dutch.

Opeens is het moment daar. Het is tijd om de laatste pagina uit dit proefschrift te schrijven, het dankwoord. Raar maar waar, maar de zinnen willen niet komen. De laatste vier jaar passeren de revue, de leuke en minder leuke momenten. En dan rijst de vraag: 'Wie ga ik bedanken?'. Ik waarschuw van te voren. Een werkplek valt en staat voor mij met de mensen die er werken. Ik heb veel plezier beleefd aan de gezellige sfeer die heerst op het CWI. Het is dan ook niet mogelijk iedereen te noemen, die de afgelopen vier jaar hebben gemaakt tot wat het zijn. Daarom vooraf, beste collega's, bedankt!

Jan, mijn promotor en begeleider, wil ik bedanken voor de zorgvuldigheid, waarmee hij mijn volledige proefschrift heeft gelezen. Daarnaast heb ik volop zijn steun gekregen om contact te maken met medewetenschappers in het veld. Dit heb ik enorm gewaardeerd. Op wetenschappelijk gebied stond Jans deur altijd voor me open.

Joke, mijn medebegeleider, deed me denken aan een definitie van wetenschap: 'Een objectief medium waarin nieuwe ideeën worden gewogen en getoetst'. Een zwak numeriek resultaat sneuvelde steevast onder haar kritische blik. Soms kon ze me tot wanhoop drijven, maar na een aantal nieuwe dagen achter mijn computer, was het eindresultaat altijd beter. Naast de wetenschap, kon ik met Joke over alles praten, van een moderne dansvoorstelling tot aan de laatste problematiek rondom de oio-salarissen.

I am also grateful to Dave Williamson, whose enthusiasm for the field of numerical methods in dynamic meteorology is a stimulus for many researchers in the field. On the several occasions we met, it was always fun talking to him. His clarifying discussions on the pole problem and the literature he mentioned have been very useful. I greatly appreciate his willingness to travel overseas in such an uncertain time.

De langste tijd op het CWI heb ik doorgebracht met Boris. Onze werkhouding lag ver uiteen, maar ons einddoel was hetzelfde, promoveren. Zo nu en dan had ik graag meer op hem geleken. Mervyn was net als ik een PIO (de P blijft geheim, de I en O zijn niet lastig te raden). Vastlopende codes, lange dagen, maar ook de goede gesprekken en pesterijen. Ik zal ze missen. Margreet en Nada verlengden regelmatig de weg naar het toilet of koffieapparaat, meiden bedankt voor de babbels. Vanuit MAS waren er ook de mensen van het eerste uur. Mijn eerste kennismaking met het CWI was met Barry. Hij bood me de mogelijkheid hier mijn stage te lopen en kon me overtuigen voor een promotie te blijven. En dan was er Jaap, mijn kamerbuurman. Zijn gesprekken 's ochtends om kwart voor acht voor we begonnen en zijn enorme optimisme hebben hem een speciale plek gegeven.

Naast onderzoek namen de sociale contacten op het werk een belangrijke plaats in. Met de personeelsvereniging, eerst als lid en uiteindelijk als voorzitter, liepen we over het wad en liet ik me verleiden tot een avond karten (of heb ik toch niet in zo'n ding gezeten?). Beste medebestuurders, dank jullie wel. Siem en Miech, jullie zien ons nog veel vaker! En Walter, als lastige secretaris, je nam altijd de tijd om naar mij te luisteren. Met Martijn, Serge en Harald startte mijn eerste promotiejaar goed. Samen zetten we de PhDays op en al snel kende ik iedere numeriek wiskundige promovendus uit Nederland en Leuven met of zonder zadelpijn of een nat pak van het kanoën.

Een paar mensen wil ik speciaal noemen: Sascha en Fleur, mijn paranimfen, die net als ik bezig zijn met hun eigen boekwerk en die ik na veel rode oren nog altijd mag bellen, mijn ouders, die dan misschien niets van mijn onderzoek begrijpen, maar als geen ander weten hoe ik in elkaar zit en bovenal Harald, die me naast een heel goed wetenschappelijk oor, veel meer heeft gegeven.

Hoofddorp, november 2001

Contents

Chapter 1

Introduction

1.1 Circulation models

The weather affects everyone. It is among the most discussed topics around the world, although its context might differ significantly. Consider, for instance, people in Africa who struggle with drought, whereas we wonder whether it is necessary to bring an umbrella to work; or thousands of people who are being evacuated because their home town is hit by a tornado, whereas others try to decide whether Corsica or Cyprus would be a better location to spend the holidays when it comes to hours of sunshine; or an Egyptian farmer who happily overlooks the Nile flooding and fertilizing its surrounding banks, whereas a Limburger hopes that the heavy rainfall will stop, so the Maas will not burst its banks. For a multitude of reasons, the ability to predict the weather and climate has fascinated people for centuries.

In 1922, Richardson was the first to use numerical modeling as a tool in weather prediction. He acknowledged that to complete a numerical weather prediction an enormous number of calculations had to be made very rapidly. He estimated that a typical global prediction would require a factory of 64.000 people equipped with calculators, see [60]. Consequently, the idea of numerical weather prediction (NWP) was discarded and it was not until the late 1940s that NWP flourished when Von Neumann used one of the first electronic computers (ENIAC) to perform these calculations.

Today, weather and climate prediction rely on so-called global circulation models, i.e., a numerical model for describing the evolution of the state of the atmosphere on a global scale. A circulation model numerically solves a set of equations which represent this evolution. It consists of three main interacting parts. These are data assimilation, numerical dynamics, and physical parametrization:

- Data assimilation involves the incorporation of data from observations into

the model. At the beginning of a forecast an initial guess of the current state of the atmosphere is required. Observations obtained over a certain period of time and at different locations, for instance, from ships, weather stations, radiosondes etc., must be quality controlled and combined to produce this initial condition. In addition, data assimilation is used to correct the global circulation model during or after a forecast simulation.

- The dynamical component is concerned with the numerical solution of the so-called primitive equations of the hydrodynamics in the atmosphere. These equations are the equations of momentum, the continuity equation, an energy equation and an equation of state.

- Physical parametrization is used to incorporate other important physical processes occurring in the atmosphere, for instance, radiation, cumulus convection, large-scale precipitation, and turbulence. Most of these processes occur on scales too small to be directly resolved by the numerical model and can differ significantly in their representation. In addition, each circulation model includes a different parametrization scheme depending on the accuracy required and the computational capacity available to solve the problem.

Figure 1.1 visualizes the components of a circulation model and their interaction pattern.

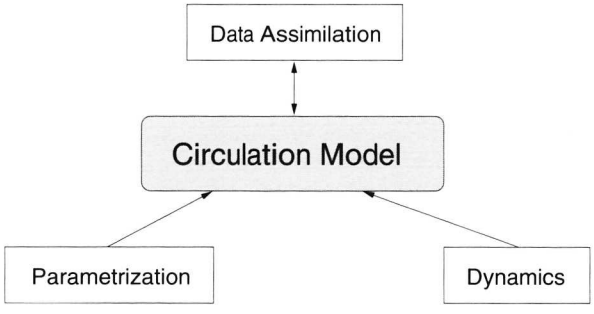


Figure 1.1: Schematic representation of the components in a circulation model.

1.1.1 Horizontal dynamics

In this thesis, we focus on the dynamical part of a circulation model. In particular, we investigate numerical methods to efficiently solve the shallow water equations (SWEs) in spherical geometry. These equations serve as a first prototype of the horizontal dynamics in a global circulation model.

The SWEs can easily be derived from the primitive equations of hydrodynamics. These primitive equations are the classic Navier Stokes equations of fluid mechanics

with the exception that atmospheric motion evolves in a rotating reference system, which introduces an additional force, the Coriolis force. This force is particularly important in large scale atmospheric motion. For a thorough derivation of the primitive and shallow water equations, we refer to [32,55]. The numerical solution of the SWEs is discussed in Chapter 3–5.

1.2 Numerical methods in circulation models

Weather prediction demands results which are as accurate as possible over a time period of a couple of days calculated within given time, say, a couple of hours. Climate simulation, on the other hand, demands that the results remain accurate over a time period which is as long as possible, e.g., several years, decades or even centuries. The accuracy of the prediction depends on the numerical method, the resolution of the space-time grid, the incorporated data and the physical parametrization scheme. Since the computations are known to be very time-consuming, much interest is directed at the development of efficient numerical methods on high-resolution grids. On these grids, the requirements of the numerical scheme for weather and climate prediction practically coincide. In Section 1.3, we summarize our achievements in that direction. First, we discuss typical considerations necessary to obtain an efficient numerical method for solving the horizontal dynamics in spherical geometry.

1.2.1 Spatial discretization schemes

A wide variety of numerical methods underlie the currently operational global circulation models. In particular, there is discussion about which spatial discretization scheme is best to discretize the horizontal dynamics. For instance, the Integrated Forecast System (IFS)-model of ECMWF and the Community Climate Model (CCM)-model operational at NCAR incorporate a spectral transform method, whereas the Global Environmental Multiscale (GEM)-model of the Canadian Meteorological Centre applies a variable-resolution cell-integrated finite element scheme. The GME model of the German Weather Service (DWD) and the Hirlam model developed by a consortium of several European weather services including the Royal Netherlands Meteorological Institute (KNMI) adopt a central finite-difference scheme. For a description of the various operational circulation models, we refer to [9–11,40,49,50,89].

For several decades, the spectral transform method has been most popular. However, over the years, its disadvantages have become more apparent. With increasing grid resolution, its computational costs increase much faster than those of a finite difference or finite element method. Second, the method suffers from Gibbs' phenomenon which occurs for strongly varying variables, such as the concentration of water vapor [59,91]. Third, the method is non-conservative. In view of the aforementioned disadvantages and with the trend toward high-resolution grids,

alternative methods are being explored. We investigate a finite volume method, viz., Osher's finite volume method with a ($\kappa = \frac{1}{3}$)-scheme for the constant state interpolation, see Section 1.3.1 and Chapter 3.

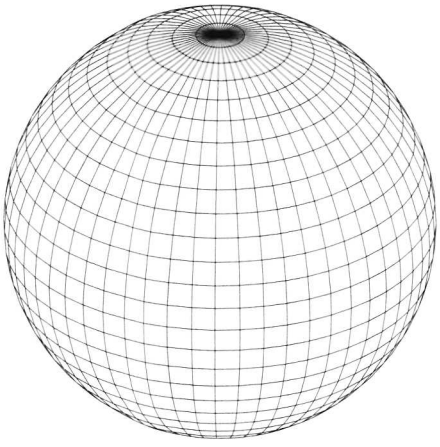
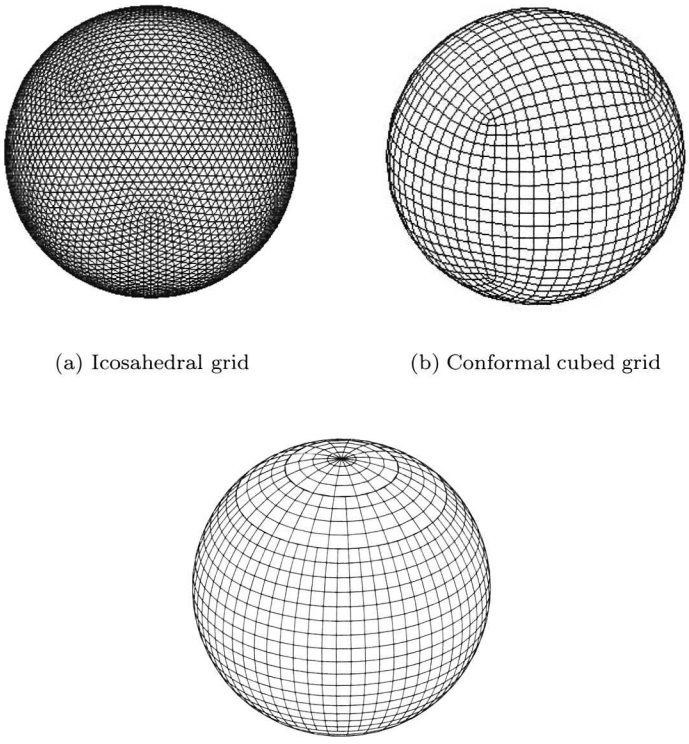


Figure 1.2: A uniform latitudinal-longitudinal grid over the sphere.

1.2.2 The pole problem

A common prejudice against finite difference and finite volume methods concerns their inefficiency due to a severe step size restriction when applied on a standard uniform latitudinal-longitudinal (lat-lon) grid with an explicit time integration method to solve the resulting semi-discrete system. A standard uniform lat-lon grid uses grid lines of constant latitude (parallels) and longitude (meridians), see Figure 1.2. The inefficiency has to do with the pole problem, which includes all problems related to the non-existence of the longitudinal unit vector in the poles and the convergence of the meridians when approaching them. The pole problem can be resolved in several ways: (1) by a filter suppressing irrelevant high-frequency waves, (2) by a different grid distribution and/or a different coordinate system, or (3) by the application of an implicit time integration method. The first and second approach have been investigated extensively. A detailed discussion of various filters is presented in, for instance, [57,70]. For a description of several grid types, we refer to [3,41,86] for the reduced grid, to [27,28,64,84] for the icosahedral or geodesic grid and to [58,62,63] for the cubic grid. These different grid types are displayed in Figure 1.3. They all aim at a redistribution of the grid cells over the sphere to obtain a quasi-uniform cell distribution to alleviate the step size restriction. In addition, they remove the singularity at the poles by introducing a non-singular coordinate system, which is necessarily composite or non-conformal.



(a) Icosahedral grid

(b) Conformal cubed grid

(c) Reduced grid

Figure 1.3: Various grid distributions on the sphere.

The third remedy, the use of an implicit time integration method, has so far only been applied efficiently in a semi-Lagrangian semi-implicit method. Its basic principle is explained below. The shallow water equations exhibit two different types of wave-like solutions. These are the slowly varying advective wave propagating with the wind velocity and the much faster low-amplitude gravity waves. The latter low energy waves have no significant role in atmospheric circulation patterns. Unfortunately, these fast waves dictate the admissible step size in explicit time integration methods. In the semi-Lagrangian semi-implicit method, the governing equations are integrated along the characteristic corresponding to the advective wave, whereas the gravity waves are solved with an implicit time integration method. For a thorough review of the semi-Lagrangian method, we refer to [71].

We investigate two possible remedies for the pole problem: (1) a combined lat-lon reduced grid with two stereocaps in the polar region and (2) a linearly-

implicit Rosenbrock time integration method (Ros3) combined with approximate matrix factorization (AMF) applied to the full Eulerian form of the shallow water equations on a uniform lat-lon grid. The application of a fully implicit scheme is commonly assumed to be too expensive. We will, however, refute this assumption.

1.3 Efficient numerical methods

In this section, we briefly introduce the various components of the numerical methods discussed in this thesis.

1.3.1 Osher's scheme

The continuous SWEs can be presented in conservation form. In this formulation, the dependent variables describing the state of the atmosphere, are directly derived from the underlying physical conservation laws, i.e., conservation of mass, momentum and energy. To guarantee the conservation of these quantities in the numerical approximation, we apply a finite volume method to spatially discretize the SWEs.

In a finite volume method the sphere is divided into a number of grid cells, finite volumes, over which the conservation form of the SWEs is integrated in space. This gives the more natural integral form of the conservation laws applied to each finite volume. To discretize the resulting integrals, we apply an upwind scheme. An upwind scheme is favored, because it incorporates information about the characteristic waves of the shallow water problem into the numerical solution process. Furthermore, an upwind scheme can be combined with a so-called limiter to ensure a smooth capturing of variables with large gradients, which makes this combination preferable to a spectral transform method.

We have chosen Osher's approximate Riemann solver for evaluating the flux between volumes. Our motivation for this choice is as follows. First, Osher's scheme is robust and second-order accurate when combined with a proper state interpolation. Second, Osher's scheme has an excellent boundary treatment, which makes Osher's solver preferable to, for instance, Roe's solver. This property might seem irrelevant for the SWEs, because they describe a pure initial value problem. However, it is valuable for the information exchange at the interface between different subgrids in a composite grid. Finally, Osher's scheme is an upwind scheme of flux difference splitting (FDS) type. Flux vector splitting (FVS) schemes are not applicable in this case, since the necessary homogeneity condition of the flux is not fulfilled. For a thorough discussion on upwind schemes we refer to [31, 82]. Osher's scheme is extensively studied in Chapter 3 when applied to the SWEs.

1.3.2 The combined grid

In addition to the standard uniform lat-lon grid, we consider a combined grid composed of a stereographic grid at the two polar caps and a reduced lat-lon grid in

the intermediate region, see Figure 1.4. Similar to other alternative grid distributions, this combined grid redistributes the grid cells over the sphere to alleviate the step size restriction for explicit time integration methods. In addition, this grid distribution has no singular points. Each of the coordinate systems is conformal, which means that the metric coefficient associated with the coordinate transformation only depends on the spatial variable and not on its direction. Consequently, the flux evaluations on the aforementioned grids are straightforward.

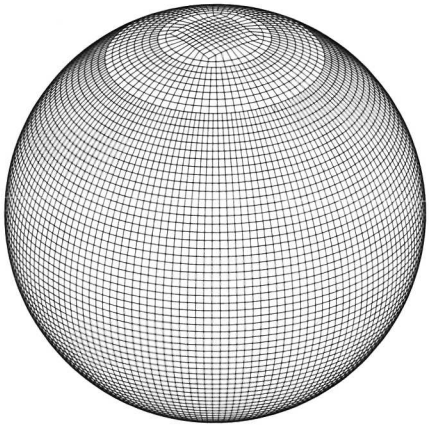


Figure 1.4: A combined grid consisting of a reduced lat-lon grid away from the poles and a stereographic grid at the two polar caps.

The idea of the combined grid originates from earlier attempts by Phillips, who suggested to cover the sphere with three different coordinate systems, viz., the mercator and two stereographic projections [56]. Unfortunately, his spatial discretization scheme required the interpolation of grid points in neighboring grids whenever a variable outside the current grid part was needed. These interpolations could lead to a loss of mass as was shown by Stoker [74]. In our case, such complications are avoided because of the mass conservation properties of the Osher's scheme and its consistent boundary treatment.

1.3.3 A third-order Rosenbrock method with approximate matrix factorization

To avoid a severe step size restriction when calculating on a uniform lat-lon grid, we also propose a third-order Rosenbrock method (Ros3) to integrate the semi-discrete SWEs in time. A Rosenbrock method is a linearly-implicit Runge-Kutta method which solves general non-linear ODE systems, $\dot{\mathbf{w}} = \mathbf{F}(\mathbf{w})$, see, e.g., [13, 25]. This method is called linearly implicit since per time step it requires the solution of

linear systems rather than non-linear ones. In this sense, the method is intermediate between an explicit and implicit Runge-Kutta method. The linear system solves are expensive, but can be simplified significantly to reduce the computational costs. For that purpose, approximate matrix factorization (AMF) is applied, see, e.g., [2, 14, 34, 54].

The combination of Ros3 with AMF leads to an efficient time integration method, while preserving the important properties of A-stability and third-order accuracy of the original Ros3 method. The A-stability property suggests that large step sizes are admissible in the numerical approximation of the evolution in the shallow water system. Ros3 with AMF is studied in Chapter 4 and Chapter 5.

1.4 Air quality models

Another important group of atmospheric models consists of air quality (or transport-chemistry) models, which are used to describe the chemical composition of the atmosphere. These models are used to study the effects of air pollution. The chemical composition of the atmosphere is altered by chemical reactions, advection, diffusion, emissions and depositions, which are all included in the model. Advection is the transport of a species by a given wind field. Diffusion represents the turbulent mixing of the species.

Global circulation and air quality models are sometimes connected. In so-called on-line air quality models the transported species are treated as additional variables in a complete circulation model. As a consequence, their concentration is directly affected by the calculated wind field. A simpler approach is provided by a so-called off-line model. In these models, the transported species are advected by a given wind field from a meteorological database. In that case, there is no direct interaction with results from a circulation model. In Chapter 2, we investigate a specific numerical technique, viz. operator splitting, and its effects when applied in air quality models. For further reading on air quality models, we refer to [20, 67, 80, 86].

1.5 Operator Splitting

A numerical technique often applied in circulation and air quality modeling is operator splitting. This technique subdivides the full problem in a number of subprocesses, which can then be solved with different numerical techniques and step sizes suitable to the specific subprocess. An air quality model is commonly subdivided, e.g., in the advection, the diffusion, and the chemistry part. The solution of the chemistry part requires a numerical method adapted to efficiently solve stiff problems, whereas the advection part requires a numerical advection scheme which respects the underlying mass-conservation laws and avoids the generation of over- or undershoots. Undershoots can lead to nonphysical negative concentrations, which

severely lowers the robustness of the solution process, as they can introduce instability. A global circulation model incorporates operator splitting at various levels. For instance, most models separately treat the physical parametrization and dynamical part. The first part is very time-consuming, requiring an efficient solution method which permits large step sizes, whereas the dynamical part requires a more frequent update. Other splittings concern the subdivision of the horizontal and vertical dynamics or the subdivision of the horizontal dynamics in the longitudinal and latitudinal direction. These splittings are often referred to as dimensional splittings, since the operators are subdivided along a specific direction of movement. Finally, we mention the subdivision of the horizontal dynamics in the advection part, and the Coriolis and pressure gradient forces.

1.5.1 The splitting error

Operator splitting significantly simplifies the numerical solution process. Unfortunately, this simplification has one disadvantage. The separate treatment of the various subprocesses creates a splitting error. The magnitude of this error must be controlled and may not lead to an unstable solution process. In Chapter 2, we investigate this error for a Strang splitting method [75] which adopts a symmetrical order of reappearance to solve the different subprocesses. This splitting error is known to be of second-order. We focus on pure initial value problems. An expression for this error is derived by the application of the Lie operator formalism which facilitates the analysis of the splitting error for a coupled non-linear system of partial differential equations. The error expressions are investigated in more detail for advection-diffusion-reaction equations as used in air quality modeling.

1.5.2 Approximate matrix factorization vs Strang splitting

Like operator splitting, approximate matrix factorization (AMF) is used to simplify a numerical solution process and to make this process cost effective. In Chapter 5, we compare both techniques when applied to the SWEs. We investigate Ros3 with AMF and Strang splitting combined with a third-order Rosenbrock method to integrate the subprocesses in time. We are interested in the local error and the numerical dispersion relations. The numerical dispersion relations demonstrate the influence of the numerical method on the characteristic waves of the shallow water problem. The advective (or Rossby) wave must be represented accurately, because it describes an important part of atmospheric dynamics.

In meteorological practice, operator splitting techniques are considered inappropriate for solving the primitive equations when they split the advective and Coriolis terms. Together, these terms generate the Rossby waves. The separate treatment of the advection and Coriolis terms appears to jeopardize a correct representation of these waves and therefore apparently obstructs a correct representation of the atmospheric tendency to geostrophic balance. Ros3 with AMF on the other hand,

accurately resolves these waves, see Chapter 5.

1.6 A future perspective

This thesis investigates efficient numerical solution methods for solving the 2D shallow water equations in spherical geometry. Our ultimate objective is to extend and apply these methods to more realistic 3D models simulating global atmospheric circulation. This extension however, requires some precaution. For instance, theoretical results from Hundsdorfer [36,37] predict that Ros3 with AMF factorized in three dimensions is no longer A-stable. This indicates that in 3D practice, it is no longer possible to take large step sizes, while maintaining a stable solution process. If necessary, these deficiencies can be resolved, for instance, by the application of a dimensional splitting method solving the horizontal and vertical dynamics, separately.

To investigate these matters and possible solutions, a 3D test case is required. For testing new numerical methods to be used in circulation models, Williamson *et al* developed a standard 2D SWEs test set [88]. Such a standard test set is not available for 3D applications, although the dynamical intercomparison project [15] provides an alternative. This test case includes a 3D dynamical part extended with two simple forcing terms simulating the effects of radiation and vertical turbulence. Held and Suarez [29], Boer and Denis [5], Williamson *et al* [90] all proposed a simple physical parametrization scheme for these processes. Unfortunately, a standard reference solution is not provided for the dynamical core test case. Therefore, we are currently investigating a 3D instationary variant of the Ekman boundary layer. Results are not presented in this thesis.

1.7 Outline of this thesis

This thesis is organized as follows.

In chapter 2, we focus on the Strang splitting method applied to arbitrary autonomous systems of differential equations. An expression is derived for the Strang splitting error using the Lie operator formalism, the concept of commutators for non-linear problems, the modified problem and the Baker-Campbell-Hausdorff formula. The error is analyzed in greater detail for the advection-diffusion-reaction equations, resulting in a theorem which shows under which conditions advection, diffusion and reaction commute. When all processes commute, no splitting error is found.

The next two chapters are closely related. Both chapters discuss efficient numerical methods for solving the SWEs in spherical geometry and for avoiding the pole-problem. Their perspectives are different.

Chapter 3 focuses on the spatial discretization of the SWEs. A combined latlon reduced grid with two stereocaps is proposed. Special attention is paid to

the connection problem at the grid interface between the stereocaps and the lat-lon reduced grid. Osher's scheme is chosen to spatially discretize the SWEs. In addition to its favorable properties inherent to a finite volume method, Osher's scheme easily resolves this connection problem. Numerical results for Test 2 of the SWEs test set support these qualities of Osher's scheme and the combined grid.

In Chapter 4, the linearly-implicit A-stable Ros3 time integration method is discussed. The SWEs are linearized to investigate stability for the combination of Ros3 with approximate matrix factorization. Calculations are performed on a uniform lat-lon grid. This combination proves to be cost-effective, while maintaining the favorable properties of the original Ros3 method. Its efficiency is demonstrated by a comparison to the third-order explicit Runge-Kutta method applied on the combined grid proposed in Chapter 3. Again, numerical results are given for the SWEs test set.

In Chapter 5, Ros3 with AMF is further explored. Its local error and numerical dispersion relations are studied for the SWEs in spherical geometry. A comparison is made between this method and an alternative method for simplifying the numerical solution process, viz. Strang splitting. Theoretical and numerical results are derived. The analysis shows that Ros3 with AMF makes a good candidate to efficiently solve the semi-discrete SWEs on a global fine resolution uniform lat-lon grid. Strang splitting on the other hand, is inadequate in view of its inefficiency due to a large local error in the polar region.

Chapter 2

Analysis of Operator Splitting for Advection-Diffusion-Reaction Problems from Air Pollution Modeling

Summary

Operator or time splitting is often used in the numerical solution of initial boundary value problems for differential equations. It is, for example, standard practice in computational air pollution modeling where we encounter systems of three-dimensional, time-dependent partial differential equations of the advection-diffusion-reaction type. For such systems little attention has been devoted to the analysis of splitting and to the question why splitting can work so well. From the theoretical point of view, the success of splitting is primarily determined by the splitting error. This paper presents an analysis of operator splitting aimed at providing insight into the splitting error. Using the Lie operator formalism, a general expression is derived for a three-term Strang splitting in the pure initial value case. For a class of advection-diffusion-reaction problems the splitting error is analyzed in greater detail. A special case is discussed in which the splitting error can be reduced. Also some attention is paid to the use of operator splitting in initial boundary value problems.

2.1 Introduction

Virtually all processes modeled by time-dependent partial differential equations (PDEs) split additively in subprocesses for which simpler PDEs exist. This greater simplicity also carries over to their numerical counterparts, which already a long time ago has led to the use of operator splitting or time splitting. Within operator splitting subprocesses are treated on their own in numerical time-stepping while adopting a certain order of reappearance. An early influential paper is Strang [75], where a symmetrical order of reappearance was proposed, which formally yields 2nd-order consistency.

In this paper we focus on this form of symmetrical Strang splitting for systems of advection-diffusion-reaction equations.

$$\frac{\partial c}{\partial t} + \nabla \cdot (\underline{u}\, c) = \nabla \cdot (K \nabla c) + R(c), \quad c = c(\underline{x}, t), \quad \underline{x} \in \mathbb{R}^3. \tag{2.1}$$

Although our findings do have a wider scope, our motivating application is atmospheric air quality modeling where PDE systems like (2.1) lie at the heart of complicated models employed in studies on the chemical composition of the atmosphere. The societal motivation for these studies concerns air pollution. Throughout we suppose that the velocity vector $\underline{u}$ and the diffusion coefficient matrix K are given. Hence the problem is linear with respect to advection and diffusion, but nonlinear in the chemical reaction term R. The dependent variable c represents a vector of chemical species concentrations, which evolve in time due to advection, diffusion, chemical interactions, emissions, and depositions, the latter three are all contained in R.

To the best of our knowledge, one of the first influential papers on computational air quality modeling discussing splitting is McRae, Goodin and Seinfeld [20]. More references specifically concerning air quality modeling can be found in Zlatev [92]. Nowadays operator splitting is standard practice in this field. However, for PDE systems like (2.1), in the literature very little attention has been devoted to the analysis of splitting and to the question why splitting can work so well. From the theoretical point of view, the success of splitting is primarily determined by the splitting error, which is introduced by solving subproblems one after another in a completely decoupled manner. In general this splitting error always exists, also when all subproblems are solved exactly. The aim of this paper is to present an analysis of operator splitting and to provide insight into the splitting error.

In Section 2.2 we derive an expression for the Strang splitting error for arbitrary autonomous systems of differential equations using the Lie operator formalism, including the notion of commutators for nonlinear problems, the notion of the modified problem and the celebrated Baker-Campbell-Hausdorff formula. Here we have made fruitful use of material from Sanz-Serna [65] and Sanz-Serna and Calvo [66]. Section 2.3 focuses on the advection-diffusion-reaction problem (2.1). The body of this section consists of a theorem, which shows under which circumstances advection, diffusion and reaction commute with one another, assuming exact integration.

This commutativity is of great importance, because when all processes commute, we have a zero splitting error. In Section 2.4 the splitting error is discussed in greater detail for a number of simplified test models. Simplifications cannot be avoided since for the general problem class (2.1) the error expressions are much too long to handle. Further we discuss ways to reduce the splitting error and address the subject of inconsistencies, which can occur if Strang splitting is used in case of initial boundary value problems. The final Section 2.5 summarizes our findings and contains a number of general remarks.

2.2 Strang splitting and the Lie operator formalism

In this section we will derive an expression for the Strang splitting error for the general, nonlinear, autonomous system of differential equations,

$$c_t = f(\underline{x}, c) \equiv f_1(\underline{x}, c) + f_2(\underline{x}, c) + f_3(\underline{x}, c), \quad t \in [t_0, T], \quad \underline{x} \in \mathbb{R}^d, \quad c(\underline{x}, t_0) = c_0(\underline{x}). \tag{2.2}$$

The solution $c(\underline{x}, t)$ is supposed to be vector-valued in $\mathbb{R}^m$ and f and its parts f_1, f_2 and f_3 can represent a nonlinear vector function in $\mathbb{R}^m$ or some spatial derivative operator. In our notation we will mostly, just for convenience, suppress the dependence on the spatial variable $\underline{x} = (x, y, z)$. The spatial dimension d is not yet fixed. To derive the splitting error expression, at this stage we merely consider an abstract initial value problem (2.2) in the function space $\mathbf{S}$ of real, sufficiently often differentiable vector-valued functions c on $\mathbb{R}^d \times [t_0, T]$. In addition we assume that all operators encountered in our derivations, are sufficiently differentiable in all their variables. Our starting problem (2.1) provides a particular example for (2.2).

2.2.1 Strang splitting

Let $S(\tau)$ denote the solution (semigroup) operator for (2.2), that is

$$c(t + \tau) = S(\tau)\, c(t),$$

and $S_k(\tau)$ the solution operator for the subproblem $c_t = f_k(c)$. Let $\tilde{S}_k(\tau)$ denote a consistent, numerical approximation to $S_k(\tau)$, for example defined by a Runge-Kutta type method. For the abstract initial value problem (2.2), we then compactly represent the celebrated Strang splitting scheme [75] by

$$\tilde{c}(t + \tau) = \tilde{S}(\tau)\, \tilde{c}(t), \quad \tilde{S}(\tau) \equiv \tilde{S}_1\left(\tfrac{1}{2}\tau\right) \tilde{S}_2\left(\tfrac{1}{2}\tau\right) \tilde{S}_3(\tau)\, \tilde{S}_2\left(\tfrac{1}{2}\tau\right) \tilde{S}_1\left(\tfrac{1}{2}\tau\right). \tag{2.3}$$

The solution $\tilde{c}(t + \tau)$ denotes the approximation to $c(t + \tau)$ resulting from approximately solving the subproblems $c_t = f_k(c)$ in the given sequential order. The

solution operator $\tilde{S}$ is the resulting splitting approximation to S. Note that $\tilde{S}_k$ is still thought to be space continuous, that is without spatial discretization. In our derivation we will not specify $\tilde{S}_k$, but instead we assume that with $\tilde{S}_k$ we may associate the modified problem [65,66][1].

$$c_t = F_k(c) \equiv f_k(c) + \tau^{p_k} E_k(c). \tag{2.4}$$

where $\tau^{p_k} E_k(c)$ represents the local truncation error of the integration method defining $\tilde{S}_k$. The integer p_k is the order of consistency. By definition, as the local error of integration schemes is normally an infinite series expansion in τ, E_k itself may still depend on the step size τ. The modified problem concept is very convenient when it is combined with the Lie operator formalism introduced below. Adopting the modified problem concept means that we act as if we apply Strang splitting to the modified problem,

$$c_t = F(c) \equiv F_1(c) + F_2(c) + F_3(c). \tag{2.5}$$

while solving the subproblems $c_t = F_k(c)$ exactly. Trivially, with $\tilde{S}_k$ one may associate the exact solution operator S_k, in which case the original subproblems $c_t = f_k(c)$ are supposed to be solved exactly, that is without time integration error.

2.2.2 The Lie operator formalism

Strang splitting always leads to a second-order approximation, at least in a formal sense. We are interested in the structure of the splitting error. Albeit tedious, local splitting errors can always be obtained by straightforward Taylor expansions (see for example [46, 75]). This, however, leads to an expression which does not reveal in a clear way the structure of the error. For its derivation we therefore adopt the Lie operator formalism. This formalism will enable use of the celebrated Baker-Campbell-Hausdorff formula. The BCH formula yields a lot of insight in the particular structure of splitting errors. The authors learned the Lie operator formalism from [65,66]. For selfcontainedness we here repeat the material from [65, 66] needed for our purpose. We also made fruitful use of a brief unpublished note of our colleague W. Hundsdorfer, who also refers to [66]. A nice introduction to Lie operators can also be found in [23].

Consider the general differential equation (2.5). With each given operator F, a Lie operator is associated, which we denote by $\mathcal{F}$. This Lie operator is a linear operator acting on the space of operators defined on $\mathbf{S}$. $\mathcal{F}$ maps each operator G into the new operator $\mathcal{F}G$, such that for any element $c \in \mathbf{S}$,

$$(\mathcal{F}G)(c) = G'(c) F(c). \tag{2.6}$$

[1] Throughout we use $c \in \mathbf{S}$ to denote the solution of any differential equation. From the context it will be clear to which equation we are referring, for example our original problem (2.2) or a different problem such as (2.4). Likewise, c can denote an arbitrary element in $\mathbf{S}$.

($'$ denotes differentiation with respect to c). For the solution $c(t)$ of (2.5) it easily follows that

$$(\mathcal{F}G)(c(t)) = \frac{\partial}{\partial t} G(c(t)), \tag{2.7}$$

and from induction to k that

$$\frac{\partial^k}{\partial t^k} G(c(t)) = \left(\mathcal{F}^k G\right)(c(t)). \tag{2.8}$$

The above relations (2.7) and (2.8) hold for any G defined on $\mathbf{S}$, in particular for the identity I. Inserting I for G and using the Taylor expansion of the true solution, we can write $c(t+\tau)$ in terms of the exponentiated Lie operator form or Lie-Taylor series,

$$c(t+\tau) = \left(e^{\tau\mathcal{F}} I\right)(c(t)).$$

The same argument concerning this exponentiated Lie operator applies to each of the subproblems $c_t = F_k(c)$. When we compose the resulting exponentiated Lie operators in the same order as the solution operators in the splitting procedure, with which they are associated, we can reveal that the Strang splitting solution (2.3) can be expressed as

$$\tilde{c}(t+\tau) = \left(e^{\frac{1}{2}\tau\mathcal{F}_1} e^{\frac{1}{2}\tau\mathcal{F}_2} e^{\tau\mathcal{F}_3} e^{\frac{1}{2}\tau\mathcal{F}_2} e^{\frac{1}{2}\tau\mathcal{F}_1} I\right)(\tilde{c}(t)). \tag{2.9}$$

At this stage the BCH formula proves to be useful. Let X, Y be linear operators. According to this formula, the product $e^X e^Y$ can then be written as the exponential e^Z of

$$Z = X + Y + \frac{1}{2}[X,Y] + \frac{1}{12}([X,X,Y] + [Y,Y,X]) + \frac{1}{24}[X,Y,Y,X] + \cdots, \tag{2.10}$$

where $[X,Y]$ is the commutator $[X,Y] = XY - YX$ and $[X,X,Y]$ is recursively defined by $[X,X,Y] = [X,[X,Y]]$, etc. Note that, if X and Y are Lie operators, Z is also a Lie operator.

We put $X = \frac{1}{2}\tau\mathcal{F}_1$ etc. and apply (2.10) four times, or Yoshida's formula [66] twice, resulting in an expression for the symmetrical Strang splitting solution (2.9),

$$\tilde{c}(t+\tau) = \left(e^{\tau\tilde{\mathcal{F}}} I\right)(\tilde{c}(t)), \quad e^{\tau\tilde{\mathcal{F}}} \equiv e^{\frac{1}{2}\tau\mathcal{F}_1} e^{\frac{1}{2}\tau\mathcal{F}_2} e^{\tau\mathcal{F}_3} e^{\frac{1}{2}\tau\mathcal{F}_2} e^{\frac{1}{2}\tau\mathcal{F}_1},$$

where the new Lie operator $\tilde{\mathcal{F}}$ is formally defined by an infinite series expansion which is even in τ. Its leading part reads

$$\begin{aligned}\tilde{\mathcal{F}} = {} & \mathcal{F}_1 + \mathcal{F}_2 + \mathcal{F}_3 - \frac{1}{24}\tau^2[\mathcal{F}_1,\mathcal{F}_1,\mathcal{F}_2] - \frac{1}{24}\tau^2[\mathcal{F}_1,\mathcal{F}_1,\mathcal{F}_3] + \\ & + \frac{1}{12}\tau^2[\mathcal{F}_2,\mathcal{F}_2,\mathcal{F}_1] - \frac{1}{24}\tau^2[\mathcal{F}_2,\mathcal{F}_2,\mathcal{F}_3] + \frac{1}{12}\tau^2[\mathcal{F}_3,\mathcal{F}_3,\mathcal{F}_1] + \\ & + \frac{1}{12}\tau^2[\mathcal{F}_3,\mathcal{F}_3,\mathcal{F}_2] + \frac{1}{12}\tau^2[\mathcal{F}_2,\mathcal{F}_3,\mathcal{F}_1] + \frac{1}{12}\tau^2[\mathcal{F}_3,\mathcal{F}_2,\mathcal{F}_1] + \mathcal{O}(\tau^4).\end{aligned} \tag{2.11}$$

If we are able to recover the operator $\tilde{F}$ corresponding with $\tilde{\mathcal{F}}$, we are led to the modified problem,

$$c_t = \tilde{F}(c),$$

associated with the symmetrical Strang splitting scheme.

We first derive the operators associated with the commutators (the so-called Lie or Poisson brackets). Direct application of (2.6) to the commutator $[\mathcal{F}_l, \mathcal{F}_m]$ yields for any G and any $c \in \mathbf{S}$,

$$[\mathcal{F}_l, \mathcal{F}_m] G(c) = (G'(c) F_m(c))' F_l(c) - (G'(c) F_l(c))' F_m(c).$$

Repeating this for $[\mathcal{F}_k, \mathcal{F}_l, \mathcal{F}_m]$ and inserting the identity I for G, gives

$$[\mathcal{F}_k, \mathcal{F}_l, \mathcal{F}_m] I(c) = (F_m' F_l)' F_k - (F_l' F_m)' F_k - (F_k' F_m)' F_l + (F_k' F_l)' F_m,$$

where all operators at the right-hand side are evaluated at c. We rewrite this expression as

$$[\mathcal{F}_k, \mathcal{F}_l, \mathcal{F}_m] I(c) = F_{lm}' F_k - F_k' F_{lm}, \quad F_{lm} \equiv F_m' F_l - F_l' F_m, \tag{2.12}$$

where, naturally, the new operator F_{lm} is called the commutator for F_l and F_m. To find $\tilde{F}$ we insert expression (2.12) for all commutators occurring in (2.11), which results in the modified problem for the Strang splitting (2.3),

$$c_t = \tilde{F}(c) \equiv F(c) + \tau^2 E_F(c) + \mathcal{O}(\tau^4), \tag{2.13}$$

where $\tau^2 E_F(c)$ is the counterpart of the τ^2-term of (2.11). Remember here equation (2.7). After rearranging the terms, to make the contribution of splitting F_1 from F_2, F_1 from F_3 and F_2 from F_3 to the splitting error more precise, E_F is written as

$$\begin{aligned} E_F \equiv & -\frac{1}{24} F_{12}' (F_1 + 2F_2 + 2F_3) + \frac{1}{24} (F_1' + 2F_2' + 2F_3') F_{12} \\ & -\frac{1}{24} F_{13}' (F_1 + 2F_2 + 2F_3) + \frac{1}{24} (F_1' + 2F_2' + 2F_3') F_{13} \\ & -\frac{1}{24} F_{23}' (F_2 + 2F_3) + \frac{1}{24} (F_2' + 2F_3') F_{23}. \end{aligned} \tag{2.14}$$

The solution of the modified problem (2.13), assuming it exists, may be interpreted as the Strang splitting solution (backward analysis interpretation [65]).

The term $\tau^2 E_F(c(t))$ represents the leading term of the local error of the Strang splitting scheme evaluated at $c(t)$. Note that the global error, $\tilde{c}(t+\tau) - c(t+\tau)$, can be directly seen to satisfy

$$\tilde{c}(t+\tau) - c(t+\tau) = \left(e^{\tau \tilde{\mathcal{F}}} I\right) (\tilde{c}(t) - c(t)) + \left(e^{\tau \tilde{\mathcal{F}}} I - e^{\tau \mathcal{F}} I\right) (c(t)),$$

where $(e^{\tau\tilde{F}}I - e^{\tau F}I)\,c(t))$ is the complete local splitting error. The local splitting error is even in τ provided that the Lie operators are independent of τ or also even in τ. The leading τ^2-term is of course equal to the τ^2-term in (2.11).

A few important aspects concerning the splitting error should already be mentioned. When the three split operators F_1, F_2, F_3 commute with one another, $\tilde{F} = F$, no splitting error occurs. When, for example, only F_1 and F_2 commute, the first and second term connected with the commutator F_{12} cancel and no error occurs due to splitting F_1 from F_2. It is the Lie operator approach that attends to this clarity. The beauty of this approach is that it can be formulated for any autonomous operator F with its split parts F_1, F_2, F_3.

What remains to be done is to identify the local splitting error for the original problem (2.2) that would arise if the substeps would be integrated exactly. For that purpose we work the modified problem expression (2.4) into (2.13) and (2.14). A straightforward computation then leads to

$$c_t = \tilde{f}(c) \equiv f(c) + \tau^2 E_f(c) + \mathcal{O}\left(\tau^{2+p_1}\right) + \mathcal{O}\left(\tau^{2+p_2}\right) + \mathcal{O}\left(\tau^{2+p_3}\right) + \mathcal{O}\left(\tau^4\right), \tag{2.15}$$

where

$$\tau^2 E_f(c) = \tau^2 E_s(c) + \tau^{p_1} E_1(c) + \tau^{p_2} E_2(c) + \tau^{p_3} E_3(c),$$

with E_s defined by

$$\begin{aligned} E_s \equiv\ & -\frac{1}{24} f_{12}' (f_1 + 2f_2 + 2f_3) + \frac{1}{24} (f_1' + 2f_2' + 2f_3') f_{12} \\ & -\frac{1}{24} f_{13}' (f_1 + 2f_2 + 2f_3) + \frac{1}{24} (f_1' + 2f_2' + 2f_3') f_{13} \\ & -\frac{1}{24} f_{23}' (f_2 + 2f_3) + \frac{1}{24} (f_2' + 2f_3') f_{23}. \end{aligned} \tag{2.16}$$

We see that in (2.15) the leading term consists of the sum of the three local integration errors introduced in (2.4) and the error term $\tau^2 E_s(c)$. The operator E_s obviously defines the leading term of the local splitting error for exact integration. That is, if all split steps would be integrated exactly, or just very accurately, then this term will dominate the local splitting error. On the other hand, if f_1, f_2, f_3 commute with one another, E_s will completely vanish. This means that the success of Strang splitting in terms of local accuracy is determined by E_s in the first place.

2.3 Advection-diffusion-reaction problems

In this section we will consider the advection-diffusion-reaction problem (2.1). In relation to (2.2) we associate f_1 with advection, f_2 with diffusion and f_3 with chemistry, that is

$$f_1(c) = -\nabla \cdot (\underline{u} c), \quad f_2(c) = \nabla \cdot (K \nabla c), \quad f_3(c) = R(c).$$

Observe that the velocity $\underline{u} = (u, v, w)$, the diffusion matrix coefficient K and the reaction term $R(c)$ do depend on the spatial variable $\underline{x} = (x, y, z)$. Also note that no component coupling exists in the advection and diffusion parts as opposed to the chemistry part $R(c)$ ($R(c) \in \mathbb{R}^m$).

2.3.1 Commutativity

First we will answer the question when true commutativity occurs between the advection, diffusion and chemistry operators. In that case no splitting error exists between the commutating processes. To find the answer we have to elaborate the commutators.

$$f_{lm}(c) = f'_m(c) f_l(c) - f'_l(c) f_m(c), \quad (l, m) = (1, 2), (1, 3), (2, 3),$$

and equate them to zero. In this elaboration the derivatives $f'_1(c)$ and $f'_2(c)$ are to be interpreted componentwise. They in fact act as diagonal matrix differential operators having equal entries. More precisely, owing to their linearity we have, for any element $s \in \mathbf{S}$,

$$f'_1(c)\, s \equiv f_1(s) = -\nabla \cdot (\underline{u} s), \quad f'_2(c)\, s \equiv f_2(s) = \nabla \cdot (K \nabla s).$$

Trivially, the derivative $f'_3(c)$ is the $m \times m$ Jacobian matrix $R'(c)$. Our elaboration leads to the following theorem.

Theorem 1

a) *Advection commutes with diffusion if $\underline{u}$ and K are independent of $\underline{x}$.*

b) *Advection commutes with chemistry if $\nabla \cdot \underline{u} = 0$ and R is independent of $\underline{x}$.*

c) *Diffusion commutes with chemistry if R is linear in c and independent of $\underline{x}$.*

d) *With exact integration no splitting error exists if R is linear in c and $\underline{u}$. K and R are independent of $\underline{x}$.*

Result d) is based on a), b), c) for which the proof is given below. Results a) and d) can also be concluded from Fourier analysis (the standard constant coefficient case). Note that the requirement R independent of $\underline{x}$ does not mean that R is independent of $c = c(\underline{x}, t)$.

Proof.

a) For commutativity of advection and diffusion we need equality of

$$f'_2(c) f_1(c) = -\nabla \cdot (K \nabla (\nabla \cdot (\underline{u} c))) \text{ and } f'_1(c) f_2(c) = -\nabla \cdot (\underline{u} (\nabla \cdot (K \nabla c))).$$

Recall that c is a vector but that $\underline{u}$ and K act componentwise. Further elaborating these two expressions trivially shows equality, if both $\underline{u}$ and K are independent of

$\underline{x}$. In general the two expressions are not equal.
b) We need to compare

$$f_3'(c) f_1(c) = -R'(c) \nabla \cdot (\underline{u}c) \text{ and } f_1'(c) f_3(c) = -\nabla \cdot (\underline{u} R(c)).$$

Let $R_x(c)$ denote the partial derivative vector of $R(\underline{x}, c)$ with respect to x. Introduce a similar meaning for $R_y(c)$ and $R_z(c)$. An elementary calculation yields

$$f_3'(c) f_1(c) = -R'(c) (\underline{u} \cdot \nabla c) - R'(c) (\nabla \cdot \underline{u}) c,$$

and

$$\begin{aligned} f_1'(c) f_3(c) &= -(u R(c))_x - (v R(c))_y - (w R(c))_z \\ &= -R'(c) (\underline{u} \cdot \nabla c) - (\nabla \cdot \underline{u}) R(c) - (u R_x(c) + v R_y(c) + w R_z(c)). \end{aligned}$$

The two expressions are equal if the velocity field is divergence-free and R is independent of x, y and z. This proves part b) of the theorem. Note that in this case R is allowed to depend on c.
c) For commutativity of diffusion and chemistry we need equality of

$$f_3'(c) f_2(c) = R'(c) (\nabla \cdot (K \nabla c)) \text{ and } f_2'(c) f_3(c) = (\nabla \cdot (K \nabla)) R(c).$$

Introduce the vectors,

$$X = R_x(c) + R'(c) c_x, \quad Y = R_y(c) + R'(c) c_y, \quad Z = R_z(c) + R'(c) c_z.$$

Then we can write

$$f_2'(c) f_3(c) = \frac{\partial}{\partial x} (K_{11} X + K_{12} Y + K_{13} Z) + \frac{\partial}{\partial y} (K_{21} X + K_{22} Y + K_{23} Z) + \\ + \frac{\partial}{\partial z} (K_{13} X + K_{23} Y + K_{33} Z),$$

and

$$f_3'(c) f_2(c) = R'(c) \left[\frac{\partial}{\partial x} (K_{11} c_x + K_{12} c_y + K_{13} c_z) + \right. \\ \left. + \frac{\partial}{\partial y} (K_{21} c_x + K_{22} c_y + K_{23} c_z) + \frac{\partial}{\partial z} (K_{31} c_x + K_{32} c_y + K_{33} c_z) \right].$$

It immediately follows that in general the two expressions will differ in value. However, in the special case that R is linear in c and explicitly independent of $\underline{x}$, we do have equality and hence commutativity. Note that in this case dependence of K on $\underline{x}$ is permitted. □

We have to conclude that in almost every practical situation splitting errors arise, since the case of a space independent velocity field $\underline{u}$ and diffusion matrix K combined with a space independent and linear chemistry process R, hardly occurs. On the other hand, the extended use of Strang splitting in computational air pollution modeling leads to the conjecture that in this field splitting errors are kept within reasonable bounds, something which is confirmed for the examples presented in [81]. The following interpretation of the results of Theorem 1, based on relevant practical properties of $\underline{u}$, K and R, is in further support of this conjecture.

An important feature for air pollution models of the state of the atmosphere [21] is the diurnal cycle of sunsets and sunrises. This cycle obviously introduces a space-time dependency which manifests itself in two ways relevant to operator splitting errors, viz. through the photochemical reactions and the vertical transport. Let us first consider the photochemistry. After sunset, photochemical reactions are switched off. This not only simplifies the chemistry, but also strongly diminishes the spatial dependency of R. If also temperature and humidity hardly vary in $\underline{x}$, then at nightly periods R is often totally independent of x. Hence, if $\nabla \cdot \underline{u} = 0$, advection will commute with chemistry according to result b) of Theorem 1, diminishing the splitting error. The vertical transport is modeled by parameterized turbulent diffusion through the coefficient K. Since at night the stability of the atmosphere often increases, in many models K decreases to very small values after sunset. This means that the commutators f_{12} and f_{13} between diffusion and advection and diffusion and chemistry strongly decrease, which will lead to a strong decrease of the splitting error. It also often occurs that the velocity field $\underline{u}$ and the diffusion coefficient K vary slowly in $\underline{x}$, so that even during day time f_{12} can get small in large parts of the space domain.

Summarizing, the diurnal cycle strongly influences the commutators leading to a relatively small local splitting error over nightly periods. During these periods the global splitting error will also decrease owing to stability. In other words, the splitting error will oscillate with the diurnal cycle and not amplify beyond bound for evolving time. Specific circumstances will of course determine actual values.

2.4 Illustrations

We now proceed with simplified test models from class (2.1) so as to further study the local splitting error, in particular the leading error term $\tau^2 E_s$ defined in equation (2.16). Furthermore, we look at ways to reduce the splitting error in these cases and we pay attention to initial boundary value problems. Simplified models are used to avoid error terms too long to handle.

2.4.1 Examples of commutators

First, consider the 3D problem,

$$c_t + u\,c_x + v\,c_y = (\kappa\,c_z)_z + R(c)\,, \quad u_x + v_y = 0, \tag{2.17}$$

in which the transport is based on a divergence-free, horizontal velocity field, $\underline{u} = (u, v, 0)$, and on vertical diffusion with diffusion coefficient κ. This problem is relevant to many practical studies in the field of atmospheric air quality modeling where horizontal wind patterns dominate advection by wind and one-dimensional parameterized turbulent diffusion is used to simulate transport in the vertical direction. Putting

$$f_1(c) = -u\,c_x - v\,c_y, \quad f_2(c) = (\kappa\,c_z)_z\,, \quad f_3(c) = R(c)\,,$$

we derive the commutators,

$$\begin{aligned}
f_{12}(c) &= -\left(\kappa\,(u\,c_x + v_y)_z\right)_z + u\,(\kappa\,c_z)_{xz} + v\,(\kappa\,c_z)_{yz}\,,\\
f_{13}(c) &= u\,R_x(c) + v\,R_y(c)\,,\\
f_{23}(c) &= -\kappa_z R_z(c) - \kappa\,R_z z(c) - 2\kappa\,R'_z(c)\,c_z - \kappa\,R''(c)\,c_z c_z.
\end{aligned}$$

Despite the simplifications introduced in (2.17), these commutators still turn out to be rather complicated. The associated splitting error term E_s becomes too long to provide even little insight. Therefore a further simplification is introduced below. In passing we note that f_{12}, rewritten as

$$\begin{aligned}
f_{12}(c) &= -\kappa_z u_z c_x - \kappa_z v_z c_y - 2\kappa\,u_z c_{xz} - 2\kappa\,v_z c_{yz} - \kappa\,u_{zz} c_x - \kappa\,v_{zz} c_y\\
&\quad + \kappa_x u\,c_{zz} + \kappa_{xz} u\,c_z + \kappa_y v\,c_{zz} + \kappa_{yz} v\,c_z,
\end{aligned}$$

reveals that when u and v are constant in z and κ is constant in x and y, the commutator f_{12} vanishes yielding a zero advection-diffusion splitting error.

We now proceed with the 2D problem,

$$c_t + u\,c_x = \kappa\,c_{zz} + R(c)\,, \quad u \text{ constant}, \quad \kappa = \kappa(x)\,, \quad R(c) = R(x, c)\,, \tag{2.18}$$

with x and z as the independent space variables. Only a constant velocity in the x-direction exists, the diffusion coefficient κ is restricted to a x-dependent function, and the reaction term R may only depend on x, but not on z. For this model the split functions read

$$f_1(c) = -u\,c_x, \quad f_2(c) = \kappa\,c_{zz}, \quad f_3(c) = R(c)\,.$$

Of importance is that all three commutators,

$$f_{12}(c) = u\,\kappa_x c_{zz}. \quad f_{13}(c) = u\,R_x(c)\,, \quad f_{23}(c) = -\kappa\,R''(c)\,c_z c_z,$$

are unequal to zero, with the exception of special cases of course. In this sense sufficient generality is maintained compared to (2.17). According to (2.16), after a long calculation, $\tau^2 E_s(c)$ is found equal to

$$\tau^2 E_s(c) = \tau^2 \left(E_{12}(c) + E_{13}(c) + E_{23}(c) \right), \tag{2.19}$$

where

$$E_{12}(c) = -\frac{1}{24} u^2 \kappa_{xx} c_{zz} - \frac{1}{12} u \kappa_x R''(c) c_z c_z, \tag{2.20}$$

$$E_{13}(c) = -\frac{1}{24} u^2 R_{xx}(c) + \frac{1}{12} u \left(R'(c) R_x(c) - R'_x(c) R(c) \right) + \frac{1}{12} u \kappa R''_x(c) c_z c_z, \tag{2.21}$$

$$E_{23}(c) = \frac{\kappa}{24} \left(\left(R''(c) c_z c_z \right)' \left(\kappa c_{zz} + 2R(c) \right) - \left(\kappa \frac{\partial^2}{\partial z^2} + 2R'(c) \right) \left(R''(c) c_z c_z \right) \right). \tag{2.22}$$

Even for the simplified model problem (2.18) E_s is still a rather complicated expression, providing again little insight into the splitting error. We have to reckon with stiff chemistry, in which case R and its derivatives can possess extremely large entries. Whether these large entries will actually diminish the accuracy, depends in part on the size of $R''(c) c_z c_z$, being present in E_{12}, E_{13} and E_{23}. Observe here that $R''(c)$ is a tensor, $R''(c) c_z$ a matrix and c_z a vector, so that componentwise

$$\left(R''(c) c_z c_z \right)^{(i)} = \sum_{j,k=1}^{m} \frac{\partial^2 R^{(i)}(c)}{\partial c^{(j)} \partial c^{(k)}} c_z^{(j)} c_z^{(k)}.$$

If the chemistry is based on at most second order reactions, which is normal in atmospheric chemistry, the second derivative operator R'' is constant. Further, many of the entries will be zero since chemistry normally gives rise to very sparse Jacobian matrices (species react with only a few others). However, at least a few large entries will always remain and the coupling between fast (stiff) and slowly (non-stiff) reacting species will determine how these large entries enter the local error.

Observe also that, in accordance with Theorem 1, E_{12} vanishes if κ is constant and E_{13} vanishes if R is independent of x. In general, E_{23} vanishes if and only if all entries of R'' are zero. This is the case for linear chemistry, that is for

$$R(c) = G c + B(x, z),$$

with G a constant matrix. The source and sink vector B can still be space dependent. However, in contrast to the diffusion-chemistry error, in this case the advection-chemistry error E_{13} does not vanish as it is given by

$$E_{13}(c) = -\frac{1}{24} u^2 B_{xx} + \frac{1}{12} u G B_x. \tag{2.23}$$

The advection-diffusion error reads

$$E_{12}(c) = -\frac{1}{24}u^2\kappa_{xx}c_{zz}.$$

As the error (2.23) illustrates, strong spatial variations in the sources and sinks contribute to the splitting error.

2.4.2 Splitting advection and diffusion

We next examine the effect of only Strang splitting advection and diffusion for the 2D model problem (2.18). In this case we are able to say more about the splitting error in relation to spatial and time integration errors. So we consider the model problem,

$$c_t + u\,c_x = \kappa\,c_{zz}, \quad u \text{ constant}, \quad \kappa = \kappa(x). \tag{2.24}$$

According to (2.20), the modified equation for (2.24) reads

$$c_t + u\,c_x = \kappa\,c_{zz} - \frac{1}{24}\tau^2u^2\kappa_{xx}c_{zz} + \mathcal{O}\left(\tau^4\right).$$

The error $-\frac{1}{24}\tau^2u^2\kappa_{xx}c_{zz}$ can be seen as artificial diffusion due to splitting. To keep the local splitting error sufficiently small, it turns out to be necessary that in first approximation

$$\frac{1}{24}\tau^2u^2|\,\kappa(x)_{xx}\,| \ll \kappa(x). \tag{2.25}$$

The explicit quadratic dependence on τu is clarifying as it reveals that in an actual application the Strang splitting should work well, as long as for the numerical advection integration a normal CFL-condition holds and the split step size τ is taken equal to the advection step size Δt .

Let Δx denote a mesh width in the x-direction. A normal CFL-condition then is

$$\frac{\Delta t\,|u|}{\Delta x} \leq C_{\mathrm{CFL}} \approx 1.$$

Inserting this condition and the equality $\tau = \Delta t$ in (2.25) gives

$$\frac{1}{24}C_{\mathrm{CFL}}^2\,(\Delta x)^2\,|\,\kappa(x)_{xx}\,| \ll \kappa(x).$$

If $C_{\mathrm{CFL}} \approx 1$ and $|\kappa(x)_{xx}|$ is of moderate size compared to $\kappa(x)$, the leading local splitting error contribution will behave like $\mathcal{O}(\Delta x)^2$. This order of accuracy is satisfactory in the sense that many numerical advection schemes also generate $\mathcal{O}(\Delta x)^2$ errors by the spatial discretization of the advection operator and

$\mathcal{O}(\Delta t)^2 = \mathcal{O}(\tau^2) = \mathcal{O}(\Delta x)^2$ errors by the temporal integration. On the other hand, if very large values for τu are allowed, as for example made possible by the use of an implicit unconditionally stable advection integrator, or by many successive steps within split intervals with a conditionally stable explicit one, then large splitting errors can arise.

Would we allow κ in (2.24) to also depend on z, the modified equation is given by

$$\begin{aligned} c_t + u\,c_x &= (\kappa\,c_z)_z - 1/24\,\tau^2 u^2\,(\kappa_{xx}c_z)_z + 1/12\,\tau^2 u\{(-(\kappa_x\kappa_{zz})_z + (\kappa\,\kappa_{xzz})_z)\,c_z \\ &\quad + (-3\kappa_x\kappa_{zz} + 3\kappa\,\kappa_{xzz})\,c_{zz} + (-2\kappa_x\kappa_z + 2\kappa\,\kappa_{xz})\,c_{zzz}\} + \mathcal{O}(\tau^4)\,. \end{aligned}$$

Obviously, with appropriate modifications the above statements also hold for the case $\kappa = \kappa(x, z)$.

2.4.3 Reducing splitting errors

The error expressions (2.20) to (2.22) once again show that in general splitting errors will exist, because they depend on very different solution and problem properties. However, in actual applications it is sometimes possible to eliminate at least part of the splitting error. In this paragraph we will consider some of these possibilities.

For problem (2.17) one sometimes decides to solve chemistry and vertical diffusion coupled [22, 78, 79] so as to avoid error terms like E_{23} resulting from splitting diffusion and chemistry. This coupled solving involves the solution of a 1D diffusion-reaction system for every vertical column in a 3D grid. Unfortunately, when the number of chemical species is large [78], in spite of the 1D nature, a direct solution method using a standard band-solver in the linear algebra is costly. An iterative tridiagonal Gauss-Seidel type process is a very competitive alternative though, but this type of solution process only works for gas-phase chemistry [79]. Coupling between diffusion and chemistry yields in some, but not in every case, an acceptable possibility to reduce the splitting error.

Part of the splitting error can be truly eliminated for problems of the form,

$$c_t + u\,c_x = f(x, c)\,, \quad u \text{ constant}. \tag{2.26}$$

We restrict ourselves to the 1D case, but the theory can easily be extended to 2D and 3D problems with a non-constant velocity field. Although f can represent any arbitrary nonlinear vector function in $\mathbb{R}^m$, we shall associate with f a chemical process. Note that our following derivation can also be applied to problems like (2.17), where $f(x, y, z, c)$ stands for vertical diffusion and chemistry. Observe at last, as proved in Theorem 1, that the dependence of f on x in (2.26) is essential, because otherwise no splitting error exists and our derivation is redundant.

We consider a special splitting technique for equation (2.26) similar to a semi-Lagrangian method. The underlying idea has been discussed previously in [39] and

in [46,47]. A Lagrangian method solves

$$\frac{dc}{dt} = f(x(t), c), \quad \dot{x} = u, \tag{2.27}$$

along the characteristics, using a moving grid to keep track of them. In case of a semi-Lagrangian method one still solves (2.27) along the characteristics, but with this difference that no moving grid is used and the solutions $c(x^* - u\tau, 0)$, needed as initial values for integration along the characteristics to calculate the solutions $c(x^*, \tau)$ in the gridpoints x^*, are found by interpolation between known solutions in neighbouring gridpoints. Hence, within each time step a semi-Lagrangian method maps the Lagrangian solution to an Eulerian grid.

Our splitting variant of this semi-Lagrangian method over an interval $[0, \tau]$ is described as

$$\begin{aligned}
&\frac{\partial c_1}{\partial t} + u\frac{\partial c_1}{\partial x} = 0, && c_1(x,0) = \tilde{c}(x,0) \quad (a)\\
&\frac{dc_2}{dt} = f(x(t), c_2), \quad \dot{x} = u, && c_2(x - u\tau, 0) = c_1(x,\tau) \quad (b)\\
&\tilde{c}(x,\tau) = c_2(x,\tau).
\end{aligned} \tag{2.28}$$

First, the advection step (2.28a) is carried out on an Eulerian grid. Then the second equation (2.28b) is integrated on the same grid, but using $x = x(t)$, with as initial value the solution obtained from the proceeding advection step. Note here the resemblance with the semi-Lagrangian method. The initial values needed for integration along the characteristics are determined in a proceeding step apart from the actual integration. If the advection step is solved exactly on the grid, no splitting error occurs between advection and chemistry. When no exact advection step is achieved, the errors, which arise in an actual Eulerian advection step, resemble the interpolation errors of the semi-Lagrangian method.

The way in which we obtain the solution to (2.28b) is not prescribed. One can think for instance of applying a splitting scheme to split diffusion from chemistry or in case of gas-phase chemistry one can decide to use the earlier mentioned iterative tridiagonal Gauss-Seidel solution method.

2.4.4 Strang splitting in initial boundary value problems.

Till now, we restricted ourselves to pure initial value problems. In practical applications though, we mostly encounter initial boundary value problems. When we use operator splitting in these situations, we have to reckon with boundary errors. We will now focus on the subject of prescribing boundary conditions in the intermediate steps of the Strang splitting and on the resulting possibility of inconsistencies between these boundary conditions and the solutions calculated in the proceeding intermediate steps. These inconsistencies can lead to numerical errors.

We consider once more the 2D autonomous problem (2.17) ($v=0$) now described over a bounded domain $\{(x,z) \mid 0 \le x \le 2\pi,\ 0 \le z \le z_H\}$.

$$c_t + u\,c_x = (\kappa\,c_z)_z + R(c), \tag{2.29}$$

where u is constant in x and κ and R can depend on x and z. As boundary conditions we prescribe 2π-periodicity in x-direction, and on $z=0$ (the earth surface) and $z=z_H$ we prescribe

$$\kappa\,c_z = d(x)\,c + E(x), \quad d(x) < 0 \quad \text{at } z = 0, \tag{2.30}$$

$$\kappa\,c_z = 0, \qquad \text{at } z = z_H. \tag{2.31}$$

The first condition describes the flux $\kappa\,c_z$ at the earth surface in terms of deposition $d\,c$ and emission E. The second condition describes a no flux condition at the upper boundary of our domain. Our boundary conditions are chosen in close relation with boundary conditions found in practical applications. κ, d, E and R are assumed 2π-periodic in x, which occurs in true global models if x is associated with the longitudinal direction [81].

We apply Strang splitting to system (2.29) over the interval $[0,\tau]$, which yields

$$\begin{aligned}
&\frac{\partial c_1}{\partial t} + u\frac{\partial c_1}{\partial x} = 0, && c_1(x,z,0) = c(x,z,0) && \text{(a)}\\
&\frac{\partial c_2}{\partial t} = \frac{\partial}{\partial z}\left(\kappa\frac{\partial c_2}{\partial z}\right) + \text{b.c.} && c_2(x,z,0) = c_1(x,z,\frac{\tau}{2}) && \text{(b)}\\
&\frac{\partial c_3}{\partial t} = R(c_3) && c_3(x,z,0) = c_2(x,z,\frac{\tau}{2}) && \text{(c)}\\
&\frac{\partial c_4}{\partial t} = \frac{\partial}{\partial z}\left(\kappa\frac{\partial c_4}{\partial z}\right) + \text{b.c.} && c_4(x,z,\frac{\tau}{2}) = c_3(x,z,\tau) && \text{(d)}\\
&\frac{\partial c_5}{\partial t} + u\frac{\partial c_5}{\partial x} = 0, && c_5(x,z,\frac{\tau}{2}) = c_4(x,z,\tau) && \text{(e)},
\end{aligned} \tag{2.32}$$

where the initial value $c(x,z,0)$ in (2.32) satisfies the boundary conditions. Note that the boundary conditions are prescribed in step (2.32b) and (2.32d), so the solutions $c_2(x,z,\frac{\tau}{2})$ and $c_4(x,z,\tau)$ always satisfy the given conditions.

Consider the initial value for step (2.32b) delivered after exact time and space integration of step (2.32a),

$$c_2(x,z,0) = c_1(x,z,\frac{\tau}{2}) = c(x - u\frac{\tau}{2}, z, 0).$$

If $u_z \neq 0$ for $z=0$ and $z=z_H$, then at time $t=0$ in step (2.32b) the boundary conditions (2.30) and (2.31) are not met, as can be seen from

$$\frac{\partial}{\partial z}\left(c_2(x,z,0)\right) = c_z(x - u\frac{\tau}{2}, z, 0) - u_z\frac{\tau}{2}c_x(x - u\frac{\tau}{2}, z, 0). \tag{2.33}$$

The initial value for step (2.32b) is inconsistent with the boundary conditions prescribed in this step. Numerical errors will exist if we don't choose the time step τ large enough to damp out the initial error due to this inconsistency. Note however that at the end of step (2.32b) the boundary conditions are always met.

Now take $u_z = 0$, then $c_{2_z}(x, z, 0) = 0$ holds when $c_z(x - u\frac{\tau}{2}) = 0$ as can be concluded from (2.33). At a large distance from the earth surface $u_z = 0$ is likely to occur, thus no boundary condition inconsistency will exist at $z = z_H$, when z_H is chosen large enough. However, at the earth surface we must satisfy

$$\kappa(x,0)\, c_{2_z}(x,0,0) = d(x)\, c_2(x,0,0) + E(x), \tag{2.34}$$

or, inserting (2.33) into (2.34), where still $u_z = 0$, we must satisfy

$$\kappa(x,0)\, c_z\left(x - u\frac{\tau}{2}, 0, 0\right) = d(x)\, c\left(x - u\frac{\tau}{2}, 0, 0\right) + E(x).$$

In general this relation will only hold if κ, d and E are independent of x.

Similarly we can show that in general the solution of the chemistry step (2.32c) used as initial value in step (2.32d) introduces an inconsistency with the prescribed boundary conditions in this step. If at $z = z_H$

$$c_z = 0 \quad \text{and} \quad R_z(x, z, c) = 0, \tag{2.35}$$

no inconsistency is obtained, because the solution of step (2.32c) satisfies

$$\frac{\partial c_3}{\partial t\, \partial z} = R'\left(x, z, c_3\right) \frac{\partial c_3}{\partial z} + R_z\left(x, z, c_3\right).$$

For $z = z_H$ large enough, the assumptions (2.35) represent the realistic case. On the earth surface, however, we expect an inconsistency, for $R_z(x, z, c) = 0$ and also $c_z = 0$ may be violated there. Further, it is possible that due to the prescribed emission and deposition condition (2.30) in step (2.32b) strong transient exists, which can lead to a disturbance from the chemical equilibrium solution.

In [81] a comparison was made between solving the 3D problem (2.17) with a Rosenbrock method in combination with approximate factorization, and with the Strang splitting method. Approximate factorization can be seen as a form of splitting performed at the numerical algebra level rather than at the operator level as is done in Strang splitting. As boundary conditions were used

$$\kappa\, c_z = 0, \quad \text{at } z = 0 \text{ and } z = z_H,$$

while for the imposed wind field, $u_z = v_z = 0$. In [81] was argued that due to this form of splitting at the numerical algebra level, operator splitting errors as well as errors, arising from inconsistencies between the boundary conditions and the initial values prescribed in the intermediate steps in Strang splitting, could be avoided. This should lead to more accurate solutions in favor of the Rosenbrock method with

approximate factorization. Results proved them right, but the gain in accuracy was not as great as was expected. However, the results in [81] might have been too positive where the Strang splitting method was concerned. The specific choice of the boundary conditions led to no inconsistencies, while also the property $u_z = v_z = 0$ contributed to reduction of the splitting error between advection and diffusion. In other words, in a more realistic situation, where boundary conditions such as (2.30) and (2.31) can occur, the Rosenbrock method with approximate factorization might be a good alternative to Strang splitting. Future research has to throw light on this aspect.

2.5 Conclusions

In this paper we focussed on operator splitting, where we mainly restricted ourselves to three-term symmetrical Strang splitting primarily applied to time-dependent advection-diffusion-reaction (ADR) problems. For pure initial value problems the Lie operator formalism proves to be very useful to derive the structure of the splitting error. Through the notion of commutativity we are able to state in which cases the usage of Strang splitting leads to no splitting error. Application of a three-term symmetrical Strang splitting to pure initial value problems of the ADR-type leads to no splitting error between advection, diffusion and chemistry, when, with exact integration of the intermediate steps in the Strang splitting, the chemistry $R(c)$ is linear in c, and the wind field $\underline{u}$, the diffusion coefficient matrix K and R are independent of the spatial variable $\underline{x}$.

However, in most applications splitting errors will occur. By relating the physics of the problem with the commutators, we have conjectured that in air pollution models the splitting error will oscillate with the diurnal cycle and will not grow beyond bound for evolving time. Unfortunately, the splitting error expression is too complicated for real insight into its actual magnitude.

To avoid or reduce the splitting error several techniques can be applied. One concerns problems of the form (2.17), where diffusion and chemistry can be solved coupled, so only a 1D diffusion-reaction system has to be solved for every vertical column in 3D, avoiding an error due to splitting diffusion and chemistry. Secondly, for problems of the form (2.26) an alternative splitting technique exists, similar to a semi-Lagrangian method. A chemistry step is integrated along the characteristics proceeded by an advection step on an Eulerian grid, leaving no splitting error if the advection step is solved exactly and else resulting in an error similar to the interpolation errors of the semi-Lagrangian method.

Several questions concerning operator splitting remain. A good start for further research is the analysis of the splitting error in practical situations by using global Richardson extrapolation to estimate the splitting error for evolving time.

Chapter 3

Spatial Discretization of the Shallow Water Equations in Spherical Geometry using Osher's Scheme

Summary

The shallow water equations in spherical geometry provide a first prototype for developing and testing numerical algorithms for atmospheric circulation models. Since the seventies these models have often been solved with spectral methods. Increasing demands on grid resolution combined with massive parallelism and local grid refinement seem to offer significantly better perspectives for gridpoint methods. In this paper we study the use of Osher's finite volume scheme for the spatial discretization of the shallow water equations on the rotating sphere. This finite volume scheme of upwind type is well suited for solving a hyperbolic system of equations. Special attention is paid to the pole problem. To that end Osher's scheme is applied on the common (reduced) latitude-longitude grid and on a stereographic grid. The latter is most appropriate in the polar region as in stereographic coordinates the pole singularity does not exist. The latitude-longitude grid is preferred on lower latitudes. Therefore, across the sphere we apply Osher's scheme on a combined grid connecting the two grids at high latitude. We will show that this provides an attractive spatial discretization for explicit integration methods, as it can greatly reduce the time step limitation incurred by the pole singularity when using a latitude-longitude grid only. When time step limitation plays no significant role, the standard (reduced) latitude-longitude grid is advocated provided that the grid is kept sufficiently fine in the polar region to resolve flow over the poles.

3.1 Introduction

People have long tried to forecast the weather, first by observation of current and historical meteorological data and later by numerical simulation with circulation models based on atmospheric primitive equations [12, 26, 32, 48]. Today, circulation models are widespread. In addition to being used in weather forecasting, they are applied as climate simulation models and provide meteorological input data needed in air pollution descriptions.

During the sixties the field of frequently used approximation methods in circulation models consisted mainly of gridpoint methods. When Orszag and Eliasen *et al* [17, 51] introduced the spectral transform method in global atmospheric modeling, this accent shifted. Because spectral methods proved to be very accurate and cost efficient, they started to dominate the field of approximation methods used in global atmospheric modeling. Recently, the discussion on numerical methods applicable in circulation models has been renewed. Spectral methods are no longer considered ideal. Progression in atmospheric modeling, on the meteorological as well as on the computational side, demands higher grid resolutions than in the past. The workload of a spectral method grows very fast when the number of grid points is increased. Therefore, the relevant question can be posed whether at high resolutions an improved gridpoint method can compete with a spectral method. This is also stated in [8, 16]. In addition, the global property of a spectral method has some other drawbacks. Although this property contributes highly to the accuracy of the found solution, it leads to inconveniences when one tries to parallelize spectral codes on parallel machines with distributed memory. Furthermore, a spectral method can suffer from Gibbs' phenomenon (spectral ringing) when applied in areas where flow patterns with strong gradients are encountered, for example, in front simulation.

In this paper, we develop a new numerical gridpoint method. We apply a finite volume method of upwind type. We decided on this method, because it is conservative and respects the characteristic directions associated with the hyperbolic character of our equations. In addition, compared to a spectral method, it behaves well in areas where flow patterns with strong gradients are expected. From the class of finite volume methods, Osher's approximate Riemann solver makes a good choice. First, it is robust and second-order accurate when combined with the right state interpolation. Second, from a future perspective, it has a logical extension to more realistic primitive equations and it has a consistent boundary treatment, which makes Osher's solver preferable to, for instance, Roe's solver. Finally, our upwind scheme is a scheme of flux difference splitting type (FDS). Schemes of flux vector splitting type (FVS) do not provide an alternative in this case, since the necessary condition for these schemes, i.e., that the Jacobian of the flux vector is homogeneous of degree one, is not fulfilled. For a detailed description of FDS and FVS methods we refer to [31].

To avoid the well-known pole problem [69], which arises when a gridpoint method is applied on a full uniform latitude-longitude (lat-lon) grid, we study a reduced lat-

lon grid and a combined grid composed of a (reduced) lat-lon grid away from the poles and a stereographic grid at the two polar caps. The combined grid consists of three computational domains with a rectangular grid almost everywhere. All three mappings used to map the physical domain onto the computational domain are conformal. These qualities yield flux calculations that are simple and straightforward. The use of a stereographic grid has been proposed before by Phillips [56] and Browning *et al* [7].

To validate our discretization scheme and grid, we consider the 2D shallow water equations (SWEs) on the rotating sphere, which serve as a first prototype for a circulation model. The SWEs describe the behavior of a shallow homogeneous incompressible and inviscid fluid layer. Although in comparison to the full set of atmospheric primitive equations, the SWEs are incomplete, they present some of the major difficulties associated with the horizontal dynamical aspects of circulation models on the Earth.

In Section 3.2, we focus on the formulation of the SWEs in the two different coordinate systems. In Section 3.3.1, we attend to the construction of our combined grid. The spatial discretization of the equations, i.e., a description of our finite volume method, is given in Section 3.3.2. Special attention is paid to the connection problem, which occurs at the grid interface, when coupling the spherical grid part with the stereocaps. Numerical results from calculations on combined grids and on fully lat-lon grids are given in Section 3.4. Calculations are done on test case 2 of the test set in [88], which is standard for testing new numerical methods for solving the SWEs in spherical geometry. Test case 2 provides us with a good non-linear test to evaluate the scheme's ability to handle the poles. Since the test set consists of problems with smooth flow patterns, it does not provide a test to reveal all favorable features of our scheme. Therefore, the objective of this paper can best be summarized as a first validation of whether Osher's scheme applied on a combined grid yields an appropriate candidate to solve the SWEs in spherical geometry. The main conclusions of our investigations are formulated in Section 3.5.

3.2 The shallow water equations

Since they cover important aspects of the horizontal dynamical behavior of the atmosphere, the SWEs on the sphere suffice as a first prototype of a circulation model. Through the laws of conservation of mass and momentum, the SWEs on the sphere can be derived to describe the behavior (velocities and fluid depth) of a shallow homogeneous incompressible and inviscid fluid layer on the Earth. In other words, we assume that the atmosphere can be regarded as a thin layer of air in which the density is uniform and constant, and viscous effects can be ignored. By using the SWEs, it is further assumed that the velocity component normal to the earth surface, the vertical component, can be neglected compared to the horizontal velocity component. Furthermore, the vertical component of the Coriolis

acceleration is neglected in comparison with gravity. The acceleration of gravity, g, is assumed to be constant, containing both the effects related to the centrifugal force and the gravitational attraction of the Earth. The pressure gradient force is considered to be hydrostatic. The SWEs then follow from the Navier Stokes equations on the rotating sphere by integration over the fluid depth (depth-averaging); for details see [26]. A derivation of more realistic atmospheric primitive equations can be found in [26, 32].

3.2.1 The shallow water equations in spherical coordinates

Let (λ, ϕ, t) denote the independent variables longitude ($\lambda \in [0, 2\pi)$), latitude ($\phi \in [-\frac{\pi}{2}, +\frac{\pi}{2}]$), and time ($t \geq 0$). Let u be the velocity in the longitudinal direction, v the velocity in the latitudinal direction, and H the depth of the fluid layer. Let h be the height of the free surface above the sphere at sea level, $h = H + h_s$, where h_s accounts for the orography of the Earth associated with the height of mountains. Further, let $\underline{u}$ denote the horizontal velocity field (u, v) defined by $u = a\cos\phi\frac{\mathrm{d}\lambda}{\mathrm{d}t}$ and $v = a\frac{\mathrm{d}\phi}{\mathrm{d}t}$. Let f denote the Coriolis parameter, $2\Omega\sin\phi$, with Ω the angular velocity of the Earth, a the radius of the Earth, and g the gravitational constant. The SWEs on the sphere in flux form can then be formulated as

$$\frac{\partial H}{\partial t} + \nabla\cdot(H\underline{u}) = 0, \tag{3.1}$$

$$\frac{\partial Hu}{\partial t} + \nabla\cdot(Hu\,\underline{u}) = (f + \frac{u}{a}\tan\phi)\,Hv - \frac{gH}{a\cos\phi}\frac{\partial h}{\partial\lambda}. \tag{3.2}$$

$$\frac{\partial Hv}{\partial t} + \nabla\cdot(Hv\,\underline{u}) = -(f + \frac{u}{a}\tan\phi)\,Hu - \frac{gH}{a}\frac{\partial h}{\partial\phi}, \tag{3.3}$$

where the divergence operator is defined by

$$\nabla\cdot\underline{u} \equiv \frac{1}{a\cos\phi}\left[\frac{\partial u}{\partial\lambda} + \frac{\partial v\cos\phi}{\partial\phi}\right].$$

The right-hand sides in the momentum equations (3.2) and (3.3) represent, respectively, the Coriolis force, the hydrostatical pressure gradient force, and an additional term due to the relative motion in the rotating coordinate system in longitudinal and latitudinal direction, see [32].

3.2.2 The shallow water equations in stereographic coordinates

The spherical formulation of the SWEs (3.1)-(3.3) has the disadvantage that it is singular at the poles. To circumvent this problem, the SWEs can be formulated in the stereographic coordinate system using a different stereographic projection on each hemisphere. Since these projections are only singular in opposite poles, no

singularity problem arises. We note that the stereographic projection is conformal, so the general form of the equations is preserved.

The stereographic projection in terms of the latitude-longitude coordinates is defined by

$$x_{\mathrm{st}} = a\, m \cos\phi \cos\lambda, \tag{3.4}$$

$$y_{\mathrm{st}} = a\, m \cos\phi \sin\lambda, \tag{3.5}$$

where m is the map factor

$$m = \frac{2}{1 + \alpha \sin\phi}, \tag{3.6}$$

with α distinguishing between the northern ($\alpha = 1$) and the southern hemisphere projection ($\alpha = -1$). The poles are directly projected onto the origin of the stereographic planes. The northern hemisphere is projected from the south pole onto the northern stereographic plane, which is the plane locally tangent to the sphere at the north pole, see Figure 3.1. Likewise, the southern hemisphere is projected from the north pole onto the southern stereographic plane, which is locally tangent to the sphere at the south pole. A description of the construction of the stereographic projection can be found in Appendix 6.1. Note that the positive stereographic x_{st}-axis for both the northern and the southern hemisphere corresponds with the intersection of the half-plane $S_{\lambda=0}$ and the corresponding stereographic plane. Likewise, the positive stereographic y_{st}-axis corresponds, for both hemispheres, with the intersection of the half-plane $S_{\lambda=\pi/2}$ and the corresponding stereographic plane.

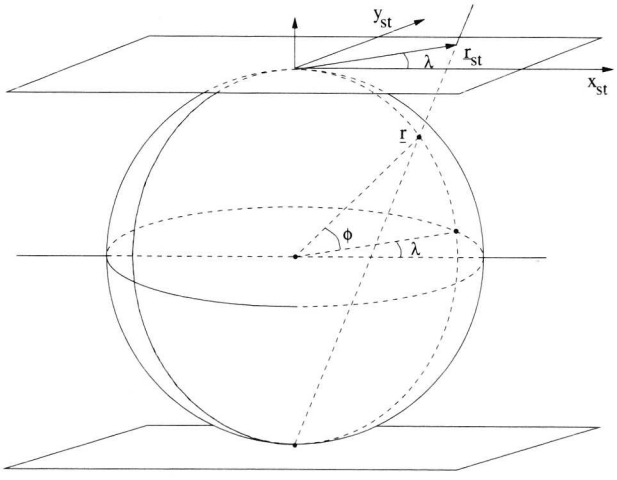


Figure 3.1: The stereographic planes for the northern (southern) hemisphere projections.

Before we give the SWEs in the stereographic formulation, as found, for instance, in [7,56,87], we need to define the velocity field in the new stereographic coordinate

system. Let $\underline{U} = (U, V)$ be the velocity field in stereographic coordinates with U the velocity in the $x_{\rm st}$-direction and V the velocity in $y_{\rm st}$-direction. We have

$$\underline{U} = \begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} m^{-1}\frac{\mathrm{d}x_{\rm st}}{\mathrm{d}t} \\ m^{-1}\frac{\mathrm{d}y_{\rm st}}{\mathrm{d}t} \end{pmatrix}.$$

where $\frac{\mathrm{d}x_{\rm st}}{\mathrm{d}t}$, $\frac{\mathrm{d}y_{\rm st}}{\mathrm{d}t}$ are the usual total derivatives and $\frac{1}{m}$ is a scale factor with m as given in (3.6). When we now consider the momentum equations in the stereographic $x_{\rm st}$- and $y_{\rm st}$-direction, the stereographic formulation of the SWEs in flux form reads

$$\frac{\partial H}{\partial t} + \nabla \cdot (H\underline{U}) = 0. \tag{3.7}$$

$$\frac{\partial HU}{\partial t} + \nabla \cdot (HU\underline{U}) = \left[\alpha f - \frac{(x_{\rm st}V - y_{\rm st}U)}{2a^2}\right] HV - mgH\frac{\partial h}{\partial x_{\rm st}}, \tag{3.8}$$

$$\frac{\partial HV}{\partial t} + \nabla \cdot (HV\underline{U}) = -\left[\alpha f - \frac{(x_{\rm st}V - y_{\rm st}U)}{2a^2}\right] HU - mgH\frac{\partial h}{\partial y_{\rm st}}. \tag{3.9}$$

where the divergence operator is defined by

$$\nabla \cdot (A\underline{U}) \equiv m^2 \frac{\partial}{\partial x_{\rm st}}\left(\frac{AU}{m}\right) + m^2 \frac{\partial}{\partial y_{\rm st}}\left(\frac{AV}{m}\right). \tag{3.10}$$

This formulation is derived in Appendix 6.2. To complete the discussion on the two different coordinate systems, we here give the relations between the stereographic and spherical velocity components. These relations, which of course are valid only outside the poles, are needed in Section 3.3.2.

$$U = -u\sin\lambda - \alpha v\cos\lambda, \tag{3.11}$$

$$V = u\cos\lambda - \alpha v\sin\lambda. \tag{3.12}$$

3.3 Spatial discretization

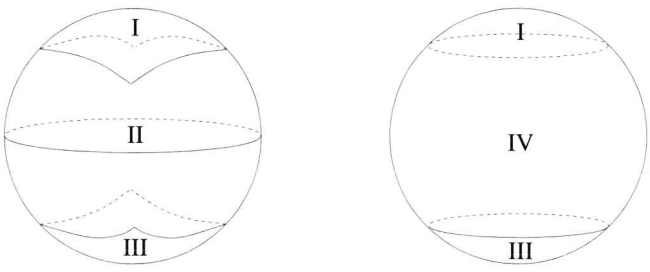
In the past, several types of grids have been proposed to circumvent the problems related to solving the SWEs on a global lat-lon grid. Two examples are the composite cubic grid [62,63] and the icosahedral grid [83]. The first yields a non-conformal mapping of the sphere onto a cube. The latter grid consists of triangles.

In this section we introduce another grid. Our motivation is to provide a grid on which calculations are simple and straightforward. Therefore, we aim at a grid distribution which can be conformally mapped onto a rectangular computational domain without any singular points.

3.3.1 Using stereographic grids

Over the years several suggestions have been made to circumvent the singularity problem which arises at the poles when one tries to solve the SWEs in spherical

coordinates. In 1956, Phillips [56] studied this problem. He suggested covering the sphere with three different coordinate systems. On part of the northern as well as on the southern hemisphere he used a stereographic coordinate system centered at the poles. In between those two regions he chose a mercator projection. His distribution of the coordinate systems is illustrated in Figure 3.2(b). To couple the different coordinate systems, Phillips had to interpolate from points in neighboring grids whenever a variable outside the current grid part was needed. In 1975 Stoker [74] showed that these interpolations could contribute to loss of mass.



(a) Our combined grid (b) Phillips' combined grid

Figure 3.2: Two combined grids and their applied coordinate systems,
(I) northern hemisphere stereographic projection,
(II) spherical coordinate system,
(III) southern hemisphere stereographic projection,
(IV) mercator projection.

In 1977 Starius [72] introduced the composite mesh method. Like Phillips, he used multiple coordinate systems, but he avoided interpolations within neighboring grids by letting the grids, corresponding with the different coordinate systems, overlap. To prosper from both methods, Browning *et al* [7] combined the ideas of Starius and Phillips. They applied the composite mesh method to the SWEs by using two stereographic coordinate systems centered respectively at the north and south pole and extended beyond the equator.

Our approach is also based on the ideas of Phillips, that is, we use three different non-overlapping coordinate systems, where stereographic coordinate systems are applied in the arctic and antarctic regions. In the intermediate region, however, our choice of the coordinate system differs from Phillips'. Since spherical coordinates are natural and easily implemented in regions away from the poles, we prefer a spherical coordinate system in the intermediate region. In addition, lat-lon grids are still standard in meteorological applications. A further differentiation from Phillips' method concerns the coupling of the different coordinate systems. Although this subject is not addressed until Section 3.3.2, we state here that with our choice of

a finite volume method we are able to avoid the interpolation problems found by Phillips. Our distribution of the coordinate systems is shown in Figure 3.2(a).

In this paragraph we discuss the exact distribution of the three different coordinate systems across the sphere. As mentioned before, we prefer to use a lat-lon grid in a region away from the poles. We define this region as $R_{II} = \{(\lambda, \phi, a) : \lambda \in [0, 2\pi),\ \phi \in [-\overline{\phi}, \overline{\phi}]$ with $\overline{\phi} < \frac{\pi}{2}\}$. From an illustrative point of view we assume that our lat-lon grid has a uniform distribution. Note that more advanced grid distributions are possible. In Section 3.4, for instance, we apply a reduced lat-lon grid. To find a suitable grid distribution in the stereographic regions, we project the uniform lat-lon grid of region R_{II} onto the stereographic planes, as illustrated for one hemisphere in Figure 3.3. Note that meridians and parallels correspond with respectively dashed and solid lines.

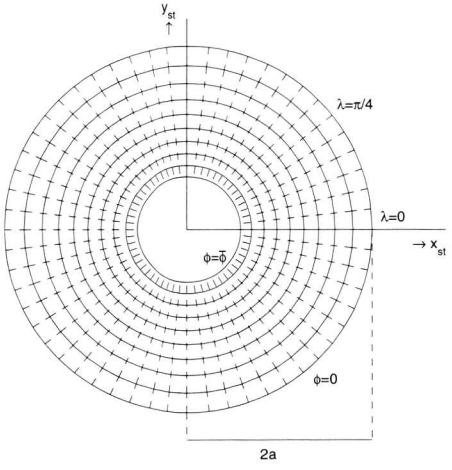
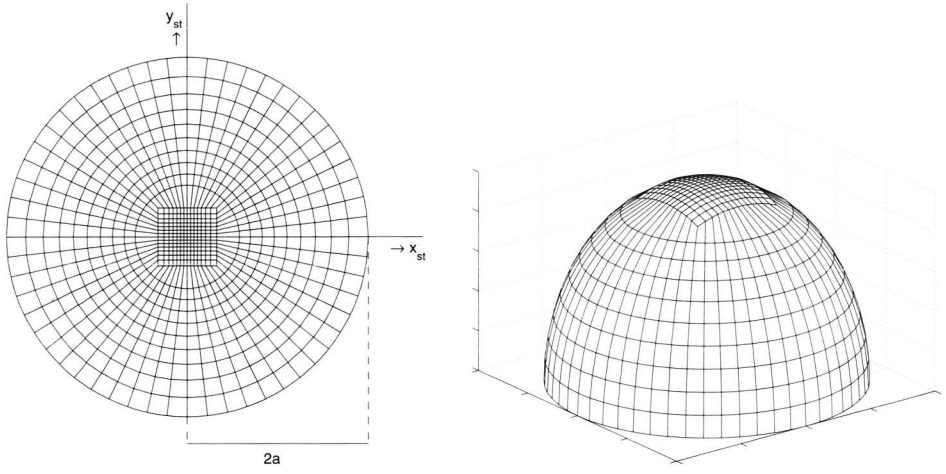


Figure 3.3: Northern hemisphere projection from the south pole of a uniform lat-lon grid. Dashed lines correspond with meridians (λ constant). Solid lines correspond with parallels (ϕ constant).

In the middle of the resulting projection we place a square with bottom left-hand corner $(x_{\text{st}}, y_{\text{st}}) = (-x_r, -x_r)$ and top right-hand corner $(x_{\text{st}}, y_{\text{st}}) = (x_r, x_r)$, $x_r > 0$. The corresponding regions on the sphere are denoted by region I (northern hemisphere) and III (southern hemisphere). To secure a proper fit between the grids on regions I, III, and R_{II}, we extend the projected meridians until they intersect with the squares. The resulting cells between these regions are added to region R_{II} giving the region II shown in Figure 3.2(a). The solid lines in Figure 3.4(a) and Figure 3.4(b) correspond with the cell edges. We then demand that N_λ, defined as $N_\lambda = \frac{1}{\Delta\lambda}$, is a multiple of eight. Under this condition the intersection points have mirror images on the opposite edge. After these points are connected, a non-uniform rectangular grid distribution on the square results, see Figure 3.4(a). The total grid

distribution over the sphere is now fully known, see Figure 3.4(b). Finally, we remark that x_r, N_λ and $\overline{\phi}$ are still free parameters. Exact values are given for each test case. These values affect, for instance, the CFL-number, the meshwidth factors, and the accuracy. For visualization purposes we used $N_\lambda = 56$, $x_r = 0.32279\, a$, and $\overline{\phi} = 57.8°$.



(a) Northern hemisphere stereographic projection of the grid.

(b) Side view of the grid over the northern hemisphere.

Figure 3.4: Details of a combined grid.

3.3.2 The semi-discrete system in general terms

Without the Coriolis and additional forces, the SWEs closely resemble the Euler equations, which can be found in, for instance, gas dynamic applications. For the full set of primitive equations this resemblance is even more explicit. Much theory concerning the space discretization of the Euler equations has already been developed, see, for instance, [31]. In our approximation method we gratefully adopt existing ideas from this theory. In this section, we will describe the semi-discrete system for the SWEs (3.1)-(3.3) and (3.7)-(3.9) with special attention to the coupling between the spherical and stereographic grids.

Main outline of the finite volume method

We begin this section with a main outline of our method. To guarantee conservation of mass and momentum in our semi-discrete system or, in other words, to

respect the underlying physical conservation laws, we use the finite volume method, which is standard practice for the Euler equations. We focus on the stereographic region I. Similar results can be derived for the spherical region II and for region III. Calculations are done in the computational domain, which results after projection of regions I, II, and III on the regions associated with the corresponding coordinate systems. In the computational domains regular, (non-)uniform rectangular grids occur.

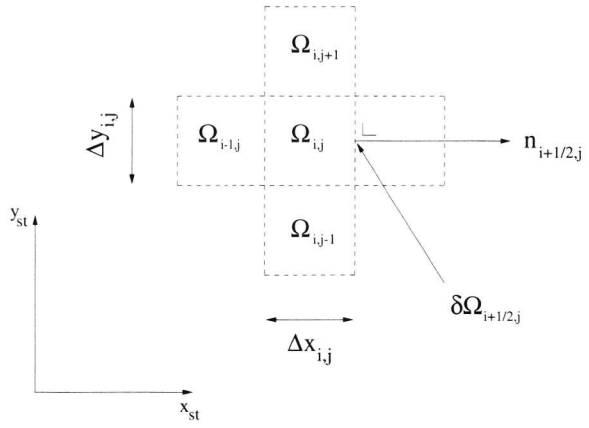


Figure 3.5: The grid cell $\Omega_{i,j}$ in the stereographic coordinate system.

Let $\Omega_{i,j}$ be a grid cell with boundary $\delta\Omega_{i,j}$. We denote its four neighbors by $\Omega_{i\pm1,j}$ and $\Omega_{i,j\pm1}$. The boundary between two neighboring cells, for instance, between $\Omega_{i+1,j}$ and $\Omega_{i,j}$, is denoted by $\delta\Omega_{i+1/2,j}$, $\underline{n}_{i+1/2,j} = (n_{x_{\rm st}}, n_{y_{\rm st}})$ is the outwardly directed unit normal along this boundary, $\Delta x_{i,j}$ and $\Delta y_{i,j}$ are respectively the lengths of $\delta\Omega_{i,j\pm1/2}$ and $\delta\Omega_{i\pm1/2,j}$, see Figure 3.5. We associate with each grid cell its cell center $\underline{x_{\rm st}}_{i,j} = (x_{{\rm st}_{i,j}}, y_{{\rm st}_{i,j}})$ with state variable $\underline{q}_{i,j} = (H_{i,j}, H_{i,j}U_{i,j}, H_{i,j}V_{i,j})$ and we assume that the state variable is constant over each cell. The finite volume method now gives

$$\frac{\partial \underline{q}_{i,j}}{\partial t} + \frac{m_{i,j}^2}{\Delta x_{i,j}\Delta y_{i,j}} \oint_{\delta\Omega_{i,j}} \frac{1}{m}\underline{F}\, n_{x_{\rm st}} + \frac{1}{m}\underline{G}\, n_{y_{\rm st}}\, dS = - \left(\underline{f}_{x_{\rm st}}\left(\underline{q}_{i,j}, \underline{x_{\rm st}}_{i,j}\right) + \underline{f}_{y_{\rm st}}\left(\underline{q}_{i,j}, \underline{x_{\rm st}}_{i,j}\right)\right), \quad (3.13)$$

where $\underline{F}$ and $\underline{G}$ are the fluxes in stereographic $x_{\rm st}$- and $y_{\rm st}$-direction,

$$\begin{aligned} \underline{F}(\underline{q}) &= \left(HU, HU^2 + \frac{1}{2}gH^2, HUV\right)^T, \\ \underline{G}(\underline{q}) &= \left(HV, HUV, HV^2 + \frac{1}{2}gH^2\right)^T, \end{aligned}$$

and

$$\underline{f}_{x_{\rm st}}(\underline{q},\underline{x_{\rm st}}) = \left(0,-[\alpha f-\frac{(x_{\rm st}V-y_{\rm st}U)}{2a^2}]HV+mgH\frac{\partial h_s}{\partial x_{\rm st}}+\frac{1}{4a^2}gH^2x_{\rm st},0\right)^T,$$

$$\underline{f}_{y_{\rm st}}(\underline{q},\underline{x_{\rm st}}) = \left(0,0,[\alpha f-\frac{(x_{\rm st}V-y_{\rm st}U)}{2a^2}]HU+mgH\frac{\partial h_s}{\partial y_{\rm st}}+\frac{1}{4a^2}gH^2y_{\rm st}\right)^T.$$

To respect the characteristic directions associated with the hyperbolic character of our equations, we apply an upwind scheme to discretize the integral in (3.13). Within the group of finite volume upwind methods we distinguish two different categories, concerning flux vector splitting (FVS) and flux difference splitting (FDS) methods. For a detailed description of both methods we refer to [31]. Methods from the first category do not suffice as discretization schemes for the SWEs. The condition that the Jacobian of the flux vector $\underline{F}$ with respect to $\underline{q}$ is homogeneous of degree one (see [31]) is not fulfilled. We apply Osher's approximate Riemann solver [52, 53], which makes an excellent choice from the group of FDS methods. Osher's scheme is robust and second-order accurate, when combined with the right state interpolation [76]. Furthermore, from a future perspective, it has a logical extension to more realistic primitive equations and a consistent boundary treatment. The last argument made us decide in favor of Osher's approximate Riemann solver before Roe's, which is often used in gas dynamics applications.

The semi-discrete system reads

$$\begin{aligned}\frac{\partial \underline{q}_{i,j}}{\partial t}+\frac{m_{i,j}^2}{\Delta x_{i,j}\Delta y_{i,j}}\Bigg[\;& T^{-1}(0)\,\underline{F}_{(\mathrm{O})}\left(T(0)\,\underline{q}^{\mathrm{L}}_{i+\frac{1}{2},j},T(0)\,\underline{q}^{\mathrm{R}}_{i+\frac{1}{2},j}\right)\frac{\Delta y_{i,j}}{m_{i+\frac{1}{2},j}}\\ &+T^{-1}(\tfrac{\pi}{2})\,\underline{F}_{(\mathrm{O})}\left(T(\tfrac{\pi}{2})\,\underline{q}^{\mathrm{L}}_{i,j+\frac{1}{2}},T(\tfrac{\pi}{2})\,\underline{q}^{\mathrm{R}}_{i,j+\frac{1}{2}}\right)\frac{\Delta x_{i,j}}{m_{i,j+\frac{1}{2}}}\\ &+T^{-1}(\pi)\,\underline{F}_{(\mathrm{O})}\left(T(\pi)\,\underline{q}^{\mathrm{L}}_{i-\frac{1}{2},j},T(\pi)\,\underline{q}^{\mathrm{R}}_{i-\frac{1}{2},j}\right)\frac{\Delta y_{i,j}}{m_{i-\frac{1}{2},j}}\\ &+T^{-1}(\tfrac{3\pi}{2})\,\underline{F}_{(\mathrm{O})}\left(T(\tfrac{3\pi}{2})\,\underline{q}^{\mathrm{L}}_{i,j-\frac{1}{2}},T(\tfrac{3\pi}{2})\,\underline{q}^{\mathrm{R}}_{i,j-\frac{1}{2}}\right)\frac{\Delta x_{i,j}}{m_{i,j-\frac{1}{2}}}\;\Bigg]\\ =\;&-\left(\underline{f}_{x_{\rm st}}\left(\underline{q}_{i,j},\underline{x_{\rm st}}_{i,j}\right)+\underline{f}_{y_{\rm st}}\left(\underline{q}_{i,j},\underline{x_{\rm st}}_{i,j}\right)\right),\end{aligned} \tag{3.14}$$

where $T(\theta)$ is a rotation matrix defined by

$$T(\theta)=\begin{pmatrix}1&0&0\\0&\cos\theta&\sin\theta\\0&-\sin\theta&\cos\theta\end{pmatrix} \tag{3.15}$$

and $\underline{F}_{(\mathrm{O})}$ is the Osher flux given as

$$\underline{F}_{(\mathrm{O})}(\underline{q}^{\mathrm{L}}, \underline{q}^{\mathrm{R}}) = \frac{1}{2}\left(\underline{F}(\underline{q}^{\mathrm{L}}) + \underline{F}(\underline{q}^{\mathrm{R}})\right) - \frac{1}{2}\int_{\underline{q}^{\mathrm{L}}}^{\underline{q}^{\mathrm{R}}} |A(\underline{q})|\, d\underline{q}. \tag{3.16}$$

A is here defined as the Jacobian of the fluxvector $\underline{F}$ with respect to $\underline{q}$, $A = \partial \underline{F}/\partial \underline{q}$. The absolute value of this Jacobian is defined by

$$|A(\underline{q})| = P(\underline{q})\,|\Lambda|\,P^{-1}(\underline{q}).$$

where P and Λ result from diagonalizing the Jacobian matrix as $A = P\Lambda P^{-1}$. Note that the Osher fluxes in (3.14) describe local fluxes, i.e., they point in the direction of the outwardly directed unit normal on the corresponding boundary. The Osher flux (3.16) approximates the local flux across a boundary $\delta\Omega$, which results when at the left and the right of this boundary the constant states $\underline{q}^{\mathrm{L}}$ and $\underline{q}^{\mathrm{R}}$ are found.

So far, we have not mentioned the evaluation of the constant states. It is through these evaluations that we are able to properly couple the different grids. Furthermore, the state evaluations determine the accuracy of our scheme. On a uniform grid, second-order accuracy can be proven [68]. We attend to this topic in the next section. It remains to say that the Osher scheme is special for its choice of the integration path in its flux (3.16). Using the Osher flux boils down to a maximum of five flux evaluations, $\underline{F}(\underline{q})$, per cell boundary. In case of the most common atmospheric flow patterns, i.e., flows where we have $|u| \leq \sqrt{gH}$, we find that the Osher flux requires only one flux evaluation per cell boundary, when we use the P-variant Osher path suggested by Hemker and Spekreijse [30]. Details of the construction of the integration path and the Osher flux can be found in Appendix 6.3.

Determination of the constant states

In this section we define the constant states. We still focus on the stereographic region zooming in on the state evaluation in the x_{st}-direction. The states in the y_{st}-direction are defined in a similar way. We apply 1D state interpolation, i.e., the state $\underline{q}^{\mathrm{L}}_{i+\frac{1}{2},j}$ only depends on the states of neighboring cells in the x_{st}-direction. For the remaining part of this subsection, we suppress the index j in our notation. To define the constant states, we use the $(\kappa = \frac{1}{3})$-scheme [76]. On a uniform grid it reads

$$\begin{aligned} \underline{q}^{\mathrm{L}}_{i+\frac{1}{2}} &= \underline{q}_i + \tfrac{(1-\kappa)}{4}(\underline{q}_i - \underline{q}_{i-1}) + \tfrac{(1+\kappa)}{4}(\underline{q}_{i+1} - \underline{q}_i), \\ \underline{q}^{\mathrm{R}}_{i+\frac{1}{2}} &= \underline{q}_{i+1} + \tfrac{(1-\kappa)}{4}(\underline{q}_{i+1} - \underline{q}_{i+2}) + \tfrac{(1+\kappa)}{4}(\underline{q}_i - \underline{q}_{i+1}). \end{aligned} \tag{3.17}$$

Unfortunately, our grid in the projected stereographic region is non-uniform. When the grid is sufficiently smooth, this discrepancy is often circumvented by simply applying the existing κ-scheme (3.17). Although this condition holds for our grid,

we do not adopt this approach. We wish to avoid any additional errors which might prevent us from properly identifying the influence of the coupling between the different grids. Therefore, we have applied a modification of the κ-scheme (3.17) for non-uniform grids. The general form of this modified κ-scheme can be found in Appendix 6.4 for different values of κ. The general form is defined as a function, I_κ, with the states and cell widths of neighboring grid cells in the interpolation direction as arguments. The standard non-uniform state interpolation is represented in Table 3.1.

	Left	Right
A	$\underline{q}^{\mathrm{L}}_{\frac{1}{2}} =$ Transformation $\underline{q}^{\mathrm{R}}_{\frac{1}{2}} = I_{-1}(\underline{q}_2, \underline{q}_1, \ell_2, \ell_1)$	$\underline{q}^{\mathrm{L}}_{N+\frac{1}{2}} = I_{-1}(\underline{q}_{N-1}, \underline{q}_N, \ell_{N-1}, \ell_N)$ $\underline{q}^{\mathrm{R}}_{N+\frac{1}{2}} =$ Transformation
B	$\underline{q}^{\mathrm{L}}_{\frac{3}{2}} = I_1(\underline{q}_1, \underline{q}_2, \ell_1, \ell_2)$ $\underline{q}^{\mathrm{R}}_{\frac{3}{2}} = I_{\frac{1}{2}}(\underline{q}_3, \underline{q}_2, \underline{q}_1, \ell_3, \ell_2, \ell_1)$	$\underline{q}^{\mathrm{L}}_{N-\frac{1}{2}} = I_{\frac{1}{2}}(\underline{q}_{N-2}, \underline{q}_{N-1}, \underline{q}_N, \ell_{N-2}, \ell_{N-1}, \ell_N)$ $\underline{q}^{\mathrm{R}}_{N-\frac{1}{2}} = I_1(\underline{q}_N, \underline{q}_{N-1}, \ell_N, \ell_{N-1})$
C	$\underline{q}^{\mathrm{L}}_{\frac{5}{2}} = I_{\frac{1}{2}}(\underline{q}_1, \underline{q}_2, \underline{q}_3, \ell_1, \ell_2, \ell_3)$	$\underline{q}^{\mathrm{R}}_{N-\frac{3}{2}} = I_{\frac{1}{2}}(\underline{q}_N, \underline{q}_{N-1}, \underline{q}_{N-2}, \ell_N, \ell_{N-1}, \ell_{N-2})$
D	$\underline{q}^{\mathrm{L}}_{i+\frac{1}{2}} = I_{\frac{1}{3}}(\underline{q}_{i-1}, \underline{q}_i, \underline{q}_{i+1}, \ell_{i-2}, \ell_{i-1}, \ell_i, \ell_{i+1}, \ell_{i+2})$, $\underline{q}^{\mathrm{R}}_{i+\frac{1}{2}} = I_{\frac{1}{3}}(\underline{q}_{i+2}, \underline{q}_{i+1}, \underline{q}_i, \ell_{i+3}, \ell_{i+2}, \ell_{i+1}, \ell_i, \ell_{i-1})$	

Table 3.1: The different state interpolation methods used near the grid boundary. The indices A, B, C and D here correspond with the different cell boundary situations illustrated in Figure 3.6.

Near the grid interface between the stereographic and spherical region, see Figure 3.4(a), the stencil of the non-uniform ($\kappa = \frac{1}{3}$)-scheme is too large, demanding state variables from outside the stereographic region. To avoid transformations and difficulties associated with the kink in the grid cells, we regard the grid interface as a real boundary. This means that locally we have to reduce the size of our stencil. To that end we have also formulated the non-uniform equivalents of the 2-point central ($\kappa = 1$)-scheme, the 2-point upwind ($\kappa = -1$)-scheme, and the 3-point upwind ($\kappa = \frac{1}{2}$)-scheme. Figure 3.6 shows which interpolation scheme is applied on each cell boundary. The associated state interpolations are given in Table 3.1. Note that although it is a 3-point interpolation scheme, the ($\kappa = \frac{1}{3}$)-scheme, as opposed to the ($\kappa = \frac{1}{2}$)-scheme, cannot be applied at the cell boundaries $\delta\Omega_{5/2}$ and $\delta\Omega_{N-3/2}$, because in these cases a cell width from outside the stereographic region is needed. In the next section, we will discuss the Transformation entry in Table 3.1.

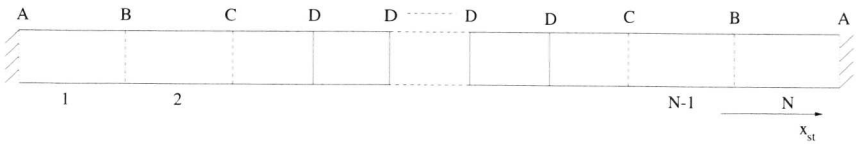


Figure 3.6: Illustration of the cell boundaries, where another interpolation scheme than standard is needed.

The finite volume method and the constant states on the spherical computational domain.

The same line of semi-discretization as described in Section 3.3.2 is applied to derive the semi-discrete system for the region II, see Figure 3.2(a). Note that for this region calculations are done on the (λ, ϕ)-plane. The semi-discrete system easily follows from equations (3.14)-(3.16), when we replace $m_{i,j}$, $\Delta x_{i,j}$, $\Delta y_{i,j}$, $\underline{f}_{x_{\rm st}}$, $\underline{f}_{y_{\rm st}}$, and $\underline{q}$ successively by $1/(a\cos\phi_{i,j})$, $\Delta\lambda_{i,j}$, $\Delta\phi_{i,j}$, $\underline{f}_{\lambda}$, $\underline{f}_{\phi}$, and $\underline{q} = (H,\, Hu,\, Hv)$, where

$$\begin{aligned}
\underline{f}_{\lambda}(\underline{q}, \underline{r}) &= \left(0, -(f + \frac{u}{a}\tan\phi)Hv + \frac{gH}{a\cos\phi}\frac{\partial h_s}{\partial\lambda}, 0\right)^T, \\
\underline{f}_{\phi}(\underline{q}, \underline{r}) &= \left(0, 0, (f + \frac{u}{a}\tan\phi)Hu + \frac{gH^2\sin\phi}{2a\cos\phi} + \frac{gH}{a}\frac{\partial h_s}{\partial\phi}\right)^T.
\end{aligned}$$

Note that the form of the flux vectors $\underline{F}$ and $\underline{G}$ remains the same, since both coordinate systems are conformal.

To evaluate the constant states on region II we again use 1D state interpolation. This time it concerns interpolation in the λ- or ϕ-direction depending on the cell boundary under consideration. As standard interpolation scheme the $(\kappa = \frac{1}{3})$-scheme is applied. In the λ-direction this scheme can be applied everywhere, because, in that direction, our grid is uniform and has no grid boundaries. In the ϕ-direction we have to account for the grid interface between the spherical and the stereographic grids. We treat this interface as if it concerns a piecewise constant real boundary approximating the cell boundaries by the lines $\phi = \phi_{i,N_\phi+1/2}$, see Figure 3.7. The resulting, partially non-uniform grid distribution resembles the one in the stereographic direction. Therefore, the associated state interpolations easily follow by applying Table 3.1 in the ϕ-direction.

Interaction between the different computational domains.

It remains to discuss the Transformation entry in Table 3.1. We again turn to the stereographic computational domain associated with region I and focus on the $x_{\rm st}$-direction, see Figure 3.6. At the grid interface between region I and II the computational domains of these regions interact. To find the states $\underline{q}^{\rm L}_{1/2}$ and $\underline{q}^{\rm R}_{N+1/2}$

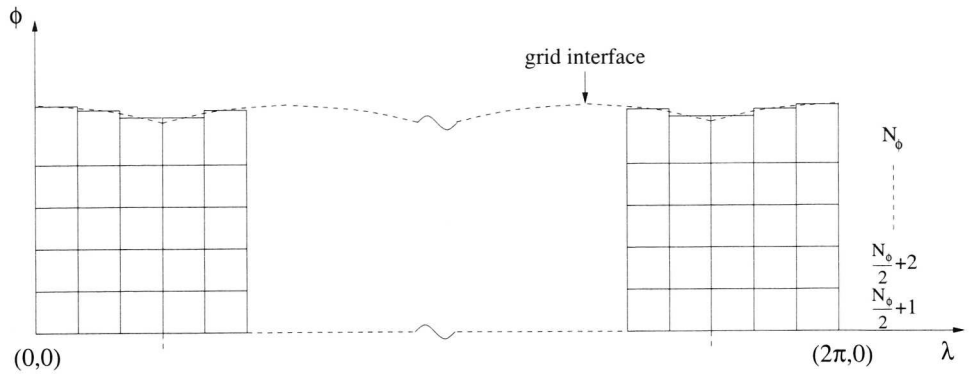


Figure 3.7: Projection of the northern hemisphere part of region II on the (λ, ϕ)-plane in combination with the approximated cell distribution at the grid boundary.

in stereographic variables, we transform the states in spherical variables found at the same cell interface boundary in the computational domain associated with region II. The word "transform" here indicates that we must convert the velocity field $\underline{u} = (u, v)$ into its stereographic representation. Note that the constant states in spherical variables are calculated by one-sided ($\kappa = -1$)-state interpolation in the ϕ-direction. This way of state evaluation yields that at every cell interface boundary, the 1D state interpolation to obtain $\underline{q}^{\mathrm{L}}_{1/2}$ and $\underline{q}^{\mathrm{R}}_{N+1/2}$ is performed in a different direction, i.e., in the direction of the projected meridians $\lambda_{1/2}$ and $\lambda_{N+1/2}$. In case of interpolations in the ϕ-direction, the Transformation entries, i.e., $\underline{q}^{\mathrm{L}}_{1/2}$ and $\underline{q}^{\mathrm{R}}_{N_\phi+1/2}$ in spherical variables, follow after transformation of the corresponding constant states in stereographic variables found at the same cell boundaries in the computational domain of region I. Here the word "Transformation" means that we must convert the velocity field $\underline{U} = (U, V)$ into its spherical equivalent. Note that, depending on the cell's position, the constant state in stereographic variables concerns a constant state calculated by one-sided ($\kappa = -1$)-state interpolation in x_{st}- or y_{st}-direction.

We conclude this section with some remarks on accuracy. In more dimensional problems a finite volume method is at most second-order accurate. To provide an order estimate we cite Spekreijse [68]. For a uniform grid, he proved, that a scheme like (3.14) is second-order accurate for interpolations based on the κ-scheme. On a large part of our domain, i.e., almost everywhere on the spherical region, see Section 3.3.2, his estimate is valid, because our grid is uniform. However, since we combine different grids, it is difficult to give the exact order of our scheme across the whole sphere. It is obvious that we endure some accuracy loss around the interface, which will be referred to as the connection problem. To be conclusive about its impact, we will give a numerical order estimate in Section 3.4.2.

3.4 Numerical tests

In this section we focus on two main objectives. First, we wish to establish to what extent the introduction of the stereographic grid resolves the problems related to the use of a global spherical coordinate system. Second, we wish to validate our spatial discretization scheme or, in other words, how Osher's scheme behaves, when applied to the SWEs on the sphere, and how accurate its results are.

To meet the necessity of a good benchmark to test new numerical methods for solving the SWEs in spherical geometry, Williamson *et al* [88] developed a test set, containing seven different test cases of increasing complexity. We concentrate on test case 2 of this test set, i.e., on the global steady state non-linear zonal geostrophic flow. Test case 2 provides us with a good test to examine the scheme's ability to handle the poles. Furthermore, it serves as a test for our Osher scheme, because it includes non-linear aspects of the SWEs. As holds for the whole test set, test case 2 is not entirely appropriate to demonstrate all favorable features of our scheme, i.e., its behavior around strong gradients. The problems in the test set have solutions with rather smooth flow patterns. Hence it is suitable for a first assessment of accuracy behavior. Besides test case 2, we also successfully solved test cases 1 and 6, i.e., advection of a cosine bell over the pole and the Rossby-Haurwitz wave. To save space we present only results for test case 2. In future work we will attend to the other cases.

3.4.1 Test case 2: Global steady state non-linear zonal geostrophic flow

Test case 2 concerns a steady state analytic solution to the non-linear SWEs. It consists of a solid body rotation with the corresponding geostrophic height field H. A parameter α is used to specify the angle between the axis of the solid body rotation and the polar axis of the spherical coordinate system: $\alpha = 0$ indicates equatorial flow and $\alpha = \pi/2$ yields flow across the pole. The analytic solution of test case 2 reads

$$H = h_0 - \left(\frac{a\Omega u_0}{g} + \frac{u_0^2}{2g}\right)(-\cos\lambda\cos\phi\sin\alpha + \sin\phi\cos\alpha)^2, \tag{3.18}$$

$$u = u_0\left(\cos\phi\cos\alpha + \sin\phi\cos\lambda\sin\alpha\right), \tag{3.19}$$

$$v = -u_0\sin\lambda\sin\alpha, \tag{3.20}$$

where the Coriolis parameter $f = 2\Omega\left(-\cos\lambda\cos\phi\sin\alpha + \sin\phi\cos\alpha\right)$ and $u_0 = 38.61$ m/s, $h_0 = 3.00 \times 10^3$ m. To be consistent with the article of Williamson *et al* [88], we tested our code for $\alpha = 0,\ 0.05,\ \pi/2 - 0.05,$ and $\pi/2$, where the second and third parameter values were added to avoid symmetries. In this article we will not present all the results, as our code produced good results for either value. We will concentrate on tests with parameter value $\alpha = \pi/2$, since for these tests the

corresponding velocity components initiate the strongest flow across the poles. We remark that these kind of flows can indeed be encountered in practical situations.

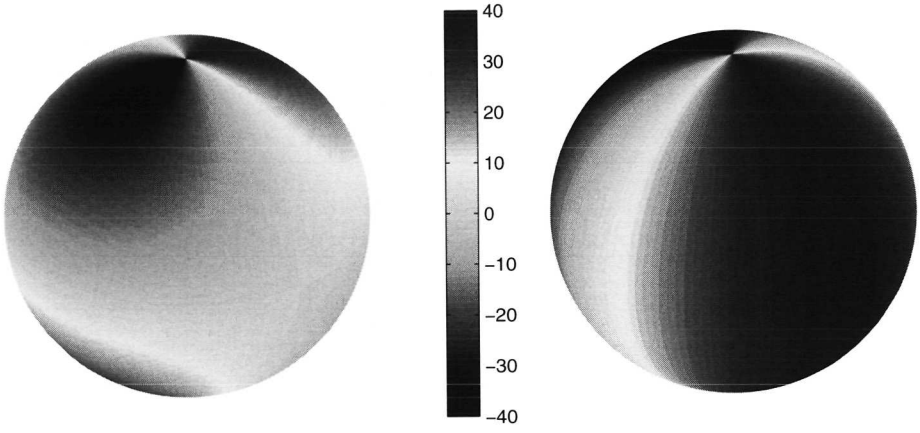


Figure 3.8: Representation of the analytic longitudinal velocity component u (left) and latitudinal velocity component v (right) on a global uniform lat-lon grid in case of global steady-state non-linear zonal geostrophic flow across the pole ($\alpha = \pi/2$).

In addition to the fact that we encounter a singularity problem when we apply the spherical formulation of the SWEs in the poles, we have to deal with some problems when approaching the poles. Figure 3.8 clearly illustrates the demand for additional caution near the poles. This figure represents the analytic longitudinal and latitudinal velocity components, u and v, found in the cell centers of an underlying uniform lat-lon grid in case of flow across the poles ($\alpha = \pi/2$). To emphasize our point we give the velocity components u and v, which follow from (3.18)-(3.20)

$$u = u_0 \sin\phi \cos\lambda, \tag{3.21}$$

$$v = -u_0 \sin\lambda. \tag{3.22}$$

The figure shows that the spherical velocity components strongly vary in the polar region, bringing about difficulties in numerical approximation methods. To properly represent these velocity components, a fine grid resolution, especially in the longitudinal direction, is necessary. However, too many grid cells can lead to problems for integration methods related to stability.

We discuss two remedies to these approximation and stability problems. First, we can decide to solve the SWEs on a stereographic grid. On a stereographic grid no severe resolution problems arise, as the velocity components U and V vary much less than the spherical ones, see Figure 3.9. Second, we can consider the reduced grid approach. In that case, the lat-lon grid is coarsened in the longitudinal direction

at given latitudes. For details we direct to [3] and [86]. Both remedies suffer some problems though. On a stereographic grid, we are confronted with a connection problem at the equator when we try to combine the stereographic grids on the northern and southern hemispheres, see Figure 3.9. On a (nearly) global lat-lon grid, we are not allowed to apply the reduced grid approach to its fullest extent. Repeated reductions to arrive, for instance, at four remaining grid cells next to the poles, are inadmissible, since in that case the grid near the poles is too coarse to represent the strongly varying velocity components. With a combination of both remedies, i.e., a combined grid with a reduced lat-lon grid away from the poles and a stereographic grid at the two polar caps, we can avoid these problems and benefit from either advantages, see Figure 3.10.

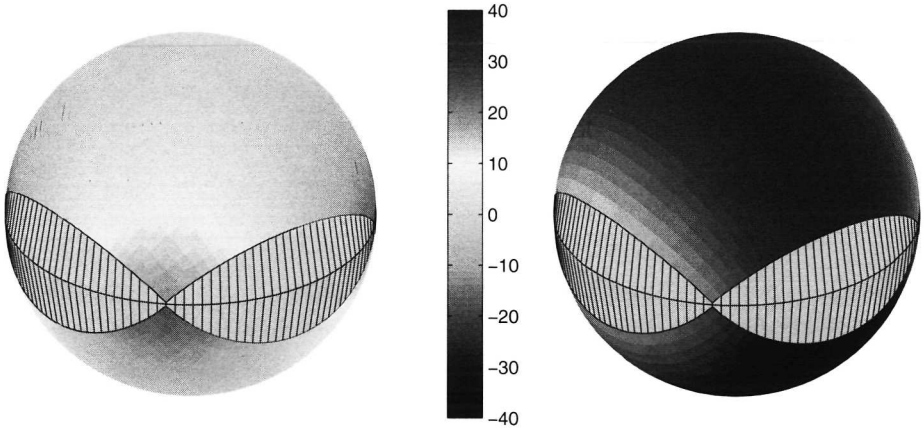


Figure 3.9: Representation of the analytic stereographic velocity components U (left) and V (right) on a "global" uniform stereographic grid in case of global steady state non-linear zonal geostrophic flow across the pole.

In the remaining part of this section we will address the following questions concerning our grid. Do the numerical results confirm the problems suggested when calculating on a global reduced lat-lon grid? Which factors determine the actual form of a combined grid, or in other words, how large should the stereocap be and how many reductions are allowed? And, how accurate are the results when calculated on a combined grid with realistic refinement?

3.4.2 Experiments on global lat-lon grids

The pole singularity

For tests on a global lat-lon grid to make sense, we must account for the non-existence of the spherical fluxes $\underline{F}$ and $\underline{G}$ in the poles. In practice, this problem is

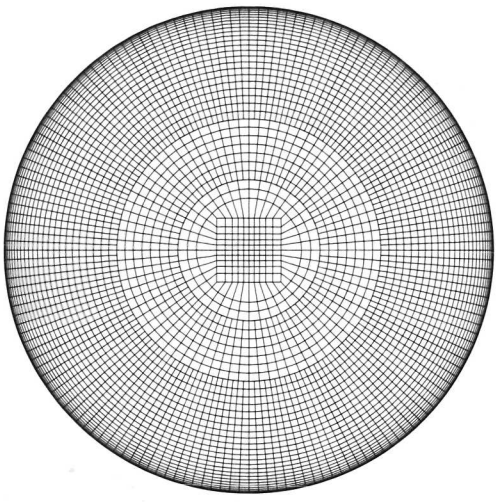


Figure 3.10: Projection of a combined grid consisting of a reduced lat-lon grid away from the poles and a stereographic grid at the two polar caps onto the cartesian (x,y)-plane (z=0). Two reductions were applied.

overcome by assuming a total zero flux across the boundaries corresponding to the poles. The question is whether the results significantly suffer from this assumption, both near and away from the poles. In fact, when the results do suffer from this assumption, we should reconsider investigating the global reduced lat-lon grid, since the results would be inadequate without an accurate resolution of the singularity problem in the pole.

We first ran a set of tests on a rectangular global lat-lon grid, where we varied the amount of gridpoints in the ϕ-direction, thus moving the neighboring cell centers closer to the pole with each test. Let nP define the amount of gridpoints in the ϕ-direction and let $\Delta\phi = \pi/\text{nP}$. In comparison with other tests, our grid distribution in the λ-direction is rather coarse ($\text{nL} = 72$). We must only make sure that the solution can be properly represented in that direction. In this way we are able to reduce computing time and avoid problems related to stability. The error measures on H are shown in Table 3.2. For time stepping we used the fourth-order Runge-Kutta method with small steps, such that the error $E_\text{r}(H)$ represents the spatial discretization error. $E_\text{r}(H)$ is defined as a maximum relative error,

$$E_\text{r}(H) = \max_{(i,j)} \left| \frac{H_{i,j} - H(\lambda_i, \phi_j)}{H(\lambda_i, \phi_j)} \right|,$$

where $H(\lambda_i, \phi_j)$ gives the analytic solution of H in cell center (i, j). The max-norm is taken over a specified region. Note that since $H \gg 1$, the relative error provides a good indication of the accuracy of our results.

	$E_r(H)_{\text{pole band}}$	$E_r(H)_{\text{whole}}$
nP = 36	$2.1 \cdot 10^{-3}$	$9.8 \cdot 10^{-3}$
nP = 72	$1.1 \cdot 10^{-3}$	$5.7 \cdot 10^{-3}$
nP = 180	$8.7 \cdot 10^{-4}$	$5.3 \cdot 10^{-3}$

Table 3.2: Error measures on H for different values of nP taken over the volumes located next to the poles and over the whole domain on a rectangular lat-lon grid (nL = 72).

Table 3.2 clearly shows that in the band next to the poles the zero flux assumption does not lead to an error increase when approaching the poles. We even observe a minor decrease and the (relative) error certainly is sufficiently small for practical purposes. Moreover, the error in the pole band is smaller than the error over the whole domain. Note that since nL is fixed, convergence of the Osher scheme is not examined in these tests.

Pole resolution problem

As mentioned before and as discussed by Williamson and Browning in [87], we encounter representation problems when we try to approximate the spherical velocity components on a too coarse grid around the poles. The following tests have been chosen to show the severity of this problem. We tested four different reduced rectangular lat-lon grids, all having nL(0) = 64 grid cells in the longitudinal direction and nP = 192 cells in latitudinal direction. nL(0) is here defined as the amount of grid cells in the longitudinal direction on the unreduced grid part. When approaching the poles, we halve the amount of grid cells in the longitudinal direction, whenever the cell width in that direction projected onto the sphere, i.e., $a \cos\phi \Delta\lambda$, is reduced with a factor of 2 following the last reduction. The specific values for nL(0) = 64 and nP = 192 are chosen so that we can arrive on a coarse grid within a few reductions and for each grid part, containing the same amount of grid cells in longitudinal direction, enough grid cells in latitudinal direction are guaranteed. Successively, we apply 1, 2, 3 or 4 reductions at the latitudes $\phi = 60°$, $75.9375°$, $82.5°$, and $86.25°$. The errors are displayed in Table 3.3. This time we concentrate on the absolute error, $E_a(u)$, found for the velocity component u instead of for H, since this component suffers the most from the inadequacy to represent the flux on a coarse lat-lon grid. Furthermore, the absolute error is shown, because the velocity component may vanish in certain points of the globe, see (3.21) and (3.22). $E_a(u)$ is defined as the maximum absolute error

$$E_a(u) = \max_{i,j} |u_{i,j} - u(\lambda_i, \phi_j)| ,$$

where $u(\lambda_i, \phi_j)$ represents the analytic velocity component u in cell center (λ_i, ϕ_j). The maximum is taken over the whole grid, where the second column entry indicates on which grid part m the maximum error is found. The index m denotes the grid part found between the $|m|$-th and $|m|+1$-th reduction. We indicate the different grid parts at the northern hemisphere with positive values of m and at the southern hemisphere with negative values of m.

	$E_a(u)$	grid part m
0 reductions,	0.32	0
1 reduction at $\phi = 60°$	1.03	-1/1
2 reductions resp. at $\phi = 60°$, 75.9375°	3.67	-2/2
3 reductions resp. at $\phi = 60°$, 75.9375°, 82.5°	15.18	-3/3
4 reductions resp. at $\phi = 60°$, 75.9375°, 82.5°, 86.25°	23.99	-4/4

Table 3.3: Error measures on u taken over the whole domain on a global reduced lat-lon grid with different levels of reduction (nL(0) = 64, nP = 192). The second column displays on which grid part m the maximum error is located.

Giving that the analytic longitudinal velocity component u has a maximum of 38.61 m/s, the results speak for themselves. It is obvious that a significant number of cells next to the poles are needed to properly represent the velocity components. For example, in this case and starting from nL(0) = 64, two reductions giving 16 cells next to the poles, already result in a maximum relative error in the longitudinal velocity component u of about 10%. Note that the maximum errors are found in the grid part closest to the pole.

Order tests

In this part, we provide a numerical order estimate for our spatial discretization scheme. As described in Section 3.3.2, we expect to find second-order accuracy on a uniform grid. To verify this, we ran some tests on a global uniform lat-lon grid. We only performed calculations on a band between latitudes $\phi = -60°$ and $\phi = 60°$ to avoid small steps related to stability. On the other areas of the sphere we prescribed the analytic solution. Note that in this way accuracy losses due to the zero flux assumption across the poles are circumvented. Successively, we applied a uniform lat-lon grid with nL = 72, 144, 288, and 576. Table 3.4 shows the relative error measures on H. We consider the max-norm over the band.

The order factor between two successive grids is given in the third column of Table 3.4. In case of second-order accuracy this factor should be 4. For the higher orders observed, we have two possible explanations. First, the theoretical order

	$E_r(H)_{\text{band}}$	$\frac{E_r(H)_{\text{band}_{nL/2}}}{E_r(H)_{\text{band}_{nL}}}$
nL = 72	$2.05 \cdot 10^{-3}$	
nL = 144	$2.69 \cdot 10^{-4}$	7.6
nL = 288	$3.65 \cdot 10^{-5}$	7.4
nL = 576	$7.17 \cdot 10^{-6}$	5.1

Table 3.4: Error measures on H for different values of nL taken over a band between the latitudes $\phi = -60°$ and $\phi = 60°$ on a global uniform lat-lon grid, where we prescribed the analytic solution outside the band.

estimate holds in the asymptotic case, i.e., when nL approaches infinity. The order factor between the grids with nL = 576 and nL = 288 already approaches four. Second, on the band between the latitudes $\phi = -60°$ and $\phi = 60°$, the flow has a strongly one-dimensional character which coincides with the meridians. For a uniform grid Spekreijse [68] proved, that a scheme like (3.14) is third-order accurate for interpolations based on the $(\kappa = \frac{1}{3})$-scheme in the 1D case. This might explain why on the coarser grids our order factors are close to eight. Note that the value 5.1 can then be attributed to the fact that on finer grids the volumes move closer to the boundary of the band, where the one-dimensional character of our flow diminishes.

In case of a non-uniform grid we provide a numerical order estimate. We evaluate the results found after calculations on a global reduced lat-lon grid. We ran four tests, each time doubling the value of nL(0) defined as the amount of grid cells in the longitudinal direction on the unreduced grid part. We begin with nL(0) = 72. The cell distribution in the unreduced grid part is uniform. We again coarsen our grid each time the cell width in the longitudinal direction projected onto the sphere is reduced by a factor of 2 as compared to the preceding reduction. In case of our grids, this rule yields three or four reductions. To make sure that our grid is not too coarse in regions close to the poles, we also ran test on grids with nL(0) = 288 and nL(0) = 576 where three instead of four reductions were applied as was originally prescribed by the reduction rule. The error measures on H, $E_r(H)$, are shown in Table 3.5. This time the max-norm is taken over the whole domain. The entries in the third column yield the order factor. Per grid we give the amount of reductions and their corresponding latitudes.

First, the results show that the reduced grid approach leads to first order accuracy. It should be noted though, that the error estimate is calculated in the max-norm over the whole domain. At the interface between the reduced grid parts we suffer from order reduction. Along the rest of our domain nearly second-order accuracy is found. Again, the grid must not be too coarse in the polar region. In case of nL(0) = 288 with four reductions, this condition is obviously not fulfilled

		$E_r(H)$	$\frac{E_r(H)_{\frac{nL}{2}(0)}}{E_r(H)_{nL(0)}}$
nL(0) = 72,	3 reductions at $\phi = 60°, 70°, 80°$	$1.10 \cdot 10^{-2}$	
nL(0) = 144,	3 reductions at $\phi = 60°, 75°, 82.5°$	$3.66 \cdot 10^{-3}$	3.0
nL(0) = 288,	4 reductions at $\phi = 60°, 75°, 82.5°, 86.25°$	$3.40 \cdot 10^{-3}$	1.1
nL(0) = 576,	4 reductions at $\phi = 60°, 75°, 82.5°, 86.25°$	$1.74 \cdot 10^{-3}$	2.0
nL(0) = 288,	3 reductions at $\phi = 60°, 75°, 82.5°$	$1.77 \cdot 10^{-3}$	2.1
nL(0) = 576,	3 reductions at $\phi = 60°, 75°, 82.5°$	$8.81 \cdot 10^{-4}$	2.0

Table 3.5: Error measures on H for different values of nL(0) taken over the whole domain on a global reduced lat-lon grid (nP = nL(0)/2), where grid coarsening is performed at the given latitudes.

resulting in almost no error reduction. Compared to unreduced grids, see, for instance, the entry 9.82×10^{-3} in Table 3.2 and 1.09×10^{-2} in Table 3.5, the reduced grid approach results in a small accuracy loss on coarse grids. The accuracy loss on finer grids will be larger since we find first order accuracy on a reduced lat-lon grid. However, its positive influence on the stability restriction for explicit time stepping compromises its use. As long as we take special care to guarantee an acceptable amount of grid cells next to the poles, the errors are sufficiently small for practical purposes.

We here omit an order estimate for calculations on a combined grid. As we will later show, the results mimic the accuracy behavior found on the reduced lat-lon grids. Investigations related to the connection problem are reported in the next section.

3.4.3 Experiments on combined grids

Placement of the stereocap

As nicely illustrated by Figure 3.9, in stereographic coordinates velocities over the poles behave normal and smoothly and hence can be approximated with much greater accuracy using a stereocap. However, we have also concluded that to cover the whole sphere a stereographic grid must be combined with, for instance, a lat-lon grid, creating a connection problem as examined in Section 3.3. In addition to the question of how this connection problem influences the accuracy, we wish to answer the question of what value we should take for $\overline{\phi}$, which we defined in Section 3.3.1 as the latitudinal boundary of the uniform lat-lon region R_{II}. We expect these questions to be related, since the larger $\overline{\phi}$, the smaller the cells in the connection band. We ran four tests on a combined unreduced grid, having nL = 144

points, i.e., with $\Delta\lambda = \Delta\phi = 2.5°$, where we gradually changed $\overline{\phi}$. Figure 3.11 shows the combined grids in case of the extreme values of $\overline{\phi}$. We coupled x_r defined in Section 3.3.1 as the x_{st}-coordinate of the top right-hand corner of the stereocap to $\overline{\phi}$, following $\phi_{x_r} = \overline{\phi} + \Delta\phi/2$. ϕ_{x_r} denotes the latitudinal coordinate corresponding to the stereographic coordinates $(x_{\mathrm{st}}, y_{\mathrm{st}}) = (x_r, y_r)$. Table 3.6 displays the different error measures on H, u and U over five different regions, i.e., over the uniform lat-lon grid part, over the cells located at the equator, over the interface cells connecting the two grids, over the stereographic grid parts and over the cells next to the poles. Note that the interface cells, the cells located at the equator and the cells next to the poles are also included in the lat-lon grid part or the stereographic parts, see Section 3.3.1. $E_{\mathrm{r}}(H)$ again describes the max-norm of the relative error on H. $E_{\mathrm{a}}(u)$ and $E_{\mathrm{a}}(U)$ describe max-norms of the absolute error on u and U, respectively.

	$E_{\mathrm{r}}(H)_{\mathrm{lat-lon}}$	$E_{\mathrm{r}}(H)_{\mathrm{equator}}$	$E_{\mathrm{r}}(H)_{\mathrm{interface}}$	$E_{\mathrm{r}}(H)_{\mathrm{stereo}}$	$E_{\mathrm{r}}(H)_{\mathrm{pole}}$
$\overline{\phi} = 47.5°$	$1.40 \cdot 10^{-1}$	$1.35 \cdot 10^{-1}$	$4.56 \cdot 10^{-2}$	$5.10 \cdot 10^{-2}$	$3.86 \cdot 10^{-2}$
$\overline{\phi} = 57.5°$	$7.58 \cdot 10^{-2}$	$3.64 \cdot 10^{-2}$	$3.27 \cdot 10^{-2}$	$1.41 \cdot 10^{-2}$	$7.14 \cdot 10^{-3}$
$\overline{\phi} = 67.5°$	$6.53 \cdot 10^{-3}$	$7.6 \cdot 10^{-5}$	$6.53 \cdot 10^{-3}$	$4.30 \cdot 10^{-3}$	$1.43 \cdot 10^{-3}$
$\overline{\phi} = 77.5°$	$2.48 \cdot 10^{-3}$	$2.30 \cdot 10^{-3}$	$2.31 \cdot 10^{-3}$	$7.42 \cdot 10^{-4}$	$3.00 \cdot 10^{-4}$
$\overline{\phi} = 87.5°$	$1.29 \cdot 10^{-3}$	$1.29 \cdot 10^{-3}$	$6.67 \cdot 10^{-4}$	$6.33 \cdot 10^{-4}$	$6.00 \cdot 10^{-4}$
	$E_{\mathrm{a}}(u)_{\mathrm{lat-lon}}$	$E_{\mathrm{a}}(u)_{\mathrm{equator}}$	$E_{\mathrm{a}}(u)_{\mathrm{interface}}$	$E_{\mathrm{a}}(U)_{\mathrm{stereo}}$	$E_{\mathrm{a}}(U)_{\mathrm{pole}}$
$\overline{\phi} = 47.5°$	43.06	17.45	43.06	37.38	$6.0 \cdot 10^{-2}$
$\overline{\phi} = 57.5°$	21.31	11.00	21.31	19.19	$3.6 \cdot 10^{-2}$
$\overline{\phi} = 67.5°$	5.23	2.39	5.23	3.43	$1.3 \cdot 10^{-2}$
$\overline{\phi} = 77.5°$	0.84	0.27	0.84	0.58	$3.8 \cdot 10^{-4}$
$\overline{\phi} = 87.5°$	0.14	0.02	0.14	0.30	$2.8 \cdot 10^{-4}$

Table 3.6: Error measures on H, u, and U for different values of $\overline{\phi}$ on four combined uniform lat-lon stereographic grids (nL=144). We give the errors $E_{\mathrm{r}}(H)$, $E_{\mathrm{a}}(u)$ and $E_{\mathrm{a}}(U)$ over five different regions, i.e., over the uniform lat-lon grid part, over the cells located at the equator, over the interface cells connecting the two grids, over the stereographic grid parts and over the cells next to the poles.

As expected, Table 3.6 shows that it is best to make the stereocap as small as possible, restricting accuracy loss due to the connection problem at the grid interface. The influence of reducing the size of the interface cells is particularly visible when concentrating on the maximum absolute error of the velocities. We encounter an accuracy reduction at the grid interface. However, on grids with a

small size stereocap this error is sufficiently small. Furthermore, both the errors on H and U are impressingly small at the poles. Comparing the overall error $E_r(H)$ for $\overline{\phi}=77.5°$ with the second entry in Table 3.5, we see that our calculations on a combined grid with a stereocap result in the same overall accuracy as the calculations on a compatible reduced grid. Note that this conclusion is true for modest and small sized stereocaps. For large stereocaps the interface cells become too distorted.

A combined grid with realistic refinement

Figure 3.11 shows that our conclusion should be handled with some consideration. When performance issues are important, the resolution increase on the stereocap due to size reduction can lead to a cut-back on the time-step caused by stability restrictions. However, this problem is easily resolved when we add the reduced grid approach to our combined grid. To show this, we end our numerical section on test case 2 of [88] by giving the results of a test on a combined reduced grid with realistic refinements. The stereocap is placed such that $\overline{\phi} = 85.625°$, $\mathrm{nL}(0) = 576$ and $\mathrm{nP}=288$. We apply three reductions, one at 60°, one at 75°, and one at 82.5°.

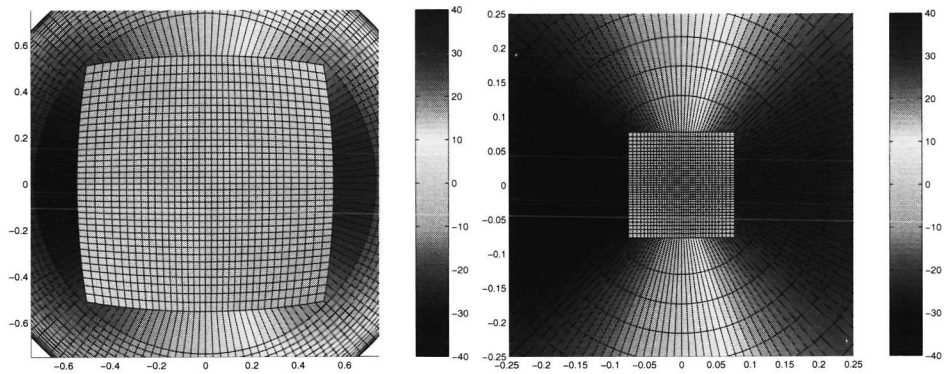


Figure 3.11: Projection of two combined grids ($\mathrm{nL}=144$) onto the cartesian (x,y)-plane ($z = 0$), where the stereocap in the right picture is moved closer to the pole. $\overline{\phi}=47.5°$ (left picture) and $\overline{\phi}=87.5°$ (right picture). Along the axes, the x- and y-coordinate are given as multiples of the earth radius.

The results confirm our expectations. We find a maximum relative error on H over our whole domain of $E_r(H) = 8.6 \cdot 10^{-4}$ and a maximum absolute error on u, U of $E_a(u, U)=0.092$. These errors show that a combined grid provides a good alternative to a global reduced lat-lon grid, see Table 3.5 case $\mathrm{nL}(0)=576$ with three reductions. This conclusion holds in particular, when the CFL-restriction demands

a too coarse lat-lon grid around the poles to maintain an acceptable time-step. This follows in comparing the smallest grid sizes found on the two different grid types. Note that in either case the smallest stepsize is found next to the poles. For the combined grid, the smallest grid size on the globe approximately reads

$$\frac{\sqrt{2}\pi a \cos\overline{\phi}}{\mathrm{nL}_{\mathrm{interface}}}. \tag{3.23}$$

On a reduced lat-lon grid, the smallest grid size reads

$$\frac{2\pi a \cos(90^\circ - \Delta\phi)}{\mathrm{nL}_{\mathrm{interface}}}. \tag{3.24}$$

Based on (3.23) and (3.24), we give the smallest grid size ratio for $\Delta\phi = 0.625^\circ$, $\mathrm{nL}_{\mathrm{interface}} = 72$ and $\overline{\phi} = 85.625^\circ$. The ratio reads

$$\frac{\frac{1}{2}\sqrt{2}\cos\overline{\phi}}{\cos(90^\circ - \Delta\phi)} \approx 4.95.$$

For explicit integration methods this ratio suggests a difference in computing time of approximately a factor of 5 in favor of the combined grid. Note that the time step restriction can indeed be encountered in practical situations, since high velocity components do occur in the polar regions.

3.5 Concluding remarks

Spectral methods currently dominate the field of approximation methods used in global circulation modeling. Since spectral methods become relatively expensive on fine grids, the demand for higher grid resolution and the better prospects for parallelization and local grid refinement have renewed interest in gridpoint methods. In this paper we have studied a sophisticated finite volume scheme for the spatial discretization of the SWEs in spherical geometry, viz. Osher's scheme [53] using the P-variant of Hemker and Spekreijse [30] for the integration path in the flux evaluation and third-order upwinding for the determination of the constant states. The scheme's second-order accuracy, its robustness, and its apprehension for the characteristic directions associated with the nonlinear equations, makes it a possible competitor to spectral methods for computations on fine grids. Note that in case of a combined grid, our method is second-order accurate in smooth regions away from the grid interface and first-order otherwise.

We have paid special attention to the pole singularity and the associated CFL-restriction. We have examined a combined grid to thoroughly alleviate the associated problems. This combined grid connects a stereographic grid in the polar regions with a lat-lon grid used at low latitudes. We have found that it is best to keep the size of the stereocap rather small to minimize connection errors at the grid

interface. Since a small stereocap involves small grid sizes at and near the cap, grid reduction in the lat-lon part can be used when it is needed to avoid very small grid sizes. In this manner the time step limitation for explicit integration methods emanating from the pole problem can be significantly reduced. Therefore, the resulting combined grid is advocated to be used together with an explicit integration scheme. In case time step stability plays a minor role, or when an implicit type integration method is used, we advocate using only a lat-lon grid, possibly reduced, because this approach is simpler. However, on lat-lon grids the singularity remains so that in case of flow over the poles the grid should be sufficiently fine.

Our findings are based on test cases 1, 2, and 6 of the standard test set from [88]. To save space we have shown results for test case 2 only. In the near future we will present results on time integration aspects using the spatial discretizations described in the current paper.

Chapter 4

Time Integration of the Shallow Water Equations in Spherical Geometry

Summary

The shallow water equations in spherical geometry provide a prototype for developing and testing numerical algorithms for atmospheric circulation models. In a previous paper we have studied a spatial discretization of these equations based on an Osher-type finite-volume method on stereographic and latitude-longitude grids. The current paper is a companion devoted to time integration. Our main aim is to discuss and demonstrate a third-order, A-stable, Rosenbrock method. Reducing the costs related to the linear algebra operations, this linearly implicit method is combined with approximate matrix factorization. Its efficiency is demonstrated by comparison with a classical third-order explicit Runge-Kutta method. For that purpose we use a known test set from literature. The comparison shows that the Rosenbrock method is by far superior.

4.1 Introduction

Present day atmospheric circulation models used in weather forecasting and climate research are often discretized by spectral transform methods. These methods are known to provide accurate solutions and to avoid the pole problem, which arises when grid-point methods are used on a standard latitude-longitude (lat-lon) grid. However, with the trend towards higher grid resolutions some of the main drawbacks of the spectral transform method become more apparent. These concern the high computational costs of the Legendre transform and the communication overhead for parallel distributed memory computers. Our investigations are directed at grid-point methods, which are expected to provide sufficient spatial accuracy for future fine-grid resolutions.

The current paper is devoted to the spherical Shallow Water Equations (SWEs), which reveal most of the major numerical difficulties associated with the horizontal dynamics found in the full set of primitive equations. The paper is a companion to [42], where we examined spatial discretizations based on an Osher-type finite-volume method [53] using the third-order upwind scheme for the constant state interpolation ($(\kappa = \frac{1}{3}$)-scheme [77]). This combination provides a solid spatial discretization for the hyperbolic SWEs.

In [42] we proposed a combined lat-lon and stereographic grid to avoid the pole problem that arises when solving the semi-discrete SWEs on a uniform lat-lon grid. In this article a different approach is adopted. Enhancing the grid resolution obviously necessitates an efficient time integration method to keep the solution costs affordable. The aim of the current paper is to demonstrate a third-order, A-stable, Runge-Kutta-Rosenbrock integration method. Rosenbrock methods are linearly implicit and hence require expensive linear system solves. We will show that this disadvantage can be overcome by the technique of approximate matrix factorization, which goes back to the early 1950s with splitting and alternating direction methods, see e.g., [54]. When combined with this technique, the Rosenbrock method does not only remain third-order consistent and A-stable, but it also becomes cost-effective. We will demonstrate its efficiency by a comparison with a classical third-order explicit Runge-Kutta method using a known SWEs test set from the literature [88]. The comparison shows that the Rosenbrock method is by far superior. In this paper the two integration methods are combined with the upwind spatial discretization from [42]. They can, of course, also be combined with the usual central spatial discretizations.

The paper is organized as follows. In Section 4.2 we briefly recall the system of SWEs and its linearization. The linearization is used as starting point to analyze stability. In Section 4.3, the third-order Rosenbrock method and the third-order explicit Runge-Kutta method are discussed. For the explicit method the time step restrictions on the uniform lat-lon and on the combined grid are derived. For the Rosenbrock method with approximate matrix factorization, A-stability is proven. Section 4.4 describes our numerical experiments, which will demonstrate the quali-

ties of the Rosenbrock method combined with approximate matrix factorization.

4.2 Preliminaries on the Shallow Water Equations

In this section, we briefly recall the system of SWEs in spherical coordinates and its linearization. Assuming Fourier-Von Neumann analysis, the linearized problem is used for the stability analysis. The spherical SWEs describe a pure initial-value problem on the rotating sphere and are defined as follows.

Let $\lambda \in [0, 2\pi)$ denote longitude, $\phi \in [-\frac{\pi}{2}, +\frac{\pi}{2}]$ latitude, and $t \geq 0$ time. Let u be the velocity in the longitudinal direction, v the velocity in the latitudinal direction, and h the height of the free surface above the sphere at sea level, i.e., $h = H + h_s$, where h_s describes the height of underlying mountains. Further, let $\underline{u}$ denote the horizontal velocity field (u, v), f the Coriolis parameter $2\Omega \sin\phi$ with Ω the angular velocity of the Earth, a the radius of the Earth, and g the gravitational constant. Using the flux-form, the two-dimensional SWEs, being composed of a continuity equation and two momentum equations, read [32, 88]

$$\frac{\partial H}{\partial t} + \nabla \cdot (H\underline{u}) = 0, \tag{4.1}$$

$$\frac{\partial Hu}{\partial t} + \nabla \cdot (Hu\underline{u}) = (f + \frac{u}{a}\tan\phi)\, Hv - \frac{gH}{a\cos\phi}\frac{\partial h_s}{\partial \lambda} - \frac{g}{a\cos\phi}\frac{\partial(\frac{1}{2}H^2)}{\partial\lambda}, \tag{4.2}$$

$$\frac{\partial Hv}{\partial t} + \nabla \cdot (Hv\underline{u}) = -(f + \frac{u}{a}\tan\phi)\, Hu - \frac{gH}{a}\frac{\partial h_s}{\partial \phi} - \frac{g}{a}\frac{\partial(\frac{1}{2}H^2)}{\partial\phi}, \tag{4.3}$$

where the divergence operator is defined by

$$\nabla \cdot \underline{u} = \frac{1}{a\cos\phi}\left[\frac{\partial u}{\partial\lambda} + \frac{\partial(v\cos\phi)}{\partial\phi}\right]. \tag{4.4}$$

The terms on the right-hand side in (4.2) and (4.3) represent forcing terms. It concerns the Coriolis force, the curvature terms, and the hydrostatic pressure gradient force. Along with the lat-lon coordinate system we apply stereographic coordinates. To save space we here omit the corresponding formulations of the SWEs. In [42] we have studied the spatial discretization of both formulations using the Osher upwind scheme.

4.2.1 The linearization

Adopting standard practice, we consider the frozen linearized system of (4.1)–(4.4) to analyze the stability properties of the integration methods. Let us linearize around a constant state vector $\overline{q} = (\overline{H}, \overline{Hu}, \overline{Hv})^T$, where the upper bar refers to frozen variables. The resulting linearized system then reads

$$q_t + A\, q_\lambda + B\, q_\phi = Cq, \tag{4.5}$$

where $q=(H, Hu, Hv)^T$,

$$A = \frac{1}{a\cos\phi}\begin{pmatrix} 0 & 1 & 0 \\ -\overline{u}^2 + g\overline{H} & 2\overline{u} & 0 \\ -\overline{u}\,\overline{v} & \overline{v} & \overline{u} \end{pmatrix}, \quad B = \frac{1}{a}\begin{pmatrix} 0 & 0 & 1 \\ -\overline{u}\,\overline{v} & \overline{v} & \overline{u} \\ -\overline{v}^2 + g\overline{H} & 0 & 2\overline{v} \end{pmatrix}. \quad (4.6)$$

and the force matrix,

$$C = \begin{pmatrix} 0 & 0 & \frac{\tan\phi}{a} \\ \frac{-g}{a\cos\phi}\frac{\partial h_s}{\partial\lambda} - \frac{2\tan\phi}{a}\overline{u}\,\overline{v} & \frac{2\tan\phi}{a}\overline{v} & \frac{2\tan\phi}{a}\overline{u} + \overline{f} \\ \frac{-g}{a}\frac{\partial h_s}{\partial\phi} + \frac{\tan\phi}{a}(\overline{u}^2 - \overline{v}^2) & -C_{23} & C_{22} \end{pmatrix}.$$

Note that the constant coefficient matrices A, B, and C do not commute, which implies that their eigensystems differ. Consequently, it is not possible to further simplify equation (4.5) to a scalar equation. For our analysis, we therefore need the eigenvalue-eigenvector decompositions of A and B. We have $A = X_A E_A X_A^{-1}$ and $B = X_B E_B X_B^{-1}$ with

$$X_A = \begin{pmatrix} 0 & 1 & -1 \\ 0 & \bar{u} + \sqrt{g\bar{H}} & -\bar{u} + \sqrt{g\bar{H}} \\ \sqrt{g\bar{H}} & \bar{v} & -\bar{v} \end{pmatrix}, \quad X_A^{-1} = \frac{1}{\sqrt{g\bar{H}}}\begin{pmatrix} -\bar{v} & 0 & 1 \\ \frac{1}{2}\left(\sqrt{g\bar{H}} - \bar{u}\right) & \frac{1}{2} & 0 \\ -\frac{1}{2}\left(\sqrt{g\bar{H}} + \bar{u}\right) & \frac{1}{2} & 0 \end{pmatrix},$$

$$X_B = \begin{pmatrix} 0 & 1 & -1 \\ \sqrt{g\bar{H}} & \bar{u} & -\bar{u} \\ 0 & \bar{v} + \sqrt{g\bar{H}} & -\bar{v} + \sqrt{g\bar{H}} \end{pmatrix}, \quad X_B^{-1} = \frac{1}{\sqrt{g\bar{H}}}\begin{pmatrix} -\bar{u} & 1 & 0 \\ \frac{1}{2}\left(\sqrt{g\bar{H}} - \bar{v}\right) & 0 & \frac{1}{2} \\ -\frac{1}{2}\left(\sqrt{g\bar{H}} + \bar{v}\right) & 0 & \frac{1}{2} \end{pmatrix},$$

and

$$E_A = \mathrm{diag}\left(\frac{\bar{u}}{a\cos\phi}, \frac{\bar{u} + \sqrt{g\bar{H}}}{a\cos\phi}, \frac{\bar{u} - \sqrt{g\bar{H}}}{a\cos\phi}\right). \quad (4.7)$$

$$E_B = \mathrm{diag}\left(\frac{\bar{v}}{a}, \frac{\bar{v} + \sqrt{g\bar{H}}}{a}, \frac{\bar{v} - \sqrt{g\bar{H}}}{a}\right). \quad (4.8)$$

Note that both decompositions exist, because our system is hyperbolic. The eigenvalue expressions for A and B are related to well-known physical features. The values containing the $\sqrt{g\bar{H}}$ term are connected with the so-called gravity waves, while the remaining values are connected with the so-called advective waves. The corresponding wave speeds differ significantly, i.e., the gravity waves run much faster than the advective ones. In practice, these gravity waves need not be resolved, because most meteorologically important motions are close to geostrophic balance which implies low amplitude gravity waves. In general, unfortunately, these waves dictate the critical time step at which stability can still be guaranteed when using explicit methods. For this reason, we focus on alternative time integration methods.

Following [42], we spatially discretize our system using Osher's scheme [53] with a higher order state interpolation, which yields a second-order method. Assuming a uniform grid, Osher's scheme applied to the constant linear system (4.5) simplifies to the third-order ($\kappa = \frac{1}{3}$)-upwind scheme [77] as given below. Consider the cell-centered grid points,

$$\lambda_j = (j - \frac{1}{2})\Delta\lambda, \quad \Delta\lambda = \frac{2\pi}{N}, \quad \phi_k = -\frac{\pi}{2} + (k - \frac{1}{2})\Delta\phi, \quad \Delta\phi = \frac{\pi}{M}, \tag{4.9}$$

and let the grid function $w_{j,k}(t)$ denote the semi-discrete approximation to the solution $q(\lambda_j, \phi_k, t)$ of (4.5) on this grid. Denote $A^+ = X_A E_A^+ X_A^{-1}$, where $E_A^+ = (|E_A| + E_A)/2$ is obtained from E_A by replacing its negative entries by zero. Introduce analogously B^+ and A^-, B^-, where the positive entries in the eigenvalue matrix are replaced by zero. The semi-discrete ($\kappa = \frac{1}{3}$)-upwind approximation to (4.5) on grid (4.9) can then be written as

$$\frac{d}{dt} w_{j,k} = L\, w_{j,k}, \quad L = L_A + L_B + C, \tag{4.10}$$

where

$$L_A = -\left(A^+ D_A^+ + A^- D_A^-\right), \quad L_B = -\left(B^+ D_B^+ + B^- D_B^-\right). \tag{4.11}$$

The operators D_A^+ and D_A^- are the upwind and downwind operators in the longitude direction, i.e.,

$$D_A^+ w_{j,k} = \frac{w_{j-2,k} - 6w_{j-1,k} + 3w_{j,k} + 2w_{j+1,k}}{6\Delta\lambda}. \tag{4.12}$$

$$D_A^- w_{j,k} = \frac{-2w_{j-1,k} - 3w_{j,k} + 6w_{j+1,k} - w_{j+2,k}}{6\Delta\lambda}. \tag{4.13}$$

D_B^+ and D_B^- denote their counterparts along latitude. A^+, B^+ etc. are evaluated in each grid cell. For convenience of notation we omit their spatial dependence.

To analyze the semi-discrete system (4.10), we introduce the harmonic wave solution,

$$w_{j,k}(t) = \hat{w}(t)\, e^{\sigma(\omega_1 \lambda_j + \omega_2 \phi_k)}, \quad \sigma = \sqrt{-1}.$$

An elementary computation yields the ordinary differential equation for the Fourier transform $\hat{w}$,

$$\frac{d}{dt} \hat{w} = \hat{L}\, \hat{w}, \quad \hat{L} = \hat{L}_A + \hat{L}_B + C, \tag{4.14}$$

where

$$\hat{L}_A = -X_A \hat{E}_A X_A^{-1}, \quad \hat{L}_B = -X_B \hat{E}_B X_B^{-1}. \tag{4.15}$$

$\hat{E}_A$ and $\hat{E}_B$ are diagonal matrices with entries,

$$\hat{e}_A = \frac{1}{3}\frac{|e_A|}{\Delta\lambda}\left((\cos\xi_1 - 1)^2 + \mathrm{sign}(e_A)\,\sigma\,(4 - \cos\xi_1)\sin\xi_1\right), \quad \xi_1 = \omega_1\Delta\lambda. \tag{4.16}$$

and

$$\hat{e}_B = \frac{1}{3}\frac{|e_B|}{\Delta\phi}\left((\cos\xi_2 - 1)^2 + \mathrm{sign}(e_B)\,\sigma\,(4 - \cos\xi_2)\sin\xi_2\right), \quad \xi_2 = \omega_2\Delta\phi. \tag{4.17}$$

e_A denotes an eigenvalue of A. Likewise, e_B denotes an eigenvalue of B. A clarifying discussion on the eigenvalues of the ($\kappa = \frac{1}{3}$)-upwind scheme, (4.16) and (4.17), can be found in [38].

The stability behavior of any integration method applied to the linear semi-discrete system (4.10) is governed by its stability behavior for the three-dimensional ODE system in Fourier space (4.14). By periodicity and symmetry, it suffices to consider ξ_1, ξ_2 in the interval $[-\pi, 0]$. Note that in our notation the dependence of $\hat{w}$ on ξ_1, ξ_2 is suppressed. For an introduction to the theory of Fourier analysis for difference schemes, we refer to [24, 61].

To analyze stability in case of calculations on a combined grid, we also need the linearization and the Fourier decomposition of the SWEs in stereographic formulation. The derivation is similar to the one above and leads to completely equivalent expressions due to the conformal character of the stereographic and lat-lon mapping. Therefore, we only list the counterparts of the eigenvalues expressions,

$$E_{A_{\mathrm{st}}} = \mathrm{diag}\,(m\bar{U},\, m(\bar{U} + \sqrt{g\bar{H}}),\, m(\bar{U} - \sqrt{g\bar{H}})), \tag{4.18}$$

$$E_{B_{\mathrm{st}}} = \mathrm{diag}\,(m\bar{V},\, m(\bar{V} + \sqrt{g\bar{H}}),\, m(\bar{V} - \sqrt{g\bar{H}})), \tag{4.19}$$

where

$$m(\phi) = \frac{2}{1 + \alpha\sin\phi}.$$

and $\bar{U}$ and $\bar{V}$ denote the 'frozen' stereographic velocity components in x_{st}- and y_{st}-direction, respectively.

4.3 The Runge-Kutta integration methods

In this section, we discuss the third-order Rosenbrock method and the third-order explicit Runge-Kutta method. Both integration methods solve general non-linear ODE systems, $\dot{w} = F(w)$. Note that the semi-discrete system of SWEs fits into this framework. We expect the Rosenbrock method to be an efficient candidate to solve this semi-discrete system, since it permits large time steps. The costs per time step are relatively high. Therefore, the third-order explicit method is included for comparison.

4.3.1 The third-order Rosenbrock method

The method is derived from the general two-stage Rosenbrock formula from the stiff ODE field [13, 25],

$$\begin{aligned} w^{n+1} &= w^n + b_1 k_1 + b_2 k_2, \\ Sk_1 &= \tau F(w^n), \\ Sk_2 &= \tau F(w^n + \alpha_{21} k_1) + \gamma_{21} \tau J k_1, \\ S &= I - \gamma \tau J, \end{aligned} \tag{4.20}$$

where b_1, b_2, α_{12}, γ_{12} and γ are free parameters which determine the methods specific properties. The numerical solution w^n approximates w at time $t = t_n$, $\tau = t_{n+1} - t_n$ denotes the step size, and $J = F'(w^n)$ is the Jacobian matrix of $F(w)$ at $w = w^n$. When low to moderate accuracy is required, methods of the Rosenbrock type have proven efficient for a variety of stiff ODE applications [25]. For method (4.20) the order of consistency p is at most 3.

We analyze the stability properties of our method by applying (4.20) to the Fourier transformed problem (4.14). The general two-stage Rosenbrock method with $p \geq 2$ then yields an amplification factor $R(\tau \hat{L})$, i.e., $\hat{w}^{n+1} = R(\tau \hat{L})\, \hat{w}^n$, with $R(z)$ defined as the stability function,

$$R(z) = 1 + \frac{2z}{1 - \gamma z} + \frac{\frac{1}{2} z^2 - z}{(1 - \gamma z)^2}. \tag{4.21}$$

The stability function $R(z)$ yields A-stability for all $\gamma \geq \frac{1}{4}$. In case of the special value $\gamma = \frac{1}{2} + \frac{1}{6}\sqrt{3}$ a third-order, A-stable function is obtained. A-stability is attractive as it implies unconditional stability in the sense of Fourier-Von Neumann for stable linear problems. However, for multi-dimensional PDE applications as ours solving twice per time step a linear system with the matrix $I - \gamma \tau F'(w^n)$ is rather expensive. Therefore, we will apply approximate matrix factorization. By this technique the numerical algebra costs are substantially reduced, while $p = 3$ and A-stability are still possible.

Approximate matrix factorization

We rewrite the semi-discrete system $\dot{w} = F(w)$ as $\dot{w} = F(w) \equiv F_A(w) + F_B(w)$, where F_A denotes the semi-discrete longitudinal operator extended with the force terms present in equation (4.2) and F_B the semi-discrete latitudinal operator extended with the force terms present in equation (4.3). Hence, F_A and F_B are one-dimensional operators defined along sets of longitudinal and latitudinal grid lines, respectively. The idea of approximate matrix factorization is to redefine S by

$$S = (I - \gamma \tau J_A)(I - \gamma \tau J_B). \quad J_A = F_A'(w^n), \quad J_B = F_B'(w^n). \tag{4.22}$$

or, equivalently, J by

$$J = F'(w^n) + \gamma\tau\tilde{J}, \quad \tilde{J} = -J_A J_B. \tag{4.23}$$

Instead of solving a huge two-dimensional linear system, we thus solve two one-dimensional linear systems, each of which is uncoupled per grid line. The costs per step then amount to two function evaluations for F, one Jacobian evaluation, and one band solve per longitudinal and latitudinal grid line. Since we use Osher's scheme on a stencil of five grid points with three solution components, each Jacobian matrix $F'_A(w^n)$ and $F'_B(w^n)$ consists of a blockband matrix with five blocks of (3×3). Note that $F'_A(w^n)$ is slightly more complex as a consequence of the periodicity in longitudinal direction. The costs per time step are still considerably higher as compared to those of a standard explicit method. However, the Rosenbrock method combined with approximate matrix factorization yields a far more efficient method, as our numerical results will show, see Section 4.4.

Approximate matrix factorization is reminiscent of the splitting technique already used in more conventional alternating direction methods during the 1950s, see e.g., [54]. The technique has been used in various other applications since then, see e.g., [2]. The authors have applied it successfully to large-scale atmospheric transport-chemistry problems, using a second-order method from class (4.20), see [4, 81]. As an iterative technique, approximate matrix factorization has been successfully applied to large-scale transport problems in surface water [35]. A recent survey can be found in [34]. In [37] and references therein, interesting theoretical stability results are given revealing some limitations of approximate matrix factorization in three-dimensional applications.

Consistency and stability properties

With J defined as in (4.23), method (4.20) is third order consistent for arbitrary $\tilde{J}$ whenever

$$b_1 + b_2 = 1, \quad b_2(\alpha_{21} + \gamma_{21}) = \frac{1}{2} - \gamma, \quad b_2\alpha_{21}^2 = \tfrac{1}{3}, \quad \gamma^2 - \gamma + \tfrac{1}{6} = 0, \quad b_2\gamma_{21} = -\gamma. \tag{4.24}$$

The fifth condition $b_2\gamma_{21} = -\gamma$ results from the matrix factorization. These conditions yield a unique solution which defines the Rosenbrock method,

$$\begin{aligned} w^{n+1} &= w^n + \tfrac{1}{4}k_1 + \tfrac{3}{4}k_2, \\ Sk_1 &= \tau F(w^n), \\ Sk_2 &= \tau F(w^n + \tfrac{2}{3}k_1) - \tfrac{4}{3}\gamma\tau J k_1, \\ S &= (I - \gamma\tau J_A)(I - \gamma\tau J_B), \end{aligned} \tag{4.25}$$

with $\gamma = \frac{1}{2} + \frac{1}{6}\sqrt{3}$. For efficiency reasons, the matrix-vector multiplication in the second stage formula is removed by redefining k_2 by $k_2 - \frac{4}{3}k_1$. This gives the

following third-order Rosenbrock method[1]

$$\begin{aligned} w^{n+1} &= w^n + \tfrac{5}{4}k_1 + \tfrac{3}{4}k_2, \\ Sk_1 &= \tau F(w^n), \\ Sk_2 &= \tau F(w^n + \tfrac{2}{3}k_1) - \tfrac{4}{3}k_1, \\ S &= (I - \gamma\tau J_A)(I - \gamma\tau J_B). \end{aligned} \tag{4.26}$$

In the remainder of this section, we will discuss stability properties of (4.26) by means of Fourier-Von Neumann analysis. To obtain the linear recurrence relation which governs stability, we apply method (4.26) to the ODE system (4.14). Using the notation introduced in Section 4.2, we find the recurrence relation $\hat{w}^{n+1} = R(\hat{Z}_A, \hat{Z}_B)\,\hat{w}^n$ where $\hat{Z}_A = \tau(\hat{L}_A + C_A)$, $\hat{Z}_B = \tau(\hat{L}_B + C_B)$, and

$$R(\hat{Z}_A, \hat{Z}_B) = I + \hat{S}^{-1}(2\hat{S} + \frac{1}{2}\hat{Z} - I)\,\hat{S}^{-1}\hat{Z}, \tag{4.27}$$

with $\hat{Z} = \hat{Z}_A + \hat{Z}_B$ and $\hat{S} = (I - \gamma\hat{Z}_A)\,(I - \gamma\hat{Z}_B)$. Suppose that $\hat{Z}_A$ and $\hat{Z}_B$ are diagonalizable and share well-conditioned eigensystems. We can then proceed with the scalar counterpart of (4.27), which reads

$$R(z_A, z_B) = 1 + \frac{2z}{(1-\gamma z_A)(1-\gamma z_B)} + \frac{\frac{1}{2}z^2 - z}{(1-\gamma z_A)^2(1-\gamma z_B)^2}, \tag{4.28}$$

with $z = z_A + z_B$ and z_A and z_B denoting eigenvalues of respectively $\hat{Z}_A$ and $\hat{Z}_B$. A convenient property of the stability function (4.28) is that it mimics the A-stability property of the original stability function (4.21). However, in this case the range of acceptable γ-values of method (4.26) for which the A-stability property holds, is smaller, as is shown in the following theorem.

Theorem 2 *The factorized stability function (4.28) satisfies* $|R(z_A, z_B)| \leq 1$ *for all* z_A, z_B *with* $Re(z_A) \leq 0$, $Re(z_B) \leq 0$ *if and only if* $\gamma \geq \frac{1}{2} + \frac{1}{6}\sqrt{3}$.

Proof By the maximum modulus theorem, it suffices to consider imaginary values $z_A = ib_1, z_B = ib_2$ for arbitrary real numbers b_1, b_2. A simple computation gives $|R(ib_1, ib_2)| \leq 1$ if and only if

$$f(b_1, b_2) \equiv \alpha_1 b_1^2 b_2^2 + \alpha_2(b_1^2 + b_2^2) + \alpha_3 b_1 b_2 \leq 0, \tag{4.29}$$

where $\alpha_1 = 3\gamma^4 - 4\gamma^5, \alpha_2 = \frac{1}{4} - 2\gamma + 5\gamma^2 - 4\gamma^3, \alpha_3 = \frac{1}{2} - 4\gamma + 8\gamma^2 - 4\gamma^3$.

An extremum of the function f is either located at a stationary interior point or at a non-interior point, i.e., for $b_1 \to \pm\infty$ or $b_2 \to \pm\infty$. We first investigate its behavior for $b_1 \to \pm\infty$. In that case f yields

$$\lim_{b_1 \to \pm\infty} \frac{f(b_1, b_2)}{b_1^2} = \left(\alpha_1 b_2^2 + \alpha_2\right), \quad \forall b_2 \in \mathbb{R}.$$

[1]This method is studied independently in [45] for integrating advection-diffusion problems on sparse grids.

This function is non-positive for all b_2 when $\alpha_1 \leq 0$ and $\alpha_2 \leq 0$, which yields

$$\gamma \geq \frac{3}{4}. \tag{4.30}$$

The same result can be derived for $b_2 \to \pm\infty$, since $f(b_1, b_2)$ is symmetric in b_1 and b_2. An extremum can also be found in a stationary point of f. Solving for $(\frac{\partial f}{\partial b_1}, \frac{\partial f}{\partial b_2}) = (0,0)$ yields

$$\begin{aligned} &b_1 = b_2 = 0. && && (a) \\ &b_1 = b_2 = b \neq 0 && \text{with} \quad b^2 = -\tfrac{2\alpha_2+\alpha_3}{2\alpha_1}. && (b) \\ &b_1 = c \neq 0 \text{ and } b_2 = -c \neq 0 && \text{with} \quad c^2 = -\tfrac{2\alpha_2-\alpha_3}{2\alpha_1}. && (c) \end{aligned} \tag{4.31}$$

We first consider the stationary point $(b_1, b_2) = (0,0)$, where $f(b_1, b_2) = 0$. Let H_f denote the Hessian determinant in a stationary point $\underline{a}$.

$$H_f(\underline{a}) = \frac{\partial^2 f}{\partial b_1^2}(\underline{a}) \frac{\partial^2 f}{\partial b_2^2}(\underline{a}) - \left(\frac{\partial^2 f}{\partial b_1 \partial b_2}(\underline{a}) \right)^2.$$

According to, e.g., [73], the function f has a local maximum in $\underline{0}$ if $H_f(\underline{0}) > 0$ and $\frac{\partial^2 f}{\partial b_1^2}(\underline{0}) < 0$. Taking into account (4.30), we thus find that f remains non-positive in a neighborhood of $(b_1, b_2) = (0,0)$, when γ satisfies

$$\gamma > \frac{1}{2} + \frac{1}{6}\sqrt{3}.$$

This condition is only sufficient. The theorem does not provide a decisive answer when $H_f(\underline{0}) = 0$. In that case a further investigation of the behavior of f in a neighborhood of $\underline{0}$ is necessary. For the γ-values at which $H_f(\underline{0}) = 0$ only $\gamma = \frac{1}{2} + \frac{1}{6}\sqrt{3}$ guarantees non-positivity of f in a neighborhood of $\underline{0}$. So, for f to be non-positive, γ should satisfy the following necessary condition,

$$\gamma \geq \frac{1}{2} + \frac{1}{6}\sqrt{3}. \tag{4.32}$$

Finally, we consider the four remaining stationary points of (4.31). These stationary points only exist when $b^2 > 0$ and $c^2 > 0$. However, these conditions contradict with the conditions (4.30) and (4.32). Therefore, in case that f is non-positive over $\mathbb{R}^2$, these points do not exist.
Summarizing, f is non-positive for all $(b_1, b_2) \in \mathbb{R}^2$ iff $\gamma \geq \frac{1}{2} + \frac{1}{6}\sqrt{3}$. □

This result is of interest in its own, as it shows that for useful values of γ the A-stability property is not lost by the matrix factorization.[2] In general, the matrices $\hat{Z}_A$ and $\hat{Z}_B$ do not commute, so that true unconditional stability for the

[2] In [37] it is pointed out that for a three-term splitting such a result does not exist.

linearized SWEs cannot be concluded from Theorem 2. Note that Theorem 2 does provide a necessary condition in this case. The following example will illustrate that for the SWEs and noncommuting matrices $\hat{Z}_A$ and $\hat{Z}_B$, Theorem 2 provides a reliable indication for unconditional stability.

Example

We have approximated the maximum value of the amplification operator (4.27) over the interval $\xi_1, \xi_2 \in [-\pi, 0]$. Calculations are performed at a location near a pole, i.e., at a location, where the longitudinal grid size $\Delta\lambda\, a \cos\phi$ on the sphere becomes very small. Locations near the poles are believed to be most critical in relation to stability (the pole problem). The example serves to identify the γ-values at which the Rosenbrock method (4.26) yields an unconditionally stable method when applied to the linearized SWEs after been spatially discretized with Osher's scheme. For comparison, the same computation will be carried out for the third-order explicit Runge-Kutta method in Section 4.3.2.

Let $\bar{u} = \bar{v} = 30$, $g\bar{H} = 10^5$, $a = 42000000/(2\pi)$ (space and time units are meters and seconds). Choose $\phi = (\pi - \Delta\phi)/2$, i.e., a location close to the north pole. Furthermore, put $\Delta\lambda = \Delta\phi = \pi/128$, which corresponds approximately to a uniform $1.4^o \times 1.4^o$ grid. Omitting the force matrix C, we have computed accurate estimates of the maximum spectral radius of $R(\hat{Z}_A, \hat{Z}_B)$ for $\tau = 10^i$, $i = 0, 1, 2, 3, 4$ and $\gamma = 0.25, 0.50, 0.75, 0.8, 0.9, 1.0$. The maxima are determined for $-\pi \leq \xi_1, \xi_2 \leq 0$ using a 100×100 grid. The following table shows these maxima for $\gamma = 0.25, 0.50, 0.75$.

τ	1	10	10^2	10^3	10^4
$\gamma = 0.25$	1.0000	1.0000	1.0008	2.2355	3.2207
$\gamma = 0.50$	1.0000	1.0000	1.0000	1.4014	1.5067
$\gamma = 0.75$	1.0000	1.0000	1.0000	1.0000	1.0000

The table reveals conditional stability for $\gamma = 0.25$ and $\gamma = 0.5$ and indicates unconditional stability for $\gamma = 0.75$. Also for $\gamma = 0.8, 0.9, 1.0$ maxima equal to 1.0 are found. This leads us to conjecture unconditional stability for all $\gamma \geq 0.75$, in line with the result of Theorem 2. We believe that the slightly larger value for $\gamma = \frac{1}{2} + \frac{1}{6}\sqrt{3} \approx 0.789$ in this theorem is due to the fact that the requirement for A-stability is more stringent. This property allows eigenvalues to lie in the whole of the left half of the complex plane, which is not the case in practice. Recall that the value $\gamma = 0.75$ also plays a special role for the stability function (4.28). Inequality (4.29) implies $\gamma \geq 0.75$ for $|b_1|, |b_2| \to \infty$.

Because the force matrix C can possess eigenvalues with a small positive real part, we have omitted C in the above computation. Note that, since A, B and C

do not share the same eigenvectors, adding the matrix C does not simply mean that the linearized SWEs become unstable. However, maxima slightly larger than 1.0 can occur, see also the example in Section 4.3.2. We assume that the matrix A dictates the stability behavior of system (4.5), because it grows with the inverse of $\cos\phi$. Note that the entries of C are comparable in size. However, A multiplies the derivative q_λ and C is only a forcing matrix multiplying q.

4.3.2 Explicit Runge-Kutta time stepping

An explicit s-stage Runge-Kutta method applied to system $\dot{w} = F(w)$ has the form,

$$w^{n+1} = w^n + \tau \sum_{i=1}^{s} b_i F(W_i), \tag{4.33}$$

$$W_i = w^n + \tau \sum_{j=1}^{i-1} a_{ij} F(W_j), \qquad i = 1, 2, \ldots, s. \tag{4.34}$$

In combination with central differences for space discretization, the most popular explicit Runge-Kutta method for hyperbolic problems is the classical four-stage method of order four. This higher order method owes its popularity to its imaginary stability boundary of $\sqrt{8}$. In comparison with other explicit methods this boundary is satisfactory and in fact close to the optimal value $s - 1 = 3$ for explicit Runge-Kutta methods [33]. However, since we employ upwinding in the space discretization, a different method is chosen.

Stability considerations

Let us consider methods of order $p = s$ for $s = 1, 2, 3, 4$. When applied to a Fourier transformed problem like (4.14), such a method yields a polynomial amplification operator $R(\hat{Z}), \hat{Z} = \tau\hat{L}$, with $R(z)$ defined by the truncated Taylor series,

$$R(z) = \sum_{i=0}^{p} \frac{1}{i!} z^i. \tag{4.35}$$

Assuming that the most severe time step restriction indeed emerges from the longitudinal operator in the polar region, it makes sense to first examine stability for the longitudinal operator alone. Hence, we take $\hat{L} = \hat{L}_A$. Since our operator is diagonalizable, we are then able to examine stability through the scalar recurrence relation $\hat{w}^{n+1} = R(z)\,\hat{w}^n$, where

$$z = \frac{\nu_A}{3}\left((\cos\xi_1 - 1)^2 + \mathrm{sign}(e_A)\,\sigma\,(4 - \cos\xi_1)\sin\xi_1\right), \quad \sigma = \sqrt{-1}, \tag{4.36}$$

with $-\pi \le \xi_1 \le 0$ and ν_A denoting the one-dimensional CFL number,

$$\nu_A = \frac{\tau\,|e_A|}{\Delta\lambda}. \tag{4.37}$$

and e_A denoting an eigenvalue of A, see (4.7). To determine the maximal value of ν_A at which each method is stable, it suffices to draw the z_A-loci which lie inside the stability region of the stability function. Accurate estimates from [38] yield

s	1	2	3	4
ν_A	0	0.87	1.62	1.74
ν_A/s	0	0.43	0.54	0.43

The scaled CFL-number, ν_A/s, is related to efficiency. Note that explicit Euler ($s=1$) is not stable. For the other three cases, the scaled CFL numbers ν_A/s are almost equal and close to 0.5. Note that the case $s=4$ includes the classical four-stage method of order four. At equal costs, third-order methods are slightly more stable.

Substitution of the maximal wave speed (maximal eigenvalue (4.16)) into ν_A yields a time step restriction for linear stability. Let $\bar{u} > 0$, then

$$\tau \leq \frac{\nu_A \, \Delta\lambda}{\max |e_A|} = \frac{a \cos(\phi) \, \nu_A \, \Delta\lambda}{\bar{u} + \sqrt{g\bar{H}}}. \tag{4.38}$$

On a uniform grid ($\Delta\lambda = \Delta\phi$) closest to the poles, $\cos(\phi) \approx \frac{1}{2}\Delta\lambda$, yielding

$$\tau \leq \frac{a \, \nu_A}{2(\bar{u} + \sqrt{g\bar{H}})} \, \Delta\lambda^2. \tag{4.39}$$

Consequently, we face a quadratic dependence on the spatial grid size instead of the usual linear one. The quadratic dependence leads to unacceptably small step sizes.

Example

To illustrate the step size restriction (4.38), we return to the example of Section 4.3.1. For the data used, (4.39) yields $\tau \leq 5.8 \, \nu_A$. Hence, we find that $\tau \leq 9.4$ for any explicit three-stage, third-order Runge-Kutta method. In our application this step size restriction is very severe.

To check the validity of expression (4.38) we again compute the maximal spectral radius (see Section 4.3.1) of the amplification operator $R(\hat{Z})$ with $R(z)$ defined by the third degree polynomial (4.35). We now distinguish between a zero and nonzero force matrix C. The table below yields the maxima for a sequence of step sizes τ. The cases Z_{ABC} and Z_{AB} refer to a nonzero and zero force matrix C, respectively. For Z_{AB} the one-dimensional expression appears to be very precise, predicting linear stability for $\tau \leq 9.4$ and error growth for larger time steps. For Z_{ABC} we see nearly equal error growth for the larger time steps. For the smaller ones, we also see a modest growth. This growth is caused by an eigenvalue of $A + B + C$ with a small positive real part.

τ	8	9	9.4	10	11
Z_{ABC}	1.015	1.015	1.015	1.201	1.728
Z_{AB}	1.000	1.000	1.000	1.209	1.737

Relaxing the step size restriction : A different grid distribution

As mentioned before, there are several ways to reduce step size limitations. We here recall the grid modifications as used in [42]. We discussed two possible remedies, i.e., longitudinal grid coarsening towards the poles [3, 42, 86], and the use of a different grid structure and coordinate system in the polar regions [42, 56]. The latter approach concerns the construction of a combined grid consisting of two stereocaps on the northern and southern hemisphere, respectively, and a (reduced) lat-lon grid in the intermediate region. Figure 4.1 visualizes such a grid distribution. In stereographic coordinates the grid distribution on either stereocap is rectangular. The same holds on the intermediate region in lat-lon coordinates.

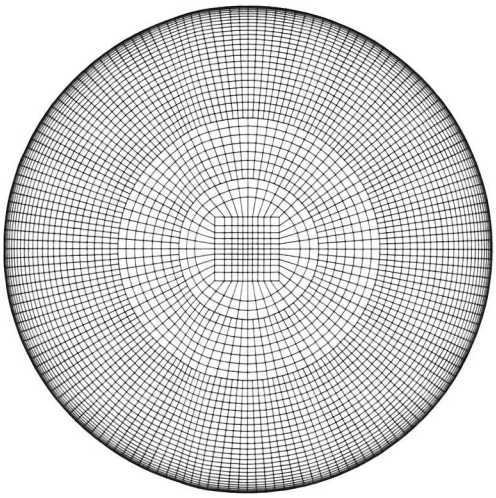


Figure 4.1: Projection of a combined grid consisting of a reduced lat-lon grid away from the poles and a stereographic grid at the two polar caps onto the cartesian (x, y)-plane $(z=0)$. Two reductions were applied.

On both grid types, we can derive a step size restriction for explicit Runge-Kutta methods similar to (4.38). We first consider a reduced grid. Such a grid is constructed from a uniform lat-lon grid around the equator by halving the amount of grid cells in the longitudinal direction when approaching the poles, whenever the

cell width in that direction projected onto the sphere is reduced by a factor 2. The distance, $a\cos\phi\,\Delta\lambda$, is called the physical cell width. Following (4.38), the step size restriction on a reduced grid yields

$$\tau \leq \frac{a\,\cos(\phi)\,\nu_A\,\Delta\lambda(\phi)}{\bar{u}+\sqrt{g\bar{H}}}, \tag{4.40}$$

where $\Delta\lambda(\phi)$ depends on the latitude ϕ, i.e., on the level of reduction. Assuming that the spherical variables, $\bar{H}$, $\bar{u}$ and $\bar{v}$, have the same order of magnitude along the whole domain, the step size restriction is most severe in the area, where the smallest physical cell width is found. On a global reduced grid this gives

$$\tau \leq \frac{2\pi}{\mathrm{nL}_{\mathrm{nRed}}}\,\frac{a\,\cos(\frac{\pi-\Delta\phi}{2})\,\nu_A}{\bar{u}+\sqrt{g\bar{H}}} = \frac{2\pi\;2^{\mathrm{nRed}}}{\mathrm{nL}_0}\,\frac{a\,\cos(\frac{\pi-\Delta\phi}{2})\,\nu_A}{\bar{u}+\sqrt{g\bar{H}}}. \tag{4.41}$$

where nRed denotes the amount of reductions on the northern hemisphere, and nL_0 and $\mathrm{nL}_{\mathrm{nRed}}$ denote the amount of cells in the longitudinal direction after 0 and nRed reductions, respectively.

On a stereographic grid, an analysis similar to Section 4.3.2 can be performed. Again assuming that the step size restriction is most severe in the area with the smallest physical cell width, we find on the combined grid,

$$\tau \leq \frac{\sqrt{2}\pi a\,\nu_A\cos\widetilde{\phi}}{\mathrm{nL}_{\mathrm{interface}}\;\max\left\{|\bar{U}+\sqrt{g\bar{H}}|,|\bar{V}+\sqrt{g\bar{H}}|\right\}}, \tag{4.42}$$

where $\widetilde{\phi}$ is the latitudinal boundary of the (reduced) lat-lon intermediate region of the combined grid and $\mathrm{nL}_{\mathrm{interface}}$ denotes the amount of longitudinal grid points on that boundary. The value $\sqrt{2}\pi a\cos\widetilde{\phi}/\mathrm{nL}_{\mathrm{interface}}$ approximates the smallest physical cell width over the sphere after projection of the stereocap onto the globe. $\bar{U}$ and $\bar{V}$ represent the linearized velocity component in x_{st}- and y_{st}-direction, respectively. Note that the stability condition (4.42) is composed of the two stability conditions found in each dimension, i.e., in the x_{st}- and y_{st}-direction, respectively. Since the matrices $A_{\mathrm{st}} = X_{A_{\mathrm{st}}}E_{A_{\mathrm{st}}}X_{A_{\mathrm{st}}}^{-1}$ and $B_{\mathrm{st}} = X_{B_{\mathrm{st}}}E_{B_{\mathrm{st}}}X_{B_{\mathrm{st}}}^{-1}$ do not share the same eigensystems, each linearized system has to be analyzed separately. In case of atmospheric applications, we expect the gravity waves to dominate the flow, i.e., the quantity $\sqrt{g\bar{H}}$ is large. Therefore, the step size restriction in stereographic variables is more or less direction independent.

To quantify the relation between the three step size restrictions (4.39), (4.41) and (4.42), we again focus on the example in Section 4.3.1. On the global uniform lat-lon grid, $\Delta\lambda{=}\Delta\phi{=}\frac{\pi}{128}$, we have

$$\tau \leq \tau_{\mathrm{uni}} = 5.8\,\nu_A. \tag{4.43}$$

On the corresponding reduced grid, $\Delta\lambda(0) = \Delta\phi = \frac{\pi}{128}$, when applying three reductions, we have

$$\tau \leq \tau_{\text{red}} = 2^{\text{nRed}}\, \tau_{\text{uni}} = 8\, \tau_{\text{uni}}. \tag{4.44}$$

Note that the number of reductions is limited by accuracy, i.e., too much reductions result in a too low grid resolution around the pole to properly represent the fast varying unit direction vectors in this area, see [42]. On the combined grid, we must first position the stereocap, i.e., we have to specify $\widetilde{\phi}$. For comparison, $\widetilde{\phi}$ is chosen such that the amount of reductions in the intermediate lat-lon region equals the amount of reductions found on the global reduced lat-lon grid, i.e., $\text{nL}_{\text{nRed}} = \text{nL}_{\text{interface}}$. In terms of τ_{uni} we find

$$\tau \leq \tau_{\text{combi}} = \frac{4\sqrt{2}\, \cos\widetilde{\phi}}{\cos\left(\frac{\pi - \Delta\phi}{2}\right)}\, \tau_{\text{uni}} \approx 34\, \tau_{\text{uni}} \quad \text{with} \quad \widetilde{\phi} = \frac{61\pi}{128}. \tag{4.45}$$

From (4.43)-(4.45), we can conclude that the step size restriction for explicit Runge-Kutta methods is considerably reduced when calculating on a global reduced or combined grid, the latter providing an even better alternative for the uniform lat-lon grid. On grids with a realistic resolution, the alleviation is even more apparent. On a global reduced grid with 3 reductions and $\Delta\lambda(0) = \Delta\phi = 2\pi/576$, and on a corresponding combined grid, $\widetilde{\phi} = \frac{137\pi}{288}$, we find

$$\tau_{\text{red}} = 8\, \tau_{\text{uni}},$$

and

$$\tau_{\text{combi}} = 40\, \tau_{\text{uni}}.$$

These are the step size restrictions for the grids on which we will evaluate the time integration methods in the following section.

The third-order explicit comparison method

In case the step size is limited by stability, a low order method, e.g., order $p=2$, will provide sufficient temporal accuracy. However, as seen in Section 4.3.2, order $p=3$ is slightly more efficient. Therefore, we use the following three-stage, third-order method for the comparison with the Rosenbrock method,

$$w^{n+1} = w^n + \tfrac{1}{6}\tau\, F(W_1) + \tfrac{1}{6}\tau\, F(W_2) + \tfrac{2}{3}\tau\, F(W_3), \tag{4.46}$$

$$W_1 = w^n, \quad W_2 = w^n + \tau F(W_1), \quad W_3 = w^n + \tfrac{1}{4}\tau\, F(W_1) + \tfrac{1}{4}\tau\, F(W_2). \tag{4.47}$$

To avoid an unacceptable workload, these experiments will be done on a combined grid.

4.4 Numerical experiments: A comparison

In the preceding section we described two Runge-Kutta methods, i.e., the third-order, A-stable, Rosenbrock method combined with approximate matrix factorization (4.26), henceforth called Ros3 with AMF, and the third-order, explicit, Runge-Kutta method (4.46), henceforth called RK3. For both methods the stability properties for the semi-discrete linearized system of SWEs (4.10) were investigated.

In this section we intend to show that Ros3 with AMF on a uniform lat-lon grid is far more efficient than RK3 even when this method is applied on a combined grid employing a stereocap to alleviate the step size restriction. We use both methods to integrate the system of ODEs resulting from spatially discretizing the SWEs with Osher's scheme. This finite volume method is discussed in [42]. To analyze whether Ros3 with AMF on a uniform lat-lon grid is more efficient than RK3 applied on a combined grid, we consider their relative workload per time step. An estimate of this relative workload is provided, which is confirmed by numerical experiments monitoring execution time.

Both methods are applied to three test cases from the widely acknowledged SWEs test set [88], which was especially developed to validate new numerical methods to be used in circulation models. It concerns Test 2, global steady-state nonlinear zonal geostrophic flow, Test 5, zonal flow over an isolated mountain, and Test 6, a Rossby-Haurwitz wave. Test 2 is chosen, because it provides a test with considerable activity in the polar region. Furthermore, it has a known analytic solution without compromising the non-linearity characteristic to the SWEs. Test 2 is a stationary test case, though. Therefore, to truly test our time integration method, we also consider two non-stationary problems, Test 5 and Test 6. For both cases, no exact solution is known and we have to rely on a high resolution spectral model for reference. These tests describe more realistic atmospheric flow patterns. For example Test 5, resolving a flow around a mountain, is challenging for most numerical solution methods. The other four tests from the SWEs test set, i.e., Tests 1, 3, 4 and 7, will be omitted, since they do not contribute additional information in relation to our efficiency question.

Calculations are performed on two different grids with related resolution. The uniform lat-lon grid has 576 grid points in longitudinal direction and 288 grid points in latitudinal direction, i.e., a $0.625° \times 0.625°$ grid. The combined grid consists of a reduced lat-lon grid for $\phi \in [-\widetilde{\phi}, \widetilde{\phi}]$ with $\widetilde{\phi} = 137\pi/288$ applying three reductions on each hemisphere and two stereocaps. Around the equator the resolution is equal to the resolution found on the uniform grid. By construction, the stereocap contains 18 grid points in x_{st}- and y_{st}-direction. Note that a combined grid has approximately 20% fewer grid points than the corresponding uniform lat-lon grid. The influence on the workload is not significant though, since some additional work is needed for the spatial coupling between the stereocap and the intermediate region. As mentioned before, efficiency mainly depends on the maximal step size allowed by the time integration method and its workload per time step.

In case of the RK3 method the step size is restricted by stability. We determine this step size by trial-and-error and denote it by τ_{RK3}. Note that the discussion on the step size restriction in Section 4.3 concerned the linearized system of SWEs and thus provides only an estimate for an upperbound for the step size. Analysis of the computational complexity of Ros3 with AMF shows that the workload per time step of the Ros3 method is approximately six times as large as the workload per time step of the RK3 method. This value is confirmed by numerical experiments on Tests 2, 5, and 6 monitoring execution time. Therefore, the Ros3 tests are run with step size $\tau_{\mathrm{Ros3}} = 6 \times \tau_{\mathrm{RK3}}$. Next the step size will be increased to determine the maximal step size at which stability is still obtained and the accuracy is still acceptable.

Besides testing on stability, we measure the accuracy of our solution for each method and step size over a prescribed time period. The accuracy is evaluated by the max-norm of the relative error of the depth of the fluid layer, $\mathrm{Rel}(H)$, and the absolute errors of the velocity components in longitudinal and x_{st}-direction, $\mathrm{Abs}(u, U)$, and latitudinal and y_{st}-direction, $\mathrm{Abs}(v, V)$, i.e.,

$$\begin{aligned}
\mathrm{Rel}(H) &= \max_{i,j} \left| \frac{H_{i,j} - H(\lambda_i, \phi_j)}{H(\lambda_i, \phi_j)} \right|, \\
\mathrm{Abs}(u) &= \max_{i,j} |u_{i,j} - u(\lambda_i, \phi_j)|, \\
\mathrm{Abs}(v) &= \max_{i,j} |v_{i,j} - v(\lambda_i, \phi_j)|,
\end{aligned}$$

and similar expressions for $\mathrm{Abs}(U)$ and $\mathrm{Abs}(V)$. $H_{i,j}$, $u_{i,j}$ etc. denote the approximate solutions. $H(\lambda_i, \phi_j)$ etc. are the reference solutions, where the solution is exact in case of Test 2 and given by a high resolution spectral method in case of Test 5 and Test 6. The high resolution spectral solutions are given on a daily basis.

Besides accuracy and stability, methods can also be tested on their abilities to conserve physical quantities, like energy and enstrophy, which are important for atmospheric flows. We monitored both quantities in the Ros3 runs. The cascade is negligible in all cases, i.e., approximately 0.1 percent over the prescribed time periods.

4.4.1 Test 2

Test 2 represents a solid body rotation, where the height field and the velocity components in longitudinal and latitudinal direction read

$$H = h_0 - \left(a\Omega u_0/g + u_0^2/(2g)\right) \left(-\cos\lambda\cos\phi\sin\alpha + \sin\phi\cos\alpha\right)^2, \tag{4.48}$$

$$u = u_0 \left(\cos\phi\cos\alpha + \sin\phi\cos\lambda\sin\alpha\right). \tag{4.49}$$

$$v = -u_0 \sin\lambda\sin\alpha, \tag{4.50}$$

where h_0 and u_0 are given. $u_0 = 38.6$ m/s, and $gh_0 = 2.94 \cdot 10^4$ m^2/s^2. Several orientations are specified, however, we use the one over the poles ($\alpha = \frac{\pi}{2}$). The

simulation period is five days. For the RK3 method $\tau_{\mathrm{RK3}} = 108$ s. To reach equal efficiency, we use Ros3 with AMF on the uniform grid with step size $\tau = 6 \times \tau_{\mathrm{RK3}} = 648$ s. The computations remain stable. For Ros3 with AMF, we then increase the step size to $\tau = 1350$ s, which still results in a stable computation. Instability is found for $\tau = 1500$ s. So, Ros3 with AMF applied on a uniform lat-lon grid is more efficient than an explicit method used on a related combined grid. We emphasize, that this grid type already significantly alleviates the step size restriction found on a uniform grid for an explicit method (recall the factor of 40 found by linear analysis). We also ran this test with the unfactorized Ros3 method. The computations with this method remained stable independent of the chosen step size.

In addition, the results on the uniform grid are more accurate than their counterparts on a combined one, as can be seen from Figure 4.3. The difference in accuracy is not caused by the time integration method, but can be attributed to the higher spatial errors found when calculating on a combined grid, see [42]. Furthermore, increasing the step size for Ros3 with AMF does not yield significant accuracy changes. Reducing the resolution on our uniform grid shows that, also in this case, the errors represent spatial ones. Note that for both methods the accuracy is satisfactory.

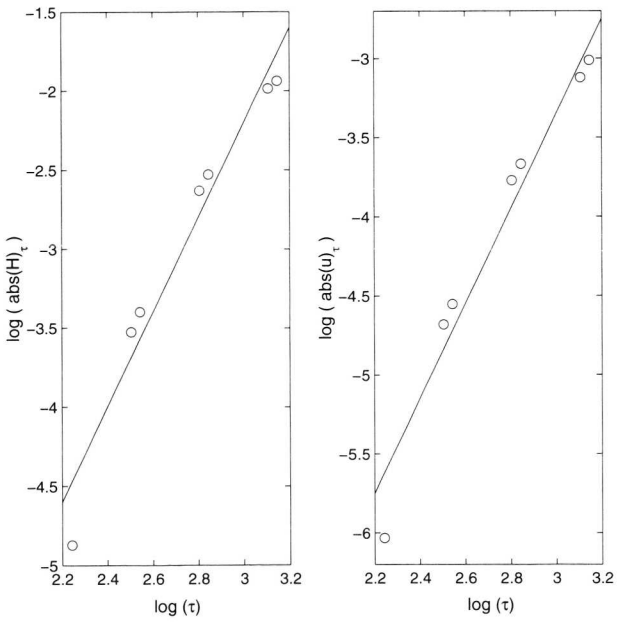


Figure 4.2: An order estimate applied to H and u respectively for Ros3 with AMF in case of test 2. The marks 'o' denote the $\log(\mathrm{abs}(H)_\tau)$ or $\log(\mathrm{abs}(u)_\tau)$, respectively. The solid lines illustrates the slope for a third order method.

A numerical order estimate for the non-linear SWE equations

Test 2 is also used to illustrate that Ros3 with AMF behaves as a third-order method. Calculations are done on a grid with resolution nL = 288 and nP = 144 for varying step sizes. As order estimate we use the l_∞-norm of the absolute error,

$$\text{abs(var)}_\tau = \max_{i,j} |\text{var}^\tau_{i,j,t} - \text{var}^{160}_{i,j,t}|.$$

where $\text{var}^\tau_{i,j,t}$ yields the approximate value of a variable var in gridpoint $\mathbf{x}_{i,j}$ at time t calculated with step size τ. We plotted this norm against the step size in a loglog-plot for respectively H and u, see Figure 4.2. The figure confirms that our method is third order consistent.

4.4.2 Test 5

Test 5 consists of a zonal flow parallel to the equator which impinges on a mountain. The initial solution is given by the solid body rotation provided for Test 2 (4.48)-(4.50) with $\alpha = 0$, $u_0 = 20$ m/s, and $h_0 = 5960$ m. The surface or mountain height is prescribed by a cone,

$$h_s = h_{s_0}\left(1 - \frac{r}{R}\right). \tag{4.51}$$

where $h_{s_0} = 2000$ m, $R = \pi/9$, $r^2 = \min[R^2, (\lambda - \lambda_c)^2 + (\phi - \phi_c)^2]$, $\lambda_c = 3\pi/2$, and $\phi_c = \pi/6$. The simulated time period is 15 days.

With regard to efficiency the results lead to conclusions similar to those found for Test 2. The RK3 method is run with a step size $\tau_{\text{RK3}} = 108$ s. The Ros3 method yields computational stability for $\tau = 648\,\text{s} = 6 \times \tau_{\text{RK3}}$. Since the reference solution is given on a daily basis, we have to secure that a one day time period can be taken in an integer number of time steps. The step size for Ros3 with AMF can be further increased. Even a step size of 2 h is possible. The results are less accurate though, see Figure 4.3. When a step size of 1 h is applied, an error in H of less than one percent is found. For the 2 h step size, we notice an error growth. Furthermore, we like to comment on the accuracy loss caused by the definition of the mountain height. To prescribe the orography, the test set introduces a cone as given by (4.51). This choice is a little unfortunate. The surface height is not continuously differentiable over the whole domain. The derivatives $\frac{\partial h_s}{\partial \lambda}$ and $\frac{\partial h_s}{\partial \phi}$ do not exist in the top and on the boundary of the cone. However, to evaluate the force terms of the SWEs (4.1)-(4.3) on the right-hand side, these derivatives are required. To circumvent this problem, we apply second-order central differences to approximate them. Results show an accuracy loss in the cells surrounding the areas, where $\frac{\partial h_s}{\partial \lambda}$ and $\frac{\partial h_s}{\partial \phi}$ are not defined. The test set does not prescribe how the undefined derivatives should be handled. Therefore, we can not be conclusive about accuracy in these areas. Figure 4.4 illustrates the relative error of H after 1 day computed with Ros3 with AMF on the uniform lat-lon grid with $\tau = 675$ s. The

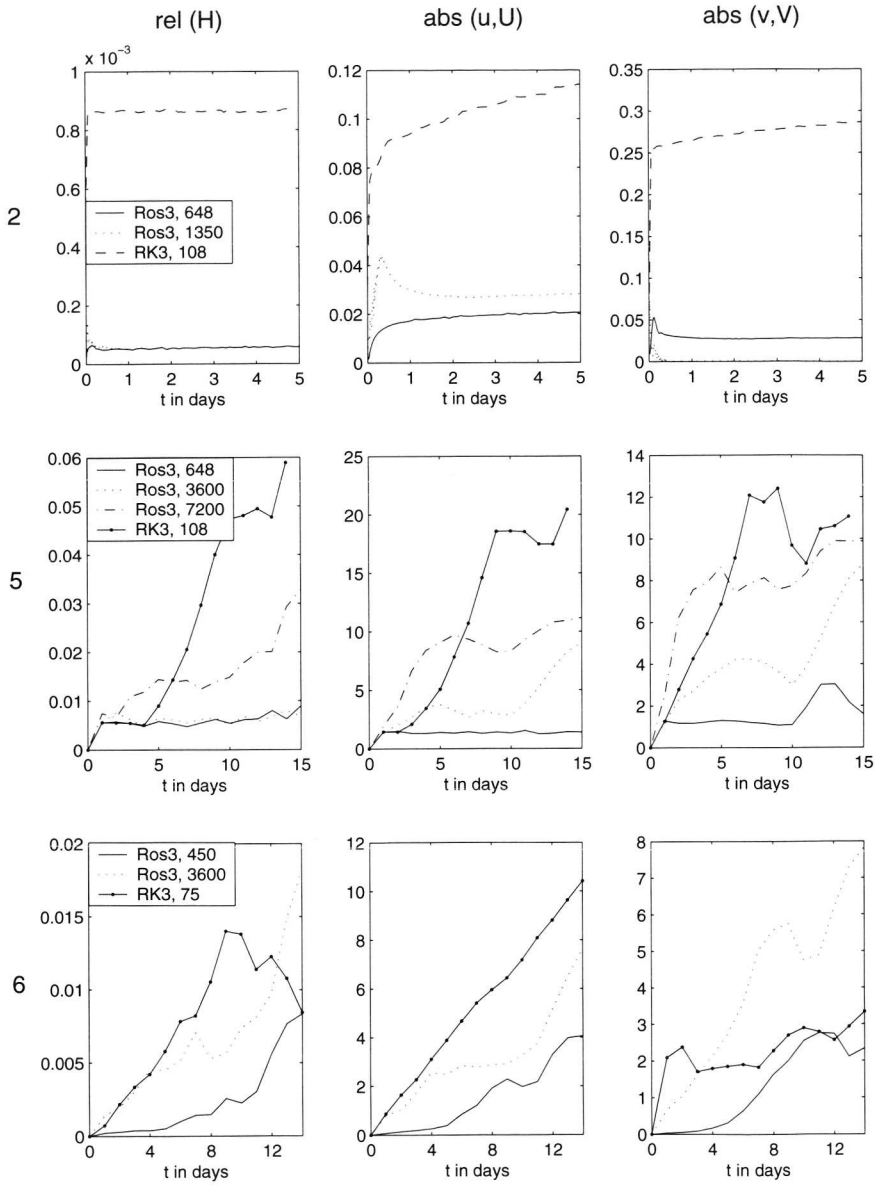


Figure 4.3: Max-norm of the relative error in H (first column), absolute error in u, U (second column) and absolute error in v, V (third column) for Test 2 (first row), Test 5 (second row) and Test 6 (third row) found for the two time integration methods (RK3 and Ros3 with AMF) with given step sizes. The errors are computed after each time step (Test 2) or on a daily basis (Test 5 and Test 6).

maximal errors are indeed located close to the circle $(\lambda - \lambda_c)^2 + (\phi - \phi_c)^2 = 0$ and close to the top $(\lambda, \phi) = (\lambda_c, \phi_c)$. Note that the errors remain local over the 1 day time period.

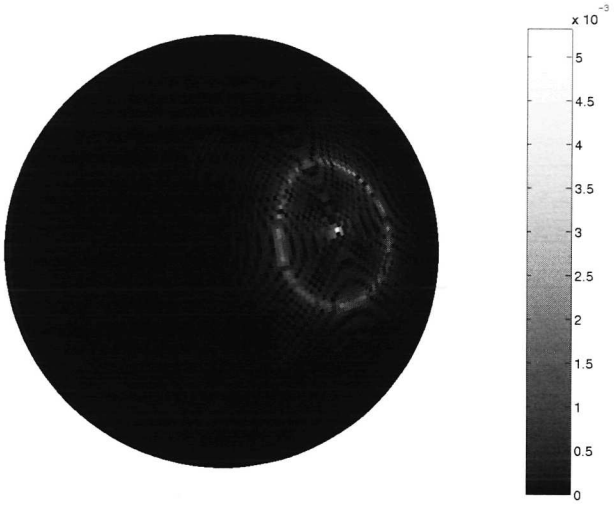


Figure 4.4: Relative error of H on a uniform grid in case of Test 5. Calculations are done with Ros3 with AMF on a uniform grid with $\tau = 675$ s.

From our results for Test 5 we again conclude that Ros3 with AMF on a uniform lat-lon grid is far more efficient than RK3 on a corresponding combined grid. We add that for Test 5 we are not really satisfied with the accuracy found in case of calculations on a combined grid. Numerical experiments show that the accuracy loss on the combined grid is mainly due to the introduction of the stereocaps. When calculating on a global reduced lat-lon grid the results are much more accurate. We assume that the vorticity waves partly intervene with the interface band and can not be represented sufficiently accurate. We could avoid this problem by moving the stereocap closer to the poles, however, this would result in a smaller step size.

4.4.3 Test 6

Test 6 is a Rossby-Haurwitz wave with a simulation period of 14 days. Again, no exact solution is known. Meteorologists consider this test standard, since similar flow patterns occur in practical applications. A reference solution is provided by a high resolution spectral circulation model.

The step size $\tau_{\text{RK3}} = 75$ s yields computational stability for the explicit RK3 method over the prescribed 14-day period. Ros3 with AMF is run for $\tau = 6 \times \tau_{\text{RK3}} = 450$ s. Increasing the step size, computational stability is still found for step size

$\tau = 3600$ s. We can conclude, that Ros3 with AMF on a uniform lat-lon grid is more efficient than the RK3 method on a corresponding combined grid. Again, the results on the uniform grid are more accurate.

4.5 Conclusion

When solving the semi-discrete SWEs on a global uniform lat-lon grid, an explicit time integration method suffers from severe restrictions on the step size (pole problem). This problem can be avoided by applying a suitable spatial grid or by choosing a more stable time integration method, viz. an implicit one. In [42] we proposed the application of a stereographic coordinate system in the polar regions combined with a reduced lat-lon grid in the intermediate region. In this article we considered an alternative time integration method, viz. the third-order Ros3 method with approximate matrix factorization.

We showed that the method is unconditionally stable, when applied to the linearized semi-discrete SWEs system on a uniform grid, provided that the Jacobian matrices of the fluxes in longitudinal and latitudinal direction commute. Furthermore, we showed that, due to the approximate matrix factorization, the method is cost effective. To verify its efficiency, we compared Ros3 with AMF on a uniform lat-lon grid to a third-order explicit RK3 method applied to the system of ODEs resulting from spatially discretizing our SWEs on a combined grid. Based on Test 2, Test 5 and Test 6 of the SWEs test set, we found that Ros3 with AMF is far more efficient than RK3 even when the latter is applied to the semi-discrete SWEs system on a combined grid, which already significantly alleviates the step size restriction.

Chapter 5

A Comparison of Operator Splitting and Approximate Matrix Factorization for the Shallow Water Equations in Spherical Geometry

Summary

The shallow water equations (SWEs) in spherical geometry provide a basic prototype for developing and testing numerical algorithms for solving the horizontal dynamics in global atmospheric circulation models. When solving the SWEs on a global fine uniform lat-lon grid, an explicit time integration method suffers from a severe stability restriction on the admissible step size. In a previous paper, we investigated an A-stable, linearly-implicit, third-order time integration method (Ros3), which we combined with approximate matrix factorization (AMF) to make it cost-effective. In this paper, we further explore this method and we compare it to a Strang-type operator splitting method. Our main focus is on the local error of the methods, their numerical dispersion relation and their accuracy and efficiency when applied to the well-known SWEs test set. The comparison shows that Ros3 with AMF accurately presents both low and mid frequency waves. Moreover, Ros3 with AMF makes a good candidate for the efficient solution of the SWEs on a global fine lat-lon grid. In contrast, Strang splitting is not advocated, in view of its inaccuracy in the polar regions and the resulting inefficiency.

5.1 Introduction

In current weather prediction and climate simulation, circulation models are used to simulate the dynamics of the atmosphere. A circulation model contains the primitive equations and a numerical solution method to solve them. Currently, there is much interest in accurate and efficient numerical methods for global circulation models. Spectral methods, long considered ideal for numerical simulation on the sphere, proved less efficient on the high resolution grids demanded to progress atmospheric modeling. In [42, 43], we therefore investigated a new grid-point method, which produced good results for the well established Shallow Water Equations (SWEs) testset [88]. This testset was developed to guide and stimulate the development of new numerical methods in circulation models and to provide a standard framework to assess them.

In [42], we discussed an Osher-type finite volume method for the spatial discretization of the SWEs on the sphere. Combined with a third-order upwind scheme for the constant state interpolation, this method is second-order accurate on uniform latitudinal-longitudinal (lat-lon) grids. In addition, we proposed an efficient time integration method in [43] for solving the resulting semi-discrete system. We applied a linearly implicit A-stable third-order Rosenbrock method (Ros3) to avoid the stability restriction associated with the well-known pole problem on uniform lat-lon grids and combined this method with approximate matrix factorization (AMF) to make it cost efficient. Ros3 with AMF produced good results for all testcases in the SWEs testset.

In this article, we further explore Ros3 with AMF and compare it to a Strang splitting method. Although both methods apply a splitting principle to simplify the solution process, their underlying techniques are very different. Strang splitting is an operator splitting technique, i.e., the original PDE problem is splitted additively in simpler PDEs which are solved separately. AMF on the other hand, factorizes the linear systems to be solved in the linearly implicit Ros3 method. In this work, we investigate the local error of both techniques, in particular, in the polar regions. Furthermore, we investigate their numerical dispersion relations to analyze their influence on the characteristic waves of the shallow water problem.

In meteorological practice, operator splitting techniques are considered unfit to solve the SWEs when they split the advection and Coriolis terms. Together these terms generate so called Rossby waves, which describe an important part of atmospheric dynamics. The separate treatment of the advection and Coriolis terms appears to jeopardize a correct representation of the Rossby waves and therefore, appears to obstruct a correct representation of the atmospheric tendency to geostrophic balance. We will show that Ros3 with AMF solves the Rossby waves accurately.

The theoretical analysis of the local error and the numerical dispersion relations serves to demonstrate that Ros3 with AMF is particularly useful to efficiently integrate the SWEs in time on high resolution grids. In addition, the results are used

to illustrate that a certain skepticism with respect to operator splitting methods is justified. The theoretical results will be confirmed by numerical experiments on the SWEs testset.

This paper is organized as follows. Section 5.2 describes the SWEs in spherical coordinates and gives a simplified formulation in a local Cartesian frame of reference. In Section 5.3, we consider the time integration methods, Ros3 with AMF and Strang splitting. Special attention is paid to accuracy and stability. Section 5.4 to Section 5.6 contain the actual comparisons between Ros3 with AMF and Strang splitting. Section 5.4 focuses on the local error of both methods when applied to the linearized SWEs in spherical coordinates. In Section 5.5, we analyze their numerical dispersion relations and demonstrate their influence on the characteristic waves associated with the original shallow water problem. In Section 5.6, we verify our theoretical results with numerical experiments. For that purpose, we concentrate on three test cases of the SWEs testset, i.e., Test 2, global steady-state non-linear zonal geostrophic flow, Test 5, zonal flow over an isolated mountain, and Test 6, the Rossby-Haurwitz wave. Finally, we formulate our conclusions in Section 5.7.

5.2 The SWEs in spherical geometry

The Shallow Water Equations on the sphere describe a flow in a shallow homogeneous incompressible and inviscid fluid layer on a rotating sphere. Since they cover important aspects of the horizontal dynamical behavior of the atmosphere, these equations serve as a first prototype of a circulation model. More specifically, they regard the atmosphere as a thin layer in which the density is uniform and constant and in which viscous effects can be ignored. In this section, we briefly recall their formulation, see also [42, 88]. For a thorough derivation, we refer to [6, 26, 55, 82].

Let (λ, ϕ, t) denote the independent variables longitude, $\lambda \in [0, 2\pi)$, latitude, $\phi \in [-\pi/2, \pi/2]$, and time, $t \geq 0$. Let u be the velocity in longitudinal direction defined by $u = a/\cos(\phi)\ \mathrm{d}\lambda/\mathrm{d}t$, v the velocity in latitudinal direction defined by $v = a\ \mathrm{d}\phi/\mathrm{d}t$ and H the depth of the fluid layer. Let h denote the height of the free surface above the sphere at sea level, i.e., $h = H + h_s$, where h_s accounts for the orography of the Earth and define $\underline{u}$ as the horizontal velocity field (u, v). Finally, let a denote the radius of the Earth, g the gravitational constant, and f the Coriolis parameter, $2\Omega \sin\phi$, with Ω the angular velocity of the Earth. The shallow water equations on the sphere are then formulated as

$$\frac{\partial Hu}{\partial t} + \nabla \cdot (Hu\underline{u}) = (f + \frac{u}{a}\tan\phi)\, Hv - \frac{gH}{a\cos\phi}\frac{\partial h_s}{\partial \lambda} - \frac{g}{a\cos\phi}\frac{\partial(\frac{1}{2}H^2)}{\partial\lambda}, \tag{5.1}$$

$$\frac{\partial Hv}{\partial t} + \nabla \cdot (Hv\underline{u}) = -(f + \frac{u}{a}\tan\phi)\, Hu - \frac{gH}{a}\frac{\partial h_s}{\partial \phi} - \frac{g}{a}\frac{\partial(\frac{1}{2}H^2)}{\partial\phi}. \tag{5.2}$$

$$\frac{\partial H}{\partial t} + \nabla \cdot (H\underline{u}) = 0, \tag{5.3}$$

where the divergence operator is defined by

$$\nabla \cdot \underline{u} = \frac{1}{a \cos\phi} \left[\frac{\partial u}{\partial \lambda} + \frac{\partial (v \cos\phi)}{\partial \phi} \right]. \tag{5.4}$$

The above equations are given in flux-form, which directly originates from the corresponding conservation laws. The first and second equation describe conservation of momentum in longitudinal and latitudinal direction, respectively. The third equation is known as the continuity equation. The source terms on the right hand side are connected to the Coriolis force, the curvature terms, and the hydrostatic pressure gradient force.

5.2.1 The locally Cartesian form of the SWEs

To facilitate the analysis of the numerical dispersion relations of our time integration methods, see Section 5.5, we also rely on a simpler version of the SWEs, viz. the SWEs in a locally Cartesian frame of reference. These equations are valid in a midlatitude synoptic system, which types of motion are common in dynamic meteorological practice. Based on midlatitude synoptic scale analysis, we are allowed to neglect the curvature terms in equations (5.1)–(5.3). In addition, we assume that the Earth is an ideal sphere, i.e., $h_s = 0$. Using the flux form, the SWEs in a locally Cartesian frame of reference are then defined as

$$\frac{\partial Hu}{\partial t} + \frac{\partial Hu^2}{\partial x} + \frac{\partial Huv}{\partial y} + g\frac{\partial \left(\frac{1}{2}H^2\right)}{\partial x} - fHv = 0, \tag{5.5}$$

$$\frac{\partial Hv}{\partial t} + \frac{\partial Huv}{\partial x} + \frac{\partial Hv^2}{\partial y} + g\frac{\partial \left(\frac{1}{2}H^2\right)}{\partial y} + fHu = 0, \tag{5.6}$$

$$\frac{\partial H}{\partial t} + \frac{\partial Hu}{\partial x} + \frac{\partial Hv}{\partial y} = 0, \tag{5.7}$$

where the x- and y-coordinate are everywhere aligned with the local east- and northward direction, respectively. u and v denote the velocity components in these directions. Note that the absence of the curvature terms does not affect any analysis concerning the impact of splitting the Coriolis force from the advection terms in numerical time integration methods.

5.3 The time integration methods

In this section we discuss the third-order Rosenbrock method (Ros3) combined with approximate matrix factorization (AMF) and a Strang splitting method. These integration methods solve general non-linear ODE systems $\dot{w} = F(w)$ with $w \in \mathbb{R}^m$. Note that any semi-discrete system of the SWEs fits into this framework, because the SWEs describe a pure initial value problem. These methods were also discussed in our earlier papers [43] and [44], respectively.

Both integration methods rely on a splitting principle, but on a different level in the solution process. Strang splitting is an operator splitting method, see [75]. It splits the different operators in the original PDE problem and solves them independently in successive substeps. Approximate matrix factorization simplifies the integration by factorizing the linear systems to be solved in the linearly implicit Ros3 method, such that these solves become less expensive.

Besides a general description of these methods, we will discuss their stability properties, which are of particular interest for meteorological applications. When calculating on a high resolution latitudinal-longitudinal grid, most time integration methods, read explicit methods, suffer from a severe restriction on the applicable time step. Since high resolution grids are the future trend, it is important to develop time integration methods which avoid such a limitation, see [86].

5.3.1 The third-order Rosenbrock method with approximate matrix factorization

We first concern ourselves with the third-order two-stage Rosenbrock method, see [13, 25, 43],

$$\begin{aligned} w^{n+1} &= w^n + \tfrac{5}{4}k_1 + \tfrac{3}{4}k_2, \\ S\,k_1 &= \tau\,F(w^n), \\ S\,k_2 &= \tau\,F(w^n + \tfrac{2}{3}k_1) - \tfrac{4}{3}k_1, \\ S &= (I - \gamma\tau J) \text{ with } \gamma = \tfrac{1}{2} + \tfrac{1}{6}\sqrt{3}, \end{aligned} \tag{5.8}$$

where $\tau = t_{n+1} - t_n$ denotes the step size, w^n denotes the numerical solution which approximates the exact solution w at time t_n, and $J = F'(w^n)$ denotes the Jacobian matrix dF/dw of $F(w)$ at $w = w^n$. This method is called linearly implicit, since it requires the solution of two linear systems with the matrix $(I - \gamma\tau J)$. In this sense, the method is intermediate between explicit and implicit Runge-Kutta methods.

The Rosenbrock method is A-stable with stability function,

$$R(z) = 1 + \frac{2z}{1 - \gamma z} + \frac{\frac{1}{2}z^2 - z}{(1 - \gamma z)^2}.$$

see [25]. A-stability is attractive as it implies unconditional stability in the sense of Fourier-Von Neumann analysis [24.82] for stable linear PDE problems.

A drawback of the Ros3 method (5.8) is that for multidimensional applications solving twice per time step a linear system with the matrix $I-\gamma\tau J$ is expensive. To reduce computational costs, while preserving A-stability and third-order accuracy, we therefore apply approximate matrix factorization. To demonstrate this technique, we rewrite the original ODE system as

$$\dot{w} = F(w) = F_\lambda(w) + F_\phi(w), \tag{5.9}$$

where F_λ and F_ϕ denote semi-discrete operators in longitudinal and latitudinal direction, respectively. In general, F also contains source terms, the distribution of which is not immediately evident from the definition of F_λ and F_ϕ. At this point we only assume that the source terms are distributed over F_λ and F_ϕ in some appropriate manner. A detailed discussion on the distribution of the source terms is presented in Section 5.5.3.

The idea of approximate matrix factorization (AMF), see e.g. [2.14.34.54], is to redefine S by

$$S = (I - \gamma\tau J_\lambda)(I - \gamma\tau J_\phi). \quad J_\lambda = F'_\lambda(w^n). \quad J_\phi = F'_\phi(w^n). \tag{5.10}$$

This significantly reduces the computational costs associated with the linear system solution. Instead of solving two huge two-dimensional linear systems per time step, we only have to solve four one-dimensional linear systems, each of which is uncoupled per grid line. While improving efficiency, Ros3 with AMF does not compromise the favorable properties of the original Ros3 method. First, Ros3 with AMF remains third-order accurate, see [43]. Second, Ros3 with AMF remains A-stable with stability function,

$$R(z_\lambda, z_\phi) = 1 + \frac{2z}{(1-\gamma z_\lambda)(1-\gamma z_\phi)} + \frac{\frac{1}{2}z^2 - z}{(1-\gamma z_\lambda)^2 (1-\gamma z_\phi)^2},$$

where $z = z_\lambda + z_\phi$, see Theorem 3.1 in [43] (Theorem 2). Theorem 3.1 implies that for matrices J_λ and J_ϕ which have a common complete system of eigenvectors, unconditional stability holds for stable linear problems in the sense of Fourier-Von Neumann analysis. Note that this is the case if these matrices commute. Although in general the matrices J_λ and J_ϕ do not commute, the theorem gives an indication for unconditional stability in practical applications.

5.3.2 The second-order Strang splitting method

Strang splitting belongs to the family of operator splitting methods. Operator splitting is based on the idea that most time-dependent ODE or PDE systems can be splitted additively in ODE or PDE systems which are simpler to solve.

We can think for instance of the earlier subdivision of F in a longitudinal and a latitudinal part, respectively. In each time step of the operator splitting method the subprocesses are treated separately using a certain order of reappearance. We adopt the symmetrical order of reappearance proposed by Strang [75], for which he proved second-order consistency.

We demonstrate this form of symmetrical Strang splitting for system (5.9). Let the numerical solution w^n approximate w at time t_n and let $\tau = t_{n+1} - t_n$ denote the step size. Furthermore, let $w_1(t)$ denote the solution of the subprocess $\dot{w_1} = F_\lambda(w_1)$ etc. Solving the substeps sequentially, one Strang splitting step from time t_n to t_{n+1} is given by

$$\dot{w_1} = F_\lambda(w_1), \quad w_1(t_n) = w^n, \qquad \text{for } t_n \le t \le t_{n+\frac{1}{2}}, \tag{5.11}$$

$$\dot{w_2} = F_\phi(w_2), \quad w_2(t_n) = w_1(t_n + \tfrac{\tau}{2}), \qquad \text{for } t_n \le t \le t_{n+1}, \tag{5.12}$$

$$\dot{w_3} = F_\lambda(w_3), \quad w_3(t_n + \tfrac{\tau}{2}) = w_2(t_n + \tau), \qquad \text{for } t_{n+\frac{1}{2}} \le t \le t_{n+1}, \tag{5.13}$$

$$\Rightarrow \quad w^{n+1} = w_3(t_n + \tau). \tag{5.14}$$

This process is second order in time under the assumption that the subprocesses are solved exactly or numerically with an integration method of at least order two. The error introduced by the splitting is called the splitting error. In case of commuting operators, i.e., $F'_\lambda F_\phi - F'_\phi F_\lambda \equiv 0$, this splitting error is zero, see [44, 65, 66]. In practice, most systems do not commute, so we always have a splitting error.

5.4 The local error

In this section, we focus on the structure of the local error for both integration methods. Our interest is in these errors in the polar regions. In actual applications, the local error of the Strang splitting method appears to increase significantly towards the poles as opposed to Ros3 with AMF.

We analyze the local error for the 'frozen' linearized system of equations derived from (5.1)–(5.4). Let us linearize around a constant state vector $\overline{q} = (\overline{Hu}, \overline{Hv}, \overline{H})^T$, where the upper bar refers to 'frozen' variables. Substituting $q = \overline{q} + q'$ in (5.1)–(5.4), the resulting linearized system reads

$$q_t + A\, q_\lambda + B\, q_\phi = C_{\text{cur}}\, q + C_{\text{cor}}\, q, \tag{5.15}$$

with the matrices A and B,

$$A = \frac{1}{a\cos\phi} \begin{pmatrix} 2\overline{u} & 0 & -\overline{u}^2 + g\overline{H} \\ \overline{v} & \overline{u} & -\overline{u}\,\overline{v} \\ 1 & 0 & 0 \end{pmatrix}, \quad B = \frac{1}{a} \begin{pmatrix} \overline{v} & \overline{u} & -\overline{u}\,\overline{v} \\ 0 & 2\overline{v} & -\overline{v}^2 + g\overline{H} \\ 0 & 1 & 0 \end{pmatrix}, \tag{5.16}$$

and the force matrices C_{cur} and C_{cor}.

$$C_{\mathrm{cur}} = \begin{pmatrix} \frac{2\tan\phi}{a}\,\overline{v} & \frac{2\tan\phi}{a}\,\overline{u} & -\frac{2\tan\phi}{a}\,\overline{u}\,\overline{v} \\ -C_{\mathrm{cur}_{12}} & C_{\mathrm{cur}_{11}} & \frac{\tan\phi}{a}\,(\overline{u}^2 - \overline{v}^2) \\ 0 & \frac{\tan\phi}{a} & 0 \end{pmatrix}, \quad C_{\mathrm{cor}} = \begin{pmatrix} 0 & f & 0 \\ -f & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{5.17}$$

where we omitted the apostrophes in equation (5.15) and assumed the Earth to be ideal, i.e., $h_s = 0$. Next, we define a uniform lat-lon grid with cell-centered grid points (λ_i, ϕ_j),

$$\lambda_i = \left(i - \frac{1}{2}\right)\Delta\lambda, \qquad \Delta\lambda = \frac{2\pi}{\mathrm{nL}}, \qquad i = 1, \ldots, \mathrm{nL},$$

$$\phi_j = -\frac{\pi}{2} + \left(j - \frac{1}{2}\right)\Delta\phi, \quad \Delta\phi = \Delta\lambda = \frac{\pi}{\mathrm{nP}}, \quad j = 1, \ldots, \mathrm{nP},$$

and let the grid function $q_{i,j}(t)$ denote the semi-discrete approximation to the solution $q(\lambda_i, \phi_j, t)$ of (5.15) on this grid. Spatially discretizating system (5.15) then yields the following ODE system,

$$\frac{\mathrm{d}q_{i,j}}{\mathrm{d}t} = L\,q_{i,j}, \; L = L_A + L_B + C_{\mathrm{cur}} + C_{\mathrm{cor}}, \tag{5.18}$$

where $L_A = -A\,\mathrm{D}_\lambda$ and $L_B = -B\,\mathrm{D}_\phi$. The matrices A and B are evaluated in each grid cell. D_λ and D_ϕ are linear difference operators in longitudinal and latitudinal direction, respectively. For instance, for a second order central discretization, they read

$$\mathrm{D}_\lambda q_{i,j} = \frac{q_{i+1,j} - q_{i-1,j}}{\Delta\lambda},$$

$$\mathrm{D}_\phi q_{i,j} = \frac{q_{i,j+1} - q_{i,j-1}}{\Delta\phi}.$$

Let L_λ and L_ϕ denote the linear splitting operators.

$$L_\lambda\, q_{i,j} = \left[L_A + C_{\mathrm{cur}_\lambda} + C_{\mathrm{cor}_\lambda}\right] q_{i,j}, \tag{5.19}$$

$$L_\phi\, q_{i,j} = \left[L_B + C_{\mathrm{cur}_\phi} + C_{\mathrm{cor}_\phi}\right] q_{i,j}, \tag{5.20}$$

with $C_{\mathrm{cur}_\lambda} + C_{\mathrm{cur}_\phi} = C_{\mathrm{cur}}$ and $C_{\mathrm{cor}_\lambda} + C_{\mathrm{cor}_\phi} = C_{\mathrm{cor}}$. System (5.18) can then be written as

$$\frac{\mathrm{d}q_{i,j}}{\mathrm{d}t} = L_\lambda\, q_{i,j} + L_\phi\, q_{i,j}. \tag{5.21}$$

The distribution of the source terms over L_λ and L_ϕ is partly fixed. The linearized curvature terms C_{cur_λ} and C_{cur_ϕ} read

$$C_{\mathrm{cur}_\lambda} = \begin{pmatrix} 0 & 0 & 0 \\ -\frac{2\tan\phi}{a}\,\overline{u} & 0 & \frac{\tan\phi}{a}\,\overline{u}^2 \\ 0 & 0 & 0 \end{pmatrix}, \; C_{\mathrm{cur}_\phi} = \begin{pmatrix} \frac{2\tan\phi}{a}\,\overline{v} & \frac{2\tan\phi}{a}\,\overline{u} & -\frac{2\tan\phi}{a}\,\overline{u}\,\overline{v} \\ 0 & \frac{2\tan\phi}{a}\,\overline{v} & -\frac{\tan\phi}{a}\,\overline{v}^2 \\ 0 & \frac{\tan\phi}{a} & 0 \end{pmatrix} \tag{5.22}$$

The first matrix, C_{cur_λ}, is exclusively connected to the curvature terms in the original SWEs system associated with a change in orientation of the unit vector in longitudinal direction, see [32]. Similarly, the second matrix, C_{cur_ϕ}, contains matrix entries related to the linearized curvature terms associated with a change in orientation of the unit vector in latitudinal direction. However, this matrix also contains part of the divergence operator. Of course, other splittings are possible, but only splitting (5.22) is natural. With respect to the Coriolis terms, no additional assumptions are made.

Observe that system (5.21) fits into the framework (5.9) with the additional advantage that $F_\lambda(w)$ and $F_\phi(w)$ are linear functions. Therefore, we can analyze the local Strang splitting and Ros3 with AMF error for the general linear ODE system,

$$\dot{w} = F(w) = F_\lambda(w) + F_\phi(w) \text{ with } F'_\lambda = \text{constant}.\ F'_\phi = \text{constant}. \tag{5.23}$$

Let $w(t_n)$ denote the exact solution of system (5.23) at time t_n and let $\tilde{w}^{n+1}$ denote the numerical solution after one time step with a particular time integration method from initial condition $w^n = w(t_n)$. The local error is then defined as

$$E_{\mathrm{loc}}^{n+1} = \left\| w(t_{n+1}) - \tilde{w}^{n+1} \right\|,$$

where $\|\cdot\|$ denotes a suitable norm, e.g., the L_∞- or L_2-norm. Assume that each of the substeps in the Strang splitting method is solved exactly in time. Omitting higher order terms, Taylor expansion of $w(t_{n+1})$ and $\tilde{w}^{n+1}$ around the exact solution $w(t_n)$ then yields

$$E_{\mathrm{loc}_{\mathrm{Strang}}}^{n+1} \approx \frac{1}{24} \Big\| \ F'_\lambda F'_\lambda F_\phi\left(w(t_n)\right) - 2\, F'_\lambda F'_\phi \left(\ F_\lambda\left(w(t_n)\right) + F_\phi\left(w(t_n)\right) \ \right) +$$

$$F'_\phi F'_\lambda \left(\ F_\lambda\left(w(t_n)\right) + 4\, F_\phi\left(w(t_n)\right) \ \right) - 2\, F'_\phi F'_\phi\, F_\lambda\left(w(t_n)\right) \Big\| \tau^3. \tag{5.24}$$

Similarly, we obtain the following local error for Ros3 with AMF,

$$E_{\mathrm{loc}_{\mathrm{Ros}}}^{n+1} \approx \Big\| \ \left(\frac{1}{24} + \frac{1}{36}\sqrt{3}\right) F' F' F'\, F\left(w(t_n)\right) +$$

$$\left(\frac{1}{12} + \frac{1}{18}\sqrt{3}\right) \left(F'_\lambda F'_\phi F' + F' F'_\lambda F'_\phi\right) F\left(w(t_n)\right) \Big\| \tau^4. \tag{5.25}$$

In the polar regions, the linear splitting operators (5.19) and (5.20) are dominated by the curvature terms, i.e., $L_\lambda \sim C_{\mathrm{cur}_\lambda}$ and $L_\phi \sim C_{\mathrm{cur}_\phi}$. The largest matrix entries of C_{cur_λ} and C_{cur_ϕ} behave as $\overline{u}^2/(a\cos\phi)$ or $\overline{u}/(a\cos\phi)$, respectively, which rapidly increases towards the poles. We here assume that $\overline{u}$ and $\overline{v}$ behave similarly. Consequently, the largest entries of F'_λ and F'_ϕ behave as $\overline{u}/(a\cos\phi)$ or $\overline{u}^2/(a\cos\phi)$ in the polar regions. Given (5.24) and (5.25), we then find

$$E_{\mathrm{loc}_{\mathrm{Strang}}}^{n+1} \sim \left(\overline{u}\left(\tau\,\overline{u}/(a\cos|\phi|)\right)^3\right) \text{ and } E_{\mathrm{loc}_{\mathrm{Ros}}}^{n+1} \sim \left(\overline{u}\left(\tau\,\overline{u}/(a\cos|\phi|)\right)^4\right). \tag{5.26}$$

Note that these estimates are based on a Taylor expansion and the ommittance of higher order terms, which is only valid, when the quotient $\tau\,\overline{u}/(a\cos\phi_{\mathrm{nP}})$ is sufficiently small.

For realistic values of τ, $\overline{u}$, and grid resolution $\Delta\phi$, the expressions in (5.26) demonstrate that the local Strang splitting error becomes much larger in the polar regions than the local Ros3 with AMF error. This is exemplified in Figure 5.1, where we assume a typical fine grid resolution, i.e., $\Delta\lambda = \Delta\phi = \pi/\mathrm{nP}$ with $\mathrm{nP} = \frac{1}{2}\,\mathrm{nL} = 180$, $\overline{u} = 10\,\mathrm{m/s}$, and a step size $\tau = 300\,\mathrm{s}$ ($\tau \ll a\cos\phi_{\mathrm{nP}}$). In this figure, the quotients from equations (5.26) are plotted over a latitudinal range $\phi \in [\pi/2 - 9\,\pi/(2 * \mathrm{nP}),\ \pi/2 - \pi/(2 * \mathrm{nP})]$, viz. the last five latitudinal grid points next to north pole. For Strang splitting, the local error increases rapidly in a band of three grid cells away from the pole. The increase of the local error of Ros3 with AMF on the other hand, is minor, and this error is significantly smaller as opposed to Strang splitting.

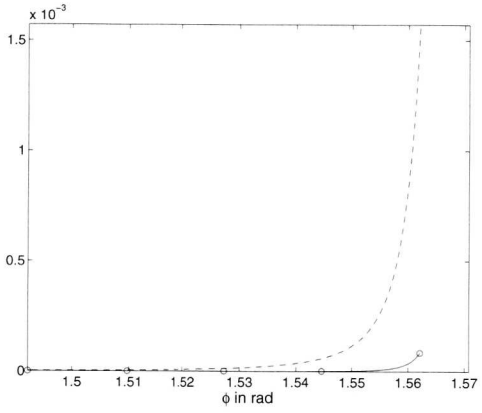


Figure 5.1: The quotient $\overline{u}\,(\tau\overline{u}/\,(a\cos\phi))^3$ for Strang splitting (*dashed*) and the quotient $\overline{u}\,(\tau\overline{u}/\,(a\cos\phi))^4$ for Ros3 with AMF (*solid*) over a latitudinal range $\phi \in [\pi/2 - 9\,\pi/(2 * \mathrm{nP}),\ \pi/2 - \pi/(2 * \mathrm{nP})]$, ($\mathrm{nP} = 180$, $\tau = 300\,\mathrm{s}$).

5.5 The dispersion relations

In meteorological practice, splitting methods are approached with a certain skepticism. It is considered unwise to split the process associated with advection waves from the Coriolis terms. Together, these processes generate so called Rossby waves, which describe an important part of atmospheric dynamics. Treating these processes separately appears to jeopardize a correct representation of these waves and, therefore, apparently obstructs a correct representation of the atmospheric tendency to geostrophic balance. To investigate this matter, we focus on the dispersion rela-

tions of the time integration methods and compare them to the dispersion relation of the original problem. This analysis will show how the time integration method affects the amplitude and propagation velocity of the waves which build up the original solution.

5.5.1 The exact dispersion relation

Since, in this section, we are primarily interested in the effects of different splittings of the advection term from the Coriolis term, it is sufficient to consider the SWEs in a local Cartesian frame of reference, (5.5)–(5.7). To derive the exact dispersion relation, we first linearize system (5.5)–(5.7) around a constant state vector $\bar{q} = (\bar{U}, \bar{V}, \bar{H})^T$, where the upper bar refers to frozen variables. We substitute $u = \bar{U} + u'$, $v = \bar{V} + v'$ and $H = \bar{H} + h'$ in the equations, which gives

$$\frac{\partial u}{\partial t} + U\frac{\partial u}{\partial x} + V\frac{\partial u}{\partial y} + g\frac{\partial h}{\partial x} - fv = 0, \tag{5.27}$$

$$\frac{\partial v}{\partial t} + U\frac{\partial v}{\partial x} + V\frac{\partial v}{\partial y} + g\frac{\partial h}{\partial y} + fu = 0, \tag{5.28}$$

$$\frac{\partial h}{\partial t} + U\frac{\partial h}{\partial x} + V\frac{\partial h}{\partial y} + H\frac{\partial u}{\partial x} + H\frac{\partial v}{\partial y} = 0, \tag{5.29}$$

where we omitted the upper bars and apostrophes for clarity. We then assume the harmonic wave solution,

$$q = (u, v, h)^T = \breve{q}(t)\, e^{i(k_1 x + k_2 y)} \text{ with } \breve{q}(t) = \hat{q}\, e^{-i\omega t}, \tag{5.30}$$

where $k = (k_1, k_2)^T \in \mathbb{R}$, $\omega \in \mathbb{C}$ and $\hat{q} =$ constant denote the wave number, the frequency and the amplitude of the wave, respectively. The frequency ω can be broken down into an imaginary part $\mathrm{Im}(\omega)$, which corresponds to damping or amplification, and a real part $\mathrm{Re}(\omega)$, which corresponds to propagation. With propagation, we associate the phase velocity c_p defined by

$$c_p = \frac{\mathrm{Re}(\omega)}{|k|}.$$

which says that any particular phase surface, i.e., a surface with a constant phase $\theta = k_1 x + k_2 y - \mathrm{Re}(\omega)\, t$, moves with normal velocity c_p in the direction of k. When the phase velocity depends on the wavenumber k, the wave is called dispersive.

Substituting the harmonic wave solution into equations (5.27)-(5.29) yields

$$\begin{pmatrix} -i\omega + Uik_1 + Vik_2 & -f & gik_1 \\ f & -i\omega + Uik_1 + Vik_2 & gik_2 \\ Hik_1 & Hik_2 & -i\omega + Uik_1 + Vik_2 \end{pmatrix} \hat{q} = 0. \tag{5.31}$$

A non-trivial harmonic wave solution of (5.27)-(5.29) exists when system (5.31) is singular. In that case, the determinant of the matrix should be zero, i.e.,

$$\det = \tilde{\omega}^3 + \tilde{\omega}\left(f^2 + gH\left(k_1^2 + k_2^2\right)\right) = 0 \tag{5.32}$$

with $\tilde{\omega} = -i\omega + Uik_1 + Vik_2$. Equation (5.32) relates the frequency ω to the wavenumber $k = (k_1, k_2)^T$. This relation is called the dispersion relation. The dispersion relation (5.32) allows three different harmonic wave solutions with frequencies,

$$\omega_{\mathrm{ex}_j}(k_1, k_2) = \begin{cases} Uk_1 + Vk_2, & \text{for} \quad j = 1, \\ Uk_1 + Vk_2 - \sqrt{f^2 + gH\left(k_1^2 + k_2^2\right)}, & \text{for} \quad j = 2, \\ Uk_1 + Vk_2 + \sqrt{f^2 + gH\left(k_1^2 + k_2^2\right)}, & \text{for} \quad j = 3, \end{cases} \tag{5.33}$$

and corresponding amplitudes,

$$\hat{q}_1 = \begin{pmatrix} -gk_2 \\ gk_1 \\ -if \end{pmatrix}, \qquad \hat{q}_{2,3} = \begin{pmatrix} igk_2 f \mp gk_1\sqrt{f^2 + gH\left(k_1^2 + k_2^2\right)} \\ -igk_1 f \mp gk_2\sqrt{f^2 + gH\left(k_1^2 + k_2^2\right)} \\ gH \end{pmatrix}. \tag{5.34}$$

The first family of waves are known as the vorticity or advection waves, which are slow waves. The second and the third family of waves are called Poincaré waves, which imply pure gravity waves when $f^2 \ll gH|k|^2$. These waves are considered to be fast. Note that none of these waves involves damping.

5.5.2 The numerical dispersion relations

In this section we derive the numerical dispersion relations of our time integration methods. The numerical dispersion relation is obtained in a similar manner as for the exact problem, i.e., by assuming a harmonic wave solution for the numerical scheme associated with the time integration method. The resulting frequencies differ from the original ones in both the imaginary and real part. The first leads to a wave with a different amplitude, which is called dissipation or accumulation. The second leads to a wave with a different propagation or phase velocity, which is called dispersion.

The Ros3 method combined with approximate matrix factorization

We first focus on the numerical dispersion relation associated with the third-order Rosenbrock method combined with approximate matrix factorization. Normally, a numerical dispersion relation is discussed in connection to a difference scheme, which is the result of a certain discretization in space and integration in time [82]. Below, we analyze the numerical dispersion relation associated with the time integration method for the continuous form of the linearized SWEs.

We write the linearized equations (5.27)-(5.29) in matrix form,

$$\frac{\partial q}{\partial t} = -\begin{pmatrix} U & 0 & g \\ 0 & U & 0 \\ H & 0 & U \end{pmatrix} \frac{\partial q}{\partial x} + \begin{pmatrix} fv \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} V & 0 & 0 \\ 0 & V & g \\ 0 & H & V \end{pmatrix} \frac{\partial q}{\partial y} - \begin{pmatrix} 0 \\ fu \\ 0 \end{pmatrix}. \quad (5.35)$$

Substitution of (5.30) into equation (5.35) yields the following ODE system for the Fourier transform $\breve{q}(t)$,

$$\begin{aligned} \frac{\mathrm{d}\breve{q}}{\mathrm{d}t} &= -i\,k_1 \begin{pmatrix} U & 0 & g \\ 0 & U & 0 \\ H & 0 & U \end{pmatrix} \breve{q} + \begin{pmatrix} 0 & f & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \breve{q} + \\ & -i\,k_2 \begin{pmatrix} V & 0 & 0 \\ 0 & V & g \\ 0 & H & V \end{pmatrix} \breve{q} + \begin{pmatrix} 0 & 0 & 0 \\ -f & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \breve{q}. \end{aligned} \quad (5.36)$$

Next, we apply Ros3 with AMF to system (5.36), where we divide the right-hand side of (5.36) into a part depending on the wavenumber k_1 and a part depending on the wavenumber k_2, i.e.,

$$\frac{\mathrm{d}\breve{q}}{\mathrm{d}t} = A(k_1)\,\breve{q} + B(k_2)\,\breve{q}. \quad (5.37)$$

This distribution corresponds to a dimensional splitting similar to (5.9). Note that with the specification of $A(k_1)$ and $B(k_2)$ the distribution of the Coriolis terms over these matrices is not yet fixed. At this point. we assume that $A(k_1)$ contains the first Coriolis matrix of equation (5.36) and $B(k_2)$ contains the second. In Section 5.5.3, other distributions will be considered.

The application of Ros3 with AMF to (5.36) yields

$$\breve{q}^{\,n+1} = R(\tau, A(k_1), B(k_2))\,\breve{q}^{\,n}, \quad (5.38)$$

where $\breve{q}^{\,n}$ denotes the approximation of the Fourier transform $\breve{q}(t)$ at time $t = t_n$ and the amplification factor $R(\tau, A(k_1), B(k_2))$ is defined by the stability function,

$$R_{\mathrm{ros}}(\tau, A, B) = I + 2\,\tau S^{-1}(A + B) + \tau S^{-1}\left(\frac{1}{2}\tau\,(A + B) - I\right) S^{-1}(A + B)$$

with the matrix $S = (I - \gamma\tau A)\,(I - \gamma\tau B)$. For a further discussion on this stability function, we refer to our earlier paper [43].

To derive the numerical dispersion relation. we then substitute the numerical harmonic wave solution,

$$\breve{q}^{\,n} = \hat{q}\,e^{-i\,\omega_{\mathrm{ros}} t_n}, \quad (5.39)$$

into (5.38) to obtain

$$M_{\text{ros}}\,\hat{q} = e^{-i\,\omega_{\text{ros}}\tau}\,\hat{q} \quad \text{with} \quad M_{\text{ros}} = R_{\text{ros}}(\tau, A(k_1), B(k_2)).$$

This gives the following numerical dispersion relation.

$$\omega_{\text{ros}_j} = \frac{\ln\left(\lambda_{M_{\text{ros}_j}}\right)}{\tau}\, i. \tag{5.40}$$

where $\lambda_{M_{\text{ros}_j}}$ denotes the j-th eigenvalue of the matrix M_{ros}. Note that these eigenvalues can be complex, allowing both dispersion and dissipation or accumulation. A thorough analysis of the frequencies given by (5.40) will show how the corresponding waves relate to the waves of the original problem, see Section 5.5.3.

The Strang splitting method

Next, we derive the numerical dispersion relation associated with the Strang splitting method. For its derivation, we adopt the same approach as above. So, we commence from system (5.37) to which we apply the Strang splitting method. In this case, the amplification factor $R(\tau, A(k_1), B(k_2))$ is defined by

$$R_{\text{str}}(\tau, A, B) = \exp\left(A\frac{\tau}{2}\right)\exp\left(B\tau\right)\exp\left(A\frac{\tau}{2}\right).$$

Postulating the harmonic wave solution (5.39) for the Fourier transform $\check{q}(t)$ and following the same reasoning as above, we arrive at the following dispersion relation,

$$\omega_{\text{str}_j} = \frac{\ln\left(\lambda_{M_{\text{str}_j}}\right)}{\tau}\, i, \tag{5.41}$$

where $\lambda_{M_{\text{str}_j}}$ denotes the j-th eigenvalue of the matrix $M_{\text{str}} = R_{\text{str}}(\tau, A(k_1), B(k_2))$.

5.5.3 An evaluation of the dispersion relations

In this section, we compare the exact and numerical dispersion relations (5.33), (5.40) and (5.41) to examine how well the numerical methods represent the characteristic waves of the original problem. The numerical method can damp or amplify these waves and change their phase velocity. Furthermore, the relations (5.40) and (5.41) can be used to investigate the effects of specific splittings of the advection from the Coriolis terms. The question is whether these splittings significantly influence the accuracy and/or stability of the resulting numerical method. We can easily redo the analysis of Section 5.5.2 to provide the correct numerical dispersion relations for a particular redistribution of the forces over the subprocesses in longitudinal and latitudinal direction.

In order to analyze the dispersion relations, we choose a typical setting of the parameters U, V, H, g and f. Since our original system (5.5)–(5.7) is based on midlatitude synoptic scale analysis, we apply synoptic scale values for these quantities, i.e., $U = V = 10\,\mathrm{m/s}$, $H = 10^4\,\mathrm{m}$, and $f = 2\Omega\sin(\pi/4)$, see [32]. The gravitational constant g is given as $g = 9.8\,\mathrm{m/s^2}$. Furthermore, we must specify the range of wave numbers in which we are interested. For convenience, we write the wave number vector $k = (k_1, k_2)^T$ in terms of its length $|k|$ and its direction β, so $k = (k_1, k_2)^T = (|k|\cos\beta, |k|\sin\beta)^T$. We focus on wave number vectors with length $|k| = 1$. These wave numbers include the family of advective waves with velocity $U\cos\beta + V\sin\beta$ and the two families of gravity waves with velocities $U\cos\beta + V\sin\beta \pm \sqrt{gH}$, where we used $f \ll \sqrt{gH}$. Finally, it is important to notice that we are calculating in a *local* Cartesian frame of reference. Observe that the distance Δx in the local frame of reference corresponds to a radial change of the longitude, $\Delta\lambda$. The corresponding distance on the sphere then reads $a\cos\phi\,\Delta\lambda$. Therefore, the stepsizes for mid-latitudinal motion in the local and global frame of reference, τ_{local} and τ_{global}, are related as

$$\tau_{\mathrm{global}} = a\cos(\pi/4)\;\tau_{\mathrm{local}} \approx 4.5 \cdot 10^6\,\tau_{\mathrm{local}}.$$

We elaborate the numerical dispersion relations for increasing step sizes. The minimum and maximum value of the imaginary parts of the corresponding frequencies are given in Table 5.1. The minimum and maximum values are calculated over $\beta \in [0, 2\pi)$. Observe that the corresponding frequencies of the original waves have no imaginary part. The positive imaginary parts of the frequencies in Table 5.1 then illustrate that a Strang splitting method tends to amplify both advection and gravity waves.

In case of Ros3 with AMF no such behavior is found. Each wave is either damped by the numerical method or propagates with a constant amplitude. Note that this behavior characterizes the A-stability property. Furthermore, the results indicate that Ros3 with AMF damps the various waves more rigorously than Strang splitting. For all step sizes considered, the minimum values of the imaginary parts are smaller for Ros3 with AMF than for Strang splitting. In addition, for Ros3 with AMF, the fast gravity waves are more strongly damped than the slow advective wave. The damping of Strang splitting does not distinguish between slow and fast waves.

Finally, we focus on the imaginary parts of the frequencies for a common step size τ_{local}. Assume $\Delta x = 2\pi/360$, which corresponds to a fine uniform lat-lon grid with $\Delta\lambda = \Delta\phi = 2\pi/360$. For the given synoptic values, we can then derive a maximal step size τ_{local} prescribed by the CFL-restriction, when solving the SWEs by means of a third-order Runge-Kutta method, see [43],

$$\tau_{\mathrm{local}} = \Delta x/(U + \sqrt{gH}) = 5.4 \cdot 10^{-5}\,\mathrm{s}.$$

For this step size, both methods behave excellently. In particular, their influence on the important advective wave is negligible.

τ_{local}	Strang		Ros3 with AMF	
	min	max	min	max
10^{-5}	$-0.78 \cdot 10^{-10}$	$0.67 \cdot 10^{-10}$	$-0.18 \cdot 10^{-8}$	0
	$-0.89 \cdot 10^{-10}$	$0.11 \cdot 10^{-9}$	$-0.98 \cdot 10^{-6}$	$-0.39 \cdot 10^{-6}$
	$-0.12 \cdot 10^{-9}$	$0.13 \cdot 10^{-9}$	$-0.98 \cdot 10^{-6}$	$-0.39 \cdot 10^{-6}$
10^{-4}	$-0.21 \cdot 10^{-8}$	$0.21 \cdot 10^{-8}$	$-0.18 \cdot 10^{-5}$	0
	$-0.11 \cdot 10^{-8}$	$0.11 \cdot 10^{-8}$	$-0.98 \cdot 10^{-3}$	$-0.39 \cdot 10^{-3}$
	$-0.11 \cdot 10^{-8}$	$0.11 \cdot 10^{-8}$	$-0.98 \cdot 10^{-3}$	$-0.39 \cdot 10^{-3}$
10^{-3}	$-0.21 \cdot 10^{-6}$	$0.21 \cdot 10^{-6}$	$-0.17 \cdot 10^{-2}$	$-0.90 \cdot 10^{-6}$
	$-0.11 \cdot 10^{-6}$	$0.11 \cdot 10^{-6}$	$-0.86 \cdot 10^{0}$	$-0.37 \cdot 10^{0}$
	$-0.11 \cdot 10^{-6}$	$0.11 \cdot 10^{-6}$	$-0.86 \cdot 10^{0}$	$-0.37 \cdot 10^{0}$
10^{-2}	$-0.26 \cdot 10^{-4}$	$0.26 \cdot 10^{-4}$	$-0.60 \cdot 10^{-1}$	$-0.12 \cdot 10^{-1}$
	$-0.13 \cdot 10^{-4}$	$0.13 \cdot 10^{-4}$	$-0.24 \cdot 10^{2}$	$-0.20 \cdot 10^{2}$
	$-0.13 \cdot 10^{-4}$	$0.13 \cdot 10^{-4}$	$-0.24 \cdot 10^{2}$	$-0.21 \cdot 10^{2}$

Table 5.1: Minimum and maximum values of the imaginary part of the frequencies for Strang splitting and Ros3 with AMF. The results are presented for splitting (5.42). The maxima are calculated for wave numbers $\mathbf{k} = (\sin(\beta), \cos(\beta))$, with $\beta \in [0, 2\pi)$. For each step size τ_{local}, the extrema associated with the numerical advection ($j = 1$) and the numerical gravity waves ($j = 2, 3$) are listed.

The relative errors in the phase velocities are displayed in Table 5.2. The relative error is defined as follows.

$$E_{c_p} = \frac{\text{Re}\,(\omega_{\text{num}}) - \text{Re}\,(\omega_{\text{exact}})}{\text{Re}\,(\omega_{\text{exact}})}.$$

Table 5.2 illustrates that the Strang splitting method does not affect the phase velocity of the advection wave. The gravity waves are changed by this method. Ros3 with AMF on the other hand, affects both phase velocities, although its effect on the gravity waves is minor compared to Strang splitting for $\tau_{\text{local}} \leq 10^{-3}$ s. In meteorological practice, however, numerical methods are assessed by their capability to represent the advective wave. At large step sizes, $\tau_{\text{local}} = 10^{-3}$ and $\tau_{\text{local}} = 10^{-2}$, Ros3 with AMF poorly represents the advective wave phase velocities as opposed to Strang splitting. For common step sizes though, $\tau_{\text{local}} = 10^{-4}$ and $\tau_{\text{local}} = 10^{-5}$, Ros 3 with AMF has almost no effect on the advective wave phase velocity.

The effect of a specific splitting of the Coriolis and advection term on the stability properties is studied by a comparison of the imaginary parts of the frequencies for

τ_{local}	$\max\left(E_{c_p}\right)$	
	Strang	Ros3 with AMF
10^{-5}	0	$0.58 \cdot 10^{-13}$
	$0.11 \cdot 10^{-6}$	$0.11 \cdot 10^{-10}$
	$0.11 \cdot 10^{-6}$	$0.11 \cdot 10^{-10}$
10^{-4}	0	$3.09 \cdot 10^{-10}$
	$0.11 \cdot 10^{-4}$	$0.11 \cdot 10^{-6}$
	$0.11 \cdot 10^{-4}$	$0.11 \cdot 10^{-6}$
10^{-3}	0	$0.26 \cdot 10^{-4}$
	$0.11 \cdot 10^{-2}$	$0.10 \cdot 10^{-3}$
	$0.11 \cdot 10^{-2}$	$0.10 \cdot 10^{-3}$
10^{-2}	0	0.35
	0.13	0.33
	0.13	0.33

Table 5.2: Maximum value of the relative errors in the phase velocities for Strang splitting and Ros3 with AMF. The maxima are calculated for wave numbers $k = (\sin(\beta), \cos(\beta))$, with $\beta \in [0, 2\pi)$. For each step size τ_{local}, the extrema associated with the numerical advection ($j = 1$) and the numerical gravity waves ($j = 2, 3$) are listed.

three different splittings. These are

$$A(k_1) = -ik_1 \begin{pmatrix} U & i\frac{f}{k_1} & g \\ 0 & U & 0 \\ H & 0 & U \end{pmatrix}, \qquad B(k_2) = -ik_2 \begin{pmatrix} V & 0 & 0 \\ -i\frac{f}{k_2} & V & g \\ 0 & H & V \end{pmatrix}, \tag{5.42}$$

$$A(k_1) = -ik_1 \begin{pmatrix} U & i\frac{f}{k_1} & g \\ -i\frac{f}{k_1} & U & 0 \\ H & 0 & U \end{pmatrix}, \quad B(k_2) = -ik_2 \begin{pmatrix} V & 0 & 0 \\ 0 & V & g \\ 0 & H & V \end{pmatrix}, \tag{5.43}$$

$$A(k_1) = -ik_1 \begin{pmatrix} U & 0 & g \\ 0 & U & 0 \\ H & 0 & U \end{pmatrix}, \qquad B(k_2) = -ik_2 \begin{pmatrix} V & i\frac{f}{k_2} & 0 \\ -i\frac{f}{k_2} & V & g \\ 0 & H & V \end{pmatrix}. \tag{5.44}$$

The first splitting is already examined above. The second and third splitting involve no directional separation of the Coriolis terms. In Table 5.1, the minimum and maximum value of the imaginary parts of the numerical frequencies are given for the first splitting for Strang splitting and Ros3 with AMF, respectively. For Strang

splitting, the minimal and maximal values for the second and third splitting (5.43)–(5.44) are close to zero, so only very small machine representation errors were visible. The difference of these values as opposed to the values for splitting (5.42) indicate that amplification or damping by Strang splitting indeed depends on the specific splitting. The method behaves significantly better in case of the second and third splitting, because they involve almost no damping or amplification. Ros3 with AMF on the other hand, is almost indifferent to the details of the splitting. The entries of the minimum and maximum value for the second and third splitting were almost identical to the entries of splitting (5.42).

5.6 Numerical experiments

In this section, we continue our comparison between Ros3 with AMF and Strang splitting by an assessment of these methods when applied to test cases of the well-established SWEs test set [88]. In addition, the numerical experiments serve to verify the theoretical results found in Section 5.4 and 5.5.

Both methods are used to integrate the system of ODEs resulting from spatially discretizing the full non-linear system of SWEs on the rotating sphere (5.1) (5.4). Calculations are done on a uniform lat-lon grid. As spatial discretization scheme, we apply a finite volume method, viz. an Osher scheme combined with the ($\kappa = \frac{1}{3}$)-scheme for the constant state interpolation, which proved to be well suited for solving the SWEs in spherical geometry, see [42]. Since the resulting ODE system is too difficult to be solved exactly, we have to specify the integration methods which are used to solve the substeps in the Strang splitting method. In our earlier paper, Ros3 with AMF proved far more efficient than the RK3 explicit method. Consequently, Strang splitting can only be cost effective when it is combined with an implicit time integration method. We therefore apply the Ros3 method (5.8). In addition, this method is third-order accurate, which ensures that the splitting error dominates the total error, and it is A-stable.

We concentrate on three different test cases from the well-known SWEs test set [88], viz. Test 2, global steady-state non-linear zonal geostrophic flow, Test 5, zonal flow over an isolated mountain, and Test 6, a Rossby-Haurwitz wave. All three test cases were discussed in earlier work [43]. Test 2 is used to provide an order estimate for the Strang splitting method similar to the one for Ros3 with AMF found in [43]. Test 5 and 6 are chosen, because they describe 'realistic' instationary flow patterns, and are therefore suitable to truly assess our time integration methods. In addition, they form an excellent framework to investigate the influence of the integration methods on various wave-like solutions. Test 5 involves high-amplitude gravity waves. Test 6 describes a slow Rossby-Haurwitz wave, whose flow pattern is very common in practical applications. A correct representation of this last wave is therefore of great importance.

The presentation of the numerical experiments is divided in two parts.

- First, we investigate the accuracy and efficiency of both methods for a specific splitting. As reference splitting, we use the splitting suggested in our previous paper. The results are used to identify the extent of the errors and their location on the sphere. The calculations are performed on a high resolution grid, viz. a uniform lat-lon grid with 180 grid points in latitudinal direction (nP = 180) and 360 grid points in longitudinal direction (nL = 360). The step sizes of each method are determined by trial and error. For Strang splitting, they will be the maximal step sizes at which stability is obtained and accuracy is still acceptable. For Ros3 with AMF, the step sizes are chosen such that its results are equally accurate as these of Strang splitting.
- Second, we investigate the effects of various splittings on the accuracy and efficiency of both methods. Calculations are done on a uniform lat-lon grid of 90×180 grid points in case of Test 5 and on a uniform lat-lon grid of 144×288 grid points in case of Test 6. These grids are coarser than the previous to confine the error in the polar regions, see Section 5.4. The step size is fixed for all splittings.

In both parts, the accuracy is expressed by the l_2- or l_∞-norm of the relative error of the depth of the fluid layer and the absolute errors of the velocity components in longitudinal- and latitudinal direction. In spherical geometry the discrete l_∞-norm and l_2-norms are defined as follows,

$$l_\infty(H) = \max_{i,j} \left| \frac{H_{i,j} - H(\lambda_i, \phi_j)}{H(\lambda_i, \phi_j)} \right|. \tag{5.45}$$

$$l_\infty(u) = \max_{i,j} | u_{i,j} - u(\lambda_i, \phi_j) |. \tag{5.46}$$

and

$$l_2(H) = \sqrt{\sum_{i,j} \left(H_{i,j} - H(\lambda_i, \phi_j)\right)^2 \cos\phi_j} \Bigg/ \sqrt{\sum_{i,j} \left(H(\lambda_i, \phi_j)\right)^2 \cos\phi_j}, \tag{5.47}$$

$$l_2(u) = \frac{\sqrt{\pi}}{\mathrm{nL}} \sqrt{\sum_{i,j} \left(u_{i,j} - u(\lambda_i, \phi_j)\right)^2 \cos\phi_j}, \tag{5.48}$$

where $H_{i,j}$ etc. denote the approximated solution H etc. at gridpoint (λ_i, ϕ_j) and $H(\lambda_i, \phi_j)$ etc. denote the reference solution H etc. at gridpoint (λ_i, ϕ_j), which is exact in case of Test 2 and given by a high resolution spectral method in case of Test 5 and Test 6. Note that $l_2(H)$ and $l_2(u)$ are the high-resolution finite volume equivalents of the continuous l_2-norm defined by Williamson *et al* in [88].

5.6.1 The three test cases from the SWEs test set

First, we summarize the three considered test cases from the SWE test set, viz. Test 2, Test 5, and Test 6. Test 2 represents a solid body rotation of which the

height field and the velocity components in longitudinal and latitudinal direction are defined as follows

$$H = h_0 - \left(\frac{a\Omega u_0}{g} + \frac{u_0^2}{2g}\right)\left(-\cos\lambda\cos\phi\sin\alpha + \sin\phi\cos\alpha\right)^2, \quad (5.49)$$

$$u = u_0\left(\cos\phi\cos\alpha + \sin\phi\cos\lambda\sin\alpha\right), \quad (5.50)$$

$$v = -u_0\sin\lambda\sin\alpha, \quad (5.51)$$

with h_0 and u_0 given, $u_0 = 38.6$ m/s and $gh_0 = 2.94 \cdot 10^4$ m^2/s^2. α denotes the angle between the axis of the solid body rotation and the polar axis of the spherical coordinate system. We consider flow over the poles, i.e., $\alpha = \pi/2$. Test 2 extends over a 5-days interval.

Test 5 represents a zonal flow which impinges on a mountain. The mountain height is prescribed by a cone.

$$h_s = h_{s_0}\left(1 - \frac{r}{R}\right), \quad (5.52)$$

where $h_{s_0} = 2000$ m, $R = \pi/9$, $r^2 = \min[R^2, (\lambda - \lambda_c)^2 + (\phi - \phi_c)^2]$, $\lambda_c = 3\pi/2$, and $\phi_c = \pi/6$. The initial zonal flow is given by a solid body rotation parallel to the equator. The initial height and velocity components result from equation (5.49)–(5.51) with $\alpha = 0$, $u_0 = 20$ m/s, and $h_0 = 5960$ m. The reference solution is determined by a high resolution spectral method. The simulated time period is 15 days.

Test 6 consists of a Rossby-Haurwitz wave with a simulation period of 14 days. The initial condition is provided in [88]. Meteorologists consider this test as standard, since similar flow patterns occur in practical applications. A reference solution over a fourteen-day interval is provided by a high resolution spectral circulation model.

5.6.2 Experiments with the reference splitting

The reference splitting

In this section, we specify the reference splitting for which we assess Strang splitting and Ros3 with AMF on a high-resolution grid. Similar to Section 5.4, this splitting is defined by

$$\frac{\partial q}{\partial t} = f_\lambda(q) + f_\phi(q) \quad (5.53)$$

with

$$f_\lambda(q) = \frac{-1}{a\cos\phi}\frac{\partial}{\partial\lambda}\begin{pmatrix} Hu^2 + \frac{1}{2}gH^2 \\ Huv \\ Hu \end{pmatrix} + \begin{pmatrix} -\frac{gH}{a\cos\phi}\frac{\partial h_s}{\partial\lambda} + fHv \\ -\frac{Hu^2}{a}\tan\phi \\ 0 \end{pmatrix}. \tag{5.54}$$

$$f_\phi(q) = \frac{-1}{a\cos\phi}\frac{\partial}{\partial\phi}\begin{pmatrix} Huv\cos\phi \\ \left(Hv^2 + \frac{1}{2}gH^2\right)\cos\phi \\ Hv\cos\phi \end{pmatrix} - \begin{pmatrix} -\frac{Huv}{a}\tan\phi \\ \frac{gH}{a}\frac{\partial h_s}{\partial\phi} + fHu + \frac{\tan\phi}{2a}gH^2 \\ 0 \end{pmatrix} \tag{5.55}$$

The curvature terms are distributed over f_λ and f_ϕ respecting their association with a change of orientation of the corresponding unit vector. This distribution is natural. The Coriolis forces are assigned according to the direction of the momentum equations from which they originate. With a minor difference in the distribution of the curvature terms, this splitting was successfully applied in [43] for Ros3 with AMF.

An order estimate for Strang splitting

First, we illustrate the order behavior of the Strang splitting method. Similar to the order estimate for Ros3 with AMF given in [43], calculations are done on a uniform lat-lon grid with resolution nL = 288 and nP = 144 for varying step sizes. We concentrate on Test 2. As order estimate, we use the l_∞-norm of the absolute error of H and u, defined as

$$\text{abs(var)}_\tau = \max_{i,j}\left|\text{var}^\tau_{i,j,t} - \text{var}^{\tau_{\text{ref}}}_{i,j,t}\right| \text{ with } \tau_{\text{ref}} = 80\,\text{s},$$

where $\text{var}^\tau_{i,j,t}$ yields the approximate value of a variable var in gridpoint (λ_i, ϕ_j) at time t calculated with step size τ. Figure 5.2 pictures these norms against the step size τ in a log-log plot. Note that we march to the steady-state of the semi-discrete problem. The figure illustrates that the order of the Strang splitting method is slightly higher than two in this case. By theory, second-order consistency is expected as is visualized by the slope of the solid line, which is two.

Results on Test 5 and Test 6

In this section, Ros3 with AMF and Strang splitting are applied to Test 5 and Test 6 of the SWEs test set. Our interest is in their accuracy and efficiency when used on a high resolution grid.

Calculations are done on a uniform lat-lon grid with nL = 360 and nP = 180. The step size is found by trial and error depending on the test case and the integration method. For Strang splitting we apply the following step sizes, $\tau = 216\,\text{s}$ in case of Test 5, and $\tau = 450\,\text{s}$ in case of Test 6. These step sizes are chosen such that the results are sufficiently accurate and the computation is stable. For Ros3 with AMF,

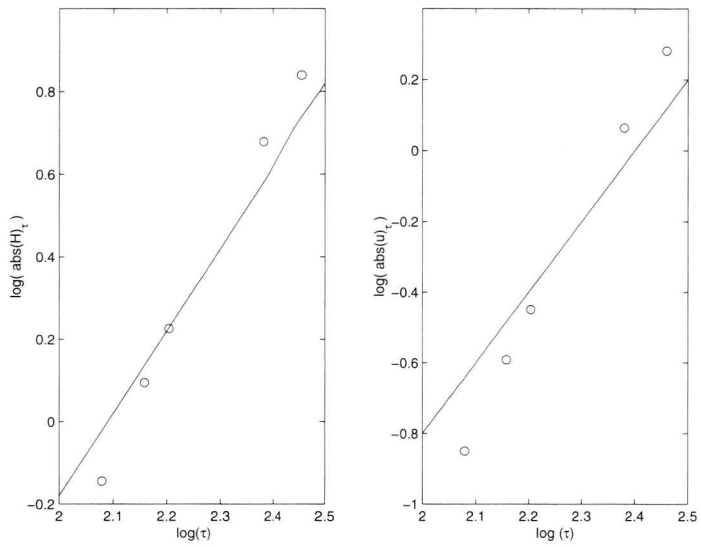


Figure 5.2: An order estimate for H and u: $\log(\mathrm{abs}(H)_\tau)$ and $\log(\mathrm{abs}(u)_\tau)$ versus $\log(\tau)$ for the Strang splitting method for Test 2 (markers). The solid lines illustrate formal second-order accuracy.

the step sizes are chosen such that its results are equally accurate as these of Strang splitting. This yields $\tau = 900\,\mathrm{s}$ in case of Test 5, and $\tau = 1200\,\mathrm{s}$ in case of Test 6. Consequently, Ros3 with AMF is far more efficient than Strang splitting. Strang splitting involves three linear system solves and three flux evaluations per time step, where we accounted the flux evaluations of F_λ and F_ϕ as two flux evaluations. Ros3 with AMF involves four linear system solves and four flux evaluations per time step. Therefore, if the step size for Ros3 with AMF is more than $4/3$ times as large as the step size of Strang splitting its workload is lower. For $\tau = 900\,\mathrm{s}$ in case of Test 5 and $\tau = 1200\,\mathrm{s}$ in case of Test 6, these ratios are 4.17 and 2.67, respectively. Finally, we comment that for Ros3 with AMF, results can be obtained for much larger step sizes. In contrast to Strang splitting, Ros3 with AMF does not suffer from a severe step size restriction. Note that, eventually, the step size for Ros3 with AMF is limited by accuracy. For very large step sizes, viz. several hours, Ros3 with AMF involves too much damping to correctly represent the solution, see Table 5.1 with $\tau_{\mathrm{local}} \geq 10^{-3}\,\mathrm{s}$.

Figure 5.3 represents the errors (5.45)–(5.48) for Test 5. The errors are sufficiently small, although the sudden increase of the $l_\infty(u)$ and $l_\infty(v)$ for Strang splitting is remarkable. This increase is caused by an interaction of the propagated spatial error, initially caused at the foot of the mountain, and the mountain itself. The spatial error is rotated over the sphere in approximately 10 days before it

again impinges on the mountain. As a result, in case of Strang splitting, a sudden increase of the local error is observed. For Ros3 with AMF, this increase is not that apparent. The spatial errors involve high-frequency waves, which are strongly damped by Ros3 with AMF. Observe that this explanation agrees with the results from Section 5.5. The considered step sizes $\tau = 216\,\mathrm{s}$ and $\tau = 900\,\mathrm{s}$ in case of Strang splitting and Ros3 with AMF, respectively, correspond to $\tau_{\mathrm{local}} = 4.8 \cdot 10^{-5}\,\mathrm{s}$ and $\tau_{\mathrm{local}} = 2.0 \cdot 10^{-4}\,\mathrm{s}$, respectively. The fact that we do not observe a sudden increase of the $l_2(u)$ and $l_2(v)$-norm for Strang splitting, shows that the error increase is local in space.

Figure 5.4 represents the errors (5.45)–(5.48) for Strang splitting and Ros3 with AMF in case of Test 6. Again, similar accuracy is obtained, but for different step sizes in favor of Ros3 with AMF. The step size applied for Ros3 with AMF is again larger than 4/3 times the stepsize applied for Strang splitting. The Rossby-Haurwitz wave represents a low frequency wave and is therefore of particular interest to meteorologists. According to Section 5.5, both methods do not significantly affect the advective wave phase velocity.

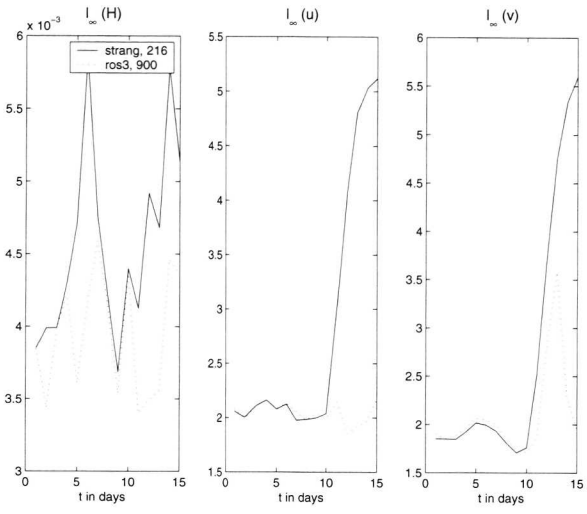
Finally, Figure 5.5 visualizes the relative error of H on the northern hemisphere projected onto the equatorial plane. This picture clearly demonstrates that the Strang splitting error is large in the polar region as opposed to Ros3 with AMF. This result confirms the results of Section 5.4, where we found that on current high resolution grids Strang splitting suffers more strongly from the pole singularity in the spherical SWEs, observable by large local errors in the polar region.

5.6.3 Experiments with several other splittings

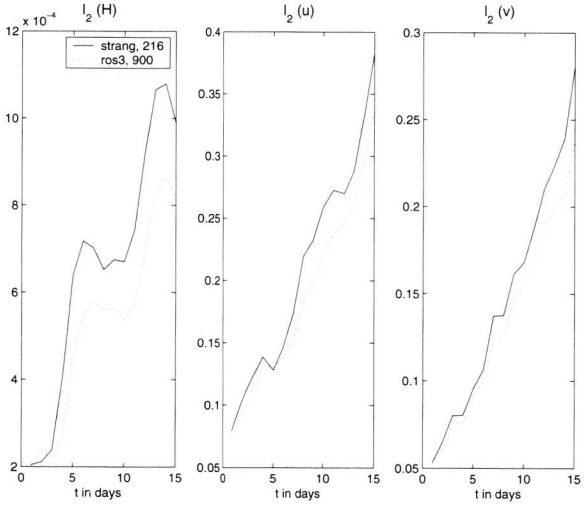
In this section, we consider several splittings of the SWEs in spherical geometry. The splittings differ in their distribution of the Coriolis forces over the flux functions f_λ and f_ϕ. Since the advection- and curvature terms are strongly connected to a specific direction, their distribution is fixed. So, we have

$$f_\lambda(q) = \frac{-1}{a\cos\phi}\frac{\partial}{\partial\lambda}\begin{pmatrix} Hu^2 + \frac{1}{2}gH^2 \\ Huv \\ Hu \end{pmatrix} - \begin{pmatrix} \frac{gH}{a\cos\phi}\frac{\partial h_s}{\partial\lambda} \\ \frac{Hu^2}{a}\tan\phi \\ 0 \end{pmatrix} + f_{\lambda_{\mathrm{Cor}}}(q). \quad (5.56)$$

$$f_\phi(q) = \frac{-1}{a\cos\phi}\frac{\partial}{\partial\phi}\begin{pmatrix} Huv\cos\phi \\ (Hv^2 + \frac{1}{2}gH^2)\cos\phi \\ Hv\cos\phi \end{pmatrix} + \\ - \begin{pmatrix} -\frac{Huv}{a}\tan\phi \\ \frac{gH}{a}\frac{\partial h_s}{\partial\phi} + \frac{\tan\phi}{2a}gH^2 \\ 0 \end{pmatrix} + f_{\phi_{\mathrm{Cor}}}(q). \quad (5.57)$$

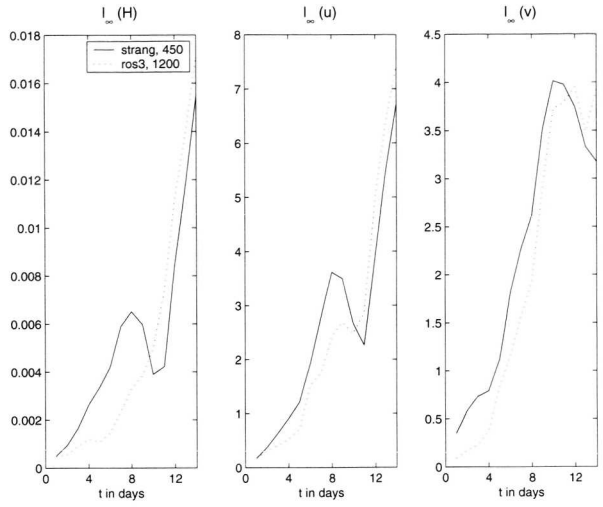


(a) l_∞-norm

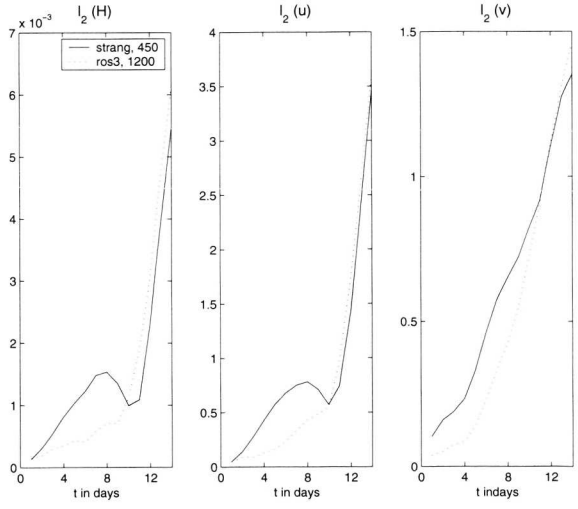


(b) l_2-norm

Figure 5.3: The l_∞-norm (fig(a)) and l_2-norm (fig(b)) of the relative error in H (first column), and absolute errors in u and v (second and third column) for Test 5 for Strang splitting (solid) and Ros3 with AMF (dotted) in case of the reference splitting.

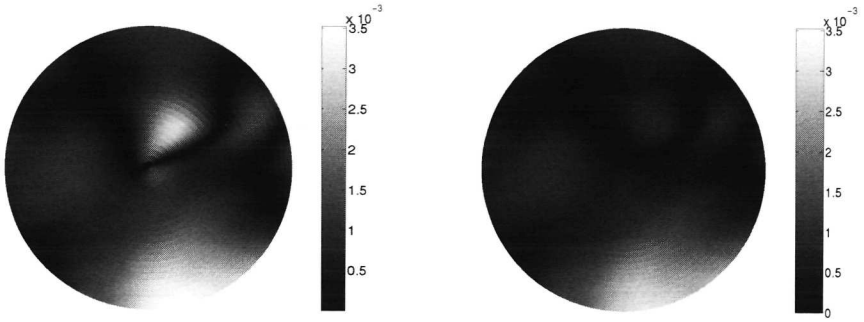


(a) l_∞-norm



(b) l_2-norm

Figure 5.4: The l_∞-norm (fig(a)) and l_2-norm (fig(b)) of the relative error in H (first column), and absolute errors in u and v (second and third column) for Test 6 for Strang splitting (solid) and Ros3 with AMF (dotted) in case of the reference splitting.



(a) Strang splitting ($\tau = 216\,\mathrm{s}$) (b) Ros3 with AMF ($\tau = 900\,\mathrm{s}$)

Figure 5.5: Polar view of the relative error in H for Test 5 for Strang splitting (fig(a)) and Ros3 with AMF (fig(b)), in case of the reference splitting. $\phi \in [\frac{7}{8}\pi, \frac{1}{2}\pi]$

where $f_{\lambda_{\mathrm{Cor}}}(q) + f_{\phi_{\mathrm{Cor}}}(q) = (fHv, -fHu, 0)^T$. The different splittings are

$$f_{\lambda_{\mathrm{Cor}}}(q) = (fHv, -fHu, 0)^T, \quad f_{\phi_{\mathrm{Cor}}}(q) = (0,0,0)^T, \tag{f12f}$$

$$f_{\lambda_{\mathrm{Cor}}}(q) = (0,0,0)^T, \quad f_{\phi_{\mathrm{Cor}}}(q) = (fHv, -fHu, 0)^T, \tag{ff12}$$

$$f_{\lambda_{\mathrm{Cor}}}(q) = (fHv, 0, 0)^T, \quad f_{\phi_{\mathrm{Cor}}}(q) = (0, -fHu, 0)^T, \tag{f1f2}$$

$$f_{\lambda_{\mathrm{Cor}}}(q) = (0, -fHu, 0)^T, \quad f_{\phi_{\mathrm{Cor}}}(q) = (fHv, 0, 0)^T, \tag{f2f1}$$

$$f_{\lambda_{\mathrm{Cor}}}(q) = \tfrac{1}{2}(fHv, -fHu, 0)^T, \quad f_{\phi_{\mathrm{Cor}}}(q) = \tfrac{1}{2}(fHv, -fHu, 0)^T, \tag{fhalf}$$

where the first two splittings involve the complete assignment of the Coriolis forces to one direction. The third splitting is the reference splitting investigated in Section 5.6.2. Splitting four and five, (f2f1) and (fhalf), are artificial. Note that the first three splittings were considered before in Section 5.5.3 for the linearized local Cartesian SWEs.

We focus on Test 5 and Test 6 of the SWEs test set. Calculations are done on a 90×180 uniform lat-lon grid over a fifteen days time period for Test 5, and on 144×288 uniform lat-lon grid over a fourteen days time period for Test 6, respectively. For Test 5, the Strang splitting method uses a fixed step size $\tau = 900\,\mathrm{s}$ for all splittings, Ros3 with AMF uses a step size $\tau = 1800\,\mathrm{s}$. The results of Test 6 are computed with a step size $\tau = 150\,\mathrm{s}$ for Strang splitting and $\tau = 450\,\mathrm{s}$ for Ros3 with AMF. The step sizes are chosen such that the results satisfy a given accuracy requirement for the reference splitting.

Since we are mainly interested in the qualitative difference between the results

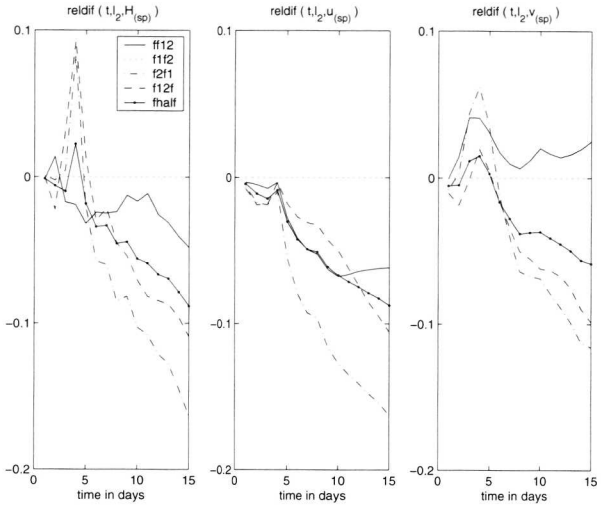
for the various splittings and in the impact of these splittings on the two integration methods, we introduce the following monitor.

$$\text{reldif}\left(t, E, H_{(\text{sp})}(t)\right) = \frac{E\left(H_{(\text{sp})}(t)\right) - E\left(H_{(\text{refsp})}(t)\right)}{E\left(H_{(\text{refsp})}(t)\right)}.$$

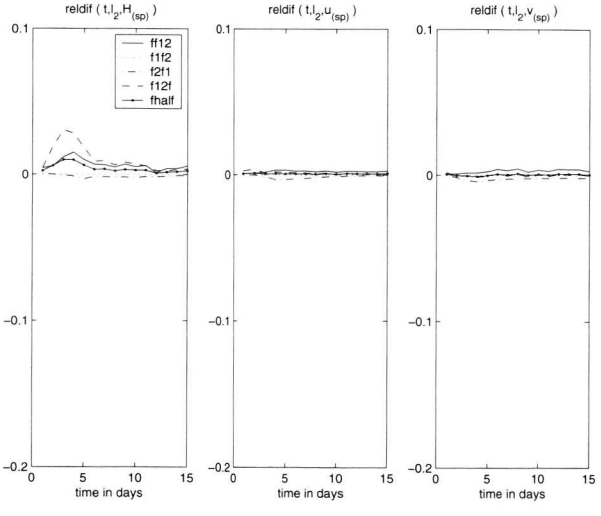
where $E(H)$ denotes the l_∞- or l_2-norm defined in (5.45) and (5.47), t denotes the time at which the solution $H_{(\text{sp})}$ is approximated and (refsp) denotes the reference splitting (f1f2). Similar expressions can be derived for the longitudinal and latitudinal velocity components.

Figure 5.6–5.7 represent the relative differences, $\text{reldif}(t, l_2, H)$, $\text{reldif}(t, l_2, u)$, and $\text{reldif}(t, l_2, v)$ for the several splittings when applied to Test 5 and Test 6. These figures demonstrate that it is difficult to identify a best splitting, because such a splitting depends on the specific test case. For instance, for Strang splitting, the reference splitting (f1f2) is not a good choice in case of Test 5. After 15-days, the l_2-norms of the relative error in H, and absolute errors in u and v are smaller for almost all other splittings, viz. $\text{reldif}(t, l_2, H) < 0$ etc., see Figure 5.6(a). Splitting (f2f1) appears better suited. For Test 6 on the other hand, the reference splitting is less accurate over the first seven days, but performs better than the other splittings on the seven days remaining, see Figure 5.7(a).

Compared to Ros3 with AMF, Strang splitting is more sensitive to the chosen splitting. For this method, the relative differences vary over a range of $[-0.16, 0.10]$ in case of Test 5 and over a range of $[-0.18, 0.11]$ in case of Test 6, see Figure 5.6(a) and 5.7(a). For Ros3 with AMF on the other hand, these differences vary over a range of $[-0.003, 0.03]$ in case of Test 5 and over a range of $\left[-2.5 \cdot 10^{-3}, 3.6 \cdot 10^{-3}\right]$ in case of Test 6, see Figure 5.6(b) and Figure 5.7(b), respectively. Ros3 with AMF is almost indifferent to the applied splitting, which agrees with our results in Section 5.5.3. For Ros3 with AMF, the reference splitting is sufficiently accurate for both test cases, although splitting (f2f1) is slightly better.

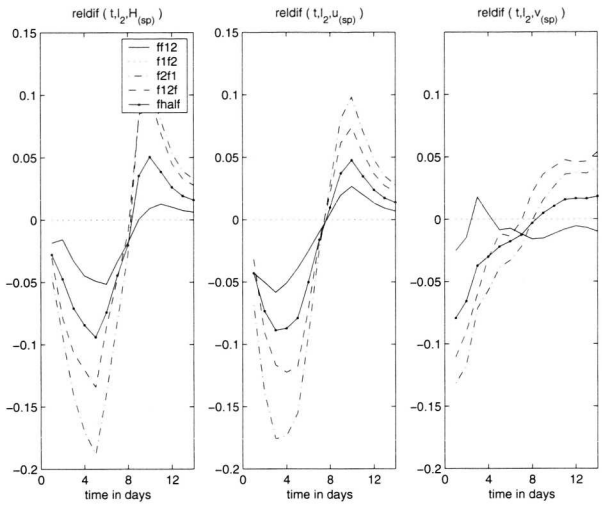


(a) Strang splitting ($\tau = 900\,\mathrm{s}$)

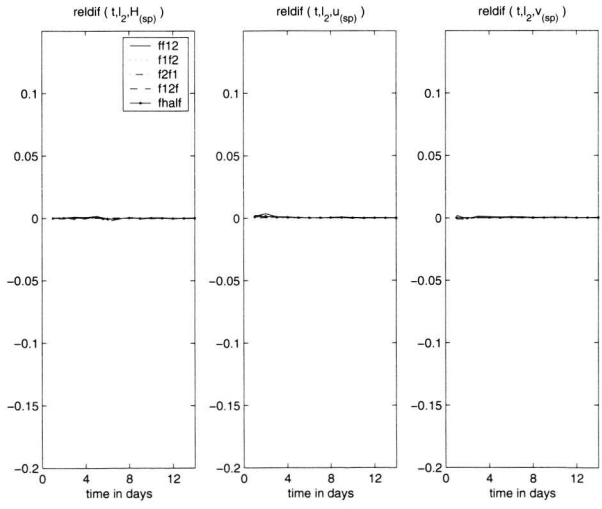


(b) Ros3 with AMF ($\tau = 1800\,\mathrm{s}$)

Figure 5.6: The relative differences, reldif(t, l_2, H), reldif(t, l_2, u) and reldif(t, l_2, v), in case of Test 5 for the splittings (ff12), (f1f2) etc. Splitting (f1f2) is used as the reference splitting. Results are presented for Strang splitting (fig(a)) and Ros3 with AMF (fig(b)).



(a) Strang splitting ($\tau = 150$ s)



(b) Ros3 with AMF ($\tau = 450$ s)

Figure 5.7: The relative differences, reldif(t, l_2, H), reldif(t, l_2, u) and reldif(t, l_2, v), in case of Test 6 for the splittings (ff12), (f1f2) etc. Splitting (f1f2) is used as the reference splitting. Results are presented for Strang splitting (fig(a)) and Ros3 with AMF (fig(b)).

5.7 Conclusion

When solving the semi-discrete SWEs on a global uniform lat-lon grid, an explicit time integration method suffers from a severe restriction on the step size (the pole problem). This problem can be avoided by the application of an implicit time integration method. In [43], we therefore investigated an A-stable linearly implicit third-order time integration method, which we combined with approximate matrix factorization to make it cost effective, viz., Ros3 with AMF.

In this article, we further explored this method and compared it to a Strang-type splitting method. First, we focused on the local error of both methods for the linearized SWEs in spherical geometry. Strang splitting is showed to suffer from a large local error in the polar region as opposed to Ros3 with AMF. Second, we investigated the numerical dispersion relations for the local Cartesian SWEs to analyze their influence on the characteristic waves of the shallow water problem. Our main focus was on the advective wave, which is most important in meteorological applications. For characteristic step sizes, both methods did not significantly affect the advective wave phase velocities. Their influence on the gravity waves, however, was very different. Ros3 with AMF damped these waves more rigorously than Strang splitting, but better represented their phase velocities. In addition, Strang splitting could lead to amplification of these waves, which makes it unsuitable for long time integration periods. Third, we applied both methods to Test 2, Test 5 and Test 6 of the SWEs test set. The numerical results agreed with the theoretical results for the local error and the numerical dispersion relations. Furthermore, they showed that Ros3 with AMF is unaffected by the chosen splitting and, most important, Ros3 with AMF is far more efficient than Strang splitting.

In conclusion, Ros3 with AMF makes a good candidate to efficiently solve the semi-discrete SWEs on a global fine resolution lat-lon grid. Strang splitting on the other hand, is not advocated in view of its inefficiency and large local error in the polar regions.

Chapter 6

Appendix

6.1 The construction of the stereographic projection for the northern hemisphere

In this appendix we construct the transformation relations between the stereographic $(x_{\mathrm{st}}, y_{\mathrm{st}})$ and spherical coordinates (λ, ϕ). Consider a point $\underline{r} = (\lambda, \phi, a)$ on the northern hemisphere, and define the half-plane S_λ as the plane λ=constant and the stereographic plane as the plane located at and locally tangent to the sphere at the pole. Then, project the point $\underline{r} = (\lambda, \phi, a)$ from the south pole onto the intersection of the plane S_λ and the stereographic plane, see figure 6.1. This projection point is denoted as $\underline{r}_{\mathrm{st}} = (x_{\mathrm{st}}, y_{\mathrm{st}})$, see Figure 6.1.

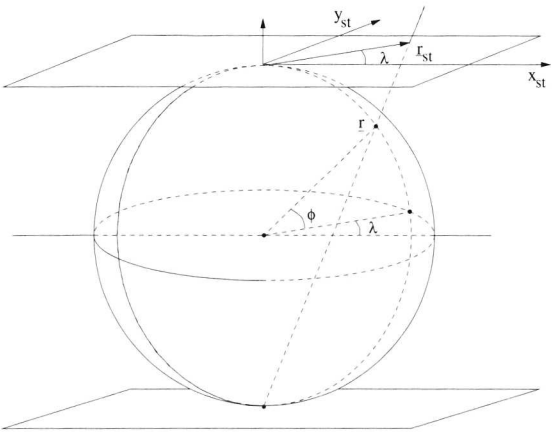


Figure 6.1: The projection of the northern hemisphere onto the stereographic plane.

From figure 6.2(a) it can then be derived that $|r_{st}| = a\,m\cos\phi$ with $m = 2/(1+\sin\phi)$. Next, let the positive stereographic x_{st}-axis correspond with the intersection of the stereographic plane and the half-plane $S_{\lambda=0}$ and let the stereographic y_{st}-axis correspond with the intersection of the stereographic plane and the half-plane $S_{\lambda=\pi/2}$. From figure 6.2(b) then follows that $\underline{r}_{st} = (\,|r_{st}|\cos\lambda,\ |r_{st}|\sin\lambda\,)$.
The transformation relations between the stereographic and spherical coordinates then yield

$$\begin{aligned} x_{st} &= a\,m\cos\phi\cos\lambda, \qquad -\pi/2 < \phi < \pi/2,\ \ 0 \le \lambda < 2\pi, \\ y_{st} &= a\,m\cos\phi\sin\lambda, \qquad -\pi/2 < \phi < \pi/2,\ \ 0 \le \lambda < 2\pi, \end{aligned}$$

where m is the map factor

$$m = \frac{2}{1+\alpha\sin\phi},$$

with α distinguishing between the northern ($\alpha = 1$) and the southern ($\alpha = -1$) hemisphere projection. A thorough description of the stereographic projection can be found in [85].

6.2 Construction of the stereographic formulation of the SWEs from the spherical formulation of the SWEs

In this appendix we construct the SWEs in stereographic coordinates from the SWEs in spherical coordinates. Note that this construction is valid on the whole sphere with exception of the poles. To derive the stereographic formulation of the SWEs on the whole sphere, tensor analysis is required. The SWEs are then derived from their description in general coordinates. For an introduction to tensor analysis, we refer to [1]. Wesseling [82] provides an excellent summary of the necessary principles for formulating the physical conservation laws in general coordinates.

Let (λ, ϕ) and (x_{st}, y_{st}) denote the spherical and stereographic coordinates, respectively. We first derive a few useful relations between these coordinates.

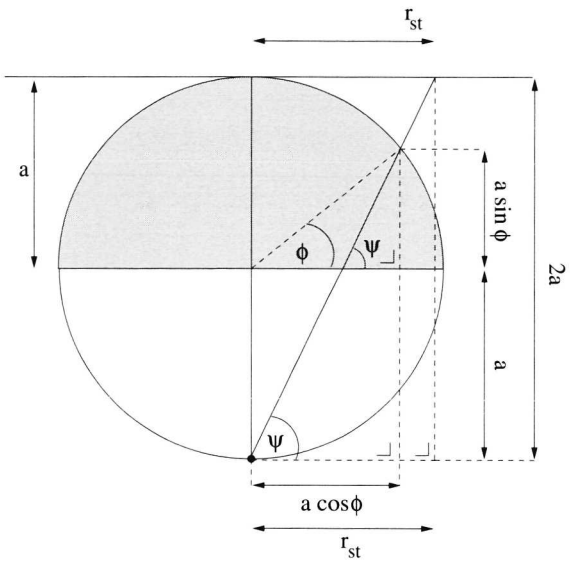
$$x_{st} = a\,m\cos\phi\cos\lambda, \tag{6.1}$$

$$y_{st} = a\,m\cos\phi\sin\lambda \tag{6.2}$$

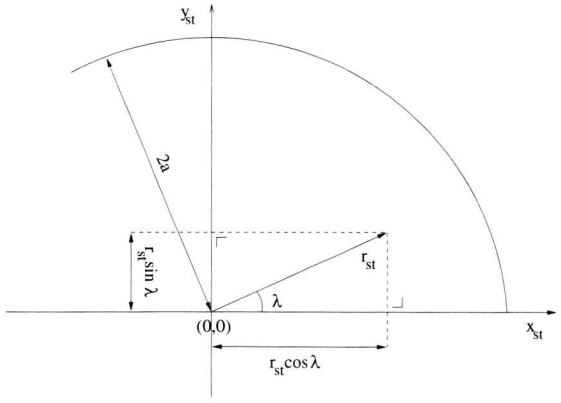
with $m = \frac{2}{1+\alpha\sin\phi}$. For the velocity fields in spherical components, $\underline{u} = (u, v)$, and stereographic components, $\underline{U} = (U, V)$, we have

$$U = -u\sin\lambda - \alpha\,v\cos\lambda, \tag{6.3}$$

$$V = u\cos\lambda - \alpha\,v\sin\lambda. \tag{6.4}$$



(a) Cross-section of the sphere with radius a and the surfaces S_λ and $S_{\lambda+\pi}$.



(b) The stereographic plane for the northern hemisphere projection in the polar case.

Figure 6.2: The geometry of the stereographic mapping (northern hemisphere).

We derive the inverses of the relations (6.1)-(6.4).

$$\varphi = \alpha \arcsin\left[\frac{4a^2 - x_{\rm st}^2 - y_{\rm st}^2}{4a^2 + x_{\rm st}^2 + y_{\rm st}^2}\right]. \tag{6.5}$$

$$\lambda = \arctan\left(\frac{y_{\rm st}}{x_{\rm st}}\right). \tag{6.6}$$

$$u = -U\sin\lambda + V\cos\lambda. \tag{6.7}$$

$$v = -\alpha U\cos\lambda - \alpha V\sin\lambda. \tag{6.8}$$

Remember that the spherical velocity field components (u, v) are defined as

$$\begin{aligned} u &= a\cos\varphi\,\dot{\lambda}, \\ v &= a\,\dot{\phi}, \end{aligned} \tag{6.9}$$

where $\dot{\lambda}$, $\dot{\phi}$ denote the substantial or total time derivatives $\frac{\mathrm{d}\lambda}{\mathrm{d}t}$, $\frac{\mathrm{d}\phi}{\mathrm{d}t}$. For the stereographic velocity field components (U, V) defined as $(m^{-1}\,\dot{x}_{\rm st}, m^{-1}\,\dot{y}_{\rm st})$, we have

$$\begin{aligned} U &= -a\,\dot{\lambda}\cos\phi\sin\lambda - \alpha\, a\,\dot{\phi}\cos\lambda, \\ V &= a\,\dot{\lambda}\cos\phi\cos\lambda - \alpha\, a\,\dot{\phi}\sin\lambda, \end{aligned}$$

and their inverses,

$$\begin{aligned} \dot{\lambda} &= \frac{1}{a\cos\phi}\left(-U\sin\lambda + V\cos\lambda\right), \\ \dot{\phi} &= \frac{\alpha}{a}\left(-U\cos\lambda - V\sin\lambda\right). \end{aligned} \tag{6.10}$$

In combination with relations (6.1)-(6.10), we are able to derive the SWEs in stereographic coordinates from their counterpart in spherical coordinates. In this last coordinate system the SWEs in advective form are given by

$$\frac{\mathrm{d}H}{\mathrm{d}t} + H\,\nabla\cdot\underline{u} = 0,$$

$$\frac{\mathrm{d}u}{\mathrm{d}t} - \left(f + \frac{u}{a}\tan\phi\right)v + \frac{g}{a\cos\phi}\frac{\partial h}{\partial\lambda} = 0,$$

$$\frac{\mathrm{d}v}{\mathrm{d}t} + \left(f + \frac{u}{a}\tan\phi\right)u + \frac{g}{a}\frac{\partial h}{\partial\phi} = 0,$$

where

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \frac{\partial H}{\partial t} + \underline{u}\cdot\nabla H,$$

$$\nabla\cdot\underline{u} = \frac{1}{a\cos\phi}\left[\frac{\partial u}{\partial\lambda} + \frac{\partial(v\cos\phi)}{\partial\phi}\right] \quad \text{and} \quad \nabla H = \left(\frac{1}{a\cos\phi}\frac{\partial H}{\partial\lambda}, \frac{1}{a}\frac{\partial H}{\partial\phi}\right).$$

We are interested in the stereographic formulation of the SWEs in flux form. The derivation steps are described below. We start with the equation of motion in x_{st}-direction.

$$\begin{aligned}
\dot{U} &\overset{\text{step 1}}{=} \left(\left(f + \tfrac{u}{a}\tan\phi\right)v - \tfrac{g}{a\cos\phi}\tfrac{\partial h}{\partial\lambda}\right)\cdot -\sin\lambda + \\
&\qquad + \left(\left(f + \tfrac{u}{a}\tan\phi\right)u + \tfrac{g}{a}\tfrac{\partial h}{\partial\phi}\right)\cdot \alpha\cos\lambda - V\dot{\lambda} \\
&\overset{\text{step 2a/2b}}{=} \alpha f V + (\alpha\sin\phi - 1)\,\dot{\lambda}V - m\,g\tfrac{\partial h}{\partial x_{\mathrm{st}}} \\
&\overset{\text{step 3}}{=} \alpha f V - \tfrac{(x_{\mathrm{st}}V - y_{\mathrm{st}}U)V}{2a^2} - m\,g\tfrac{\partial h}{\partial x_{\mathrm{st}}}.
\end{aligned}$$

Step 1 From (6.3), (6.4), (6.7) and (6.8) we can derive that

$$\begin{aligned}
\dot{U} &= -\dot{u}\sin\lambda - \alpha\,\dot{v}\cos\lambda - u\cos\lambda\,\dot{\lambda} + \alpha\,v\sin\lambda\,\dot{\lambda} \\
&= -\dot{u}\sin\lambda - \alpha\,\dot{v}\cos\lambda - V\dot{\lambda}, \\
\dot{V} &= \dot{u}\cos\lambda - \alpha\,\dot{v}\sin\lambda - u\sin\lambda\,\dot{\lambda} - \alpha\,v\cos\lambda\,\dot{\lambda} \\
&= \dot{u}\cos\lambda - \alpha\,\dot{v}\sin\lambda + U\dot{\lambda}.
\end{aligned}$$

□

Step 2a Using (6.4) and (6.9), we first rewrite part of the resulting equation from step 1 in terms of the stereographic coordinates,

$$\begin{aligned}
&\left(f + \frac{u}{a}\tan\phi\right)v\cdot -\sin\lambda + \left(f + \frac{u}{a}\tan\phi\right)u\cdot\alpha\cos\lambda \\
&\qquad = \left(f + \frac{u}{a}\tan\phi\right)(-v\sin\lambda + \alpha\,u\cos\lambda) \;=\; \alpha\,f\,V + \alpha\sin\phi\,\dot{\lambda}\,V.
\end{aligned}$$

Step 2b Using (6.1)-(6.2), we then rewrite $\frac{g}{a\cos\phi}\sin\lambda\,\frac{\partial h}{\partial\lambda} + \frac{\alpha g}{a}\cos\lambda\,\frac{\partial h}{\partial\phi}$ in terms of the stereographic coordinates,

$$\begin{aligned}
&\frac{g\sin\lambda}{a\cos\phi}\frac{\partial h}{\partial\lambda} + \frac{\alpha\,g\cos\lambda}{a}\frac{\partial h}{\partial\phi} \\
&= \frac{g}{a}\left(\frac{\sin\lambda}{\cos\phi}\left(\frac{\partial h}{\partial x_{\mathrm{st}}}\frac{\partial x_{\mathrm{st}}}{\partial\lambda} + \frac{\partial h}{\partial y_{\mathrm{st}}}\frac{\partial y_{\mathrm{st}}}{\partial\lambda}\right) + \alpha\cos\lambda\left(\frac{\partial h}{\partial x_{\mathrm{st}}}\frac{\partial x_{\mathrm{st}}}{\partial\phi} + \frac{\partial h}{\partial y_{\mathrm{st}}}\frac{\partial y_{\mathrm{st}}}{\partial\phi}\right)\right) \qquad (6.11)
\end{aligned}$$

with

$$\begin{pmatrix} \frac{\partial x_{\mathrm{st}}}{\partial\lambda} & \frac{\partial x_{\mathrm{st}}}{\partial\phi} \\ \frac{\partial y_{\mathrm{st}}}{\partial\lambda} & \frac{\partial y_{\mathrm{st}}}{\partial\phi} \end{pmatrix} = \begin{pmatrix} -a\,m\cos\phi\sin\lambda & a\,\frac{\mathrm{d}m}{\mathrm{d}\phi}\cos\phi\cos\lambda - a\,m\sin\phi\cos\lambda \\ a\,m\cos\phi\cos\lambda & a\,\frac{\mathrm{d}m}{\mathrm{d}\phi}\cos\phi\sin\lambda - a\,m\sin\phi\sin\lambda \end{pmatrix} \qquad (6.12)$$

and

$$\frac{\mathrm{d}m}{\mathrm{d}\phi} \;=\; -\frac{2\alpha\cos\phi}{(1+\alpha\sin\phi)^2} = -\frac{m\,\alpha\cos\phi}{(1+\alpha\sin\phi)}. \qquad (6.13)$$

Combining (6.11)-(6.13). we find

$$\begin{aligned}
&\frac{g}{a\cos\phi}\sin\lambda\frac{\partial h}{\partial\lambda}+\frac{\alpha\, g}{a}\cos\lambda\frac{\partial h}{\partial\phi} = \\
&= g\left(-m\sin^2\lambda-\alpha\, m\sin\phi\cos^2\lambda+\frac{-m\cos^2\phi\,\cos^2\lambda}{(1+\alpha\sin\phi)}\right)\frac{\partial h}{\partial x_{\rm st}}+ \\
&\quad + g\left(m\cos\lambda\sin\lambda-\alpha\, m\sin\phi\cos\lambda\sin\lambda-\frac{m\cos^2\phi}{1+\alpha\sin\phi}\sin\lambda\cos\lambda\right)\frac{\partial h}{\partial y_{\rm st}} \\
&= -m\, g\frac{\partial h}{\partial x_{\rm st}}. \qquad \square
\end{aligned}$$

Step 3 We focus on the total derivative $\dot{\lambda}$. Multiply $\dot{\lambda}$ with $\alpha\sin\phi-1$. From (6.1). (6.2) and (6.10) then follows

$$\begin{aligned}
(\alpha\sin\phi-1)\,\dot{\lambda} &= \left(-\frac{(1-\alpha\sin\phi)}{a^2\, m\cos^2\phi}\,(x_{\rm st}V-y_{\rm st}U)\right) \\
&= -\frac{(1-\alpha\sin\phi)\,(1+\alpha\sin\phi)}{2a^2\cos^2\phi}\,(x_{\rm st}V-y_{\rm st}U) \\
&= -\frac{1}{2a^2}\,(x_{\rm st}V-y_{\rm st}U)\,. \qquad \square
\end{aligned}$$

In a similar way we can derive the equation of motion in the $y_{\rm st}$-direction. This equation reads

$$\dot{V} = -\alpha f U+\frac{U}{2a^2}\,(x_{\rm st}V-y_{\rm st}U)-mg\frac{\partial h}{\partial y_{\rm st}}.$$

The continuity equation in terms of the stereographic coordinates is described as

$$\dot{H}+H\,\nabla\cdot\underline{U} = 0$$

with

$$\nabla\cdot\underline{U} \overset{\text{step 4}}{=} m^2\left[\frac{\partial}{\partial x_{\rm st}}\left(\frac{U}{m}\right)+\frac{\partial}{\partial y_{\rm st}}\left(\frac{V}{m}\right)\right],$$

where we applied Step 4.

Step 4 By definition, we have

$$(\nabla_{\rm sphere}\cdot\underline{u}) = (\nabla_{\rm st}\cdot\underline{U})\,, \tag{6.14}$$

where $(\nabla_{\rm sphere}\cdot\underline{u})$ is defined as the divergence operator in spherical coordinates,

$$\nabla_{\rm sphere}\cdot\underline{u}\equiv\frac{1}{a\cos\phi}\left[\frac{\partial u}{\partial\lambda}+\frac{\partial v\cos\phi}{\partial\phi}\right].$$

We need to derive the divergence operator in terms of the stereographic coordinates,

$$\nabla_{\text{sphere}} \cdot (A\underline{u}) \equiv \frac{1}{a\cos\phi}\left[\frac{\partial}{\partial\lambda}(Au) + \frac{\partial}{\partial\phi}(Av\cos\phi)\right].$$

In combination with equations (6.7), (6.8), (6.12) and (6.13), this yields

$$\nabla_{\text{sphere}} \cdot (A\underline{u}) = m\frac{\partial}{\partial x_{\text{st}}}(AU) + m\frac{\partial}{\partial y_{\text{st}}}(AV) + \frac{\alpha\sin\phi - 1}{a\cos\phi}(\cos\lambda\, AU + \sin\lambda\, AV). \quad (6.15)$$

To further explore the last term in this equation, observe that

$$\frac{\partial\phi}{\partial x_{\text{st}}} = -\frac{\alpha\cos\lambda}{a\,m} \quad \text{and} \quad \frac{\partial\phi}{\partial y_{\text{st}}} = -\frac{\alpha\sin\lambda}{a\,m},$$

where we applied equation (6.5). Together with (6.13), we then find

$$\frac{\partial m}{\partial x_{\text{st}}} = \frac{\cos\lambda\,(1-\alpha\sin\phi)}{a\cos\phi} \quad \text{and} \quad \frac{\partial m}{\partial y_{\text{st}}} = \frac{\sin\lambda\,(1-\alpha\sin\phi)}{a\cos\phi}.$$

So, the last term in equation (6.15) yields

$$-\frac{(1-\alpha\sin\phi)}{a\cos\phi}(\cos\lambda\, AU + \sin\lambda\, AV) = -\left(AU\frac{\partial m}{\partial x_{\text{st}}} + AV\frac{\partial m}{\partial y_{\text{st}}}\right).$$

Combining this equation with the equations (6.14) and (6.15), we find

$$\nabla_{\text{st}} \cdot (A\underline{U}) \equiv m^2\frac{\partial}{\partial x_{\text{st}}}\left(\frac{AU}{m}\right) + m^2\frac{\partial}{\partial y_{\text{st}}}\left(\frac{AV}{m}\right). \qquad \square$$

Summarizing, the advective form of the SWEs in stereographic coordinates reads

$$\dot{H} = -H\,\nabla\cdot\underline{U}, \quad (6.16)$$

$$\dot{U} = \alpha f V - \frac{(x_{\text{st}}V - y_{\text{st}}U)\,V}{2a^2} - m\,g\frac{\partial h}{\partial x_{\text{st}}}, \quad (6.17)$$

$$\dot{V} = -\alpha f U + \frac{U}{2a^2}(x_{\text{st}}V - y_{\text{st}}U) - m\,g\frac{\partial h}{\partial y_{\text{st}}}, \quad (6.18)$$

where, by definition, the total derivative is given by

$$\dot{H} = \frac{\partial H}{\partial t} + \dot{x_{\text{st}}}\frac{\partial H}{\partial x_{\text{st}}} + \dot{y_{\text{st}}}\frac{\partial H}{\partial y_{\text{st}}} = \frac{\partial H}{\partial t} + m\,U\frac{\partial H}{\partial x_{\text{st}}} + m\,V\frac{\partial H}{\partial y_{\text{st}}}, \quad (6.19)$$

and the divergence operator is

$$\nabla\cdot\underline{U} \equiv m^2\left[\frac{\partial}{\partial x_{\text{st}}}\left(\frac{U}{m}\right) + \frac{\partial}{\partial y_{\text{st}}}\left(\frac{V}{m}\right)\right]. \quad (6.20)$$

Finally, we combine the equations (6.16)–(6.20) to find the SWEs in flux form,

$$\begin{aligned}
\frac{\partial H}{\partial t} + \nabla \cdot (H\underline{U}) &= 0, \\
\frac{\partial HU}{\partial t} + \nabla \cdot (HU\underline{U}) &= \left[\alpha f - \frac{(x_{\mathrm{st}}V - y_{\mathrm{st}}U)}{2a^2}\right] HV - m\,g\,H\frac{\partial h}{\partial x_{\mathrm{st}}}, \\
\frac{\partial HV}{\partial t} + \nabla \cdot (HV\underline{U}) &= -\left[\alpha f - \frac{(x_{\mathrm{st}}V - y_{\mathrm{st}}U)}{2a^2}\right] HU - m\,g\,H\frac{\partial h}{\partial y_{\mathrm{st}}}.
\end{aligned}$$

Remember that this derivation is valid on the whole sphere with exception of the poles.

6.3 Construction of the Osher flux

In this appendix we describe the construction of the Osher flux. First, we investigate the Osher flux for a general hyperbolic system of equations in $\mathbb{R}^3$. Second, we zoom in on the shallow water equations. The first part of this appendix is based on the article of Osher and Solomon [53].

Consider a general hyperbolic system of conservation laws in one dimension,

$$\frac{\partial \underline{q}}{\partial t} + \frac{\partial \underline{f}(\underline{q})}{\partial x} = 0, \tag{6.21}$$

where $\underline{q}$ defines the state variable $\underline{q} = (q_1, q_2, q_3)^T \in \mathbb{R}^3$ and $\underline{f}$ defines the flux in x-direction. The system (6.21) is called hyperbolic, when the eigenvalues λ_k of the Jacobian matrix A of the flux $\underline{f}$ with respect to $\underline{q}$, $A = \partial \underline{f}/\partial \underline{q}$, are real and the corresponding eigenvectors $\underline{r}_k$ span the state space $\mathbb{R}^3$. Note that the Jacobian matrix A can depend on the state variable $\underline{q}$.

In a finite volume discretization of system (6.21), an approximation of the flux $\underline{f}(\underline{q})$ across each cell boundary is required.

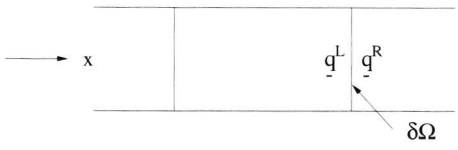


Figure 6.3: Situation at a cell boundary $\delta\Omega$.

Let $\delta\Omega$ be such a cell boundary (x=constant) and assume that at the left and right of this boundary, constant states $\underline{q}^{\mathrm{L}}$ and $\underline{q}^{\mathrm{R}}$ are defined, respectively, see Figure 6.3. We then approximate the resulting flux $\underline{f}$ across this boundary with Osher's flux

defined as

$$\underline{F}_{(\mathrm{O})}(\underline{q}^{\mathrm{L}}, \underline{q}^{\mathrm{R}}) = \frac{1}{2}\left(\underline{f}(\underline{q}^{\mathrm{L}}) + \underline{f}(\underline{q}^{\mathrm{R}})\right) - \frac{1}{2}\int_{\underline{q}^{\mathrm{L}}}^{\underline{q}^{\mathrm{R}}} |A(\underline{q})|\, d\underline{q}. \tag{6.22}$$

The absolute value of the Jacobian matrix A is here defined by $|A| = P|\Lambda|P^{-1}$, where P and Λ result from diagonalizing the Jacobian matrix as $A = P\Lambda P^{-1}$. Note that, because of the system's hyperbolic character, the matrices P and P^{-1} exist.

The structure of the Osher flux (6.22) originates from a generalization of the Engquist and Osher flux [18,19], which was developed for non-linear scalar conservation laws. In contrast to this flux, the Osher flux (6.22) is not uniquely determined. The path of integration between $\underline{q}^{\mathrm{L}}$ and $\underline{q}^{\mathrm{R}}$ in the state space $\mathbb{R}^3$ can be chosen in different ways, significantly influencing the properties of the resulting scheme. Osher made a natural choice for his path of integration, leading to his famous both elegant and well-applicable flux.

Osher's path Γ is composed of subcurves Γ_k which are based on the eigenvectors $\underline{r}_k$ of the Jacobian matrix A, i.e.,

$$\Gamma = \bigcup_{k=1}^{3} \Gamma_k,$$

where Γ_k is parameterized as

$$\Gamma_k = \left\{ \underline{q}^k(s) : \frac{d\underline{q}^k}{ds} = \underline{r}_k \quad \text{with} \quad 0 \le s \le s_k \right\}. \tag{6.23}$$

and $\underline{q}^k_{\mathrm{b}}(0)$ and $\underline{q}^k_{\mathrm{e}}(s_k)$ denote respectively the begin and end point of this subcurve. Subcurves defined in this way correspond to rarefaction or compression wave solutions of system (6.21). The subcurves Γ_k are passed in order of increasing corresponding eigenvalues λ_k, following the P(hysical)-variant proposed by Hemker and Spekreijse [30] to improve efficiency. Originally, Osher proposed to move along the subcurves Γ_k in order of decreasing corresponding eigenvalues λ_k (O(sher)-variant). Using hyperbolicity and the implicit function theorem, it can be shown that exactly one Osher path exists [53] for both the P- and O-variant. A schematic representation of the P-variant Osher path is given in Figure 6.4. The states $\underline{q}^{1/3}$ and $\underline{q}^{2/3}$ denote the unique intersection points between the different subcurves Γ_k. At the end of this section their exact value is given.

Along a subcurve Γ_k the evaluation of the integral in (6.22) turns out to be very simple. First, we rewrite equation (6.22). For that purpose, let us introduce the eigenvalues λ_k^+ and λ_k^-,

$$\lambda_k^+ = \begin{cases} \lambda_k & \text{if} \quad \lambda_k > 0 \\ 0 & \text{if} \quad \lambda_k \le 0 \end{cases} \quad \text{and} \quad \lambda_k^- = \begin{cases} 0 & \text{if} \quad \lambda_k > 0 \\ \lambda_k & \text{if} \quad \lambda_k \le 0 \end{cases}$$

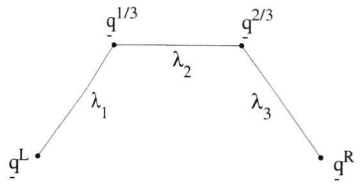


Figure 6.4: A schematic representation of the Osher path Γ in case of the P-variant.

together with the diagonal matrices $\Lambda^+ = \text{diag}\{\lambda_k^+\}$, $\Lambda^- = \text{diag}\{\lambda_k^-\}$, which give $|\Lambda| = \Lambda^+ - \Lambda^-$ and $\Lambda = \Lambda^+ + \Lambda^-$. In relation to these diagonal matrices we define $A^+ = P\Lambda^+P^{-1}$ and $A^- = P\Lambda^-P^{-1}$, which yields $|A| = P|\Lambda|P^{-1}$. By constructing the Osher flux from its scalar Engquist and Osher equivalent [53], it can be written as

$$\underline{F}_{(\mathrm{O})}(\underline{q}^{\mathrm{L}}, \underline{q}^{\mathrm{R}}) = \underline{f}(\underline{q}^{\mathrm{L}}) + \int_{\underline{q}^{\mathrm{L}}}^{\underline{q}^{\mathrm{R}}} A^- d\underline{q} \tag{6.24}$$

$$= \underline{f}(\underline{q}^{\mathrm{R}}) - \int_{\underline{q}^{\mathrm{L}}}^{\underline{q}^{\mathrm{R}}} A^+ d\underline{q}. \tag{6.25}$$

These representations reveal the upwind character of the Osher flux. More precisely, expression (6.24) states that the flux $\underline{f}(\underline{q}^{\mathrm{L}})$, corrected with the characteristic information moving in from the right side of the boundary, approximates the flux at this boundary. Note that this characteristic information corresponds with the matrix A^-. Conversely, the flux $\underline{f}(\underline{q}^{\mathrm{R}})$ corrected with the characteristic information moving in from the left side of the boundary, also approximates the flux at this boundary. In that case the characteristic information corresponds with the matrix A^+. Henceforth, we will work with representation (6.25) instead of (6.22), which amounts to evaluation of the following integral along each subcurve Γ_k,

$$\int_{\Gamma_k} A^+ d\underline{q}. \tag{6.26}$$

Let us simplify the integral (6.26) by using the parameterization of subcurve Γ_k. This yields

$$\int_{\Gamma_k} A^+ \, d\underline{q} = \int_0^{s_k} P\Lambda^+P^{-1}\underline{r}_k \, ds = \int_0^{s_k} \lambda_k^+ \, \underline{r}_k \, ds.$$

Through this formulation we can show that calculation of the Osher flux requires no more than a few flux evaluations. However, we first need to identify on which parts of the subcurves Γ_k the corresponding eigenvalues λ_k are positive. For that purpose, we make some assumptions about the eigenvalues λ_k. These assumptions are valid for most physical systems of equations. They also hold for the shallow

water equations, as we will prove in subsection 6.3.1. The eigenvalue λ_k is supposed to be either linearly degenerate, which means that along subcurve Γ_k

$$\frac{d}{ds}\lambda_k = \nabla\lambda_k \cdot \underline{r}_k \equiv 0, \tag{6.27}$$

or genuinely non-linear, which means that along subcurve Γ_k

$$\frac{d}{ds}\lambda_k = \nabla\lambda_k \cdot \underline{r}_k \neq 0. \tag{6.28}$$

The first case indicates that the eigenvalue λ_k is constant on Γ_k, i.e., λ_k^+ is either zero or λ_k on Γ_k. In the second case the eigenvalue is strictly monotone on Γ_k, which indicates that λ_k changes sign at most once on Γ_k. We will call the point, $\underline{q}_{\mathrm{s}}^k$, where this possible change occurs, a sonic point. Note that under the assumptions (6.27) and (6.28), λ_k is positive on at most one part of the subcurve Γ_k. When $\underline{q}_{b+}^k$ and $\underline{q}_{e+}^k$ denote the begin and end points of this part, we find in terms of our Osher path that these points are either $\underline{q}^{\mathrm{L}}$, $\underline{q}^{\mathrm{R}}$, $\underline{q}^{1/3}$, $\underline{q}^{2/3}$, or $\underline{q}_{\mathrm{s}}^k$.

Sofar we have not defined the exact values of the intersection states, $\underline{q}^{1/3}$, $\underline{q}^{2/3}$, and possible $\underline{q}_{\mathrm{s}}^k$. We will now attend to this topic. Therefore, we need the concept of Riemann invariants. For each k, these invariants ψ_ν^k, $\nu \neq k$, are defined as the two independent solutions of the equation,

$$\left(\frac{\partial\psi}{\partial q_1}, \frac{\partial\psi}{\partial q_2}, \frac{\partial\psi}{\partial q_3}\right) \cdot \underline{r}_k = 0, \tag{6.29}$$

This definition implies that the invariants ψ_ν^k are constant on Γ_k. It is this property that provides us with just enough equations to determine the 6 unknown state variables of $\underline{q}^{1/3}$ and $\underline{q}^{2/3}$. On subcurve Γ_1 we have

$$\begin{aligned} \psi_2^1(\underline{q}^{1/3}) &= \psi_2^1(\underline{q}^{\mathrm{L}}), \\ \psi_3^1(\underline{q}^{1/3}) &= \psi_3^1(\underline{q}^{\mathrm{L}}). \end{aligned} \tag{6.30}$$

on subcurve Γ_2,

$$\begin{aligned} \psi_1^2(\underline{q}^{1/3}) &= \psi_1^2(\underline{q}^{2/3}), \\ \psi_3^2(\underline{q}^{1/3}) &= \psi_3^2(\underline{q}^{2/3}), \end{aligned} \tag{6.31}$$

and on subcurve Γ_3,

$$\begin{aligned} \psi_1^3(\underline{q}^{2/3}) &= \psi_1^3(\underline{q}^{\mathrm{R}}), \\ \psi_2^3(\underline{q}^{2/3}) &= \psi_2^3(\underline{q}^{\mathrm{R}}). \end{aligned} \tag{6.32}$$

When a sonic point occurs, we also need the aid of the Riemann invariants. Assume that the eigenvalue λ_k is genuinely non-linear and remember that $\underline{q}_{\mathrm{b}}^k$ and

$\underline{q}_e^k$ denote respectively the begin and end point of subcurve Γ_k. Note that along our Osher path, the begin and end points of each subcurve Γ_k are known. Further, assume that for subcurve Γ_k the following inequality holds

$$\lambda_k(\underline{q}_b^k) \cdot \lambda_k(\underline{q}_e^k) \leq 0.$$

In other words, on Γ_k, a sonic point $\underline{q}_s^k$ will be found. To determine the state variable $\underline{q}_s^k$, we need at least three equations. The first two equations are provided by the Riemann invariants. We have

$$\psi_\nu^k(\underline{q}_b) = \psi_\nu^k(\underline{q}_s^k),$$

where $\nu \neq k$. The third equation follows through the definition of a sonic point,

$$\lambda_k(\underline{q}_s^k) = 0.$$

Now, we have enough information to calculate integral (6.26) for each subcurve Γ_k. When our system fulfills the conditions on the eigenvalues λ_k, the begin and end state of the part of each subcurve Γ_k along which the corresponding eigenvalue λ_k remains positive, i.e., $\underline{q}_{b+}^k(s_{b+}^k)$ and $\underline{q}_{e+}^k(s_{e+}^k)$, are known. In that case, the evaluation of the integral (6.26) reduces to at most two flux evaluations per subcurve,

$$\begin{aligned}\int_{\Gamma_k} A^+ \, d\underline{q} &= \int_0^{s_k} \lambda_k^+ \, \underline{r}_k \, ds = \int_{s_{b+}^k}^{s_{e+}^k} \lambda_k \, \underline{r}_k \, ds \\ &= \int_{s_{b+}^k}^{s_{e+}^k} A \, \underline{r}_k \, ds = \int_{\underline{q}_{b+}^k}^{\underline{q}_{e+}^k} \frac{d\underline{f}}{d\underline{q}} \, d\underline{q} = \underline{f}(\underline{q}_{e+}^k) - \underline{f}(\underline{q}_{b+}^k). \quad (6.33)\end{aligned}$$

6.3.1 The Osher flux for the Shallow Water Equations.

We have constructed the Osher flux and its P-variant Osher path for a general hyperbolic system of equations in one dimension. We now concentrate on the 2D Shallow Water Equations in spherical coordinates. Observe that, though with different variables, the construction of the Osher flux in case of the stereographic formulation runs along the same lines.

It suffices to approximate the flux $\underline{f}$ on a boundary in the (local) longitudinal direction, i.e.,

$$\underline{f}(\underline{q}) = \begin{pmatrix} q_2 \\ q_2^2/q_1 + \frac{1}{2}gq_1^2 \\ (q_2q_3)/q_1 \end{pmatrix} = \begin{pmatrix} Hu \\ Hu^2 + \frac{1}{2}gH^2 \\ Huv \end{pmatrix},$$

where $\underline{q} = (H, Hu, Hv)$ denotes the state variable. We apply the Osher flux (6.25) in combination with its P-variant Osher path, see Figure 6.4. Following the foregoing, we thus use the steps described below for the construction of the Osher flux,

1. Check whether or not the system of equations is hyperbolic. If so, determine the Riemann invariants and construct the P-variant Osher path, i.e., find $\underline{q}^{1/3}$ and $\underline{q}^{2/3}$.
2. Check whether or not the eigenvalues are linearly degenerate or genuinely non-linear. If so, relate these properties to their corresponding subcurves on the Osher path.
3. Check whether or not a sonic point is located on the subcurves corresponding to the genuinely non-linear eigenvalues. If so, calculate the corresponding states.
4. Determine along which parts of the subcurves the corresponding eigenvalues remain positive.
5. The Osher flux can then be found by combining equation (6.25), the P-variant Osher path and the parts found in step 4.

Step 1 The Jacobian matrix A of the flux $\underline{f}$ with respect to $\underline{q}=(H, Hu, Hv)$ reads

$$A = \frac{d\underline{f}}{d\underline{q}} = \begin{pmatrix} 0 & 1 & 0 \\ \frac{-q_2^2}{q_1^2} + gq_1 & \frac{2q_2}{q_1} & 0 \\ \frac{-q_2 q_3}{q_1^2} & \frac{q_3}{q_1} & \frac{q_2}{q_1} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ -u^2 + gH & 2u & 0 \\ -u\,v & v & u \end{pmatrix}.$$

Its eigenvalues are given by

$$\begin{aligned} \lambda_1 &= u - \sqrt{gH}, \\ \lambda_2 &= u, \\ \lambda_3 &= u + \sqrt{gH}, \end{aligned}$$

with corresponding eigenvectors,

$$\begin{aligned} \underline{r}_1 &= (1, u - \sqrt{gH}, v)^T, \\ \underline{r}_2 &= (0, 0, 1)^T, \\ \underline{r}_3 &= (1, u + \sqrt{gH}, v)^T, \end{aligned}$$

establishing that our system of equations is hyperbolic. Note that the eigenvalues are numbered in increasing order.
The Riemann invariants follow after solving equation (6.29) for each subcurve Γ_k,

$$\begin{aligned} \psi_2^1 &= v, & \psi_3^1 &= u + 2\sqrt{gH}, \\ \psi_1^2 &= H, & \psi_3^2 &= Hu, \\ \psi_1^3 &= v, & \psi_2^3 &= u - 2\sqrt{gH}. \end{aligned}$$

The P-variant Osher path is then illustrated in Figure 6.5. The eigenvalues indicate the propagation speeds along the corresponding characteristic directions, i.e., along the corresponding eigenvectors $\underline{r}_k$.
$\underline{q}^{1/3}$ and $\underline{q}^{2/3}$ result after solving system (6.30)–(6.32),

$$\underline{q}^{\frac{1}{3}} = \begin{pmatrix} H_{\frac{1}{3}} \\ H_{\frac{1}{3}} u_{\frac{1}{3}} \\ H_{\frac{1}{3}} v_{\mathrm{L}} \end{pmatrix}, \quad \underline{q}^{\frac{2}{3}} = \begin{pmatrix} H_{\frac{1}{3}} \\ H_{\frac{1}{3}} u_{\frac{1}{3}} \\ H_{\frac{1}{3}} v_{\mathrm{R}} \end{pmatrix},$$

where

$$\begin{aligned} H_{\frac{1}{3}} &= \tfrac{1}{16}\tfrac{1}{g}\left((u_{\mathrm{L}} - u_{\mathrm{R}}) + 2\left(\sqrt{gH_{\mathrm{L}}} + \sqrt{gH_{\mathrm{R}}}\right)\right)^2, \\ u_{\frac{1}{3}} &= \tfrac{1}{2}\left(u_{\mathrm{L}} + u_{\mathrm{R}}\right) + \sqrt{gH_{\mathrm{L}}} - \sqrt{gH_{\mathrm{R}}}. \end{aligned}$$

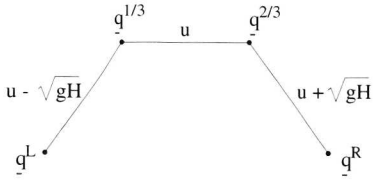


Figure 6.5: A schematic representation of the P-variant Osher path Γ.

Step 2 Elaborating the expressions (6.27) and (6.28), we then find that the eigenvalue λ_2 is linearly degenerate and the eigenvalues λ_1 and λ_3 are genuinely nonlinear. We have

$$\nabla_{\underline{q}} \lambda_2 \cdot \underline{r}_2 = -\frac{u}{H} \cdot 0 + \frac{1}{H} \cdot 0 = 0, \quad \forall \underline{q} \in S,$$

where S defines the state space $S = \{\underline{q} : q_1 \in \mathbb{R}^+,\ q_2 \in \mathbb{R},\ q_3 \in \mathbb{R}\}$ and

$$\begin{aligned} \nabla_{\underline{q}} \lambda_1 \cdot \underline{r}_1 &= \left(-\tfrac{u}{H} - \tfrac{1}{2}\sqrt{\tfrac{g}{H}}\right) \cdot 1 + \tfrac{1}{H} \cdot \left(u - \sqrt{gH}\right) = -\tfrac{3}{2}\sqrt{\tfrac{g}{H}} \neq 0, \quad \forall \underline{q} \in S, \\ \nabla_{\underline{q}} \lambda_3 \cdot \underline{r}_3 &= \left(-\tfrac{u}{H} + \tfrac{1}{2}\sqrt{\tfrac{g}{H}}\right) \cdot 1 + \tfrac{1}{H} \cdot \left(u + \sqrt{gH}\right) = \tfrac{3}{2}\sqrt{\tfrac{g}{H}} \neq 0, \quad \forall \underline{q} \in S. \end{aligned}$$

Step 3 A sonic point can occur on subcurve Γ_1 or on subcurve Γ_3. When a sonic point is located on subcurve Γ_1, or in other words, when the inequality $\lambda_1(\underline{q}^{\mathrm{L}}) \cdot \lambda_1(\underline{q}^{1/3}) \leq 0$ holds, the sonic point is given as

$$\underline{q}^1_{\mathrm{s}} = \begin{pmatrix} H_{\mathrm{s}} \\ H_{\mathrm{s}}\sqrt{gH_{\mathrm{s}}} \\ H_{\mathrm{s}} v_{\mathrm{L}} \end{pmatrix} \quad \text{with} \quad H_{\mathrm{s}} = \frac{1}{9\,g}\left(u_{\mathrm{L}} + 2\sqrt{gH_{\mathrm{L}}}\right)^2.$$

When a sonic point is located on subcurve Γ_3, i.e., when the inequality $\lambda_3(\underline{q}^{2/3}) \cdot \lambda_3(\underline{q}^{\mathrm{R}}) \leq 0$ holds. the sonic point reads

$$\underline{q}^3_{\mathrm{s}} = \begin{pmatrix} H_{\mathrm{s}} \\ -H_{\mathrm{s}}\sqrt{gH_{\mathrm{s}}} \\ H_{\mathrm{s}}v_{\mathrm{R}} \end{pmatrix} \quad \text{with} \quad H_{\mathrm{s}} = \frac{1}{9\,g}\left(u_{\mathrm{R}} - 2\sqrt{gH_{\mathrm{R}}}\right)^2.$$

Step 4 The parts on the subcurves Γ_k along which the corresponding eigenvalues are positive can be found by the signs of the eigenvalues,

$$\begin{aligned} \lambda_{\mathrm{L}} &= u_{\mathrm{L}} - \sqrt{gH_{\mathrm{L}}}, \\ \lambda_{\frac{1}{3}} &= u_{\frac{1}{3}} - \sqrt{gH_{\frac{1}{3}}}, \\ \lambda_{\frac{1}{2}} &= u_{\frac{1}{2}} = u_{\frac{1}{3}}, \\ \lambda_{\frac{2}{3}} &= u_{\frac{2}{3}} + \sqrt{gH_{\frac{2}{3}}} = u_{\frac{1}{3}} + \sqrt{gH_{\frac{1}{3}}}, \\ \lambda_{\mathrm{R}} &= u_{\mathrm{R}} + \sqrt{gH_{\mathrm{R}}}. \end{aligned} \tag{6.34}$$

Given that $\lambda_{1/3} \leq \lambda_{1/2} \leq \lambda_{2/3}$, we can write down all 16 possible sign combinations of the eigenvalues (6.34) along the Osher path, see Figure 6.6. The plus and minus signs along the Osher path in clockwise direction indicate the signs of respectively the eigenvalues λ_{L}, $\lambda_{1/3}$, $\lambda_{1/2}$, $\lambda_{2/3}$, and λ_{R}. A crossbar on Γ_1 or Γ_3 indicates the existence of a sonic point. Note that these points are also related to the sign of the eigenvalues (6.34). For each different sign combination, the required parts along the Osher path are known, respecting the properties of the eigenvalues λ_1, λ_2, and λ_3.

Step 5 We demonstrate the evaluation of the Osher flux for sign combination $(2,1)$, i.e., $\lambda_{\mathrm{L}} < 0$, $\lambda_{1/3} < 0$, $\lambda_{1/2} > 0$, $\lambda_{2/3} > 0$, $\lambda_{\mathrm{R}} \leq 0$. The eigenvalues $\lambda_{\mathrm{L}} < 0$ and $\lambda_{1/3} < 0$ indicate that $\lambda_1(\underline{q}) < 0$ along subcurve Γ_1. On Γ_2 we have $\lambda_2(\underline{q}) > 0$, because $\lambda_{1/2} = u_{1/2} > 0$ and λ_2 is a linearly degenerate eigenvalue. thus constant along Γ_2. On Γ_3 a sonic point occurs. In combination with the fact that λ_3 is linearly degenerate and $\lambda_{2/3} > 0$, this indicates that along Γ_3, between the states $\underline{q}^{2/3}$ and $\underline{q}^3_{\mathrm{s}}$, $\lambda_3(\underline{q}) > 0$. Consequently, the Osher flux reads

$$\begin{aligned} \underline{F}\left(\underline{q}^{\mathrm{L}}, \underline{q}^{\mathrm{R}}\right) &= \underline{f}\left(\underline{q}^{\mathrm{R}}\right) - \left(\int_{\Gamma_1} A^+ d\underline{q} + \int_{\Gamma_2} A^+ d\underline{q} + \int_{\Gamma_3} A^+ d\underline{q} \right) \\ &= \underline{f}\left(\underline{q}^{\mathrm{R}}\right) - \left[0 + \left(\underline{f}(\underline{q}^{\frac{2}{3}}) - \underline{f}(\underline{q}^{\frac{1}{3}}) \right) + \left(\underline{f}(\underline{q}^3_{\mathrm{s}}) - \underline{f}(\underline{q}^{\frac{2}{3}}) \right) \right] \\ &= \underline{f}(\underline{q}^{\mathrm{R}}) - \underline{f}(\underline{q}^3_{\mathrm{s}}) + \underline{f}(\underline{q}^{\frac{1}{3}}). \end{aligned}$$

Elaboration of the Osher flux for the remaining sign combinations yields Table 6.1.

$\underline{f}(\underline{q}^{\mathrm{R}}) - \underline{f}\left(\underline{q}_{\mathrm{s}}^{3}\right) + \underline{f}\left(\underline{q}_{\mathrm{s}}^{1}\right)$	$\underline{f}\left(\underline{q}_{\mathrm{s}}^{1}\right)$	$\underline{f}(\underline{q}^{\mathrm{L}}) + \underline{f}(\underline{q}^{\mathrm{R}}) - \underline{f}\left(\underline{q}_{\mathrm{s}}^{3}\right)$	$\underline{f}(\underline{q}^{\mathrm{L}})$	$u_{\frac{1}{3}} - \sqrt{gH_{\frac{1}{3}}} \geq 0$
$\underline{f}(\underline{q}^{\mathrm{R}}) - \underline{f}\left(\underline{q}_{\mathrm{s}}^{3}\right) + \underline{f}\left(\underline{q}^{\frac{1}{3}}\right)$	$\underline{f}\left(\underline{q}^{\frac{1}{3}}\right)$	$\underline{f}(\underline{q}^{\mathrm{L}}) + \underline{f}(\underline{q}^{\mathrm{R}}) - \underline{f}\left(\underline{q}_{\mathrm{s}}^{3}\right) + \underline{f}\left(\underline{q}^{\frac{1}{3}}\right) - \underline{f}\left(\underline{q}_{\mathrm{s}}^{1}\right)$	$\underline{f}(\underline{q}^{\mathrm{L}}) + \underline{f}\left(\underline{q}^{\frac{1}{3}}\right) - \underline{f}\left(\underline{q}_{\mathrm{s}}^{1}\right)$	$0 < u_{\frac{1}{3}} < \sqrt{gH_{\frac{1}{3}}}$
$\underline{f}(\underline{q}^{\mathrm{R}}) - \underline{f}\left(\underline{q}_{\mathrm{s}}^{3}\right) + \underline{f}\left(\underline{q}^{\frac{2}{3}}\right)$	$\underline{f}\left(\underline{q}^{\frac{2}{3}}\right)$	$\underline{f}(\underline{q}^{\mathrm{L}}) + \underline{f}(\underline{q}^{\mathrm{R}}) - \underline{f}\left(\underline{q}_{\mathrm{s}}^{3}\right) + \underline{f}\left(\underline{q}^{\frac{2}{3}}\right) - \underline{f}\left(\underline{q}_{\mathrm{s}}^{1}\right)$	$\underline{f}(\underline{q}^{\mathrm{L}}) + \underline{f}\left(\underline{q}^{\frac{2}{3}}\right) - \underline{f}\left(\underline{q}_{\mathrm{s}}^{1}\right)$	$-\sqrt{gH_{\frac{1}{3}}} < u_{\frac{1}{3}} < 0$
$\underline{f}(\underline{q}^{\mathrm{R}})$	$\underline{f}\left(\underline{q}_{\mathrm{s}}^{3}\right)$	$\underline{f}(\underline{q}^{\mathrm{L}}) + \underline{f}(\underline{q}^{\mathrm{R}}) - \underline{f}\left(\underline{q}_{\mathrm{s}}^{1}\right)$	$\underline{f}(\underline{q}^{\mathrm{L}}) + \underline{f}\left(\underline{q}_{\mathrm{s}}^{3}\right) - \underline{f}\left(\underline{q}_{\mathrm{s}}^{1}\right)$	$u_{\frac{1}{3}} + \sqrt{gH_{\frac{1}{3}}} \leq 0$
$u_L - \sqrt{gH_L} < 0$ $u_R + \sqrt{gH_R} \leq 0$	$u_L - \sqrt{gH_L} < 0$ $u_R + \sqrt{gH_R} > 0$	$u_L - \sqrt{gH_L} \geq 0$ $u_R + \sqrt{gH_R} \leq 0$	$u_L - \sqrt{gH_L} \geq 0$ $u_R + \sqrt{gH_R} > 0$	

Table 6.1: The Osher flux depending on the signs of the eigenvalues (6.34).

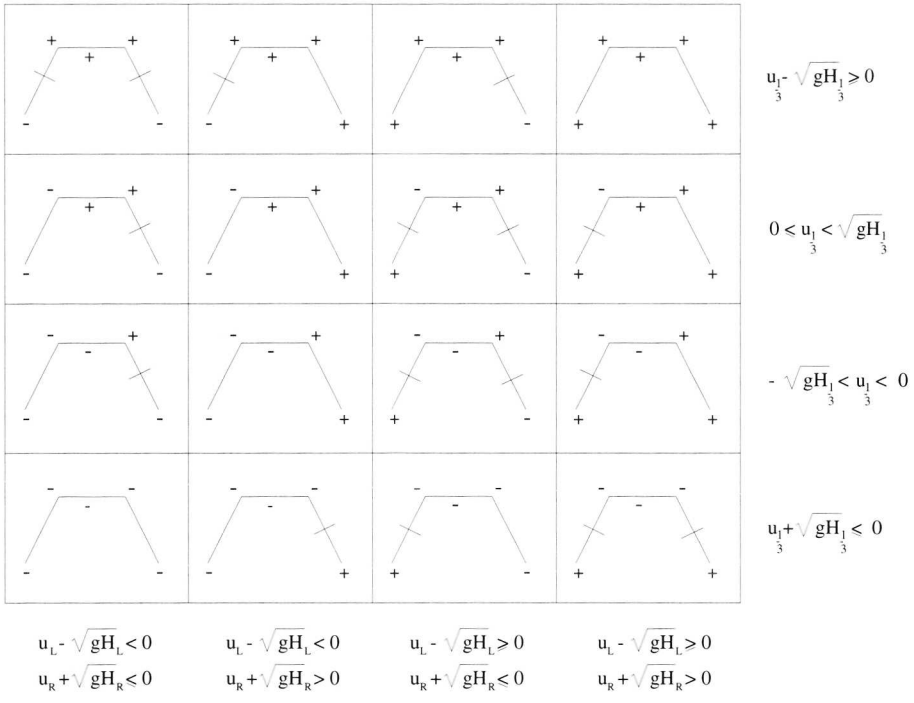


Figure 6.6: The different sign combinations of the eigenvalues along the Osher path Γ.

In relation to Table 6.1, we remark the following. Implementation of the Osher flux leads to a succession of different conditional statement evaluations, which is not very efficient. In practice though, we can discard most of the possible sign combinations. In practical flow patterns, $|\underline{u}| < \sqrt{gH}$, where H represents the depth of the atmosphere, which is always close to 10^4 m. In that case, Table 6.1 reduces to Table 6.2,

6.4 General formulation of the modified κ-scheme for non-uniform grids

In this appendix we give the general formulation of the non-uniform κ-scheme for different values of κ. It concerns the non-uniform equivalents of the 3-point ($\kappa = \frac{1}{3}$)-scheme, the 2-point central ($\kappa = 1$)-scheme, the 2-point upwind ($\kappa = -1$)-scheme, and the 3-point upwind ($\kappa = \frac{1}{2}$)-scheme. Let $\underline{q}_?$ be the unknown state variable to be found by 1D state interpolation in a certain direction, say x. Let ℓ_i denote the cell width of a cell i in x-direction and let $\underline{q}_i$ denote the state variable in its cell center, see Figure 6.7, where we use ℓ_1, ℓ_2 etc. for convenience of notation.

$\underline{f}\left(\underline{q}^{\frac{1}{3}}\right)$	$0 < u_{\frac{1}{3}} < \sqrt{gH_{\frac{1}{3}}}$
$\underline{f}\left(\underline{q}^{\frac{2}{3}}\right)$	$-\sqrt{gH_{\frac{1}{3}}} < u_{\frac{1}{3}} < 0$
$u_L - \sqrt{gH_L} < 0$	
$u_R + \sqrt{gH_R} > 0$	

Table 6.2: Reduction of Table 6.1 under the assumption $|\underline{u}| < \sqrt{gH}$.

The modified κ-scheme is now given as a function I_κ with arguments $\underline{q}_1, \ell_1, \underline{q}_2, \ell_2$ etc. based on Figure 6.7. The modified $(\kappa = \frac{1}{3})$-scheme then reads

$$I_{\frac{1}{3}}(\underline{q}_1, \underline{q}_2, \underline{q}_3, \ell_0, \ell_1, \ell_2, \ell_3, \ell_4) = \alpha\, \underline{q}_1 + \beta\, \underline{q}_2 + \gamma\, \underline{q}_3$$

with

$$\alpha = -2 * \frac{\ell_3(\ell_2\ell_3+\ell_2\ell_4+\ell_3^2+\ell_3\ell_4)}{(\ell_1^3+\ell_1^2(5\ell_2+3\ell_3+\ell_4)+\ell_1(8\ell_2^2+9\ell_2\ell_3+3\ell_2\ell_4+2\ell_3^2+\ell_3\ell_4)+4\ell_2^3+\ell_2^2(6\ell_3+2\ell_4)+\ell_2(2\ell_3^2+\ell_3\ell_4))},$$

$$\beta = 1 - \alpha - \gamma,$$

$$\gamma = 2 * \frac{\ell_2(\ell_0\ell_1+\ell_0\ell_2+2\ell_1^2+3\ell_1\ell_2+\ell_2^2)}{(\ell_0+2\ell_1+2\ell_2+\ell_3)(\ell_1\ell_2+\ell_1\ell_3+2\ell_2^2+3\ell_2\ell_3+\ell_3^2)}.$$

The modified 2-point central $(\kappa = 1)$-scheme, the 2-point upwind $(\kappa = -1)$-scheme, and the 3-point upwind $(\kappa = \frac{1}{2})$-scheme are

$$\begin{aligned}
I_1(\underline{q}_2, \underline{q}_3, \ell_2, \ell_3) &= \tfrac{\ell_3}{\ell_2+\ell_3}\, \underline{q}_2 + \tfrac{\ell_2}{\ell_2+\ell_3}\, \underline{q}_3, \\
I_{-1}(\underline{q}_1, \underline{q}_2, \ell_1, \ell_2) &= \tfrac{-\ell_2}{\ell_1+\ell_2}\, \underline{q}_1 + \tfrac{\ell_1+2\ell_2}{\ell_1+\ell_2}\, \underline{q}_2, \\
I_{\frac{1}{2}}(\underline{q}_1, \underline{q}_2, \underline{q}_3, \ell_1, \ell_2, \ell_3) &= -\tfrac{\ell_2\ell_3}{(\ell_1+2\ell_2+\ell_3)(\ell_1+\ell_2)}\, \underline{q}_1 + \tfrac{(\ell_1+2\ell_2)\ell_3}{(\ell_2+\ell_3)(\ell_1+\ell_2)}\, \underline{q}_2 \\
&\quad + \tfrac{\ell_2(\ell_1+2\ell_2)}{\ell_1\ell_2+\ell_1\ell_3+2\ell_2^2+3\ell_2\ell_3+\ell_3^2}\, \underline{q}_3.
\end{aligned}$$

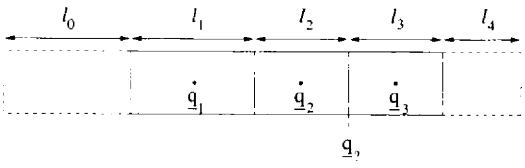


Figure 6.7: General situation around a cell boundary. $\underline{q}_?$ is the unknown state variable to be found by interpolation.

Bibliography

[1] R. Aris. *Vectors, Tensors, and the Basic Equations of Fluid Mechanics.* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1962. Reprinted by Dover, New York, 1989.

[2] 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.

[3] J.G. Blom, W. Hundsdorfer, and J.G. Verwer. Vectorization aspects of a spherical advection scheme on a reduced grid. Technical Report NM-R9418, CWI, Amsterdam, 1994.

[4] J.G. Blom and J.G. Verwer. A comparison of integration methods for atmospheric transport-chemistry problems. *J. Comput. Appl. Math.*, 126:381–396, 2000.

[5] G.J. Boer and B. Denis. Numerical convergence of the dynamics of a GCM. *Climate Dynamics*, 13:359–374, 1997.

[6] R.A. Brown. *Fluid Mechanics of the Atmosphere.* Academic Press, San Diego, 1991.

[7] G.L. Browning, J.J. Hack, and P.N. Swarztrauber. A comparison of three numerical methods for solving differential equations on the sphere. *Mon. Wea. Rev.*, 117:1058–1075, 1989.

[8] K. Cassirer, R. Hess, C. Jablonowski, and W. Joppich. The shallow water test cases for a global model with documentation of the results. Arbeitspapiere der GMD 999, GMD, 1996.

[9] The NCAR community climate model. `http://www.cgd.ucar.edu/cms/ccm3`.

[10] J. Côté, J.G. Desmarais, S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth. The operational CMC-MRB global environmental multiscale (GEM) model: Part II - Results. *Mon. Wea. Rev.*, 126:1397–1418, 1998.

[11] J. Côté, S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth. The operational CMC-MRB global environmental multiscale (GEM) model: Part I - Design considerations and formulation. *Mon. Wea. Rev.*, 126:1373–1395, 1998.

[12] T. Davies and J.C.R. Hunt. New developments in numerical weather prediction. In K.W. Morton and M.J. Baines, editors, *Numerical Methods for Fluid Dynamics V*. Clarendon Press, Oxford, 1995.

[13] K. Dekker and J.G. Verwer. *Stability of Runge-Kutta methods for Stiff Nonlinear Differential Equations*. North-Holland, 1984.

[14] E.G. D'yakonov. Difference systems of second order accuracy with a divided operator for parabolic equations without mixed derivatives. *USSR Comput. Math. Math. Phys.*, 4(5):206–216, 1964.

[15] Dynamical core intercomparison. `http://www-pcmdi.llnl.gov/dc/`.

[16] ECMWF. *ECMWF forecast model documentation*. European Centre for Medium-Range Weather Forecasts, Shinefield Park, England, ECMWF Research Manual edition, 1988.

[17] E.B. Eliasen, B. Machenhauer, and E. Rasmussen. On a numerical method for integration of the hydrodynamical equations with a spectral representation of the horizontal fields. Report No. 2, Institut for Teoretisk Meteorologi, University of Copenhagen, 1970.

[18] B. Engquist and S. Osher. Stable and entropy satisfying approximations for transonic flow calculations. *Math. Comp.*, 34:45–75, 1980.

[19] B. Engquist and S. Osher. One sided difference approximations for nonlinear conservation laws. *Math. Comp.*, 36:321–352, 1981.

[20] W.R. Goodin G.J. McRae and J.H. Seinfeld. Numerical solution of the atmospheric diffusion equation for chemically reacting flows. *J. Comput. Phys.*, 45:1–42, 1982.

[21] T.E. Graedel and P.J. Crutzen. *Atmosphere, Climate and Change*. Scientific American Library, Freeman and Company, New York-Oxford, 1995.

[22] J. Graf and N. Moussiopoulos. Intercomparison of two models for the dispersion of chemically reacting pollutants. *Beitr. Phys. Atmosph.*, 64:13–25, 1991.

[23] W. Gröbner. *Die Lie-Reihen und ihre Anwendungen*. VEB Deutscher Verlag der Wissenschaften, Berlin, 1960.

[24] B. Gustafsson, H-O. Kreiss, and J. Oliger. *Time Dependent Problems and Difference Methods*. John Wiley & Sons, Inc., New York, 1995.

[25] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems.* Springer-Verlag, Berlin, 2nd edition, 1996.

[26] G.J. Haltiner and R.T. Williams. *Numerical Prediction and Dynamic Meteorology.* Wiley, 2nd edition, 1980.

[27] R. Heikes and D.A. Randal. Numerical integration of the SWEs on a twisted icosahedral grid. Part I. *Mon. Wea. Rev.*, 123:1862–1880, 1995.

[28] R. Heikes and D.A. Randal. Numerical integration of the SWEs on a twisted icosahedral grid. Part II. *Mon. Wea. Rev.*, 123:1881–1887, 1995.

[29] I.M. Held and M.J. Suarez. A proposal for the intercomparison of the dynamical cores of atmospheric general circulation models. *Bull. Am. Meteorol. Soc.*, 73:1825–1830, 1994.

[30] P.W. Hemker and S.P. Spekreijse. Multiple grid and Osher's scheme for the efficient solution of the steady state Euler equations. *Appl. Numer. Math.*, 2:475–493, 1986.

[31] C. Hirsch. *Numerical Computation of Internal and External Flows*, volume 2: Computational Methods for Inviscid and Viscous Flow. John Wiley & Sons, Chichester, 1990.

[32] J.R. Holton. *An Introduction to Dynamic Meteorology.* Academic Press, San Diego, 3rd edition, 1992.

[33] P.J. van der Houwen. The development of Runge-Kutta methods for partial differential equations. *Appl. Numer. Math.*, 20:261–272, 1996.

[34] 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.

[35] P.J. van der Houwen, B.P. Sommeijer, and J. Kok. The iterative solution of fully implicit discretizations of three-dimensional transport problems. *Appl. Numer. Math.*, 25:243–256, 1999.

[36] W. Hundsdorfer. Stability of approximate factorizations with θ-methods. *BIT*, 39:473–483, 1997.

[37] W. Hundsdorfer. Accuracy and stability of splitting with stabilizing correction. Technical Report MAS-R9935, CWI, Amsterdam, 1999.

[38] W. Hundsdorfer, B. Koren, M. van Loon, and J.G. Verwer. A positive finite-difference advection scheme. *J. Comput. Phys.*, 117:35–46, 1995.

[39] W. Hundsdorfer and J.G. Verwer. A note on splitting errors for advection-reaction equations. *Appl. Numer. Math.*, 18:191–199, 1995.

[40] IFS Documentation. `http://www.ecmwf.int/research/ifsdocs`. (eds.) P.W. White.

[41] Y. Kurihara. Numerical integration of the primitive equations on a spherical grid. *Mon. Wea. Rev.*, 93:399–415, 1965.

[42] D. Lanser. J.G. Blom. and J.G. Verwer. Spatial discretization of the shallow water equations in spherical geometry. *J. Comput. Phys.*, 165(2):542–565, 2000.

[43] D. Lanser, J.G. Blom, and J.G. Verwer. Time integration of the shallow water equations in spherical geometry. *J. Comput. Phys.*, 171:373–393, 2001.

[44] D. Lanser and J.G. Verwer. Analysis of operator splitting for advection-diffusion-reaction problems from air pollution modelling. *J. Comp. Appl. Math.*, 111:201 216, 1999.

[45] B. Lastdrager, B. Koren, and J.G. Verwer. Solution of time-dependent advection-diffusion problems with the sparse-grid combination technique and a Rosenbrock solver. *J. Comput. Appl. Math.*, 1:86–98, 2001.

[46] R.J. LeVeque. *Time-Split Methods for Partial Differential Equations.* PhD thesis, Stanford University, 1982. Report Dept. of Comp. Science CS-82-904.

[47] R.J. LeVeque and H.C. Yee. A study of numerical methods for hyperbolic conservation laws with source terms. *J. Comput. Phys.*, 86:187–210, 1990.

[48] J.L. Lions, R. Temam, and S. Wang. New formulations of the primitive equations of atmosphere and applications. *Nonlinearity*, 5:237–288, 1992.

[49] D. Majewski. Documentation of the new global model GME. Technical report, Deutscher Wetterdienst, 1996.

[50] D. Majewski. The new global icosahedral-hexagonal grid point model GME of the Deutscher Wetterdienst. In *Recent Developments in Numerical Methods for Atmospheric Modelling.* ECMWF, 1998.

[51] S.A. Orszag. Transform method for calculation of vector coupled sums: Application to the spectral form of the vorticity equation. *J. Atmos. Sci.*, 27:890–895, 1970.

[52] S. Osher and S. Chakravarthy. Upwind schemes and boundary conditions with applications to Euler equations in general geometries. *J. Comput. Phys.*, 50:447-481, 1983.

[53] S. Osher and F. Solomon. Upwind difference schemes for hyperbolic systems of conservation laws. *Math. Comp.*, 38:339–374, 1982.

[54] D.W. Peaceman and H.H. Rachford Jr. The numerical solution of parabolic and elliptic differential equations. *J. Soc. Indust. Appl. Math.*, 3:28–41, 1955.

[55] J. Pedlosky. *Geophysical Fluid Dynamics.* Springer-Verlag, New York, 2nd edition, 1987.

[56] N.A. Phillips. A map projection system suitable for large-scale numerical weather prediction. *J. Meteor. Soc. Japan*, 75:262–267, 1957.

[57] R.J. Purser. Degradation of numerical differencing caused by Fourier filtering at high latitudes. *Mon. Wea. Rev.*, 116:1057–1066, 1988.

[58] M. Rancic, R.J. Purser, and F. Mesinger. A global shallow-water model using an expanded spherical cube: Gnomonic versus conformal coordinates. *Quart. J. Roy. Meteor. Soc.*, 122:959–982, 1996.

[59] P.J. Rasch and D.L. Williamson. The sensitivity of a general circulation model climate to the moisture transport formulation. *J. Geophys. Res.*, 96:123–137, 1991.

[60] L.F. Richardson. Weather prediction by numerical process. Cambridge University Press, 1922. Reprinted by Dover, New York, 1965.

[61] R.D. Richtmyer and K.W. Morton. *Difference Methods for Initial-Value Problems.* Interscience Publishers, New York, 2nd edition, 1967.

[62] C. Ronchi, R. Iacono, and P. S. Paolucci. The cubed sphere: A new method for the solution of partial differential equations in spherical geometry. *J. Comput. Phys.*, 124:93–114, 1996.

[63] R. Sadourny. Conservative finite-difference approximations of the primitive equations on quasi-uniform spherical grids. *Mon. Wea. Rev.*, 100:136–144, 1972.

[64] R. Sadourny, A. Arakawa, and Y. Mintz. Integration of the non-divergent barotropic vorticity equation with an icosahedral-hexagonal grid for the sphere. *Mon. Wea. Rev.*, 96:351–356, 1968.

[65] J.M. Sanz-Serna. Geometric integration. In I.S. Duff and G.A. Watson, editors, *The State of the Art in Numerical Analysis*, pages 121–143. Clarendon Press, Oxford, 1997.

[66] J.M. Sanz-Serna and M.P. Calvo. *Numerical Hamiltonian Problems.* Chapman and Hall, 1994.

[67] E.J. Spee. *Numerical methods in global transport-chemistry models.* PhD thesis, University of Amsterdam, 1998.

[68] S.P. Spekreijse. *Multigrid Solution of the Steady Euler Equations.* Number 46 in CWI Tracts. CWI, Amsterdam, 1988.

[69] B. Spotz. M. Taylor, and P. Swartztrauber. Shallow water equations on the sphere. `http://www.scd.ucar.edu/css/staff/spotz/research/swell.html`.

[70] W.F. Spotz, M.A. Taylor, and P.N. Swarztrauber. Fast shallow-water equation solvers in latitude-longitude coordinates. *J. Comput. Phys.*, 145:432–444, 1998.

[71] A. Staniforth and J. Côté. Semi-Lagrangian integration schemes for atmospheric models - A review. *Mon. Wea. Rev.*, 119:2206–2223, 1991.

[72] G. Starius. Composite mesh difference methods for elliptic and boundary value problems. *Numer. Math.*, 28:243–258, 1977.

[73] R. T. Steeley. *Calculus of Several Variables.* Scott, Foresman and Company, 1970.

[74] J.J. Stoker and E. Isaacson. Final report 1. Technical Report IMM 407, Courant Institute of Mathematical Sciences, New York University, 1975.

[75] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5:506–517, 1968.

[76] B. van Leer. Towards the ultimate conservative difference scheme. III. Upstream-centered finite-difference schemes for ideal compressible flow. *J. Comput. Phys.*, 23:276–299, 1977.

[77] B. van Leer. Upwind-difference methods for aerodynamic problems governed by the Euler equations. In B.E. Engquist, S. Osher, and R.C.J. Sommerville, editors, *Large-scale Computations in Fluid Mechanics*, AMS Series, pages 327–336. American Mathematical Society, 1985.

[78] J.G. Verwer and J.G. Blom. On the coupled solution of diffusion and chemistry in air pollution models. In E. Kreuzer and O. Mahrenholtz, editors, *Procs. Third ICIAM International Congress*, volume 4, pages 454–457. ZAMM, Akademie Verlag, 1996.

[79] J.G. Verwer, J.G. Blom, and W. Hundsdorfer. An implicit-explicit approach for atmospheric transport-chemistry problems. *Appl. Numer. Math.*, 20:191–209, 1996.

[80] J.G. Verwer, W. Hundsdorfer, and J.G. Blom. Numerical time integration for air pollution models. To appear in *Surveys Math. Indust.*, 2001.

[81] J.G. Verwer, E.J. Spee, J.G. Blom, and W. Hundsdorfer. A second order Rosenbrock method applied to photochemical dispersion problems. *SIAM J. Sci. Comput.*, 20:456–480, 1999.

[82] P. Wesseling. *Principles of Computational Fluid Dynamics*, volume 29 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2001.

[83] D.L. Williamson. Integration of the barotropic vorticity equation on a spherical geodesic grid. *Tellus*, 20:642–653, 1968.

[84] D.L. Williamson. Integration of the primitive barotropic model over a spherical geodesic grid. *Mon. Wea. Rev.*, 98:512–520, 1969.

[85] D.L. Williamson. *Numerical Methods Used in Atmospheric Models*, volume II of *GARP Publication Series No. 17*, chapter Difference approximations for fluid flow on a sphere, pages 51–120. World Meteorological Organization, Geneva, 1979.

[86] D.L. Williamson. Review of numerical approaches for modeling global transport. In H. van Dop and G. Kallos, editors, *Air Pollution Modeling and its Application IX*, Nato Challenges of Modern Society, pages 377–394, New York, 1992. Plenum Press.

[87] D.L. Williamson and G.L. Browning. Comparison of grids and difference approximations for numerical weather prediction over a sphere. *J. Appl. Meteor.*, 12:264–274, 1973.

[88] D.L. Williamson, J.B. Drake, J.J. Hack, R. Jacob, and P.N. Swarztrauber. A standard test set for numerical approximations to the shallow water equations in spherical geometry. *J. Comp. Phys.*, 102:211–224, 1992.

[89] D.L. Williamson and J.G. Olson. Climate simulations with a semi-Lagrangian version of the NCAR Community Climate Model. *Mon. Wea. Rev.*, 122:1594–1610, 1993.

[90] D.L. Williamson, J.G. Olson, and B.A. Boville. A comparison of semi-Lagrangian and Eulerian tropical climate simulations. *Mon. Wea. Rev.*, 126:1001–1012, 1998.

[91] D.L. Williamson and P.J. Rasch. Water vapor transport in the NCAR CCM2. *Tellus*, 46A:34–51, 1994.

[92] Z. Zlatev. *Computer Treatment of Large Air Pollution Models*. Kluwer Academic Publishers, 1995.

Summary

For a multitude of reasons, the ability to predict the weather and climate has fascinated people for centuries. Today, weather and climate prediction rely on so-called global circulation models which describe the evolution of the atmospheric circulation, i.e., the evolution of field variables like wind velocity, humidity, temperature, etc. A circulation model consists of a set of mathematical equations representing this evolution. It contains three main interacting parts, viz., a data assimilation, a numerical dynamics, and a physical parameterization part. We concentrate on the second part, which is concerned with the numerical solution of the so-called primitive equations of hydrodynamics in the atmosphere.

Circulation models are rather complex and their numerical solution requires much computational effort. In addition, a prediction demands accurate results calculated within a reasonable amount of time. The accuracy of the prediction depends on the specific model in use, the applied numerical method, the resolution of the considered space-time grid, the incorporated data and the physical parameterization scheme. Since the computations are known to be very time-consuming, much interest is directed at the development of efficient numerical methods on high-resolution grids. In this thesis, we therefore investigate numerical methods to efficiently solve the shallow water equations (SWEs) in spherical geometry on fine-resolution grids. These equations serve as a first prototype of the horizontal dynamics in a global circulation model.

We study a finite volume method for the spatial discretization of the SWEs in spherical geometry, viz., Osher's finite volume method using the P-variant integration path in the flux evaluation and third-order upwind for the determination of the constant states. This scheme is preferred, because it is second-order accurate, robust and apprehensive for the characteristic directions associated with the nonlinear equations. In addition, it has an excellent boundary treatment, its compact stencil facilitates computational speed up by domain decomposition, and the resulting semi-discrete system respects the physical conservation laws underlying the original shallow water problem. Moreover, this method combined with a so-called limiter ensures a smooth capturing of field variables with large gradients as opposed to the often applied spectral transform method.

A common prejudice against a finite volume method concerns its inefficiency due to a severe stability step size restriction when combined with an explicit time integration method for solving the resulting semi-discrete system on a uniform latitudinal-longitudinal (lat-lon) grid. This grid is standard in atmospheric applications and uses the lines of constant longitude (meridians) and latitude (parallels). This inefficiency has to do with the pole problem which includes all problems related to the non-existence of the longitudinal unit vector in the poles and the convergence of the meridians when approaching them.

We discuss two ways to resolve this pole-problem: (1) a combined lat-lon reduced grid with two stereocaps in the polar region, and (2) a linearly-implicit Rosenbrock time integration method (Ros3) combined with approximate matrix factorization (AMF) applied to the full Eulerian form of the SWEs on a uniform lat-lon grid.

The combined grid has no singular points. Furthermore, it alleviates the step size restriction by redistributing the grid cells over the sphere to obtain a more uniform cell distribution. This grid is advocated in combination with Osher's scheme and an explicit time integration method. Osher's scheme is used, because its boundary treatment facilitates the information transfer necessary at the grid interface between the different grid parts. Locally, the resulting scheme becomes first-order accurate in space.

The linearly-implicit method removes the stability step size restriction. Ros3 is A-stable and third-order accurate. A-stability is attractive, as it implies unconditional stability in the sense of Fourier-Von Neumann for stable linear problems. In practice, this indicates that much larger step sizes can be taken than with an explicit time integration method. However, Ros3 involves several expensive linear system solves per time step. To reduce these computational costs, we combine this method with approximate matrix factorization. This combination is also proven unconditionally stable when applied to the linearized SWEs on a uniform lat-lon grid. In addition, it remains third-order accurate.

Both remedies are validated by numerical experiments on a well-established test set from the literature. These results show that Ros3 with AMF is far more efficient than an explicit time integration method, viz., a third-order explicit Runge-Kutta method, even when the latter is applied to the semi-discrete SWEs on a combined grid.

Another important group of global atmospheric models consists of air quality models, which are used to describe the chemical composition of the atmosphere. These models are used to study the effects of air pollution. The chemical composition of the atmosphere is altered by chemical reactions, advection, diffusion, emissions and depositions, which are all included in the model. A numerical technique often applied in circulation and air quality models is operator splitting. This technique subdivides the full problem in a number of subprocesses, which can then be solved with different numerical techniques and step sizes suitable to the specific subprocess.

Unfortunately, the separate treatment of the subprocesses creates a splitting er-

ror. The magnitude of this error must be controlled and may not lead to an unstable solution process. To investigate these matters, an error expression is derived for a Strang splitting method which adopts a symmetrical order of reappearance to solve the different subprocesses. We focus on pure initial value problems. The analysis of the splitting error for coupled non-linear systems of partial differential equations is facilitated by the application of the Lie operator formalism. The error expressions are investigated in more detail for advection-diffusion-reaction (ADR) equations as used in air quality modeling. A theorem is derived which states under which circumstances the application of operator splitting to the ADR equations leads to a zero splitting error between advection, diffusion and chemistry. In practice, a splitting error is likely to occur.

Finally, a comparison is made between operator splitting and Ros3 with AMF. Both techniques simplify a numerical solution process to make it cost-effective. For the SWEs in spherical geometry, we investigate Ros3 with AMF and Strang splitting combined with a third-order Rosenbrock method to integrate the subprocesses in time. We are interested in the local error and the numerical dispersion relations. The latter demonstrate the influence of the numerical method on the characteristic waves of the shallow water problem. In meteorological practice, the correct representation of the advective (Rossby) waves is valued, because they describe an important part of the atmospheric dynamics. For characteristic step sizes, both methods do not significantly affect these waves. However, these results concern the local Cartesian form of the SWEs which is only valid in mid-latitude analysis. In a full spherical geometry, numerical results show that Ros3 with AMF is far more efficient than Strang splitting. The inefficiency of the latter method is caused by a severe accuracy step size restriction in the polar region.

Samenvatting

Sinds jaar en dag is de mens geïnteresseerd in weer en klimaat. Tegenwoordig geschieden weers- en klimaatvoorspellingen met behulp van circulatiemodellen. Atmosferische circulatie betreft de stroming van lucht in de atmosfeer gegeven via toestandsvariabelen zoals windsnelheid, luchtvochtigheid, temperatuur, etc. Een circulatiemodel bestaat uit een set van wiskundige vergelijkingen, die dit circulatieproces beschrijven. Naast data-assimilatie en fysische parametrisering vormt de numerieke dynamica een belangrijk onderdeel van dit soort modellen. Dit onderdeel houdt zich bezig met het numeriek oplossen van de primitieve vergelijkingen uit de hydrodynamica van de atmosfeer. De eisen aan een circulatiemodel zijn hoog. Zo vereist weersvoorspelling zo nauwkeurig mogelijke resultaten over een vaste tijdsperiode berekend binnen een zo kort mogelijk tijdsbestek. De nauwkeurigheid van een voorspelling hangt af van het specifieke model, de numerieke oplosmethode, de resolutie van het plaats-tijd rooster, de opgenomen data en het fysische parametriseringsschema. Circulatiemodellen zijn bovendien zeer complex en rekenintensief. De belangstelling voor efficiënte numerieke methoden op fijne roosters is dan ook groot. In dit proefschrift richt ik mij op de ontwikkeling van efficiënte numerieke methoden voor het oplossen van de ondiepwatervergelijkingen (SWEs) in een bolgeometrie op fijne roosters. Deze SWEs voldoen als een eerste beschrijving van de horizontale dynamica in een globaal circulatiemodel.

Voor de plaatsdiscretisatie van de SWEs maak ik gebruik van een eindige-volume methode. Op deze wijze voldoet het resulterende semi-discrete systeem aan de fysische behoudswetten die ten grondslag liggen aan de SWEs. Voor de fluxevaluatie pas ik Oshers upwindschema toe gecombineerd met een P-variant Osherpad en een derde-orde 1D-toestandsinterpolatie ($(\kappa = \frac{1}{3})$-schema). De resulterende methode heeft een aantal voordelen. Oshers upwindschema is een flux-difference-splitting schema implicerend dat lokaal een correcte informatieverspreiding langs de karakteristieken van het probleem plaatsvindt. Het schema is tweede-orde consistent en robuust. Het heeft een nette randbehandeling en een compact stencil, dat bijvoorbeeld de toepassing van een rekentijdbeperkende techniek zoals domeindecompositie toelaat. Bovendien, garandeert een combinatie van dit schema met een limiter een gladde representatie van variabelen met sterke gradiënten in tegen-

stelling tot de veelvuldig toegepaste spectrale transformatiemethode.

Een veelgenoemd nadeel van een eindige-volume methode betreft zijn verwachte inefficiëntie indien toegepast op een uniform breedtegraad-lengtegraad (lat-lon) rooster in combinatie met een expliciete tijdsintegratiemethode. In dat geval beperkt stabiliteit de toelaatbare tijdstap. Een uniform lat-lon rooster is standaard in atmosferische toepassingen en berust op roosterlijnen van constante breedte- (breedtecirkel) en lengtegraad (meridiaan). Deze inefficiëntie valt onder het poolprobleem. Deze verzamelnaam omvat alle problemen gerelateerd aan de singulariteit van de longitudinale eenheidsvector in de pool en de convergentie van de meridianen in de richting van de pool. Twee mogelijke oplossingen van het poolprobleem zijn onderzocht: (1) Een gecombineerd lat-lon gereduceerd rooster met twee stereokappen in de poolstreek. (2) Een lineair-impliciete Rosenbrock tijdsintegratiemethode (Ros3) gecombineerd met benaderende matrixfactorisatie (AMF) toegepast op een uniform lat-lon rooster.

Het gecombineerde rooster bevat geen singuliere punten. De roosterverdeling is zodanig dat een opeenhoping van cellen in de poolstreek wordt voorkomen. De tijdstaprestrictie is zo aanzienlijk gereduceerd. Dit rooster wordt gecombineerd met Oshers schema en een expliciete tijdsintegratiemethode. De nette randbehandeling van Oshers schema waarborgt het behoud van massa en impuls aan de interface tussen de verschillende roosterdelen. De impliciete Ros3-methode is A-stabiel en derde-orde nauwkeurig. A-stabiliteit is aantrekkelijk, omdat deze eigenschap onvoorwaardelijke stabiliteit impliceert in de zin van Fourier-Von Neumann voor stabiele lineaire problemen. In de praktijk impliceert deze eigenschap veelal dat de toelaatbare tijdstap aanzienlijk groter is dan met een expliciete tijdsintegratiemethode. De combinatie met AMF is essentieel voor efficiëntie. Deze techniek reduceert de rekenkosten geassocieerd met de oplossing van de lineaire systemen per tijdstap. Ros3 met AMF behoudt derde-orde nauwkeurigheid en A-stabiliteit. Beide remedies zijn beoordeeld op basis van numerieke experimenten op een gerenommeerde testset uit de literatuur. Deze resultaten tonen aan dat Ros3 met AMF vele malen efficiënter is dan een expliciete tijdsintegratiemethode zoals bijvoorbeeld een derdeorde expliciete Runge-Kutta methode, zelfs wanneer deze methode wordt toegepast op een gecombineerd rooster.

Circulatiemodellen vormen veelal de motor achter luchtkwaliteitsmodellen. Deze modellen beschrijven de chemische samenstelling van de atmosfeer. Deze is onderhevig aan chemische reacties, transport door wind, diffusie, emissies en deposities. Een veelvuldig toegepaste numerieke techniek in circulatie- en luchtkwaliteitsmodellen is operator splitting. Operator splitting vereenvoudigt het numerieke oplosproces door het op te splitsen in subproblemen en die vervolgens in een voorgeschreven volgorde op te lossen met een numerieke techniek toegespitst op het specifieke deelprobleem. De aparte behandeling van de subproblemen creëert echter een splitfout. De grootte van deze fout moet beperkt blijven en mag geen aanleiding geven tot een instabiel oplosproces. Een foutuitdrukking is afgeleid voor Strang-splitting, dat een symmetrisch oplospatroon voor de deelproblemen geeft. De analyse betreft gekop-

pelde niet-lineaire pure beginwaardeproblemen. Lie-operatoren vereenvoudigen de afleiding aanzienlijk. De foutuitdrukkingen zijn nader beschouwd voor de advectie-diffusie-reactie vergelijkingen kenmerkend voor luchtkwaliteitsmodellen. Een theorema is afgeleid, dat voorschrijft onder welke omstandigheden de toepassing van operator splitting op de advectie-diffusie-reactie vergelijkingen geen aanleiding geeft tot een splitfout.

Als laatste, vergelijk ik operator splitting met Ros3 met AMF. Beide technieken simplificeren het numerieke oplosproces. Als operator splitting beschouw ik Strang-splitting gecombineerd met Ros3 voor de tijdsintegratie van de subprocessen. De vergelijking betreft de SWEs in bolgeometrie, waarvoor de lokale fout en de numerieke dispersierelatie zijn afgeleid. Deze relatie beschrijft de invloed van de numerieke methode op de representatie van de karakteristieke golven van het ondiepwaterprobleem. In de meteorologie is een correcte weergave van de advectieve (Rossby) golf vereist. Deze golf is kenmerkend voor de atmosferische dynamica. Voor karakteristieke tijdstappen is de fout van de numerieke technieken op deze golven nihil. Deze resultaten gelden echter in een midden-breedtegraad synoptische analyse. In een volledige bolgeometrie is Ros3 met AMF veel efficiënter dan Strang-splitting. De onnauwkeurigheid van Strang-splitting in de poolstreek noodzaakt tot kleine tijdstappen.