IMEX Finite Volume Methods for the Shallow Water Equations

Dissertation
zur Erlangung des
Doktorgrades der Naturwissenschaften
(Dr. rer. nat.)

am Fachbereich Physik, Mathematik und Informatik
der Johannes Gutenberg-Universität
in Mainz

von

M. Sc. Georgij Bispen

geboren am 5.12.1985 in Riga (Lettland)

01.2015

1. Gutachter:
2. Gutachter:

Tag der mündlichen Prüfung: 27.02.2015

Abstract

Shallow water is defined as a constant density fluid in hydrostatic balance that is bounded from below by a rigid surface and from above by another fluid with negligible inertia. Moreover, the horizontal scale is much larger then the vertical one, e.g. a typical ratio is 0.05. Integrating the incompressible Navier-Stokes equations along the vertical axis one can obtain the shallow water equations (SWE). The SWE belong to the class of hyperbolic systems of partial differential equations (balance laws) that provide suitable approximations to the large-scale motion of oceans, rivers and the atmosphere. Thereby mass, momentum as well as energy and potential vorticity for smooth solutions are conserved. Here, two characteristic velocities are distinguished: the advection velocity, i.e. the velocity of mass transport, and the gravity wave speed, i.e. the velocity of surface waves, which carry energy and momentum. The Froude number is a reference number and is given by the fraction of the advection and the gravity waves reference velocity. It is typically very small for large-scale flows, e.g. 0.01.

Time-explicit finite volume methods belong to the most frequently used numerical schemes to solve hyperbolic balance laws. Consequently, we have to respect the CFL stability condition and the time increment is approximately proportional to the Froude number. Thus, low Froude numbers, say below 0.2, lead to very small time steps and result in high computational costs and dissipative solutions. Typical problems arising from meteorology or oceanography however concentrate on modelling of flow phenomena. Thus, the advection is the physical quantity of interest. Consequently a desirable stability condition would only depend on the advection velocity and not on gravity waves. Then, the time steps would increase approximately with the factor of the reciprocal of the Froude number, e.g. the time steps could be 100 times larger if the Froude number is 0.01.

In the case of a constant bottom topography, it is well-known that solutions of the SWE converge to solutions of the *lake equations/ zero Froude number SWE* as the Froude number approaches zero, if suitable initial data are provided. In this limiting process, the equations change their type from hyperbolic to mixed hyperbolic-elliptic. Thus, for small Froude numbers the convergence order of standard numerical schemes may decrease or even the numerical solution may break down.

Oceanographic and atmospheric flows are typically small perturbations of an underlying equilibrium state. It is well-known that numerical schemes for balance laws have to preserve certain equilibrium states exactly, otherwise spurious motion may occur. Thus, the approximation of source terms is crucial. Numerical schemes that preserve equilibrium states are called *well-balanced*.

The aim of this thesis is to develop well-balanced numerical schemes for the SWE, so that a stability constraint on time steps is independent of the Froude number. Moreover, the convergence order should be uniform with respect to the Froude number. Since solutions of the SWE converge to solutions of the lake equations as the Froude number approaches zero, numerical schemes with uniform convergence rates with respect to the Froude number have to provide a consistent approximation of the lake equations as the Froude number approaches zero. These schemes are called *asymptotic preserving*. In order to derive such schemes we rewrite the SWE in an alternative form. Here we consider the free surface elevation with respect to a still water level as a variable instead

of the fluid depth. We split the alternative SWE in a stiff, linear part governing the fast waves and a non-stiff part governing the remaining slow flow. Both, the stiff and non-stiff subsystems are hyperbolic. Computing the stiff parts implicitly and the non-stiff ones explicitly we obtain a stability constraint that only depends on the advection velocity. We use IMEX Runge-Kutta and IMEX multi-step schemes to obtain a semi-implicit time-discretisation and show that the resulting semi-discrete schemes are well-balanced. Further, we study the asymptotic preserving property of the semi-discrete schemes. Particularly, we consider the IMEX Euler, ARS(2,2,2), RK2CN Runge-Kutta schemes and a consistent BDF-type multi-step scheme (SBDF).

We use the finite volume space discretisation and propose two approaches to solve an initial value problem: the straight and the elliptic approach. In the straight approach we approximate the surface integrals resulting from the finite volume space discretisation by numerical flux functions. To approximate the stiff part we use central finite differences or approximated evolution operators, that take into account multi-dimensional wave propagation. The non-stiff part is approximated by standard compressible solvers, e.g. the Rusanov flux. We obtain the solution at a new time step by solving the resulting linear systems. Here the source term approximation is motivated by the well-balanced property. More precisely, the "lake at rest"- equilibrium state will be preserved exactly. Using the elliptic approach, we compute the free surface elevation at a new time step by solving an elliptic equation. The elliptic equation can be derived by plugging the momentum equation into the continuity equation. Afterwards, the momentum update can be computed in an explicit way. The stiff terms are approximated by central finite differences in the elliptic approach, whereas the non-stiff terms are approximated by means of the Rusanov flux, as in the straight approach. Here, a special source term discretisation is not necessary in order to achieve a well-balanced scheme. We show that some of our derived IMEX finite volume schemes are indeed well-balanced and asymptotic preserving under certain conditions. To this end we apply the theory of circulant matrices.

The well-balanced and asymptotic preserving properties for both, first and second order methods, are verified by numerical tests. We show the uniform convergence with respect to the Froude number is only achieved when the stiff terms are approximated by central finite differences using the straight approach or elliptic approach which can be derived from the straight approach. Moreover, corresponding schemes are stable and asymptotic preserving.

Kurzdarstellung

Flachwasser ist per Definition ein Fluid mit konstanter Dichte, dass sich in hydrostatischem Gleichgewicht befindet. Ferner befindet sich unterhalb des Fluids ein fester und unbeweglicher Boden. Oberhalb des Fluids befindet sich ein anderes Fluid mit vernachlässigbarer Reibung. Zudem ist die horizontale Ausbreitung wesentlich größer als die Vertikale - beispielsweise mit dem Größenverhältnis 0.05. Durch die Integration der Navier-Stokes Gleichung entlang der vertikalen Achse erhalten wir ein vereinfachtes Modell: die Flachwassergleichungen (SWE). Die SWE sind ein hyperbolisches System von partiellen Differentialgleichungen (Bilanzgleichungen), die großskalige Strömungen in

Ozeanen, Flüssen und der Atmosphäre adäquat modellieren. Dabei wird Masse, Impuls und im Falle einer glatten Lösung die Energie und potentielle Vortizität erhalten. In dem vereinfachten System der SWE lassen sich zwei Geschwindigkeiten unterscheiden: Die Advektionsgeschwindigkeit, d.h. die Geschwindigkeit mit der Masse transportiert wird, und die Geschwindigkeit von Schwerewellen, d.h. die Geschwindigkeit der Wellen auf der Oberfläche des Fluids. Letztere transportieren Energie und Impuls. Die Froude-Zahl ist eine Kenngröße der Strömung und ergibt sich als das Verhältnis von der Referenzadvektionsgeschwindigkeit zu der Referenzgeschwindigkeit der Schwerewellen. Sie ist sehr klein für großskalige Strömungen, z.B. 0.01.

Explizite Finite-Volumen-Verfahren gehören zu den meistgenutzten numerischen Verfahren zum Lösen von hyperbolische Bilanzgleichungen. Dabei muss die CFL-Stabilitätsbedingung eingehalten werden, wodurch die Zeitschrittweite sich näherungsweise proportional zur Froude-Zahl verhält. Typische Problemstellungen in der Ozeanographie und Meteorologie sind Beschreibungen von Strömungen mit sehr kleinen Froude-Zahlen. Dies führt zu sehr kleinen Zeitschritten und resultiert in langen Rechenzeiten und dissipativen Resultaten. Ferner ist die Advektion die bedeutende Größe. Somit ist eine Stabilitätsbedingung die nur von der Advektion und nicht von der Geschwindigkeit der Schwerewellen abhängt wünschenswert. Dadurch würden für kleine Froude-Zahlen die Zeitschritte ungefähr um den Faktor des Kehrwertes der Froude-Zahl wachsen. Bei der Froude-Zahl 0.01 wären die Zeitschritte etwa 100 Mal größer.

Im Falle einer konstanten Bodentopographie ist bekannt, dass Lösungen der SWE im Grenzwert Froude-Zahl gegen Null zu Lösungen der *Froude-Zahl Null Flachwassergleichungen* konvergieren, falls die Anfangswerte adäquat sind. Hierbei wechseln die Gleichungen ihren Typ von hyperbolisch zu hyperbolisch-elliptisch. Im Übergangsregime kleiner Froude-Zahlen kann die Konvergenzgeschwindigkeit von numerischen Verfahren abnehmen oder sogar das Verfahren selbst abbrechen.

Ozeanographische und atmosphärische Strömungen sind meist kleine Störungen eines zugrunde liegenden Gleichgewichtszustandes. Es ist bekannt, dass numerische Verfahren zur Lösung von Bilanzgleichungen gewisse Gleichgewichtszustände exakt erhalten müssen damit keine Strömungen künstlich durch das numerische Verfahren erzeugt werden. Daher ist die Approximation der Quellterme bei Bilanzgleichungen wesentlich. Numerische Verfahren die gewisse Gleichgewichtszustände erhalten heißen *ausbalanciert*.

Ziel dieser Arbeit ist die Herleitung numerischer, ausbalancierter Verfahren zur Lösung der SWE, so dass eine Stabilitätsbedingung nur von der Advektion abhängt und die Konvergenzordnung der Verfahren gleichmäßig bezüglich der Froude-Zahl ist. Wegen der Konvergenz der SWE gegen die Froude-Zahl Null Flachwassergleichungen müssen gleichmäßig bezüglich der Froude-Zahl konvergente numerische Verfahren im Grenzfall Froude-Zahl gegen Null eine konsistente Approximation dieser Gleichungen sein, d.h. die Verfahren sind *asymptotisch erhaltend*. Um solche Verfahren herzuleiten schreiben wir die SWE in eine alternative Form um. Hierbei wird anstatt der Fluidtiefe die Erhebung des Fluids zu einer konstanten Wasseroberfläche als Variable betrachtet. Die alternativen SWE teilen wir in einen steifen, linearen Teil, der die schnellen Strömungen modelliert, und einen nicht-steifen Teil, der die übrigen langsamen Strömungen modelliert. Hierbei sind das steife und nicht-steife Teilsystem hyperbolisch. Durch eine implizite Behandlung des steifen Teils und eine Explizite des Nichtsteifen ist die resultierende CFL-Stabilitätsbedingung nur von der Advektion abhängig. Wir verwenden IMEX

Runge-Kutta und IMEX Mehrschrittverfahren zur semi-impliziten Zeitdiskretisierung und zeigen, dass die resultierenden semi-diskreten Verfahren ausbalanciert sind. Ferner untersuchen wir ob die Verfahren asymptotisch erhaltend sind. Die IMEX Euler, ARS(2,2,2), RK2CN Runge-Kutta Verfahren sowie ein BDF-artiges Mehrschrittverfahren (SBDF) werden speziell betrachtet.

Zur Ortsdiskretisierung nutzen wir das Finite Volumen Verfahren. Hierbei stellen wir zwei Methoden zur Lösung eines Anfangswertproblems vor: die direkte und die elliptische Methode. Bei der direkten Methode werden die aus der Finite Volumen Ortsdiskretisierung resultierenden Oberflächenintegrale durch numerische Flussfunktionen approximiert. Für den steifen Anteil verwenden wir entweder zentrale finite Differenzen oder approximative Evolutionsoperatoren, die auf der Theorie der Bicharakteristiken beruhen und alle unendlich viele Richtungen der Informationsausbreitung berücksichtigen. Der nicht-steife Teil wird durch kompressible Standardverfahren approximiert, etwa mittels des Rusanov-Flusses. Die Lösung zu einem neuen Zeitschritt erhalten wir durch Lösen der resultierenden Gleichungssysteme. Die Approximation des Quellterms wird dadurch motiviert, dass die resultierenden Verfahren ausbalanciert sein sollen. Genauer soll der Ruhe Equilibriumzustand mit Null-Geschwindigkeit exakt gelöst werden. Bei der elliptischen Methode wird die Erhebung des Fluids in einem neuen Zeitschritt durch die Lösung einer elliptischen Gleichung berechnet. Diese elliptische Gleichung kann durch Einsetzen der Impulserhaltung in die Kontinuitätsgleichung hergeleitet werden. Anschließend kann der Impuls zum neuen Zeitschritt explizit ausgerechnet werden. Bei der elliptischen Methode wird der steife Teil durch zentrale finite Differenzen approximiert, während der nicht-steife Anteil, wie bei der direkten Methode, durch den Rusanov-Fluss berechnet wird. Eine besondere Quelltermapproximation ist hier nicht notwendig damit die resultierenden Verfahren ausbalanciert sind. Wir zeigen, dass ein Teil der hergeleiteten Verfahren in der Tat unter gewissen Voraussetzungen ausbalanciert und asymptotisch erhaltend sind. Dazu benutzen wir unter anderem die Theorie zirkulanter Matrizen.

Numerische Simulationen bestätigen die theoretischen Resultate bezüglich der Ausbalanciertheit und asymptotischer Erhaltung. Es werden Verfahren erster und zweiter Ordnung untersucht. Dabei stellten wir fest, dass nur bei Diskretisierung der steifen Terme mittels zentraler finiter Differenzen gleichmäßige Konvergenz bezüglich der Froude-Zahl erreicht wird. Hierbei sind nur diejenigen Verfahren stabil und asymptotisch erhaltend, die die direkte Methode verwenden, oder in die direkte Methode umgeschrieben werden können.

Contents

δ_{ij}	$:= \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$ Kronecker delta
$\mathbf{e}_i$	canonical standard basis vector $(0, ..., 0, 1, 0, ..., 0)^T$; $(\mathbf{e}_i)_j = \delta_{ij}$
$\mathbf{n}$	outer unit normal vector
$\mathbf{u} \cdot \mathbf{v}$	the euclidean scalar product between the vectors $\mathbf{u}$ and $\mathbf{v}$
$\mathbf{1}$	the vector $(1, ..., 1)^T$
$\mathbb{1}$ or $\mathbb{1}_d$	identity matrix (in $\mathbb{R}^{d \times d}$)
$A \otimes B$	$:= \begin{bmatrix} a_{11}B & ... & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & ... & a_{mn}B \end{bmatrix}$ Kronecker product
∇f	$:= (f_{x_1}, ..., f_{x_d})^T$ gradient of f
$\nabla \cdot \mathbf{u}$	$:= (u_1)_{x_1} + ... + (u_d)_{x_d}$ divergence of $\mathbf{u}$
$\nabla \cdot \begin{bmatrix} \mathbf{u}_1^T \\ \vdots \\ \mathbf{u}_p^T \end{bmatrix}$	$:= \begin{bmatrix} \nabla \cdot \mathbf{u}_1 \\ \vdots \\ \nabla \cdot \mathbf{u}_p \end{bmatrix}$
$\nabla \cdot (\mathbf{u} \otimes \mathbf{u})$	$:= \nabla \cdot (\mathbf{u} \otimes \mathbf{u}^T)$ convective terms
$[g]_{x=a}^{x=b}$	$:= g\|_{x=b} - g\|_{x=a}$
S^{d-1}	$:= \left\{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1 \right\}$
$H^s(\Omega)$	Sobolev(-Slobodetski) space on the domain Ω for $s \geq 0$
$\langle \mathbf{u}_1, ..., \mathbf{u}_p \rangle$	$:= \left\{ \sum_{i=1}^{p} c_i \mathbf{u}_i : c_i \in \mathbb{R} \right\}$
$\mathcal{N}(A), \mathcal{R}(A)$	kernel and range of the matrix A

Notations. Here $\mathbf{u} = (u_1, ..., u_d), \mathbf{v}, \mathbf{u}_1, ..., \mathbf{u}_p$ are vector fields $\mathbb{R}^d \to \mathbb{R}^d$, $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{p \times r}$ are matrices and $f : \mathbb{R}^d \to \mathbb{R}$, $g : \mathbb{R}^m \to \mathbb{R}^n$ are functions.

1. Introduction

Free-surface[1] flows under gravity are present in a large set of problems: sloshing in fuel tanks in rocket technology, tides in ocean, breaking of waves on shallow beaches, roll waves in open channels, flood waves in rivers, dam break modelling, atmospheric flows, tsunamis, cf. [1, 2, 111, 119] and the references therein. Assume that the governing model equations for fluid flow is given, the main difficulty is to compute the free-surface. It is a boundary and therefore suitable boundary conditions have to be satisfied. But the location of the boundary is unknown and thus also the domain where the equations have to be solved is unknown, too. In the special case of shallow water[2] this problem can be treated by introducing the water depth as a new variable and integrate the conservation laws of mass and momentum along the vertical axis, cf. [111, 112], leading to the shallow water equations[3] (SWE). The SWE read

$$h_t + \nabla \cdot (h\mathbf{u}) = 0, \tag{1.1a}$$

$$(h\mathbf{u})_t + \nabla \cdot (h\mathbf{u} \otimes \mathbf{u}) + gh\nabla h = -gh\nabla \tilde{b}, \tag{1.1b}$$

where h denotes the water depth, $\tilde{b}$ the bottom topography, g the gravitation constant and $\mathbf{u} = (u_1, u_2)$ the velocity field averaged along the vertical axis, cf. Figure 2.1. Here (1.1a) is the continuity equation, i.e. mass conservation, and (1.1b) conserves the momentum. Though, the vertical velocity does not appear in the SWE it is neither zero nor constant, since otherwise the water may not cross a bottom hump. If the solution is smooth, the SWE conserve the energy $E = (h\mathbf{u} \cdot \mathbf{u} + gh^2)/2$ and the potential vorticity ω/h with the vorticity $\omega = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}$, [88], too. Shallow water models are suitable approximations of large-scale oceanographic and atmospheric flows or river flows, cf. [88, 111, 112].

From a mathematical point of view, the SWE belong to the class of hyperbolic system of balance laws, that are typically solved numerically by (time-)explicit finite volume schemes. Then, the CFL-condition [28] has to be satisfied, i.e. the numerical domain of dependence must cover the real domain of dependence. In case of the SWE the transported information is mass, momentum and energy. The mass is transported by the advection velocity $\mathbf{u}$, whereas energy and momentum are carried by the gravity waves (surface waves) with the velocity $c = \sqrt{gh}$ [73, 88, 112]. Thus the CFL-stability condition for the SWE

$$\max\left\{\frac{|u_1| + c}{\Delta x}, \frac{|u_2| + c}{\Delta y}\right\} \Delta t \leq CFL < 1 \tag{1.2}$$

[1] Typically a fluid is bounded by some other fluid from above. If we speak about a free-surface, we think of a fluid bounded from above by another fluid with negligible inertia.

[2] An incompressible fluid in hydrostatic balance, bounded from above by a free-surface and from below by a rigid surface is called shallow water, if the horizontal scale L is much larger then any vertical scale H, $H/L < 0.05$ [74].

[3] In the French scientific community the SWE are also called St. Venaint equations. Note that St. Venaint derived only the one dimensional version [39].

has to be satisfied for time-explicit approximations using a grid with widths $\Delta x, \Delta y$ in space and time increment Δt.

Typical problems arising from meteorology or oceanography however concentrate on modelling of flow phenomena, e.g. weather forecasting. Thus the advection is the physical quantity of interest. Relevant advection velocity reference values concerning large-scale motions are $10m/s$ in the atmosphere and $0.1 - 2m/s$ in the ocean, cf. [73, 88, 112]. Gravity waves may exceed the speed of $200m/s$ in the ocean [73] and $300m/s$ in the atmosphere [88]. The ratio between advection and gravity wave reference velocities is called the *Froude number* $\varepsilon = u_{ref}/c_{ref}$. It is a characteristic number of the flow with values that rarely exceed 0.2 [119] and are of order 10^{-2} for large-scale phenomena [8, 73, 88, 112]. A desirable restriction on time steps would be to consider only the propagation of information that is of concern, i.e. the advection. However, for low Froude numbers the velocity of gravity waves dictates the time increment Δt:

$$\begin{aligned} \frac{1}{\varepsilon}\frac{u_{ref}\Delta t}{\min\{\Delta x, \Delta y\}} \approx \left(1+\frac{1}{\varepsilon}\right)\frac{u_{ref}\Delta t}{\min\{\Delta x, \Delta y\}} &= \frac{(u_{ref}+c_{ref})\Delta t}{\min\{\Delta x, \Delta y\}} \\ &\approx \max\left\{\frac{|u_1|+c}{\Delta x}, \frac{|u_2|+c}{\Delta y}\right\}\Delta t \leq CFL < 1. \end{aligned} \tag{1.3}$$

Consequently the time step Δt is several orders of magnitude smaller then the advection time scale. For example a vortex that moves with the advection reference velocity u_{ref} needs around $1/\varepsilon$, e.g. 100, time steps to pass a single cell. Therefore computations in the low Froude number regime with the explicit finite volume schemes are computationally expensive, and use many small times steps that yield to dissipative results, cf. [55].

In order to overcome the strong stability condition on the time step (1.3), i.e. to use larger time steps, several *split-explicit*, *semi-implicit* and implicit schemes have been proposed in literature [14, 37, 44, 45, 54, 55, 64, 65, 66, 90, 91, 95, 98, 99, 108], as well as [4, 15, 92, 118] for recent results. Using fully implicit schemes circumvents successfully the CFL-condition, but a large nonlinear system is introduced that is computationally very expensive, cf. [55]. Another strategy is to split the SWE into a stiff subsystem governing the fast waves and another non-stiff one responsible for the remaining slow flow. Here, the stiff subsystem is preferably linear and thus easy to solve. Applying an explicit scheme with smaller time steps to the stiff subsystem one obtains a split-explicit scheme. Another possibility is to treat the stiff part implicitly and the non-stiff one explicitly. Consequently the stability condition depends only on the velocity of the slow waves. Since the problem is fully nonlinear, it is crucial and nontrivial how to split the governing equations into subsystems modelling slow and fast waves.

If the bottom topography is constant, the SWE are equivalent to the isentropic Euler equations with the pressure-density relation $p(\rho) = g\rho^2/2$ and the Froude number coincides with the Mach number. Thus, the SWE admit "compressible" effects. The Mach number is a measure of the compressibility of the flow in the isentropic Euler equations. If the Mach number $\varepsilon = \mathcal{O}(1)$ the flow is in the compressible regime. If the Mach number $0 < \varepsilon \ll 1$, then the flow is in the weakly compressible regime. For $\varepsilon \to 0$ we speak about the incompressible regime. Analogously we speak formally about the compressible, weakly compressible and incompressible regimes for the SWE if the Froude number $\varepsilon = \mathcal{O}(1)$ or $\varepsilon \ll 1$, even though the incompressible fluid model is used

for the description of the SWE. It is a well-known result due to Klainerman and Majda [62, 63] and Ebin [40] that the isentropic Euler equations converge to the incompressible Euler equations as the Mach number approaches zero, if suitable assumptions are satisfied. The equivalent counterpart of the incompressible Euler equations are the *zero Froude number SWE* that are also referred to as the *lake equations*, cf. (2.34). Thus the SWE (1.1) should converge to the zero Froude number SWE as the Froude number approaches zero. Moreover the hyperbolic SWE/ isentropic Euler equations change to the hyperbolic-elliptic type incompressible Euler equations/ zero Froude number SWE. Further, the velocity $\mathbf{u}$ is restricted by a divergence constraint, cf. [21, 62, 63]. Therefore compressible solvers may break down or experience poor convergence in the weakly compressible/ incompressible regime, cf. [14]. Consequently numerical schemes with uniform convergence order with respect to the Froude number have to behave like a compressible solver in the compressible regime, i.e. they provide non-oscillatory solutions and good resolution of shock-type discontinuities; whereas in the low Froude number limit $\varepsilon \to 0$ the numerical schemes have to approximate the limit equations, i.e. zero Froude number SWE/ incompressible Euler equations. If there is no stability condition needed, that depends on the Froude number, such schemes are *asymptotic preserving*. More precisely, we call a numerical scheme for the compressible regime asymptotic preserving, if it provides a consistent approximation of the zero Froude-number equations as the Froude-number approaches zero and there is no stability constraint that depends on the Froude-number. The concept of asymptotic preserving schemes was introduced by Jin [57] for multi-scale kinetic equations as well as for hyperbolic balance laws [37, 92, 108], see also the review paper [58].

Let us point out, that the source term approximation for balance laws is crucial. The description of oceanographic and atmospheric flows is the main application for the SWE. Here, the solution is typically a perturbation of an underlying equilibrium state. It is well-known, that numerical schemes used to solve balance laws have to preserve certain equilibrium states exactly by properly balancing the fluxes and source terms, otherwise non-physical behaviour like spurious waves may occur. Such schemes are called well-balanced or satisfying the C-property. Various techniques to design well-balanced schemes have been presented in literature [7, 20, 25, 49, 72, 76, 85]. The C-property was introduced in [13]. For the SWE, the so-called *lake at rest solution*, where the water level is constant and no flow occurs, i.e. $h + \tilde{b}$ =const., $\mathbf{u} = 0$, is an important equilibrium state. Note, that many of relevant flows are small deviations of this state. Let us point out that typically the Coriolis force plays an important role for large-scale phenomena. The corresponding equilibrium state is called the *geostrophic equilibrium*, cf. [112]. But since we want to focus on low Froude numbers we neglect the Coriolis force for the sake of simplicity.

The aim of this thesis is to develop and analyse semi-implicit, asymptotic preserving, well-balanced numerical schemes to solve the SWE for all Froude numbers, where the CFL-stability condition is independent of the Froude number. Thus the scheme should use large time steps analogous to a purely advective flows with no gravity waves. Moreover the order of convergence should be uniform with respect to the Froude number. To this end we use an alternative formulation of the SWE, where the free surface elevation above a still water level is considered as an unknown variable instead of the water depth. Similar formulations have been considered in [44, 72, 79, 80, 102, 101]. We benefit by

using the above mentioned alternative formulation in the following way: the water depth is decomposed into a given lake at rest depth $h^{(0)}$ and its small perturbation - the free surface elevation. This helps to avoid cancellation errors for very low Froude numbers, e.g. 10^{-7}, cf. [104].

Following Giraldo et al. [44] we split the alternative SWE into a linear, hyperbolic, stiff system and a hyperbolic, non-stiff one. Here, the linear system is the first order wave equation system, analogous to linear acoustic equations, with the wave velocity $\sqrt{g \cdot h^{(0)}} \approx c$. We approximate the stiff linear part implicitly and the non-stiff one explicitly. To this end *IMEX (implicit-explicit)* Runge-Kutta or IMEX multi-step schemes are used for the time discretisation, cf. [5, 6, 17, 18, 19]. Consequently we obtain a time step restriction of the type

$$\max\left\{\frac{|u_1|}{\Delta x}, \frac{|u_2|}{\Delta y}\right\}\Delta t \leq CFL_u < 1, \tag{1.4}$$

that is independent of the gravity wave speed. Since this CFL condition only depends on the advection velocity $\mathbf{u}$, we refer to the corresponding CFL number as CFL_u. We show that suitable IMEX schemes are well-balanced and enforce the asymptotic preserving property for the semi-discrete schemes. The spatial discretisation is realised by the finite volume method. In order to advance the numerical solution in time two approaches are introduced: the *straight* and the *elliptic approach*. The straightforward solving of the resulting linear systems is the straight approach, cf. [15, 44]. Another way to compute time evolution is to use an underlying elliptic equation for the surface elevation, cf. [4, 37, 92, 108]. In the straight approach, the implicit surface integrals associated with the fast waves are approximated either by central finite differences or in a genuinely multidimensional way. The multidimensional approximation is realised by applying multidimensional approximated evolution operators, which are based on the theory of bicharacteristics and take all of the infinitely many directions of wave propagation into account, cf. [3, 15, 81, 82, 83]. Suitable source term discretisations are presented to enforce the well-balanced property of the fully discrete schemes. In case of the elliptic approach, implicit terms are approximated by central finite differences. The surface integrals of the resulting non-linear part, treated explicitly in time, can be computed by standard shock-capturing schemes for both approaches. We consider first and second order discretisations in time and space. Numerical experiments demonstrate that the schemes developed in this thesis are well-balanced. We also investigate and discuss the question of accuracy, stability and the asymptotic preserving property.

The thesis is organised in the following way: In Chapter 2 we introduce the governing equations. Particularly, we derive the alternative SWE and present a suitable splitting. In Chapter 3, we introduce exact and approximated evolution operators using the theory of bicharacteristics. The IMEX schemes as well as the straight and elliptic approaches are introduced in Chapter 4. Particularly we consider the IMEX Euler, ARS(2,2,2), RK2CN IMEX Runge-Kutta schemes and the SBDF IMEX multi-step scheme. We finish this chapter by discussing the well-balanced property of the semi-discrete IMEX schemes. In Chapter 5 the asymptotic preserving property of semi-discrete IMEX schemes is shown. In Chapter 6 we introduce the finite volume method and derive the IMEX finite volumes scheme as well as suitable source term discretisations. The source term discretisation

is motivated by the well-balanced property. Then we analyse the asymptotic preserving property of the IMEX finite volume schemes in Chapter 7. The numerical results are presented in Chapter 8. Finally we conclude in Chapter 9 the thesis and formulate open problems. In the appendix A an application of the theory of bicharacteristics to the multidimensional non-hydrostatic Euler equations is shown. In particular we derive exact and approximate evolution operators for a multidimensional linear subsystem, which is stiff with respect to small Mach number.

2. Governing equations

In this chapter we introduce the governing equations used to describe the shallow water flow. We will analyse this model briefly and present its eigenstructure, as well as the so-called zero Froude number equations.

First, we recall the definition of a hyperbolic balance law in Section 2.1. The SWE in dimensional and non-dimensional form are presented in Section 2.2 and the zero Froude number SWE in Section 2.3. The numerical schemes, that will be presented in the following chapters are based on an alternative formulation of the SWE, cf. [15, 44, 72, 79, 80, 101, 102], where the free surface elevation above a still water level is considered instead of the water depth. We present this formulation in Section 2.4. In Section 2.5 we derive the low Froude number limit equations of the alternative SWE formulation. To this end we use asymptotic analysis, cf. [61]. We overcome the strong time restriction due to the CFL-condition (1.2) by splitting the flux function of the SWE into a linear part, governing the fast waves, and a nonlinear one, cf. [15, 44]. The fast waves are treated implicitly and therefore do not need to be considered in the CFL-condition. Thus, the time restriction is independent of the Froude number. We present and discuss the linear/nonlinear splitting in Section 2.6. At the end of this chapter, in Section 2.7, we comment on the eigenstructure of the shallow water systems and the stiff and non-stiff subsystems that have been previously derived. In particular we demonstrate that the stiff and non-stiff subsystems of the alternative SWE are indeed stiff and non-stiff.

2.1. Hyperbolic balance laws

Hyperbolic partial differential equations (PDEs) are used in many areas, such as meteorology, oceanography, aviation and astronautics in order to describe time evolution of systems with finite speed of information propagation. Typically, they arise from the conservation of physical quantities like mass, momentum or energy and are given in the form of a balance law

$$\mathbf{w}_t + \nabla \cdot \mathcal{F}(\mathbf{x}, t, \mathbf{w}) = K(\mathbf{x}, t, \mathbf{w}), \tag{2.1a}$$

where

$$\mathbf{w} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^p, \quad (\mathbf{x}, t) \mapsto \mathbf{w}(\mathbf{x}, t) \tag{2.1b}$$

$$\mathcal{F} : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}^{p \times d}, \quad (\mathbf{x}, t, \mathbf{w}) \mapsto [F_1(\mathbf{x}, t, \mathbf{w}), ..., F_d(\mathbf{x}, t, \mathbf{w})], \tag{2.1c}$$

$$K : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}^p, \quad (\mathbf{x}, t, \mathbf{w}) \mapsto K(\mathbf{x}, t, \mathbf{w}). \tag{2.1d}$$

Here, $\mathbf{w}$ is a vector containing conserved quantities, $\mathcal{F}$ is a flux function that consists of the columns F_i and K is a source term. We have

$$\nabla \cdot \mathcal{F} = \sum_{i=1}^{d} \left[\frac{dF_i}{d\mathbf{w}} \mathbf{w}_{x_i} + \frac{dF}{dx_i} \right], \tag{2.2}$$

if $\mathcal{F} \in (C^1(\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^p))^{p \times d}$. Denoting the Jacobians of $dF_i/d\mathbf{w}$ in (2.2) by A_i we obtain the system (2.1a) in quasi-linear form

$$\mathbf{w}_t + \sum_{i=1}^{d} A_i(\mathbf{x}, t, \mathbf{w}) \mathbf{w}_{x_i} = K(\mathbf{x}, t, \mathbf{w}) - \sum_{i=1}^{d} \frac{dF}{dx_i}(\mathbf{x}, t, \mathbf{w}) =: Q(\mathbf{x}, t, \mathbf{w}). \tag{2.3}$$

Note that not every equation of the form (2.1a) or (2.3) is hyperbolic.

Definition 2.1.1

1. *The matrix pencil of a system of the form (2.3) is the matrix valued function*

$$\mathbb{P}(\mathbf{w}, \mathbf{n}) = \sum_{i=1}^{d} \frac{dF_i}{d\mathbf{w}}(\mathbf{x}, t, \mathbf{w}) n_i, \quad \mathbf{n} = (n_1, ..., n_d)^T \in \mathbb{R}^d. \tag{2.4}$$

2. *Let $\Omega \subset \mathbb{R}^d, D \subset \mathbb{R}^p$ be domains. The system (2.3) is hyperbolic in $\Omega \times D$, if all eigenvalues of the corresponding matrix pencil $\mathbb{P}$ are real for all $\mathbf{x} \in \Omega, t \in \mathbb{R}, \mathbf{w} \in D$ and $\mathbf{n} \in \mathbb{R}^d$. We say that a system (2.3) is hyperbolic, if it is hyperbolic in $\Omega \times \mathbb{R}^p$.*

3. *System (2.1a) is hyperbolic, if the corresponding system (2.3) is hyperbolic.*

4. *A hyperbolic system (2.3) is strictly hyperbolic, if all matrix pencil eigenvalues are simple.*

5. *A hyperbolic system (2.3) is diagonally hyperbolic, if the matrix pencil is diagonalisable.*

Remark 2.1.2 *Note that some textbooks call a quasi-linear balance law (2.3) hyperbolic, if it is diagonally hyperbolic.*

It is well-known, that solutions of hyperbolic balance laws may become discontinuous or *blow up* in finite time - even for smooth initial data, cf. [12, 56]. Therefore the concept of *weak solutions* needs to be introduced.

Definition 2.1.3 *Let $\Omega = \mathbb{R}^d$, $\mathbf{w}^0 \in L^\infty_{loc}(\mathbb{R}^d)$. A function $\mathbf{w} \in L^\infty_{loc}(\mathbb{R}^d \times [0, \infty])$ is a weak solution of (2.1a) with $K = 0$ and the initial value $\mathbf{w}^0$, if*

$$\int_0^\infty \int_{\mathbb{R}^d} \mathbf{w}\phi_t + \mathcal{F}(\mathbf{w}) \nabla \phi \, d\mathbf{x} \, dt + \int_{\mathbb{R}^d} \mathbf{w}^0(\mathbf{x}) \phi(\mathbf{x}, 0) \, d\mathbf{x} = 0 \tag{2.5}$$

for all $\phi \in C_0^\infty(\mathbb{R}^d \times [0, \infty))$.

By using the concept of weak solutions we are able to describe a larger range of discontinuous phenomena, e.g. shock and contact discontinuities. But weak solutions are not unique. In order to take the "physical right" solution the concept of *entropy solutions* is used.

Definition 2.1.4 *We call $\eta \in C^1(\mathbb{R}^p)$ the entropy of the system (2.1a), if there exist functions $\Phi_1, ..., \Phi_d \in C^1(\mathbb{R}^p)$ so that*

$$\nabla_{\mathbf{w}}\eta(\mathbf{w})^T A_j(\mathbf{w}) = \nabla_{\mathbf{w}}\Phi_j(\mathbf{w}), \quad j = 1, ..., d. \tag{2.6}$$

If the function η is convex, η is a convex entropy. The functions $\Phi_1, ..., \Phi_d$ are called entropy fluxes. The pair (η, Φ) is called (convex) entropy-entropy flux pair.

Definition 2.1.5 *A weak solution of the homogeneous system (2.1a) is called an (convex) entropy solution, if for any (convex) entropy-entropy flux pair the condition*

$$(\eta(\mathbf{w}))_t + \sum_{j=1}^{d} G_s(\mathbf{w}) \leq 0 \tag{2.7}$$

is satisfied in the sense of distributions on $\mathbb{R}^d \times (0, \infty)$, i.e.

$$\int_0^\infty \int_{\mathbb{R}^d} \eta(\mathbf{w})\phi_t + \sum_{j=1}^{d} G_j(\mathbf{w})\phi_{x_j} \, d\mathbf{x} \, dt \geq 0 \tag{2.8}$$

for all non-negative test functions $\phi \in C_0^\infty(\mathbb{R}^d \times (0, \infty))$.

The entropy inequality corresponds to the second law of thermodynamics, i.e. the entropy in a thermodynamic process is non-decreasing in time.

Remark 2.1.6

- *If (2.1a) is a scalar equation, then every (convex) continuously differentiable function η is a (convex) entropy of (2.1a), where the corresponding entropy fluxes are obtained by integration of $G_j' = \eta' F_j'$, $j = 1, .., d$.*
- *Note that the mathematical entropy $\eta(\mathbf{w})$ has the negative sign in comparison with the physical entropy*

$$S = c_v \log\left(\frac{p/p_0}{(\rho/\rho^0)^\kappa}\right) \tag{2.9}$$

  of an inviscid gas.

In the following of this section, we will state some fundamental results about the existence and uniqueness of hyperbolic conservation laws. First we consider one-dimensional hyperbolic conservation laws in a quasi-linear form

$$\mathbf{w}_t + A(\mathbf{w})\mathbf{w}_x = 0, \quad \mathbf{w}(\mathbf{x}, 0) = \mathbf{w}^0(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d. \tag{2.10}$$

Definition 2.1.7 *Let the system (2.10) be strictly hyperbolic. We call an eigenvalue* $\lambda = \lambda(\mathbf{w})$ *of* $A(\mathbf{w})$ *genuinely nonlinear in* D*, if*

$$(\nabla_{\mathbf{w}}\lambda(\mathbf{w})) \cdot \mathbf{r}(\mathbf{w}) \neq 0 \tag{2.11}$$

for all $\mathbf{w} \in D$*, where* $\mathbf{r}$ *is an eigenvector corresponding to* λ*. If we have*

$$(\nabla_{\mathbf{w}}\lambda(\mathbf{w})) \cdot \mathbf{r}(\mathbf{w}) = 0 \tag{2.12}$$

instead, we call λ *linearly degenerate.*

If the eigenvalues of the matrix A from the conservation law (2.10) are either linearly degenerate or genuinely nonlinear, the Riemann problem (2.10), where

$$\mathbf{w}^0(\mathbf{x}) = \begin{cases} \mathbf{u}_L & \text{if } \mathbf{x} \leq 0 \\ \mathbf{u}_R & \text{else} \end{cases}, \tag{2.13}$$

has a unique weak entropy solution for $\|\mathbf{u}_L - \mathbf{u}_R\|$ small enough. This solution consists of at most $d+1$ constant states separated by shock waves, contact discontinuities or rarefaction waves, cf. [47]. In [46], Glimm constructed a family of approximative solutions by means of a sequence of solutions of Riemann problems. This allowed him to prove the existence of (2.10) for small data. Around 30 years later Bressan et al. [22, 23, 24] proved the uniqueness and the continuous dependence on the initial data.

Theorem 2.1.8 *Let the system in (2.10) be strictly hyperbolic and either genuinely nonlinear or linearly degenerate in a neighbourhood of a constant state* $\tilde{\mathbf{w}}$*. Then there exist two constants* $C_1, C_2 > 0$*, so that if the initial data* $\mathbf{w}^0$ *satisfies*

$$\|\mathbf{w}^0 - \tilde{\mathbf{w}}\|_{L^\infty(\mathbb{R})} \leq C_1, \quad TV_{\mathbb{R}}(\mathbf{w}^0) \leq C_2, \tag{2.14}$$

the initial value problem (2.10) has a unique global weak entropy solution $\mathbf{w} \in C([0,\infty), L^1_{loc}(\mathbb{R}))$ *and the initial value is satisfied in the sense of traces, i.e.* $\mathbf{w}(\cdot, 0) = \mathbf{w}^0$ *in* $L^1_{loc}(\mathbb{R})$*.*

Remark 2.1.9 *The abbreviation TV in Theorem 2.1.8 denotes the total variation. We refer the reader for a definition to the textbook [93]. However, one finds*

$$TV_{(a,b)}(w) = \sup \left\{ \sum_{j=1}^{k-1} |w(x_{j+1}) - w(x_j)| : a < x_1 < ... < x_k < b \right\} \tag{2.15}$$

for $w \in H^1(a,b)$*.*

Concerning scalar multi-dimensional equations (2.1a), Kruzhkov [71] showed the following theorem.

Theorem 2.1.10 *The scalar hyperbolic balance law (2.1a) with an initial value* $w^0 \in L^\infty(\mathbb{R}^d)$ *has a unique weak entropy solution* $w \in C((0,\infty), L^1_{loc}(\mathbb{R}^d))$*.*

Thus, existence and uniqueness of a weak entropy solution can be shown for scalar multi-dimensional hyperbolic balance laws and systems of hyperbolic conservation laws. However, it turns out that the entropy condition (2.1.5) is not enough to single out a weak solution for multi-dimensional systems of hyperbolic conservation laws. De Lellis and Székelyhidi Jr. proved the non-uniqueness of the weak entropy solution for the compressible Euler equations (2.16), cf. [35, 36],

$$\rho_t + \nabla \cdot (\rho \mathbf{u}) = 0, \tag{2.16a}$$

$$(\rho \mathbf{u})_t + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) + \nabla p(\rho) = 0, \tag{2.16b}$$

where the pressure p is a known function that satisfies $p' > 0$, e.g. $p(\rho) = A\rho^\gamma$ with a constant $A > 0, \gamma \geq 1$.

Theorem 2.1.11 *Let $d \geq 2$. Then, for any given function p there exist bounded initial data $(\rho^0, \mathbf{u}^0)$ with $\rho^0 \geq const. > 0$ for which there are infinitely many bounded admissible solutions $(\rho, \mathbf{u})$ of (2.16) with $\rho \geq const. > 0$.*

Hence, the well-posedness of multi-dimensional isentropic Euler and shallow water equations with constant bottom topography is an open problem.

2.2. Standard form of the SWE

In oceanography, meteorology or river flow engineering the SWE

$$h_t + \nabla \cdot (h\mathbf{u}) = 0, \tag{1.1a}$$

$$(h\mathbf{u})_t + \nabla \cdot (h\mathbf{u} \otimes \mathbf{u}) + gh\nabla h = -gh\nabla \tilde{b}, \tag{1.1b}$$

describe a thin layer of constant density fluid in hydrostatic balance bounded from below by a rigid and from above by a free surface, where the horizontal scale is much larger then the vertical one, say 20 times larger [74]. Here, h denotes the water depth, $\tilde{b}$ the bottom topography, $g \approx 9.81 m/s^2$ the gravity constant and $\mathbf{u}$ the velocity field averaged along the vertical axis. For given characteristic scales $t_{ref}, h_{ref}, L_{ref}, u_{ref} = L_{ref}/t_{ref}$ we apply the standard non-dimensionalisation procedure by rewriting the SWE (1.1) in the dimensionless variables

$$\hat{t} = \frac{t}{t_{ref}}, \quad \hat{h} = \frac{h}{h_{ref}}, \quad \hat{\mathbf{x}} = \frac{\mathbf{x}}{L_{ref}}, \quad \hat{\mathbf{u}} = \frac{\mathbf{u}}{u_{ref}}, \quad \hat{\tilde{b}} = \frac{\tilde{b}}{h_{ref}}. \tag{2.17}$$

This gives us the dimensionless SWE

$$h_t + \nabla \cdot (h\mathbf{u}) = 0, \tag{2.18a}$$

$$(h\mathbf{u})_t + \nabla \cdot (h\mathbf{u} \otimes \mathbf{u}) + \frac{1}{\varepsilon^2} h\nabla h = -\frac{1}{\varepsilon^2} h\nabla \tilde{b}, \tag{2.18b}$$

where we dropped the hats in the notation. Here,

$$\varepsilon = \frac{u_{ref}}{c_{ref}} = \frac{u_{ref}}{\sqrt{g h_{ref}}} \tag{2.19}$$

denotes the Froude number, which rarely exceeds 0.2 [119]. For large scale phenomena $\varepsilon \approx 10^{-2}$ is a typical value, [8, 73, 88, 112]. In the following we will only work with the dimensionless equations.

2.3. Zero Froude number limit of the SWE

Let us consider an initial value problem for the SWE (2.18) on a domain Ω. By varying the Froude number $\varepsilon > 0$ we obtain a family of initial value problems. Let us recall from Chapter 1, that the SWE (2.18) with constant bottom topography are the isentropic Euler equations with the pressure-density relation $p(\rho) = \rho^2/2$. Then, due to the work of Klainerman and Majda [62, 63] and Ebin [40], it is well-known that the above mentioned family of initial value problems posses unique solutions $\mathbf{w}^{(\varepsilon)} = (h^{(\varepsilon)}, h^{(\varepsilon)}\mathbf{u}^{(\varepsilon)})^T$ under suitable assumptions, e.g. the initial data are small perturbations of incompressible Euler equations initial data. Moreover the series of solutions $(\mathbf{w}^{(\varepsilon)})_\varepsilon$ converges in a weak sense to the solution $\mathbf{w}^{(0)}$ of the incompressible Euler equations. In the following of this section, we will derive the zero Froude number SWE/ lake equations using asymptotic analysis, where we follow Bresch et al. [21].

Motivated by the results for the isentropic Euler equations we assume that for every $\varepsilon > 0$ there is a unique solution of the SWE (2.18)

$$\mathbf{w}^{(\varepsilon)} = \mathbf{w}^{(0)} + \varepsilon \mathbf{w}^{(1)} + \varepsilon^2 \mathbf{w}^{(2)} + \mathcal{O}(\varepsilon^3), \tag{2.20}$$

that converges to $\mathbf{w}^{(0)}$ as the Froude number ε approaches zero. Further, we assume that $\mathbf{w}^{(i)}$, $i = 0, 1, 2$, is independent of the Froude number ε. Plugging the expansion (2.20) into the SWE (2.18) and comparing the like powers of the Froude number ε, we obtain first

$$\varepsilon^{-2}: \quad h^{(0)} \nabla (h^{(0)} + \tilde{b}) = 0, \tag{2.21}$$

$$\varepsilon^{-1}: \quad h^{(1)} \nabla (h^{(0)} + \tilde{b}) + h^{(0)} \nabla h^{(1)} = 0. \tag{2.22}$$

We assume that the water depth is always positive, thus $h^{(0)}$ is positive and $\nabla(h^{(0)} + \tilde{b}) = 0$, i.e. $H^{(0)}(t) := h^{(0)}(\mathbf{x}, t) + \tilde{b}(\mathbf{x})$ is constant in space. Thus the time derivative of $h^{(0)}$ is as well constant in space and $\nabla h^{(1)} = 0$. Further, averaging the continuity equation (2.18a) we get

$$-\frac{dh^{(0)}}{dt} = \nabla \cdot (h^{(0)}\mathbf{u}^{(0)}) = \frac{1}{|\Omega|} \int_{\partial\Omega} h^{(0)}\mathbf{u}^{(0)} \cdot \mathbf{n} \, ds, \tag{2.23}$$

$$-\frac{dh^{(1)}}{dt} = \nabla \cdot (h^{(1)}\mathbf{u}^{(0)} + h^{(0)}\mathbf{u}^{(1)}) = \frac{1}{|\Omega|} \int_{\partial\Omega} (h^{(1)}\mathbf{u}^{(0)} + h^{(0)}\mathbf{u}^{(1)}) \cdot \mathbf{n} \, ds. \tag{2.24}$$

Consequently we obtain the zero Froude number SWE/ lake equations

$$h^{(0)} + \tilde{b} = H^{(0)}(t), \tag{2.25a}$$

$$\nabla \cdot (h^{(0)}\mathbf{u}^{(0)}) = \frac{1}{|\Omega|} \int_{\partial\Omega} h^{(0)}\mathbf{u}^{(0)} \cdot \mathbf{n} \, ds = -\frac{dH^{(0)}}{dt}, \tag{2.25b}$$

$$(h^{(0)}\mathbf{u}^{(0)})_t + \nabla \cdot (h^{(0)}\mathbf{u}^{(0)} \otimes \mathbf{u}^{(0)}) + h^{(0)} \nabla h^{(2)} = 0. \tag{2.25c}$$

The zero Froude number SWE (2.25) simplify to

$$h^{(0)} + \tilde{b} = H^{(0)} = \text{ const.}, \tag{2.26a}$$

$$\nabla \cdot (h^{(0)}\mathbf{u}^{(0)}) = 0, \tag{2.26b}$$

$$(h^{(0)}\mathbf{u}^{(0)})_t + \nabla \cdot (h^{(0)}\mathbf{u}^{(0)} \otimes \mathbf{u}^{(0)}) + h^{(0)} \nabla h^{(2)} = 0, \tag{2.26c}$$

if the surface integral in (2.25b) vanishes. For example, $\int_{\partial\Omega} h^{(0)}\mathbf{u}^{(0)}\cdot\mathbf{n}\, ds = 0$ in the case of periodic boundary conditions - then also $h^{(1)}$ is constant and $\mathbf{u}^{(1)}$ divergence free. Note that the zero Froude number SWE (2.26) are the incompressible Euler equations with the equation of state $p(\rho) = \rho^2/2$, if the bottom topography is constant.

The zero Froude number SWE (2.25) model a balanced flow without gravity waves and have been described by Greenspan [50]. The initial value problem consists of given initial values for $h^{(0)}, \mathbf{u}^{(0)}$ as well as the bottom topography $\tilde{b}$. The evolution of $\mathbf{u}$ as well as the water depth component $h^{(2)}$ have to be computed. Levermore et al. [78] showed that this initial value problem is well-posed.

2.4. Alternative SWE formulation

Typically, solutions of oceanographic or atmospheric flows are small perturbations of an underlying equilibrium state. If we assume periodic boundary conditions, we can illustrate this fact by considering the formal single scale expansion (2.20) of the primitive variables

$$H(\mathbf{x},t) = h(\mathbf{x},t) + \tilde{b}(\mathbf{x}) = (h^{(0)} + \tilde{b}) + \varepsilon h^{(1)} + \varepsilon^2 h^{(2)}(\mathbf{x},t) + \mathcal{O}(\varepsilon^3), \tag{2.27a}$$

$$\mathbf{u}(\mathbf{x},t) = \mathbf{u}^{(0)}(\mathbf{x},t) + \varepsilon\mathbf{u}^{(1)}(\mathbf{x},t) + \varepsilon^2\mathbf{u}^{(2)}(\mathbf{x},t) + \mathcal{O}(\varepsilon^3), \tag{2.27b}$$

where $h^{(0)}, h^{(2)}, \mathbf{u}^{(0)}$ are solutions of (2.26). Hence, the underlying equilibrium state is the so-called lake at rest state $\mathbf{w} = (h^{(0)}, 0)^T$, where the water level $H^{(0)} = h^{(0)} + \tilde{b}$ is constant and no motion occurs. Following Giraldo and Restelli [44], we divide the water depth into a constant water level $H^{(0)}$ and the free surface elevation z

$$h(\mathbf{x},t) + \tilde{b}(\mathbf{x}) = H^{(0)} + z(\mathbf{x},t), \quad z = \varepsilon h^{(1)} + \varepsilon^2 h^{(2)}(\mathbf{x},t). \tag{2.28}$$

Here, $z = \mathcal{O}(\varepsilon)$ is a small perturbation of the underlying constant water surface $H^{(0)}$. Introducing the new bottom topography

$$b = \tilde{b} - H^{(0)} = -h^{(0)}, \tag{2.29}$$

we rewrite the SWE (2.18) using the conservative set of variables $(z, \mathbf{m} = h\mathbf{u})$, cf. Figure 2.1,

$$z_t + \nabla\cdot\mathbf{m} = 0, \tag{2.30a}$$

$$\mathbf{m}_t + \nabla\cdot\left(\frac{\mathbf{m}\otimes\mathbf{m}}{z-b} + z\frac{z-2b}{2\varepsilon^2}\mathbb{1}\right) = -\frac{z}{\varepsilon^2}\nabla b. \tag{2.30b}$$

System (2.30) was first introduced in literature by Rogers et al [101, 102]. It was also used in [44, 72, 79, 80]. However there is a slight difference: our bottom topography has the opposite sign.

Remark 2.4.1 *The information about the constant $H^{(0)}$ is necessary to define the free surface elevation z. Although, it is typically unknown. We can estimate it up to $\mathcal{O}(\varepsilon)$. This can be done for example by setting $H^{(0)}$ to be the average water level or the maximum or minimum water level.*

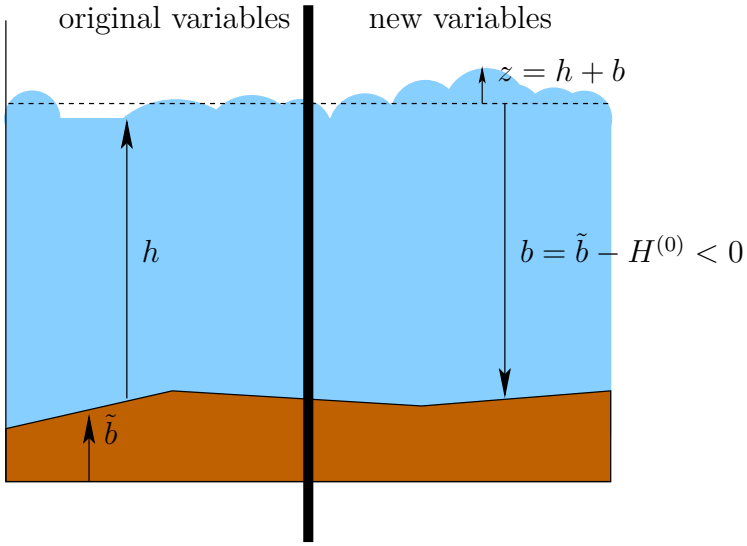

Figure 2.1.: Shallow water equations variables.

Remark 2.4.2 *For low Froude numbers, we expect water depth fluctuations of the order* $\mathcal{O}(\varepsilon^2)$ *in space. For very small Froude numbers, e.g.* $\varepsilon = 10^{-7}$*, the alternative formulation (2.30) should experience a better behaviour then the standard one (2.18) with regard to round off and cancellation errors, cf. [104].*

2.5. Zero Froude number limit of the alternative SWE formulation

We repeat the asymptotic analysis approach from Section 2.3 to derive the zero Froude number equations of the alternative shallow water formulation (2.30). To this end, we plug in the formal expansion (2.20) into the alternative SWE (2.30) and compare the like orders of ε:

$$\varepsilon^{-2}: \quad (z^{(0)} - b)\nabla z^{(0)} = 0, \tag{2.31a}$$

$$\varepsilon^{-1}: \quad (z^{(0)} - b)\nabla z^{(1)} + z^{(1)}\nabla z^{(0)} = 0, \tag{2.31b}$$

$$\varepsilon^{0}: \quad z_t^{(0)} + \nabla \cdot \mathbf{m}^{(0)} = 0, \tag{2.31c}$$

$$\mathbf{m}_t^{(0)} + \nabla \cdot \frac{\mathbf{m}^{(0)} \otimes \mathbf{m}^{(0)}}{z^{(0)} - b} + ((z - b)\nabla z)^{(2)} = 0. \tag{2.31d}$$

We assume the water depth to be positive, i.e. $z^{(0)} - b > 0$. Thus, it follows from (2.31a), (2.31b)

$$h^{(0)}(\mathbf{x}, t) + b(\mathbf{x}) = z^{(0)}(\mathbf{x}, t) = z^{(0)}(t), \quad z^{(1)}(\mathbf{x}, t) = z^{(1)}(t). \tag{2.32}$$

We average the continuity equation (2.30a) over the domain Ω and obtain by the Gauss theorem

$$z_t^{(0)} = -\frac{1}{|\Omega|} \int_{\partial\Omega} \mathbf{m}^{(0)} \cdot \mathbf{n} \; ds. \tag{2.33}$$

For certain boundary conditions, e.g. periodic or no-slip boundary conditions, the surface integral in (2.33) vanishes and we obtain the alternative zero Froude number SWE

$$z^{(0)} = \text{ const.}, \tag{2.34a}$$

$$\nabla \cdot \mathbf{m}^{(0)} = 0, \tag{2.34b}$$

$$\mathbf{m}_t^{(0)} + \nabla \cdot \frac{\mathbf{m}^{(0)} \otimes \mathbf{m}^{(0)}}{z^{(0)} - b} + (z^{(0)} - b)\nabla z^{(2)} = 0. \tag{2.34c}$$

Then we also have $z^{(1)} = const.$ and $\nabla \cdot \mathbf{m}^{(1)} = 0$.

2.6. Splitting of the alternative SWE formulation

In the introduction, we have pointed out that one may experience difficulties solving low Froude/ Mach number problems with standard compressible solvers. This may be the high computational costs and dissipative results due to the severe time step-restriction (1.2). Also slower convergence rates or even the breakdown of numerical solution may occur. Following Giraldo and Restelli [44, 45, 98, 99] we split the alternative SWE (2.30) into a hyperbolic, stiff, linear part that governs the fast waves and a non-stiff one. Using a semi-implicit time discretisation we can treat the stiff part in an implicit way, whereas the non-stiff one is treated explicitly. Consequently, the fast waves do not affect the CFL-stability condition. The purpose of this section is to present the splitting.

The linear subsystem of the alternative SWE (2.30) reads

$$z_t + \nabla \cdot \mathbf{m} = 0, \tag{2.35a}$$

$$\mathbf{m}_t - \frac{1}{\varepsilon^2}\nabla(zb) = -\frac{1}{\varepsilon^2} z \nabla b. \tag{2.35b}$$

Note that (2.35) are the equations of linear acoustics with velocity $\sqrt{-b}/\varepsilon$, if the bottom topography b is constant. Further,

$$z_t = 0 \tag{2.36a}$$

$$\mathbf{m}_t + \nabla \cdot \left(\frac{\mathbf{m} \otimes \mathbf{m}}{z-b} + \frac{z^2}{2\varepsilon^2}\mathbb{1} \right) = 0 \tag{2.36b}$$

is the corresponding nonlinear subsystem of the alternative SWE (2.30). Both subsystems can be written in the form of hyperbolic balance laws in divergence form

$$\mathbf{w}_t + \nabla \cdot \mathcal{F}_L(\mathbf{w}) = K(\mathbf{w}), \quad \mathbf{w}_t + \nabla \cdot \mathcal{F}_{NL}(\mathbf{w}) = 0, \tag{2.37a}$$

$$\mathcal{F}_L(\mathbf{w}) = \begin{bmatrix} \mathbf{m} \\ -\frac{b}{\varepsilon^2} z \mathbb{1} \end{bmatrix}, \quad \mathcal{F}_{NL}(\mathbf{w}) = \begin{bmatrix} 0 \\ -\frac{\mathbf{m} \otimes \mathbf{m}^T}{z-b} + \frac{z^2}{2\varepsilon^2}\mathbb{1} \end{bmatrix}, \tag{2.37b}$$

$$K(\mathbf{w}) = \begin{bmatrix} 0 \\ -\frac{z \nabla b}{\varepsilon^2} \end{bmatrix}. \tag{2.37c}$$

This splitting is crucial for the schemes that are presented in this thesis. First of all, it is suitable from the viewpoint of the eigenstructure in the following way: The eigenvalues of the matrix pencil corresponding to the linear flux $\mathcal{F}_L$ are of order $\mathcal{O}(1/\varepsilon)$, whereas the eigenvalues corresponding to the nonlinear system are only of order $\mathcal{O}(1)$ - see Section 2.7. Thus, the linear part is stiff and the nonlinear one is non-stiff. Let us remark, that the Froude number to the power two appears in a denominator in the nonlinear flux but there is no influence on the matrix pencil eigenvalues. Second, the linear part covers the constraints (2.34a), (2.34b) in the low Froude number limit. Therefore, we will be able to show that the IMEX schemes developed in Chapters 4, 6 are asymptotic preserving, see Chapters 5, 7.

2.7. Eigenstructure of the governing equations

We have introduced the governing equations and presented the behaviour of the standard and alternative SWE (2.18), (2.30) in the low Froude number limit by considering asymptotic expansions. In this section, we will analyse the eigenstructure of the governing equations. This is necessary for several reasons:

- prove that the governing equations are hyperbolic
- the eigenstructure is needed to derive the evolution Galerkin operator
- the linear subsystem (2.35) must have eigenvalues that are $\mathcal{O}(1/\varepsilon)$ for $\varepsilon \to 0$

- the nonlinear subsystem (2.36) must have eigenvalues that are $\mathcal{O}(1)$ for $\varepsilon \to 0$
- the spectral radius of the matrix pencil is the maximum propagation speed of information in a system, thus the eigenstructure of the nonlinear system (2.36) gives us the necessary time restriction, i.e. CFL-condition

This section is divided in four parts, where each part presents the eigenstructure of a governing equation.

2.7.1. Shallow water equations

The SWE (2.18) written in conservative variables $\mathbf{w} = (h, h\mathbf{u})^T$ read

$$\mathbf{w}_t + \nabla \cdot \mathcal{F}(\mathbf{w}) = \tilde{K} \tag{2.38}$$

with

$$\mathcal{F}(\mathbf{w}) = \begin{bmatrix} h\mathbf{u} \\ h\mathbf{u} \otimes \mathbf{u}^T + \dfrac{h^2}{2\varepsilon^2}\mathbb{1} \end{bmatrix}, \quad \tilde{K} = \begin{bmatrix} 0 \\ -c^2\nabla\tilde{b} \end{bmatrix}, \quad c = \frac{\sqrt{h}}{\varepsilon} \tag{2.39}$$

in divergence form. We can rewrite the SWE into quasilinear form

$$\mathbf{w}_t + \sum_{i=1}^{d} A_i \mathbf{w}_{x_i} = \tilde{K}, \tag{2.40a}$$

$$A_i = \begin{bmatrix} 0 & \mathbf{e}_i^T \\ -u_i\mathbf{u} + c^2\mathbf{e}_i & \mathbf{u}\mathbf{e}_i^T + u_i\mathbb{1} \end{bmatrix}, \quad i = 1, ..., d. \tag{2.40b}$$

If $d = 2$ we obtain the SWE by averaging of the three-dimensional Euler equations. However, we can formally generalise our system to any space dimension $d \geq 2$. The eigenvalues of the matrix pencil

$$\mathbb{P} := \sum_{i=1}^{d} A_i n_i = \begin{bmatrix} 0 & \mathbf{n}^T \\ -(\mathbf{u}\cdot\mathbf{n})\mathbf{u} + c^2\mathbf{n} & \mathbf{u}\cdot\mathbf{n}^T + (\mathbf{u}\cdot\mathbf{n})\mathbb{1} \end{bmatrix}, \quad \|\mathbf{n}\| = 1, \tag{2.41}$$

are

$$\lambda_1 = \mathbf{u}\cdot\mathbf{n} - c, \quad \lambda_2 = ... = \lambda_d = \mathbf{u}\cdot\mathbf{n}, \quad \lambda_{d+1} = \mathbf{u}\cdot\mathbf{n} + c, \tag{2.42}$$

if the spatial dimension $d \geq 2$. It is easy to verify that the corresponding right and left eigenvectors are

$$\mathbf{r}^1 = \begin{bmatrix} 1 \\ \mathbf{u} - c\mathbf{n} \end{bmatrix}, \quad \mathbf{r}^2 = \begin{bmatrix} 0 \\ \mathbf{t}^1 \end{bmatrix}, \quad ..., \quad \mathbf{r}^d = \begin{bmatrix} 0 \\ \mathbf{t}^{d-1} \end{bmatrix}, \quad \mathbf{r}^{d+1} = \begin{bmatrix} 1 \\ \mathbf{u} + c\mathbf{n} \end{bmatrix}, \tag{2.43a}$$

$$\mathbf{l}^1 = \frac{1}{2c}\begin{bmatrix} \mathbf{u}\cdot\mathbf{n} + c \\ -\mathbf{n} \end{bmatrix}, \quad \mathbf{l}^2 = \begin{bmatrix} -\mathbf{u}\cdot\mathbf{t}^1 \\ \mathbf{t}^1 \end{bmatrix}, \quad ..., \quad \mathbf{l}^d = \begin{bmatrix} -\mathbf{u}\cdot\mathbf{t}^{d-1} \\ \mathbf{t}^{d-1} \end{bmatrix}, \tag{2.43b}$$

$$\mathbf{l}^{d+1} = \frac{1}{2c}\begin{bmatrix} -\mathbf{u}\cdot\mathbf{n} + c \\ \mathbf{n} \end{bmatrix}.$$

Here, $\{\mathbf{t}^1, ..., \mathbf{t}^{d-1}\}$ is an orthonormal basis of the tangential space to $\mathbf{n}$. In the 1D case, $\mathbf{n} = 1$ and the eigenvalues are $u \pm c$ with corresponding eigenvectors $\mathbf{r}^1 = (1, u-c)^T$, $\mathbf{r}^2 = (1, u+c)^T$, $\mathbf{l}^1 = (u+c, -1)^T/(2c)$, $\mathbf{l}^2 = (-(u-c), 1)^T/(2c)$. Since c is non-zero, the eigenvectors $(\mathbf{r}^i)_i$ create a basis. Therefore, the matrix pencil $\mathbb{P}$ is diagonalisable and the SWE (2.18) are diagonally hyperbolic.

2.7.2. Alternative form of the shallow water equations

For the alternative SWE (2.30)

$$\mathbf{w}_t + \nabla \cdot \mathcal{F}(\mathbf{w}) = K(\mathbf{w}) \tag{2.44}$$

with

$$\mathcal{F}(\mathbf{w}) = \begin{bmatrix} \mathbf{m} \\ \dfrac{\mathbf{m} \otimes \mathbf{m}^T}{z-b} + z\dfrac{z-2b}{2\varepsilon^2}\mathbb{1} \end{bmatrix}, \quad K(\mathbf{w}) = \begin{bmatrix} 0 \\ -\dfrac{z\nabla b}{\varepsilon^2} \end{bmatrix}, \tag{2.45}$$

the quasilinear form reads

$$\mathbf{w}_t + \sum_{i=1}^{d} A_i \mathbf{w}_{x_i} = 0. \tag{2.46}$$

Here, A_i are the Jacobians given in (2.40a). Since the A_i are the same as in the previous section, the matrix $\mathbb{P}$ is also the same - as the eigenvalues and their eigenvectors. Consequently, the alternative SWE (2.30) are diagonally hyperbolic.

2.7.3. Linear subsystem of the alternative shallow water equations

The linear subsystem of the alternative SWE (2.35) reads in conservative variables $\mathbf{w} = (z, \mathbf{m})^T$

$$\mathbf{w}_t + \sum_{i=1}^{d} A_i \mathbf{w}_{x_i} = 0, \tag{2.47}$$

$$A_i = \begin{bmatrix} 0 & \mathbf{e}_i^T \\ c_b^2 \mathbf{e}_i & \end{bmatrix}, \quad c_b = \frac{\sqrt{-b}}{\varepsilon}. \tag{2.48}$$

Note that the Jacobians in (2.48) can be obtained by setting the advection velocity $\mathbf{u}$ to zero and $c = c_b$ in (2.40b). Thus eigenvalues of the matrix pencil

$$\mathbb{P} := \sum_{i=1}^{d} A_i n_i = \begin{bmatrix} 0 & \mathbf{n}^T \\ c_b^2 \mathbf{n} & 0 \end{bmatrix}, \quad \|\mathbf{n}\| = 1, \tag{2.49}$$

are

$$\lambda_1 = -c_b, \quad \lambda_2 = ... = \lambda_d = 0, \quad \lambda_{d+1} = c_b, \tag{2.50}$$

if the spatial dimension $d \geq 2$. The corresponding right and left eigenvectors are

$$\mathbf{r}^1 = \begin{bmatrix} \dfrac{1}{c_b} \\ -\mathbf{n} \end{bmatrix}, \quad \mathbf{r}^2 = \begin{bmatrix} 0 \\ \mathbf{t}^1 \end{bmatrix}, \quad ..., \quad \mathbf{r}^d = \begin{bmatrix} 0 \\ \mathbf{t}^{d-1} \end{bmatrix}, \quad \mathbf{r}^{d+1} = \begin{bmatrix} \dfrac{1}{c_b} \\ \mathbf{n} \end{bmatrix}, \tag{2.51a}$$

$$\mathbf{l}^1 = \frac{1}{2}\begin{bmatrix} c_b \\ -\mathbf{n} \end{bmatrix}, \quad \mathbf{l}^2 = \begin{bmatrix} 0 \\ \mathbf{t}^1 \end{bmatrix}, \quad ..., \quad \mathbf{l}^d = \begin{bmatrix} 0 \\ \mathbf{t}^{d-1} \end{bmatrix}, \quad \mathbf{l}^{d+1} = \frac{1}{2}\begin{bmatrix} c_b \\ \mathbf{n} \end{bmatrix}. \tag{2.51b}$$

Here, $\{\mathbf{t}^1, ..., \mathbf{t}^{d-1}\}$ is an orthogonal basis of the tangential space to $\mathbf{n}$. In the 1D case, $\mathbf{n} = 1$ and the eigenvalues are $\pm c_b$ with corresponding eigenvectors $\mathbf{r}^1 = (1/c_b, -1)^T$, $\mathbf{r}^2 = (1/c_b, 1)^T$, $\mathbf{l}^1 = (c_b, -1)^T/2$, $\mathbf{l}^2 = (c_b, 1)^T/2$. Also the right eigenvectors $(\mathbf{r}^i)_i$ are linearly independent. Therefore, the matrix pencil $\mathbb{P}$ is diagonalisable. Hence, the linear system (2.35) is diagonally hyperbolic.

2.7.4. Nonlinear subsystem of the alternative shallow water equations

The nonlinear subsystem of the SWE (2.36) reads in conservative variables $\mathbf{w} = (z, \mathbf{m})^T$

$$\mathbf{w}_t + \sum_{i=1}^{d} A_i \mathbf{w}_{x_i} = 0, \tag{2.52a}$$

$$A_i = \begin{bmatrix} 0 & 0 \\ -\dfrac{m_i}{(z-b)^2}\mathbf{m} + \dfrac{z}{\varepsilon^2}\mathbf{e}_i & \dfrac{\mathbf{m}\mathbf{e}_i^T}{z-b} + \dfrac{m_i}{z-b}\mathbb{1} \end{bmatrix}. \tag{2.52b}$$

The eigenvalues of the matrix pencil

$$\mathbb{P} := \sum_{i=1}^{d} A_i n_i = \begin{bmatrix} 0 & 0 \\ -\dfrac{\mathbf{m}\cdot\mathbf{n}}{(z-b)^2}\mathbf{m} + \dfrac{z}{\varepsilon^2}\mathbf{n} & \dfrac{\mathbf{m}\cdot\mathbf{n}^T}{z-b} + \dfrac{\mathbf{m}\cdot\mathbf{n}}{z-b}\mathbb{1} \end{bmatrix}, \quad \|\mathbf{n}\| = 1, \tag{2.53}$$

are

$$\lambda_1 = 0, \quad \lambda_2 = ... = \lambda_d = \frac{\mathbf{m}\cdot\mathbf{n}}{z-b}, \quad \lambda_{d+1} = 2\frac{\mathbf{m}\cdot\mathbf{n}}{z-b}, \tag{2.54}$$

if the spatial dimension $d \geq 2$. It is easy to verify that the corresponding right eigenvectors are

$$\mathbf{r}^1 = \begin{bmatrix} 2(\mathbf{u}\cdot\mathbf{n})^2 \\ (\mathbf{u}\cdot\mathbf{n})^2\mathbf{u} - \dfrac{z}{\varepsilon^2}\left[2(\mathbf{u}\cdot\mathbf{n})\mathbf{n} - \mathbf{u}\right] \end{bmatrix}, \quad \mathbf{r}^2 = \begin{bmatrix} 0 \\ \mathbf{t}^1 \end{bmatrix}, \quad ..., \tag{2.55a}$$

$$\mathbf{r}^d = \begin{bmatrix} 0 \\ \mathbf{t}^{d-1} \end{bmatrix}, \quad \mathbf{r}^{d+1} = \begin{bmatrix} 0 \\ \mathbf{u} \end{bmatrix}.$$

Here, $\{\mathbf{t}^1, ..., \mathbf{t}^{d-1}\}$ is an orthonormal basis of the tangential space to $\mathbf{n}$. If and only if $\mathbf{m}\cdot\mathbf{n} \neq 0$, the eigenvectors $(\mathbf{r}^i)_i$ are linearly independent. Then, the corresponding left eigenvectors are

$$\mathbf{l}^1 = \frac{\mathbf{e}_1}{2(\mathbf{u}\cdot\mathbf{n})^2}, \quad \mathbf{l}^2 = \begin{bmatrix} -\dfrac{z\mathbf{t}^1\cdot\mathbf{u}}{(\varepsilon\mathbf{u}\cdot\mathbf{n})^2} \\ \mathbf{t}^1 - \dfrac{\mathbf{t}^1\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{n}}\mathbf{n} \end{bmatrix}, \quad ..., \tag{2.55b}$$

$$\mathbf{l}^d = \begin{bmatrix} -\dfrac{z\mathbf{t}^{d-1}\cdot\mathbf{u}}{(\varepsilon\mathbf{u}\cdot\mathbf{n})^2} \\ \mathbf{t}^{d-1} - \dfrac{\mathbf{t}^{d-1}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{n}}\mathbf{n} \end{bmatrix}, \quad \mathbf{l}^{d+1} = \begin{bmatrix} \frac{1}{2}\left[\dfrac{z}{(\varepsilon\mathbf{u}\cdot\mathbf{n})^2} - 1\right] \\ \dfrac{\mathbf{n}}{\mathbf{u}\cdot\mathbf{n}} \end{bmatrix}, .$$

If $\mathbf{m}\cdot\mathbf{n} = 0$, the velocity vector $\mathbf{u}$ lies in the tangential plane to $\mathbf{n}$ and the right eigenvectors $(\mathbf{r}^i)_i$ are linearly dependent. In the 1D case, $\mathbf{n} = 1$ and the eigenvalues are $0, 2u$ with corresponding eigenvectors $\mathbf{r}^1 = (2u, u^3 - zu/\varepsilon^2)^T$, $\mathbf{r}^2 = u\mathbf{e}_2$, $\mathbf{l}^1 = \mathbf{e}_1/(2u^2)$, $\mathbf{l}^2 = ((z/(\varepsilon u)^2 - 1)/2, 1/u)^T$. The nonlinear subsystem (2.36) is hyperbolic.

3. Evolution operators

Over the last 14 years the finite volume evolution Galerkin (FVEG) schemes have been developed to approximate hyperbolic balance laws. They have been studied extensively by Lukáčová, Noelle, Arun and collaborators, where the FVEG methods have been applied to approximate solutions of wave equation [81, 82, 83], Euler equations [3, 83, 84, 86], shallow water equations [15, 16, 55, 83, 85]. We point out the recent paper [121], where a discontinuous Galerkin scheme combined with evolution operators is applied to meteorological problems using non-hydrostatic Euler equations. The derivation of the corresponding exact and approximate evolution operators with locally frozen Jacobians is presented in the appendix A. The numerical results indicate that multidimensional flow phenomena are approximated in an accurate and stable way, see the review paper [83] and the references therein for more details. The main idea of a FVEG scheme is to predict cell interface values with an evolution operator, also called EG operator, and to use the values to evaluate the flux integrals by means of numerical quadrature. Approximate evolution operators are main building blocks of a FVEG scheme.

The aim of this chapter is to derive exact and approximate evolution operators for the alternative SWE (2.30) for spatial dimensions $d \geq 2$. Required properties for the derivation are presented in Section 3.1. In Section 3.2 we present the general approach of deriving an exact evolution operator for a hyperbolic balance law. In Section 3.3 we apply the previously introduced approach to the alternative SWE (2.30). In order to evaluate the exact evolution operator we have to approximate it. This is done in Section 3.4. The so-called local evolution operator is derived in Section 3.5.

3.1. Spherical coordinates

In the following section we will present a procedure to represent evolution operators for quasi-linear hyperbolic systems, where normal and tangential vectors on the unit sphere and their properties are excessively used. Therefore, in this section we recall some standard calculus, introduce notation and provide required properties to derive exact and approximate evolution operators.

Let us first recall the parametrisation of the unit $(d-1)$-sphere S^{d-1} for $d > 1$, that

we will refer to as "sphere" for the sake of brevity. We use the spherical coordinates

$$\mathbf{n}(\boldsymbol{\phi}) = \mathbf{n}(\phi_1, ..., \phi_{d-1}) = \begin{bmatrix} \sin(\phi_1) \cdot ... \cdot \sin(\phi_{d-1}) \\ \sin(\phi_1) \cdot ... \cdot \sin(\phi_{d-2}) \cdot \cos(\phi_{d-1}) \\ \sin(\phi_1) \cdot ... \cdot \sin(\phi_{d-3}) \cdot \cos(\phi_{d-2}) \\ \vdots \\ \sin(\phi_1) \cdot \cos(\phi_2) \\ \cos(\phi_1) \end{bmatrix} = \begin{bmatrix} k_{d-1} \sin(\phi_{d-1}) \\ k_{d-1} \cos(\phi_{d-1}) \\ k_{d-2} \cos(\phi_{d-2}) \\ \vdots \\ k_2 \cos(\phi_2) \\ k_1 \cos(\phi_1) \end{bmatrix} \tag{3.1}$$

with

$$k_j = k_j(\phi_1, ..., \phi_{j-1}) := \prod_{i=1}^{j-1} \sin \phi_i = \operatorname{sgn}(k_j) \left\| \frac{d\mathbf{n}}{d\phi_j} \right\|. \tag{3.2}$$

The last equality in (3.2) can be obtained by a straightforward computation. The domain of definition of the mapping $\mathbf{n}$ is

$$O = \begin{cases} [0, \pi]^{d-2} \times [0, 2\pi) & \text{if } d > 3 \\ [0, 2\pi) & \text{if } d = 2. \end{cases} \tag{3.3a}$$

Due to the periodicity of sine and cosine functions the intervals of the set O can be shifted by multiples of π, e.g. the set

$$\tilde{O} = \begin{cases} [-\pi, 0] \times [0, \pi]^{d-3} \times [0, 2\pi) & \text{if } d > 3 \\ [-\pi, 0] \times [0, 2\pi) & \text{if } d = 3 \\ [-\pi, \pi) & \text{if } d = 2 \end{cases} \tag{3.3b}$$

is a domain of definition, too. Note that shifting a sine or cosine function by $\pm\pi$ is the same as to multiply it by -1. Thus, the mapping $\mathbf{n} : \tilde{O} \to S^{d-1}$ is the point reflection of the mapping $\mathbf{n} : O \to S^{d-1}$.

In the following we will integrate the components of $\mathbf{n}(\boldsymbol{\phi})$ over the sphere to obtain the exact evolution operator of the linear part of the SWE (2.35). One finds

$$\int_{S^{d-1}} f \; dS^{d-1} = \int_O f(\mathbf{n}(\boldsymbol{\phi})) |dS^{d-1}|(\boldsymbol{\phi}) \; d\boldsymbol{\phi}, \tag{3.4}$$

where the surface element reads

$$dS^{d-1} = |dS^{d-1}|(\boldsymbol{\phi}) d\boldsymbol{\phi} = \left(\prod_{i=1}^{d-1} |k_i| \right) d\boldsymbol{\phi} = \left(\prod_{i=1}^{d-2} |\sin^{d-1-i} \phi_i| \right) d\boldsymbol{\phi}. \tag{3.5}$$

Here, the representation of the surface element dS in (3.5) can be computed by induction over the spatial dimension d. If f is a polynomial, we can simplify calculations with the following lemma that describes the L^2-orthogonality property of the normal components:

Lemma 3.1.1 *The components of the spheric coordinates* $\mathbf{n}(\boldsymbol{\phi})$ *are* L^2*-orthogonal, i.e.*

$$\int_{S^{d-1}} x_i x_j \; dS^{d-1} = \int_O n_i(\boldsymbol{\phi}) n_j(\boldsymbol{\phi}) \; |dS^{d-1}|(\boldsymbol{\phi}) \; d\boldsymbol{\phi} = 0, \quad 1 \le i < j \le d. \tag{3.6}$$

Proof: We obtain

$$\int\limits_{S^{d-1}} x_i x_j \; dS^{d-1} = - \int\limits_{S^{d-1}} x_i x_j \; dS^{d-1} \tag{3.7}$$

using in addition another suitable parametrisation of the sphere. Thus the integral is zero. □

Definition 3.1.2 *We define the squared L^2-norm of the normal components as N_d^2, i.e.*

$$N_d^2 = \int\limits_{S^{d-1}} x_i^2 \; dS^{d-1} = \int\limits_{O} n_i(\boldsymbol{\phi})^2 |dS^{d-1}|(\boldsymbol{\phi}) \; d\boldsymbol{\phi} \tag{3.8}$$

for $i = 1, ..., d$. Note that N_d^2 is well-defined due to the symmetry of the sphere. One finds

$$N_2^2 = \pi, \quad N_3^2 = \frac{4\pi}{3}, \quad N_4^2 = \frac{\pi^2}{2}. \tag{3.9}$$

Let $\mathbf{n}(\boldsymbol{\phi})$ be a point on the sphere. The outer unit normal to $\mathbf{n}(\boldsymbol{\phi})$ is $\mathbf{n}(\boldsymbol{\phi})$ itself. An orthogonal basis of the corresponding tangential space can be obtained by differentiation of $\mathbf{n}(\boldsymbol{\phi})$ with respect to the parameters $\phi_1, ..., \phi_{d-1}$. In this way we obtain the following orthonormal basis of the tangential space

$$\mathbf{t}^j(\boldsymbol{\phi}) = \mathbf{t}^j(\phi_j, ..., \phi_{d-1}) := \frac{1}{k_j} \frac{d\mathbf{n}}{d\phi_j} = \begin{bmatrix} \cos(\phi_j)\mathbf{n}(\phi_{j+1}, ..., \phi_{d-1}) \\ -\sin(\phi_j) \\ 0 \end{bmatrix}, \quad j = 1, ..., d-1. \tag{3.10}$$

Here, we use formally that k_1 and $\mathbf{n}$ map the empty set onto the number 1.

Example 3.1.3 *For $d = 2$ and $d = 3$ we have the following orthonormal basis of $\mathbb{R}^d$*

$$\underline{d=2}: \quad \mathbf{n}(\phi) = \begin{bmatrix} \sin\phi \\ \cos\phi \end{bmatrix}, \quad \mathbf{t}(\phi) = \begin{bmatrix} \cos\phi \\ -\sin\phi \end{bmatrix}, \tag{3.11a}$$

$$\underline{d=3}: \quad \mathbf{n}(\boldsymbol{\phi}) = \begin{bmatrix} \sin\phi_1 \sin\phi_2 \\ \sin\phi_1 \cos\phi_2 \\ \cos\phi_1 \end{bmatrix}, \quad \mathbf{t}^1(\boldsymbol{\phi}) = \begin{bmatrix} \cos\phi_1 \sin\phi_2 \\ \cos\phi_1 \cos\phi_2 \\ -\sin\phi_1 \end{bmatrix}, \quad \mathbf{t}^2(\boldsymbol{\phi}) = \begin{bmatrix} \cos\phi_2 \\ -\sin\phi_2 \\ 0 \end{bmatrix}. \tag{3.11b}$$

Lemma 3.1.4 *For $d \geq 2$ we have*

$$-\sum_{j=1}^{d-1} \frac{d}{d\phi_j} \frac{\mathbf{t}^j(\boldsymbol{\phi}) |dS^{d-1}|(\boldsymbol{\phi})}{k_j} = (d-1)|dS^{d-1}|(\boldsymbol{\phi})\mathbf{n}(\boldsymbol{\phi}). \tag{3.12}$$

Proof: Without loss of generality we assume $\boldsymbol{\phi} \in O$ and proof the lemma by induction with respect to the spatial dimension d. First, note that the surface element $dS^{d-1} \geq 0$ on O, $d \geq 2$. Thus we can neglect the absolute value. The first induction step is done for $d = 2$

$$-\frac{d}{d\phi} \Big(\mathbf{t}(\phi) |dS^1|(\phi) \Big) = \mathbf{n}(\boldsymbol{\phi}) |dS^1|(\phi) \tag{3.13}$$

and is obviously true. Let us assume that the statement holds for a fixed $d \geq 2$. Then we have to show that

$$-\sum_{j=1}^{d} \frac{d}{d\phi_j} \frac{\mathbf{t}^j(\boldsymbol{\phi})|dS^d|(\boldsymbol{\phi})}{k_j} = d\ |dS^d|(\boldsymbol{\phi})\mathbf{n}(\boldsymbol{\phi}). \tag{3.14}$$

We split the sum in two parts

$$-\sum_{j=1}^{d} \frac{d}{d\phi_j} \frac{\mathbf{t}^j(\boldsymbol{\phi})|dS^d|(\boldsymbol{\phi})}{k_j} = -\frac{d}{d\phi_1}\left(\mathbf{t}^1(\boldsymbol{\phi})|dS^d|(\boldsymbol{\phi})\right) - \sum_{j=2}^{d} \frac{d}{d\phi_j} \frac{\mathbf{t}^j(\boldsymbol{\phi})|dS^d|(\boldsymbol{\phi})}{k_j}. \tag{3.15}$$

The first part satisfies

$$-\frac{d}{d\phi_1}\left(\mathbf{t}^1(\boldsymbol{\phi})|dS^d|(\boldsymbol{\phi})\right) = \mathbf{n}(\boldsymbol{\phi})|dS^d|(\boldsymbol{\phi}) - \mathbf{t}^1(\boldsymbol{\phi})(d-1)|dS^d|(\boldsymbol{\phi})\frac{\cos\phi_1}{\sin\phi_1} \tag{3.16}$$

due to (3.5) and (3.10), whereas the second part can be rewritten in the following way

$$\begin{aligned}
-\sum_{j=2}^{d} \frac{d}{d\phi_j} \frac{\mathbf{t}^j(\boldsymbol{\phi})|dS^d|(\boldsymbol{\phi})}{k_j} &= -\sum_{j=2}^{d} \frac{d}{d\phi_j}\left(\begin{bmatrix} \cos(\phi_j)\mathbf{n}(\phi_{j+1},...,\phi_d) \\ -\sin\phi_j \\ 0\cdot\mathbf{1}_{j-1} \end{bmatrix} \cdot \prod_{i=1,i\neq j}^{d} k_i\right) \\
&= -\sin(\phi_1)^{d-2}\sum_{j=2}^{d} \frac{d}{d\phi_j}\left(\begin{bmatrix} \cos(\phi_j)\mathbf{n}(\phi_{j+1},...,\phi_d) \\ -\sin\phi_j \\ 0\cdot\mathbf{1}_{j-1} \end{bmatrix} \cdot \prod_{i=2,i\neq j}^{d}\prod_{l=2}^{i-1}\sin\phi_l\right) \\
&= \sin(\phi_1)^{d-2}(d-1)|dS^{d-1}|(\phi_2,...,\phi_d)\begin{bmatrix} \mathbf{n}(\phi_2,...,\phi_d) \\ 0 \end{bmatrix} \\
&= \frac{(d-1)|dS^d|(\boldsymbol{\phi})}{\sin(\phi_1)}\begin{bmatrix} \mathbf{n}(\phi_2,...,\phi_d) \\ 0 \end{bmatrix}.
\end{aligned} \tag{3.17}$$

Here, we first plugged in (3.2) and (3.10) and then pulled the terms depending on ϕ_1 in front of the sum. Thus, we find ourselves in the situation of the spatial dimension $d-1$ with the angles $\phi_2, ..., \phi_{d-1}$ and can apply the induction hypothesis.

Putting together (3.15)-(3.17) we obtain

$$\begin{aligned}
-\sum_{j=1}^{d} \frac{d}{d\phi_j} \frac{\mathbf{t}^j(\boldsymbol{\phi})|dS^d|(\boldsymbol{\phi})}{k_j} &= \mathbf{n}(\boldsymbol{\phi})|dS^d|(\boldsymbol{\phi}) \\
&\quad + (d-1)|dS^d|(\boldsymbol{\phi})\left\{\frac{1}{\sin\phi_1}\begin{bmatrix} \mathbf{n}(\phi_2,...,\phi_d) \\ 0 \end{bmatrix} - \mathbf{t}^1(\boldsymbol{\phi})\frac{\cos\phi_1}{\sin\phi_1}\right\}.
\end{aligned} \tag{3.18}$$

Hence, we show that the bracket on the right-hand side equals $\mathbf{n}(\boldsymbol{\phi})$ to prove the lemma. Plugging in (3.10) we obtain

$$\begin{aligned}
&\frac{1}{\sin\phi_1}\begin{bmatrix} \mathbf{n}(\phi_2,...,\phi_d) \\ 0 \end{bmatrix} - \mathbf{t}^1(\boldsymbol{\phi})\frac{\cos\phi_1}{\sin\phi_1} \\
&= \frac{1}{\sin\phi_1}\begin{bmatrix} \mathbf{n}(\phi_2,...,\phi_d) \\ 0 \end{bmatrix} - \begin{bmatrix} \cos(\phi_1)\mathbf{n}(\phi_2,...,\phi_d) \\ -\sin\phi_1 \end{bmatrix}\frac{\cos\phi_1}{\sin\phi_1} \\
&= \mathbf{n}(\boldsymbol{\phi}).
\end{aligned} \tag{3.19}$$

□

Remark 3.1.5 *If $k_i(\boldsymbol{\phi}) = 0$, then $k_j(\boldsymbol{\phi}) = 0$ for all $j \geq i$. If additionally the index i is minimal with that property, i.e. $k_{i-1}(\boldsymbol{\phi}) \neq 0$, then*

1. $\phi_{i-1} \in \{0, \pi\}$
2. *the normal vector is*

$$\mathbf{n}(\boldsymbol{\phi}) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \pm k_{i-1} = k_{i-1}\cos\phi_{i-1} \\ k_{i-2}\cos\phi_{i-2} \\ \vdots \\ k_2\cos(\phi_2) \\ k_1\cos(\phi_1) \end{bmatrix} = \mathbf{n}(\phi_1, ..., \phi_{i-1}, 0, ..., 0) \tag{3.20}$$

3. *the vectors*

$$\mathbf{t}^{i-1}(\tilde{\boldsymbol{\phi}}) = \mathbf{e}_{d-i+1}\cos\tilde{\phi}_{i-1} = \pm\mathbf{e}_{d-i+1}, \quad \mathbf{t}^{j}(\tilde{\boldsymbol{\phi}}) = \mathbf{e}_{d-j}, \quad j = i, ..., d-1, \tag{3.21}$$

are tangential vectors, where $\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_1, ..., \tilde{\phi}_{d-1}) = (\phi_1, ..., \phi_{i-1}, 0, ..., 0)$. Thus, we have

$$\langle \mathbf{e}_1, ..., \mathbf{e}_{d-i+1} \rangle = \langle \mathbf{t}^{i-1}(\boldsymbol{\phi}), ..., \mathbf{t}^{d-1}(\boldsymbol{\phi}) \rangle \tag{3.22}$$

In the 3D case $k_2(\boldsymbol{\phi})$ is zero, if and only if $\mathbf{n}(\boldsymbol{\phi})$ is the north or south pole of the sphere. Then, the tangential plane is the (x, y)-plane with an orthonormal basis $\{\pm\mathbf{e}_2, \mathbf{e}_1\}$.

For a quasi-linear hyperbolic partial differential equation with spatial dependent Jacobians, the outer unit normal of a wave front, cf. Section 3.2 or [29, 96] for details, is a non-trivial continuously differentiable curve $\mathbf{n}(\boldsymbol{\phi}(t))$ on the sphere, that is given by the ordinary differential equation

$$\boldsymbol{\psi} = \frac{d}{dt}\mathbf{n}(\boldsymbol{\phi}(t)) \tag{3.23}$$

see [96]. Here, the mapping $\boldsymbol{\phi} : \mathbb{R} \to \mathbb{R}^{d-1}$ is not unique due to the periodicity of sine and cosine functions; and also due to Remark 3.1.5. Moreover, $\boldsymbol{\phi}$ may be discontinuous. Without loss of generality we assume that $\boldsymbol{\phi} : \mathbb{R} \to O$. Since $\mathbf{n} : O \backslash \partial O \to \mathbf{n}(O \backslash \partial O)$ is a diffeomorphism and $\mathbf{n}(\boldsymbol{\phi})$ is continuously differentiable, the mapping $\boldsymbol{\phi}$ is continuously differentiable on $\boldsymbol{\phi}^{-1}(O \setminus \partial O)$. Hence, we have

$$\boldsymbol{\psi} = \frac{d}{dt}\mathbf{n}(\boldsymbol{\phi}(t)) = \sum_{i=1}^{d-1} \frac{d\mathbf{n}}{d\phi_i}\frac{d\phi_i}{dt} = \sum_{i=1}^{d-1} k_i \mathbf{t}^i \frac{d\phi_i}{dt} \tag{3.24}$$

for all $t \in \boldsymbol{\phi}^{-1}(O \setminus \partial O)$.

In [3, 15], where spatial dependent Jacobians and the spatial dimension $d = 2$ were concerned, the following ODE system for the angle $\boldsymbol{\phi} = \phi_1$ is used

$$\frac{d\phi_1}{dt} = \boldsymbol{\psi} \cdot \mathbf{t}^1. \tag{3.25}$$

This is obtained by multiplying (3.24) with the tangential vector $\mathbf{t}^1$. In the case $d > 2$ multiplying (3.24) with the tangential vector $\mathbf{t}^j$, $j = 1, ..., d-1$, gives us the following ODE system

$$k_j \frac{d\phi_j}{dt} = \boldsymbol{\psi} \cdot \mathbf{t}^j \tag{3.26}$$

for the angles $\phi_j \in O \setminus \partial O$, $j = 1, ..., d-1$. But we are going to use the the ordinary differential equations (3.24) instead.

3.2. Evolution operator for a quasi-linear hyperbolic system

Consider a diagonally hyperbolic balance law of the quasi-linear form

$$\mathbf{w}_t + \sum_{i=1}^{d} A_i(\mathbf{w}, \mathbf{x}, t)\mathbf{w}_{x_i} = Q(\mathbf{w}) \in \mathbb{R}^p, \quad \mathbf{x} \in \mathbb{R}^d, \quad t \geq 0, \tag{3.27}$$

with the corresponding matrix pencil

$$\mathbb{P} = \sum_{i=1}^{d} A_i n_i, \quad \|\mathbf{n}\| = 1. \tag{3.28}$$

Let $\lambda_1, ..., \lambda_p$ be the eigenvalues of the matrix pencil $\mathbb{P}$ with the corresponding linear independent right and left eigenvectors $\mathbf{r}^1, ..., \mathbf{r}^p$, $\mathbf{l}^1, ..., \mathbf{l}^p$, so that $\mathbf{l}^i \cdot \mathbf{r}^j = \delta_{ij}$.

Applying the classic theory of hyperbolic partial differential equations [29, 96] it is known, that time evolution happens along the characteristic conoids (or surfaces) in space-time. For a point $P = (\mathbf{x}_P, t^{n+1})$ in space-time, the domain of dependence is the union of p conoid mantles, not necessarily pairwise different, with the apex P. A characteristic conoid is generated by a $(d-1)$-parameter family of bicharacteristic curves. The i-th family is generated by solutions of

$$\frac{dx_j^i}{dt} = (\mathbf{l}^i)^T A_j \mathbf{r}^i =: \chi_j^i, \quad \frac{dn_j(\boldsymbol{\phi}^i)}{dt} = -(\mathbf{l}^i)^T \left(\sum_{k,l=1}^{d} n_k n_l \left[n_k \frac{\partial A_l}{\partial x_j} - n_j \frac{\partial A_l}{\partial x_k} \right] \right) \mathbf{r}^i =: \psi_j^i \tag{3.29}$$

and initial values $\mathbf{x}^i(t^{n+1}) = \mathbf{x}_P, \mathbf{n}(\boldsymbol{\phi}^i(t^{n+1})) = \mathbf{n}(\omega) \in S^{d-1}$ for $i = 1, ..., p$, $j = 1, ..., d$, cf. [96, 97]. Here, $\mathbf{x}^i = \mathbf{x}^i(t)$ denotes the projection onto space of a bicharacteristic curve lying onto the i-th characteristic conoid - the so-called *rays*. Further, $\mathbf{n}(\boldsymbol{\phi}^i(t)) \in S^{d-1}$ is the outer normal vector of the wave front[1] corresponding to the i-th conoid at time t.

In the following of this section, we will develop an exact representation of the solution of system (3.27), that is based on rewriting the system in characteristic variables.

Let

$$R = \left[\mathbf{r}^1 \ ... \ \mathbf{r}^p \right], \quad R^{-1} = \left[\mathbf{l}^1 \ ... \ \mathbf{l}^p \right]^T, \tag{3.30}$$

be the matrices consisting of left and right eigenvectors and diagonalising the matrix pencil. Then,

$$\mathbf{v} = R^{-1}\mathbf{w} \tag{3.31}$$

[1] A wave front of a characteristic conoid consists of the points $\mathbf{x}$, so that $(\mathbf{x}(t^*), t^*)$ lies on the characteristic conoid at a fixed time t^*.

are the characteristic variables. We plug in $\mathbf{w} = R\mathbf{v}$ in system (3.27). First, we differentiate by means of the product rule. Then, we multiply the equation with the matrix R^{-1} from the left. This leads us to the system (3.27) rewritten in characteristic variables

$$\mathbf{v}_t + \sum_{i=1}^{d} B_i \mathbf{v}_{x_i} = R^{-1} \left[Q - \left(R_t + \sum_{i=1}^{d} A_i R_{x_i} \right) \mathbf{v} \right], \tag{3.32}$$

where

$$B_i = R^{-1} A_i R, \quad i = 1, ..., d. \tag{3.33}$$

Denoting the diagonal of B_i by D_i for $i = 1, ..., p$, we can quasi-diagonalise the system (3.32) and obtain

$$\mathbf{v}_t + \sum_{i=1}^{d} D_i \mathbf{v}_{x_i} = F + S, \tag{3.34}$$

with

$$F = R^{-1} \left[Q - \left(R_t + \sum_{i=1}^{d} A_i R_{x_i} \right) \mathbf{v} \right], \quad S = -\sum_{i=1}^{d} (B_i - D_i) \mathbf{v}_{x_i}. \tag{3.35}$$

Here, $(D_i)_j = \chi_j^i$ is the j-th component of the *ray velocity* corresponding to the i-th eigenvalue λ_i. Therefore, the p transport equations of the left-hand side of (3.34) are derivatives along bicharacteristics. More precisely, we have

$$\frac{dv_j(\mathbf{x}^j)}{dt} = \left(\mathbf{v}_t + \sum_{i=1}^{d} D_i \mathbf{v}_{x_i} \right)_j (\mathbf{x}^j), \quad j = 1, ..., p. \tag{3.36}$$

Integrating (3.34) along the bicharacteristic curves $(\mathbf{x}^j(t; \boldsymbol{\omega}), t)$ (3.29), $j = 1, ..., p$, from t^n to t^{n+1} and applying (3.36) we obtain

$$v_j(P) = v_j(\mathbf{x}^j(t^n; \boldsymbol{\omega}), t^n) + \int_{t^n}^{t^{n+1}} (F + S)(\mathbf{x}^j(t; \boldsymbol{\omega}), t) \, dt, \quad j = 1, ..., p. \tag{3.37}$$

We multiply (3.37) with the right-eigenvectors matrix R from the left to switch back to the conservative variables

$$\mathbf{w}(P) = R(\mathbf{x}_P, \boldsymbol{\omega}) \tilde{\mathbf{v}}^{n+1} = R(\mathbf{x}_P, \boldsymbol{\omega}) \tilde{\mathbf{v}}^n + \int_{t^n}^{t^{n+1}} R(\tilde{F} + \tilde{S})(t) \, dt, \tag{3.38}$$

where the tilde denotes that the i-th component is considered along the i-th bicharacteristic. However, only one bicharacteristic is taken into account explicitly in (3.38), namely the one determined by the condition $\mathbf{n}(\boldsymbol{\phi}(t^{n+1})) = \mathbf{n}(\boldsymbol{\omega})$. We average along all possible normal directions $\mathbf{n}(\boldsymbol{\omega})$ with $\boldsymbol{\omega} \in O$, so that all infinitely many directions of information propagation are explicitly taken into account. Thus we obtain the genuine multidimensional representation

$$\mathbf{w}(P) = \frac{1}{|S^{d-1}|} \int_O \left\{ R(\mathbf{x}_P, \boldsymbol{\omega}) \tilde{\mathbf{v}}^n + \int_{t^n}^{t^{n+1}} R(\mathbf{x}_P, \boldsymbol{\omega}) (\tilde{F} + \tilde{S})(t) \, dt \right\} |dS^{d-1}|(\boldsymbol{\omega}) \, d\boldsymbol{\omega}. \tag{3.39}$$

Remark 3.2.1 *Usually the Jacobians of system (3.27) are locally frozen, when evolution operators are derived for practical application in the literature. To the authors best knowledge there are only two publications [3, 15], where spatially varying fluxes are considered without a local linearisation of the PDE.*
The evolution operator development simplifies for linear systems in the following way: The Jacobians and the eigenvectors of the matrix pencil are constant. Hence, the ODE system (3.29) that determines the bicharacteristic curves simplifies, since $\boldsymbol{\phi}^i = \boldsymbol{\omega}$ is constant for $i = 1, ..., p$. Moreover, the term denoted by F in (3.35) simplifies to $F = R^{-1}Q$. The evolution operators and the corresponding FVEG schemes with local linearisation and no linearisation have been compared in [3] without observing significant differences.

3.3. Exact evolution operator for the linear subsystem of the alternative SWE

Let us recall a quasi-linear form of the linear subsystem for the alternative SWE (2.30)

$$\mathbf{w}_t + \sum_{i=1}^{d} A_i \mathbf{w}_{x_i} = 0, \tag{2.47}$$

$$\mathbf{w} = \begin{bmatrix} z \\ \mathbf{m} \end{bmatrix}, \quad A_i = \begin{bmatrix} 0 & \mathbf{e}_i^T \\ c_b^2 \mathbf{e}_i & 0 \end{bmatrix}, \quad c_b = \frac{\sqrt{-b}}{\varepsilon}, \tag{2.48}$$

where we assume the bottom topography b to be smooth. Our goal is to obtain the evolution operator (3.39) for the linear part (2.47), (2.48) following the procedure from the previous section. The first step is to obtain the quasi-diagonalised form

$$\mathbf{v}_t + \sum_{i=1}^{d} D_i \mathbf{v}_{x_i} = F + S, \tag{3.34}$$

with F, S from (3.35). From the eigenstructure that has been provided in Section 2.7 we get the matrices consisting of right- and left-eigenvectors

$$R = \begin{bmatrix} \frac{1}{c_b} & 0 & ... & 0 & \frac{1}{c_b} \\ -\mathbf{n} & \mathbf{t}^1 & ... & \mathbf{t}^{d-1} & \mathbf{n} \end{bmatrix}, \quad R^{-1} = \begin{bmatrix} \frac{c_b}{2} & -\frac{\mathbf{n}^T}{2} \\ 0 & (\mathbf{t}^1)^T \\ \vdots & \vdots \\ 0 & (\mathbf{t}^{d-1})^T \\ \frac{c_b}{2} & \frac{\mathbf{n}^T}{2} \end{bmatrix}. \tag{3.40}$$

Further, we compute the characteristic variables

$$\mathbf{v} = R^{-1}\mathbf{w} = \frac{1}{2} \begin{bmatrix} c_b z - \mathbf{m} \cdot \mathbf{n} \\ 2\mathbf{m} \cdot \mathbf{t}^1 \\ \vdots \\ 2\mathbf{m} \cdot \mathbf{t}^{d-1} \\ c_b z + \mathbf{m} \cdot \mathbf{n} \end{bmatrix} \tag{3.41}$$

and

$$B_i = R^{-1}A_iR = c_b \begin{bmatrix} -n_i & \frac{t_i^1}{2} & ... & \frac{t_i^{d-1}}{2} & 0 \\ t_i^1 & 0 & ... & 0 & t_i^1 \\ \vdots & \vdots & & \vdots & \vdots \\ t_i^{d-1} & 0 & ... & 0 & t_i^{d-1} \\ 0 & \frac{t_i^1}{2} & ... & \frac{t_i^{d-1}}{2} & n_i \end{bmatrix}, \quad i = 1, ..., d. \tag{3.42}$$

Due to (3.35), we get

$$S = -\frac{c_b}{2} \begin{bmatrix} \mathbf{t}^1 \cdot \nabla v_2 + ... + \mathbf{t}^{d-1} \cdot \nabla v_d \\ 2\mathbf{t}^1 \cdot \nabla(c_b z) \\ \vdots \\ 2\mathbf{t}^{d-1} \cdot \nabla(c_b z) \\ \mathbf{t}^1 \cdot \nabla v_2 + ... + \mathbf{t}^{d-1} \cdot \nabla v_d \end{bmatrix}, \tag{3.43}$$

$$F = zc_b \begin{bmatrix} -\frac{1}{2}\nabla c_b \cdot \mathbf{n} \\ \nabla c_b \cdot \mathbf{t}^1 \\ \vdots \\ \nabla c_b \cdot \mathbf{t}^{d-1} \\ \frac{1}{2}\nabla c_b \cdot \mathbf{n} \end{bmatrix} - (\mathbf{m} \cdot \mathbf{n}) \begin{bmatrix} 0 \\ \mathbf{t}^1 \cdot \frac{d\mathbf{n}}{dt} \\ \vdots \\ \mathbf{t}^{d-1} \cdot \frac{d\mathbf{n}}{dt} \\ 0 \end{bmatrix} - \sum_{j=1}^{d-1} (\mathbf{m} \cdot \mathbf{t}^j) \begin{bmatrix} \frac{1}{2}\mathbf{t}^j \cdot \frac{d\mathbf{n}}{dt} \\ \mathbf{t}^1 \cdot \frac{d\mathbf{t}^j}{dt} \\ \vdots \\ \mathbf{t}^{d-1} \cdot \frac{d\mathbf{t}^j}{dt} \\ -\frac{1}{2}\mathbf{t}^j \cdot \frac{d\mathbf{n}}{dt} \end{bmatrix}. \tag{3.44}$$

Next, we calculate the bicharacteristic curves using (3.29). Thus, we have $\chi_j^i = (B_j)_{ii}$. The right-hand side of the ODEs for the normal direction is

$$\boldsymbol{\psi}^1 = \nabla c_b - (\mathbf{n} \cdot \nabla c_b)\mathbf{n}, \quad \boldsymbol{\psi}^2 = ... = \boldsymbol{\psi}^d = 0, \quad \boldsymbol{\psi}^{d+1} = -(\nabla c_b - (\mathbf{n} \cdot \nabla c_b)\mathbf{n}) \tag{3.45}$$

by (3.29). Therefore, the ODE systems for the p families of bicharacteristic curves read

$$\begin{aligned} &\frac{d\mathbf{x}^1}{dt} = -c_b(\mathbf{x}^1)\mathbf{n}^1, && \frac{d\mathbf{x}^2}{dt} = ... = \frac{d\mathbf{x}^d}{dt} = 0, && \frac{d\mathbf{x}^{d+1}}{dt} = c_b(\mathbf{x}^{d+1})\mathbf{n}^{d+1}, \\ &\frac{d\mathbf{n}^1}{dt} = \nabla c_b - (\mathbf{n}^1 \cdot \nabla c_b)\mathbf{n}^1, && \frac{d\mathbf{n}^2}{dt} = ... = \frac{d\mathbf{n}^d}{dt} = 0, && \frac{d\mathbf{n}^{d+1}}{dt} = -(\nabla c_b - (\mathbf{n}^{d+1} \cdot \nabla c_b)\mathbf{n}^{d+1}), \end{aligned} \tag{3.46}$$

where $\mathbf{n}^i := \mathbf{n}(\boldsymbol{\phi}^i)$ for $i = 1, ..., d+1$. Note that if c_b is smooth, there exist a unique solution of (3.46) due to the Picard-Lindelöf theorem.

The next step is to compute the exact evolution operator

$$\mathbf{w}(P) = \frac{1}{|S^{d-1}|} \int_O \left\{ R(\mathbf{x}_P, \boldsymbol{\omega})\tilde{\mathbf{v}}^n + \int_{t^n}^{t^{n+1}} R(\mathbf{x}_P, \boldsymbol{\omega})(\tilde{F} + \tilde{S})(t)\ dt \right\} |dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega}. \tag{3.38}$$

Here, the integrals along all bicharacteristics appear. Due to the symmetry of the first and last bicharacteristic in (3.46), we can rewrite the integrals along the last bicharacteristic in ones along the first one, which simplifies the representation and saves computation time of discrete approximations.

Lemma 3.3.1 *Let $\mathbf{x}^i(t;\boldsymbol{\omega}), \boldsymbol{\phi}^i(t;\boldsymbol{\omega})$, $i = 1, ..., d+1$, be the solutions of the ODE system (3.46) with initial values $\mathbf{x}_P \in \mathbb{R}^d$ and $\mathbf{n}(\boldsymbol{\omega}) \in S^{d-1}, \mathbf{e}_1 = (1, 0..., 0)^T \in \mathbb{R}^{d-1}$. Then we have*

$$\mathbf{x}^1(t;\boldsymbol{\omega}+\pi\mathbf{e}_1) = \mathbf{x}^{d+1}(t;\boldsymbol{\omega}), \qquad \mathbf{n}(\boldsymbol{\phi}^1(t;\boldsymbol{\omega}+\pi\mathbf{e}_1)) = -\mathbf{n}(\boldsymbol{\phi}^{d+1}(t;\boldsymbol{\omega})), \tag{3.47a}$$

$$\mathbf{x}^{d+1}(t;\boldsymbol{\omega}+\pi\mathbf{e}_1) = \mathbf{x}^1(t;\boldsymbol{\omega}), \qquad \mathbf{n}(\boldsymbol{\phi}^{d+1}(t;\boldsymbol{\omega}+\pi\mathbf{e}_1)) = -\mathbf{n}(\boldsymbol{\phi}^1(t;\boldsymbol{\omega})), \tag{3.47b}$$

$$\mathbf{x}^2(t;\boldsymbol{\omega}) = ... = \mathbf{x}^d(t;\boldsymbol{\omega}) = \mathbf{x}_P, \quad \mathbf{n}(\boldsymbol{\phi}^2(t;\boldsymbol{\omega})) = ... = \mathbf{n}(\boldsymbol{\phi}^d(t;\boldsymbol{\omega})) = \mathbf{n}(\boldsymbol{\omega}). \tag{3.47c}$$

Proof: We obtain (3.47b) by replacing $\boldsymbol{\omega}$ with $\boldsymbol{\omega}+\pi\mathbf{e}_1$ in (3.47a).
The statement (3.47c) is obviously true due to the corresponding ODEs (3.46). Thus, it suffices to prove (3.47a). To this end, we show that $\mathbf{x}^{d+1}(t;\boldsymbol{\omega}), -\mathbf{n}(\boldsymbol{\phi}^{d+1}(t;\boldsymbol{\omega}))$ is a solution of the differential equations (3.46) for $\mathbf{x}^1(t;\boldsymbol{\omega}+\pi\mathbf{e}_1), \mathbf{n}(\boldsymbol{\phi}^1(t;\boldsymbol{\omega}+\pi\mathbf{e}_1))$. The initial values are identical. Further we have

$$\begin{aligned} \frac{d}{dt}\mathbf{x}^{d+1}(t;\boldsymbol{\omega}) &= c_b(\mathbf{x}^{d+1}(t;\boldsymbol{\omega}))\mathbf{n}(\phi^{d+1}(t;\boldsymbol{\omega})) \\ &= -c_b(\mathbf{x}^{d+1}(t;\boldsymbol{\omega}))\mathbf{n}(\phi^{d+1}(t;\boldsymbol{\omega})+\pi\mathbf{e}_1), \end{aligned} \tag{3.48a}$$

$$\begin{aligned} -\frac{d}{dt}\mathbf{n}(\phi^{d+1}(t;\boldsymbol{\omega})) &= (\nabla c_b)(\mathbf{x}^{d+1}(t;\boldsymbol{\omega})) \\ &\quad - \mathbf{n}(\boldsymbol{\phi}^{d+1}(t;\boldsymbol{\omega}))(\mathbf{n}(\boldsymbol{\phi}^{d+1}(t;\boldsymbol{\omega}))\cdot(\nabla c_b)(\mathbf{x}^{d+1}(t;\boldsymbol{\omega}))), \end{aligned} \tag{3.48b}$$

due to the ODE system (3.46) for the $(d+1)$-st bicharacteristic. Here, we obtain (3.48) using that $\sin(\phi+\pi) = -\sin(\phi), \cos(\phi+\pi) = -\cos(\phi)$ and that in each line of the parametrization $\mathbf{n}(\boldsymbol{\phi})$ either there is exactly one sine- or cosine function of the variable ϕ_1. Therefore shifting ω_1 by π equals multiplication with -1. □

Corollary 3.3.2 *The integrals along the last bicharacteristic can be rewritten as integrals along the first one via*

$$\begin{aligned} &\int_O f(\mathbf{x}^{d+1}(t;\boldsymbol{\omega}))\prod_{k=1}^{l} g_k(\boldsymbol{\phi}^{d+1}(t;\boldsymbol{\omega}))|dS^{d-1}|(\boldsymbol{\omega})\, d\boldsymbol{\omega} \\ &= (-1)^l \int_O f(\mathbf{x}^1(t;\boldsymbol{\omega}))\prod_{k=1}^{l} g_k(\boldsymbol{\phi}^1(t;\boldsymbol{\omega}))|dS^{d-1}|(\boldsymbol{\omega})\, d\boldsymbol{\omega}, \end{aligned} \tag{3.49}$$

where g_k, $k = 1, ..., l$, are components of the normal vector $\mathbf{n}$ or its derivatives with respect to ϕ_i, i.e. $d\mathbf{n}/d\phi_i = k_i\mathbf{t}^i$, $i = 1, ..., d-1$.

Proof: Note that $g_k(\phi_1, ..., \phi_{d-1}) = -g_k(\phi_1+\pi, \phi_2, ..., \phi_{d-1})$, $k = 1, ..., d-1$. Thus, the

substitution $\boldsymbol{\omega} = \boldsymbol{\omega} + \pi \mathbf{e}_1$ and application of Lemma 3.3.1 yield

$$\begin{aligned}
&\int_O f(\mathbf{x}^{d+1}(t;\boldsymbol{\omega})) \prod_{k=1}^{l} g_k(\boldsymbol{\phi}^{d+1}(t;\boldsymbol{\omega}))|dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega} \qquad (3.50)\\
&= \int_{\tilde{O}} f(\mathbf{x}^{d+1}(t;\boldsymbol{\omega}+\pi\mathbf{e}_1)) \prod_{k=1}^{l} g_k(\boldsymbol{\phi}^{d+1}(t;\boldsymbol{\omega}+\pi\mathbf{e}_1))|dS^{d-1}(\boldsymbol{\omega}+\pi\mathbf{e}_1)|\ d\boldsymbol{\omega}\\
&= (-1)^l \int_{\tilde{O}} f(\mathbf{x}^1(t;\boldsymbol{\omega})) \prod_{k=1}^{l} g_k(\boldsymbol{\phi}^1(t;\boldsymbol{\omega}))|dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega},\\
&= (-1)^l \int_{O} f(\mathbf{x}^1(t;\boldsymbol{\omega})) \prod_{k=1}^{l} g_k(\boldsymbol{\phi}^1(t;\boldsymbol{\omega}))|dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega}.
\end{aligned}$$

Here, the last equation follows since $\tilde{O}$ and O are different definition domains of parametrisation (3.1), see (3.3). □

Let us split the evolution operator (3.38) in two parts - integral of $R\tilde{\mathbf{v}}(t^n)$ and of $R(\tilde{F}+\tilde{S})$. We start with $R\tilde{\mathbf{v}}(t^n)$. Here, we apply the Corollary 3.3.2 and then the Lemma 3.1.1:

$$\begin{aligned}
&\frac{1}{|S^{d-1}|}\int_O R(\mathbf{x}_P;\boldsymbol{\omega})\tilde{\mathbf{v}}(t^n)|dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega} \qquad (3.51)\\
&= \frac{1}{|S^{d-1}|}\int_O \left[c_b z - \mathbf{m}\cdot\mathbf{n}(\boldsymbol{\phi})\right](\mathbf{x},t^n)\begin{bmatrix} \frac{1}{c_b(\mathbf{x}_P)} \\ -\mathbf{n}(\boldsymbol{\omega}) \end{bmatrix}|dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega}\\
&+ \frac{1}{|S^{d-1}|}\int_O \begin{bmatrix} 0 \\ \mathbf{m}(\mathbf{x}_P,t^n) - (\mathbf{m}(\mathbf{x}_P,t^n)\cdot\mathbf{n}(\boldsymbol{\omega}))\mathbf{n}(\boldsymbol{\omega}) \end{bmatrix}|dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega}\\
&= \frac{1}{|S^{d-1}|}\int_O \left[c_b z - \mathbf{m}\cdot\mathbf{n}(\boldsymbol{\phi})\right](\mathbf{x},t^n)\begin{bmatrix} \frac{1}{c_b(\mathbf{x}_P)} \\ -\mathbf{n}(\boldsymbol{\omega}) \end{bmatrix}|dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega} + \left(1-\frac{N_d^2}{|S^{d-1}|}\right)\begin{bmatrix} 0 \\ \mathbf{m}(\mathbf{x}_P,t^n) \end{bmatrix}
\end{aligned}$$

with $\mathbf{x} = \mathbf{x}^1(t^n;\boldsymbol{\omega})$, $\boldsymbol{\phi} = \boldsymbol{\phi}^1(t^n;\boldsymbol{\omega})$ and N_d^2 according to (3.8).

Now, let us compute the integral over $R(\tilde{F}+\tilde{S})$. The time derivatives of the tangents $\mathbf{t}^1,...,\mathbf{t}^{d-1}$ and the outer normal $\mathbf{n}$ in (3.44) vanish along the middle bicharacteristics, since $\boldsymbol{\phi}^2 = ..., \boldsymbol{\phi}^d = \boldsymbol{\omega}$ remain constant. Further, we plug the time derivatives of the outer normal vector along the first and last bicharacteristic from (3.46) into F, S, (3.43), (3.44), to obtain

$$\tilde{F}+\tilde{S} = \begin{bmatrix} -\frac{z(\mathbf{x}^1)c(\mathbf{x}^1)}{2}\mathbf{n}(\boldsymbol{\phi}^1)\cdot\nabla c_b(\mathbf{x}^1) \\ -c^2(\mathbf{x}_P)\mathbf{t}^1(\boldsymbol{\omega})\cdot\nabla z(\mathbf{x}_P) \\ \vdots \\ -c^2(\mathbf{x}_P)\mathbf{t}^{d-1}(\boldsymbol{\omega})\cdot\nabla z(\mathbf{x}_P) \\ \frac{z(\mathbf{x}^{d+1})c(\mathbf{x}^{d+1})}{2}\mathbf{n}(\boldsymbol{\phi}^{d+1})\cdot\nabla c_b(\mathbf{x}^{d+1}) \end{bmatrix} - \frac{1}{2}\sum_{j=1}^{d-1}\begin{bmatrix} \mathbf{t}^j(\boldsymbol{\phi}^1)\cdot\nabla(c_b(\mathbf{x}^1)\mathbf{t}^j(\boldsymbol{\phi}^1)\cdot\mathbf{m}(\mathbf{x}^1)) \\ 0 \\ \vdots \\ 0 \\ \mathbf{t}^j(\boldsymbol{\phi}^{d+1})\cdot\nabla(c_b(\mathbf{x}^{d+1})\mathbf{t}^j(\boldsymbol{\phi}^{d+1})\cdot\mathbf{m}(\mathbf{x}^{d+1})) \end{bmatrix}. \qquad (3.52)$$

After some tedious calculations we have

$$\frac{1}{|S^{d-1}|}\int_{t^n}^{t^{n+1}}\int_O R(\mathbf{x}_P,\boldsymbol{\omega})(\tilde{F}+\tilde{S})(t;\boldsymbol{\omega})|dS^{d-1}|(\boldsymbol{\omega})\;d\boldsymbol{\omega}\;dt \tag{3.53a}$$
$$= -\frac{1}{|S^{d-1}|}\int_{t^n}^{t^{n+1}}\int_O f(\mathbf{x},\boldsymbol{\phi})\begin{bmatrix}\frac{1}{c_b(\mathbf{x}_P)}\\ -\mathbf{n}(\boldsymbol{\omega})\end{bmatrix}|dS^{d-1}|(\boldsymbol{\omega})\;d\boldsymbol{\omega}\;dt$$
$$- c_b(\mathbf{x}_P)^2\left(1-\frac{N_d^2}{|S^{d-1}|}\right)\int_{t^n}^{t^{n+1}}\begin{bmatrix}0\\ (\nabla z)(\mathbf{x}_P,t)\end{bmatrix}dt$$

with

$$f(\mathbf{x},\boldsymbol{\phi}) = c_b(\mathbf{x})z(\mathbf{x},t)\mathbf{n}(\boldsymbol{\phi})\cdot\nabla c_b(\mathbf{x}) + \sum_{j=1}\mathbf{t}^j(\boldsymbol{\phi})\cdot\nabla\left[c_b(\mathbf{x})\mathbf{t}^j(\boldsymbol{\phi})\cdot\mathbf{m}(\mathbf{x},t)\right] \tag{3.53b}$$

and $\mathbf{x} = \mathbf{x}^1(t;\boldsymbol{\omega})$, $\boldsymbol{\phi} = \boldsymbol{\phi}^1(t;\boldsymbol{\omega})$. Here, we used Corollary 3.3.2 to rewrite the last family of bicharacteristics into the first one. The surface integrals along the middle bicharacteristics are executed, since the angles $\boldsymbol{\phi}^i(t;\boldsymbol{\omega}) = \boldsymbol{\omega}$, $i = 2,...,d$, are constant in time. Combining (3.51), (3.53) we obtain

$$\mathbf{w}(P) = \frac{1}{|S^{d-1}|}\int_O\left[c_b z - \mathbf{m}\cdot\mathbf{n}(\boldsymbol{\phi}^n)\right](\mathbf{x}^n,t^n)\begin{bmatrix}\frac{1}{c_b(\mathbf{x}_P)}\\ -\mathbf{n}(\boldsymbol{\omega})\end{bmatrix}|dS^{d-1}|(\boldsymbol{\omega})\;d\boldsymbol{\omega} \tag{3.54}$$
$$- \frac{1}{|S^{d-1}|}\int_O\int_{t^n}^{t^{n+1}} f(\mathbf{x},\boldsymbol{\phi})\;dt\begin{bmatrix}\frac{1}{c_b(\mathbf{x}_P)}\\ -\mathbf{n}(\boldsymbol{\omega})\end{bmatrix}|dS^{d-1}|(\boldsymbol{\omega})\;d\boldsymbol{\omega}$$
$$+ \left(1-\frac{N_d^2}{|S^{d-1}|}\right)\begin{bmatrix}0\\ \mathbf{m}(\mathbf{x}_P,t^n) - c_b^2\int_{t^n}^{t^{n+1}}(\nabla z)(\mathbf{x}_P,t)\;dt\end{bmatrix}$$

with $f(\mathbf{x},\boldsymbol{\phi})$ from (3.53b) and $\mathbf{x}^n = \mathbf{x}^1(t^n;\boldsymbol{\omega}), \mathbf{x} = \mathbf{x}^1(t;\boldsymbol{\omega}), \boldsymbol{\phi}^n = \boldsymbol{\phi}^1(t^n;\boldsymbol{\omega}), \boldsymbol{\phi} = \boldsymbol{\phi}^1(t;\boldsymbol{\omega})$. In order to rewrite $\int_{t^n}^{t^{n+1}}(\nabla z)(\mathbf{x}_P,t)\;dt$ we further integrate the momentum equation (2.35b) from t^n to t^{n+1} along the second bicharacteristic, which gives us

$$\mathbf{m}(P) = \mathbf{m}(\mathbf{x}_P,t^n) - c_b^2\int_{t^n}^{t^{n+1}}(\nabla z)(\mathbf{x}_P,t)\;dt. \tag{3.55}$$

Plugging (3.55) in (3.54) we obtain the so-called exact evolution operator

$$c_b(\mathbf{x}_P)z(P) = \frac{1}{|S^{d-1}|}\int_O \left[c_b z - \mathbf{m}\cdot\mathbf{n}(\boldsymbol{\phi}^n)\right](\mathbf{x}^n, t^n)|dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega} \tag{3.56a}$$

$$- \frac{1}{|S^{d-1}|}\int_O \int_{t^n}^{t^{n+1}} f(\mathbf{x}, t, \boldsymbol{\phi})\ dt|dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega},$$

$$m_i(P) = -\frac{1}{N_d^2}\int_O \left[c_b z - \mathbf{m}\cdot\mathbf{n}(\boldsymbol{\phi}^n)\right](\mathbf{x}^n, t^n) n_i(\boldsymbol{\omega})|dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega} \tag{3.56b}$$

$$+ \frac{1}{N_d^2}\int_O \int_{t^n}^{t^{n+1}} f(\mathbf{x}, t, \boldsymbol{\phi}) n_i(\boldsymbol{\omega})\ dt|dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega},$$

$$f(\mathbf{x}, t, \boldsymbol{\phi}) = \left[c_b z \mathbf{n}(\boldsymbol{\phi})\cdot\nabla c_b + \sum_{j=1}^{d-1} \mathbf{t}^j(\boldsymbol{\phi})\cdot\nabla(c_b \mathbf{t}^j(\boldsymbol{\phi})\cdot\mathbf{m})\right](\mathbf{x}, t)$$

where $\mathbf{x}^n = \mathbf{x}^1(t^n;\boldsymbol{\omega}), \mathbf{x} = \mathbf{x}^1(t;\boldsymbol{\omega}), \boldsymbol{\phi}^n = \boldsymbol{\phi}^1(t^n;\boldsymbol{\omega}), \boldsymbol{\phi} = \boldsymbol{\phi}^1(t;\boldsymbol{\omega})$. For $d = 2$ and

$$\mathbf{n}(\phi) = \begin{bmatrix} \cos(\phi) \\ \sin(\phi) \end{bmatrix}, \quad \mathbf{t}(\phi) = \begin{bmatrix} -\sin(\phi) \\ \cos(\phi) \end{bmatrix}, \tag{3.57}$$

(3.56) equals our result from [15]

$$c_b(\mathbf{x}_P)z(P) = \frac{1}{2\pi}\int_0^{2\pi} \left[c_b z - m_1\cos\phi^n - m_2\sin\phi^n\right](\mathbf{x}^n, t^n)\ d\omega \tag{3.58a}$$

$$- \frac{1}{2\pi}\int_0^{2\pi}\int_{t^n}^{t^{n+1}} \left[c_b z D_\phi^+[c_b] + D_\phi^-[m_1 c_b]\sin\phi - D_\phi^-[m_2 c_b]\cos\phi\right](\mathbf{x}, t)\ dt\ d\omega,$$

$$m_1(P) = -\frac{1}{\pi}\int_0^{2\pi} \left[c_b z - m_1\cos\phi^n - m_2\sin\phi^n\right](\mathbf{x}^n, t^n)\cos\omega\ d\omega \tag{3.58b}$$

$$+ \frac{1}{\pi}\int_0^{2\pi}\int_{t^n}^{t^{n+1}} \left[c_b z D_\phi^+[c_b] + D_\phi^-[m_1 c_b]\sin\phi - D_\phi^-[m_2 c_b]\cos\phi\right](\mathbf{x}, t)\cos\omega\ dt\ d\omega,$$

$$m_2(P) = -\frac{1}{\pi}\int_0^{2\pi} \left[c_b z - m_1\cos\phi^n - m_2\sin\phi^n\right](\mathbf{x}^n, t^n)\sin\omega\ d\omega \tag{3.58c}$$

$$+ \frac{1}{\pi}\int_0^{2\pi}\int_{t^n}^{t^{n+1}} \left[c_b z D_\phi^+[c_b] + D_\phi^-[m_1 c_b]\sin\phi - D_\phi^-[m_2 c_b]\cos\phi\right](\mathbf{x}, t)\sin\omega\ dt\ d\omega,$$

where $\mathbf{x}^n = \mathbf{x}^1(t^n;\boldsymbol{\omega}), \mathbf{x} = \mathbf{x}^1(t;\boldsymbol{\omega}), \phi^n = \phi^1(t^n;\boldsymbol{\omega}), \phi = \phi^1(t;\boldsymbol{\omega})$ and

$$D_\phi^+[f] := \mathbf{n}\cdot\nabla f, \quad D_\phi^-[f] := -\mathbf{t}\cdot\nabla f \tag{3.59}$$

denote normal and tangential derivatives.

3.4. Approximate evolution operator for the linear subsystem of the alternative SWE

In the previous section we have derived the exact evolution operator (3.56) for the linear subsystem of the alternative SWE. However, we are not able to evaluate any of the appearing integrals for two reasons: first, we do not have an analytical solution for the curves $\mathbf{x}^1, \boldsymbol{\phi}^1$. Second, we can not calculate the time integral, since the representation is implicit in time and the solution $\mathbf{w}$ is unknown for $t > t^n$.

We apply the explicit rectangle rule to approximate the time integrals and solve the ODE systems (3.46) numerically. Thus it suffices to approximate $\mathbf{x}^1(t;\boldsymbol{\omega}), \boldsymbol{\phi}^1(t;\boldsymbol{\omega})$ up to $\mathcal{O}(\Delta t^2)$-terms. Let $t \in [t^n, t^{n+1})$, then we have

$$\begin{aligned}
\mathbf{x}^1(t;\boldsymbol{\omega}) &= \mathbf{x}^1(t^{n+1};\boldsymbol{\omega}) - \int_t^{t^{n+1}} \frac{d\mathbf{x}^1}{d\tau}(\tau;\boldsymbol{\omega})\, d\tau && \text{(3.60a)}\\
&= \mathbf{x}_P + c_b(\mathbf{x}_P)\mathbf{n}(\boldsymbol{\omega})(t^{n+1} - t) + \mathcal{O}(\Delta t^2) =: \mathbf{x}(t;\boldsymbol{\omega}) + \mathcal{O}(\Delta t^2),\\
\mathbf{n}(\boldsymbol{\phi}^1(t;\boldsymbol{\omega})) &= \mathbf{n}(\boldsymbol{\phi}^1(t^{n+1};\boldsymbol{\omega})) - \int_t^{t^{n+1}} \frac{d\mathbf{n}(\boldsymbol{\phi}^1(\tau))}{d\tau}(\tau;\boldsymbol{\omega})\, d\tau && \text{(3.60b)}\\
&= \mathbf{n}(\boldsymbol{\omega}) - \{(\nabla c_b)(\mathbf{x}_P) - \mathbf{n}(\boldsymbol{\omega})\,[\mathbf{n}(\boldsymbol{\omega}) \cdot (\nabla c_b)(\mathbf{x}_P)]\}(t^{n+1} - t) + \mathcal{O}(\Delta t^2)\\
&= \mathbf{n}(\boldsymbol{\omega}) - \sum_{j=1}^{d-1} \mathbf{t}^j(\boldsymbol{\omega}) \left[\mathbf{t}^j(\boldsymbol{\omega}) \cdot (\nabla c_b)(\mathbf{x}_P)\right](t^{n+1} - t) + \mathcal{O}(\Delta t^2),\\
c_b(\mathbf{x}^1(t;\boldsymbol{\omega})) &= c_b(\mathbf{x}_P) + (t^{n+1} - t)c_b(\mathbf{x}_P)\mathbf{n}(\boldsymbol{\omega}) \cdot \nabla c_b(\mathbf{x}_P) + \mathcal{O}(\Delta t^2). && \text{(3.60c)}
\end{aligned}$$

Plugging (3.60) into the exact evolution operator (3.56) and dropping $\mathcal{O}(\Delta t^2)$ terms, we obtain the approximate evolution operator

$$\begin{aligned}
(c_b z)(P) &\approx \frac{1}{|S^{d-1}|} \int_O \left[c_b(\mathbf{x}_P) z - \mathbf{m} \cdot \mathbf{n}(\boldsymbol{\omega})\right](\mathbf{x}, t^n) |dS^{d-1}|(\boldsymbol{\omega})\, d\boldsymbol{\omega} && \text{(3.61a)}\\
&\quad - \frac{\Delta t}{|S^{d-1}|} \int_O \left[c_b(\mathbf{x}_P) \sum_{j=1}^{d-1} \mathbf{t}^j(\boldsymbol{\omega}) \cdot \nabla(\mathbf{t}^j(\boldsymbol{\omega}) \cdot \mathbf{m})\right](\mathbf{x}, t^n) |dS^{d-1}|(\boldsymbol{\omega})\, d\boldsymbol{\omega}\\
m_i(P) &\approx -\frac{1}{N_d^2} \int_O \left[c_b(\mathbf{x}_P) z - \mathbf{m} \cdot \mathbf{n}(\boldsymbol{\omega})\right](\mathbf{x}, t^n) n_i(\boldsymbol{\omega}) |dS^{d-1}|(\boldsymbol{\omega})\, d\boldsymbol{\omega} && \text{(3.61b)}\\
&\quad + \frac{\Delta t}{N_d^2} \int_O \left[c_b(\mathbf{x}_P) \sum_{j=1}^{d-1} \mathbf{t}^j(\boldsymbol{\omega}) \cdot \nabla(\mathbf{t}^j(\boldsymbol{\omega}) \cdot \mathbf{m})\right](\mathbf{x}, t^n) n_i(\boldsymbol{\omega}) |dS^{d-1}|(\boldsymbol{\omega})\, d\boldsymbol{\omega},
\end{aligned}$$

where $\mathbf{x} = \mathbf{x}(t^n;\boldsymbol{\omega})$.

The appearance of derivatives of $\mathbf{w}$ is not very suitable for the first order schemes. To this end we are going to rewrite the approximate evolution operator (3.61) without derivatives of $\mathbf{w}$ using the integration by parts.

Lemma 3.4.1 *Let $f \in C^1(\mathbb{R}^d)$, $p \in C^1(\mathbb{R}^{d-1})$ and $\mathbf{x}(t;\boldsymbol{\omega})$ be the approximation (3.60a) to the first bicharacteristic $\mathbf{x}^1(t;\boldsymbol{\omega})$. Then we have*

$$\int_\alpha^\beta \frac{dp}{d\omega_j}(\boldsymbol{\omega}) f(\mathbf{x}(t;\boldsymbol{\omega}))\ d\omega_j - p(\boldsymbol{\omega}) f(\mathbf{x}(t;\boldsymbol{\omega}))|_{\omega_j=\beta} + p(\boldsymbol{\omega}) f(\mathbf{x}(t;\boldsymbol{\omega}))|_{\omega_j=\alpha} \tag{3.62}$$

$$= -(t^{n+1}-t^n) c_b(\mathbf{x}_P) k_j(\boldsymbol{\omega}) \int_\alpha^\beta p(\boldsymbol{\omega}) \mathbf{t}^j(\boldsymbol{\omega}) \cdot (\nabla f)(\mathbf{x}(t;\boldsymbol{\omega}))\ d\omega_j$$

for $j = 1, ..., d-1$ and $\alpha, \beta \in \mathbb{R}$.

Proof: We apply the product rule

$$\frac{d}{d\omega_j}(p(\boldsymbol{\omega}) f(\mathbf{x}(t;\boldsymbol{\omega}))) = \frac{dp}{d\omega_j}(\boldsymbol{\omega}) f(\mathbf{x}(t;\boldsymbol{\omega})) \tag{3.63}$$
$$+ c_b(\mathbf{x}_P)(t^{n+1}-t) p(\boldsymbol{\omega}) k_j(\boldsymbol{\omega}) \mathbf{t}^j(\boldsymbol{\omega}) \cdot (\nabla f)(\mathbf{x}(t;\boldsymbol{\omega})).$$

Thus integrating (3.63) from α to β over ω_j proves the lemma. □

Thus, for $d = 2$ we have

$$\int_\alpha^\beta p'(\omega) f(\mathbf{x}(t;\omega))\ d\omega - p(\beta) f(\mathbf{x}(t;\beta)) + p(\alpha) f(\mathbf{x}(t;\alpha)) \tag{3.64}$$

$$= (t^{n+1}-t^n) c_b(\mathbf{x}_P) \int_\alpha^\beta p(\omega) D_\omega^-[f](\mathbf{x}(t;\omega))\ d\omega$$

with $f, p, \mathbf{x}$ as in Lemma 3.4.1 and D_ω^- from (3.59).

In a finite volume framework with hyperrectangle elements[2], we can divide the domain O into a union of subdomains O_l, $l = 1, ..., N$, so that the numerical solution $\mathbf{w}|_{O_l}$ is smooth for $l = 1, ..., N$. We apply Lemma 3.4.1 to the sum of tangential components in the approximate evolution operator (3.61) to obtain a representation that does not contain momentum derivatives. To this end, we introduce the following notation. Let

$$O_l = [\alpha_1^l, \beta_1^l] \times ... \times [\alpha_{d-1}^l, \beta_{d-1}^l] =: [\boldsymbol{\alpha}^l, \boldsymbol{\beta}^l] \tag{3.65}$$

with

$$\boldsymbol{\alpha}^l = (\alpha_1^l, ..., \alpha_{d-1}^l)^T, \quad \boldsymbol{\beta}^l = (\beta_1^l, ..., \beta_{d-1}^l)^T. \tag{3.66}$$

Further, we define

$$\boldsymbol{\alpha}(i) := (\alpha_1, ..., \alpha_{i-1}, \alpha_{i+1}, ..., \alpha_{d-1}), \quad d\boldsymbol{\omega}(i) = d\omega_1 ... d\omega_{i-1} d\omega_{i+1} ... d\omega_{d-1}, \tag{3.67}$$

for $i = 1, ..., d-1$, and

$$\int_{O_l} f(\boldsymbol{\omega})\ d\boldsymbol{\omega} = \int_{\alpha_1^l}^{\beta_1^l} ... \int_{\alpha_{d-1}^l}^{\beta_{d-1}^l} f(\boldsymbol{\omega})\ d\omega_1 ... d\omega_{d-1} =: \int_{\boldsymbol{\alpha}^l}^{\boldsymbol{\beta}^l} f(\boldsymbol{\omega})\ d\boldsymbol{\omega}. \tag{3.68}$$

[2]intervals for 1D, rectangles for 2D,...

Then we have

$$
\begin{aligned}
& c_b(\mathbf{x}_P)\Delta t \int_{O_l} \sum_{j=1}^{d-1} \mathbf{t}^j(\boldsymbol{\omega}) \cdot \nabla(\mathbf{t}^j(\boldsymbol{\omega}) \cdot \mathbf{m}(\mathbf{x}, t^n)) |dS^{d-1}|(\boldsymbol{\omega}) f(\boldsymbol{\omega})\ d\boldsymbol{\omega} \qquad (3.69) \\
&= \sum_{j=1}^{d-1} \sum_{i=1}^{d} c_b(\mathbf{x}_P)\Delta t \int_{\boldsymbol{\alpha}^l}^{\boldsymbol{\beta}^l} t_i^j(\boldsymbol{\omega}) \mathbf{t}^j(\boldsymbol{\omega}) \cdot \nabla m_i(\mathbf{x}, t^n) |dS^{d-1}|(\boldsymbol{\omega}) f(\boldsymbol{\omega})\ d\boldsymbol{\omega} \\
&= \sum_{j=1}^{d-1} \sum_{i=1}^{d} c_b(\mathbf{x}_P)\Delta t \int_{\boldsymbol{\alpha}(j)^l}^{\boldsymbol{\beta}^l(j)} \int_{\alpha_j^l}^{\beta_j^l} t_i^j(\boldsymbol{\omega}) \mathbf{t}^j(\boldsymbol{\omega}) \cdot \nabla m_i(\mathbf{x}, t^n) k_j(\boldsymbol{\omega}) \frac{|dS^{d-1}|(\boldsymbol{\omega})}{k_j(\boldsymbol{\omega})} f(\boldsymbol{\omega})\ d\omega_j\ d\boldsymbol{\omega}(j).
\end{aligned}
$$

Now, applying Lemma 3.4.1 with $p(\boldsymbol{\omega}) = -t_i^j(\boldsymbol{\omega})|dS^{d-1}|(\boldsymbol{\omega})f(\boldsymbol{\omega})/k_j(\boldsymbol{\omega})$, $j = 1, ..., d-1$, to the last line of (3.69), the product rule and afterwards Lemma 3.1.4 we obtain

$$
\begin{aligned}
& c_b(\mathbf{x}_P)\Delta t \int_{O_l} \sum_{j=1}^{d-1} \mathbf{t}^j(\boldsymbol{\omega}) \cdot \nabla(\mathbf{t}^j(\boldsymbol{\omega}) \cdot \mathbf{m}(\mathbf{x}, t^n)) |dS^{d-1}|(\boldsymbol{\omega}) f(\boldsymbol{\omega})\ d\boldsymbol{\omega} \qquad (3.70) \\
&\overset{3.4.1}{=} \sum_{j=1}^{d-1} \sum_{i=1}^{d} \int_{\boldsymbol{\alpha}^l(j)}^{\boldsymbol{\beta}^l(j)} \int_{\alpha_j^l}^{\beta_j^l} \left(-\frac{d}{d\omega_j} \frac{t_i^j(\boldsymbol{\omega}) f(\boldsymbol{\omega}) |dS^{d-1}|(\boldsymbol{\omega})}{k_j(\boldsymbol{\omega})} \right) m_i(\mathbf{x}, t^n)\ d\omega_j\ d\boldsymbol{\omega}(j) \qquad (A) \\
&+ \sum_{j=1}^{d-1} \sum_{i=1}^{d} \int_{\boldsymbol{\alpha}^l(j)}^{\boldsymbol{\beta}^l(j)} \left[\frac{t_i^j(\boldsymbol{\omega}) m_i(\mathbf{x}, t^n) f(\boldsymbol{\omega}) |dS^{d-1}|(\boldsymbol{\omega})}{k_j(\boldsymbol{\omega})} \right]_{\omega_j = \alpha_j}^{\omega_j = \beta_j} d\boldsymbol{\omega}(j) \qquad (B) \\
&= \sum_{j=1}^{d-1} \sum_{i=1}^{d} \int_{\boldsymbol{\alpha}^l(j)}^{\boldsymbol{\beta}^l(j)} \left\{ \int_{\alpha_j^l}^{\beta_j^l} \left(-f(\boldsymbol{\omega}) \frac{d}{d\omega_j} \frac{t_i^j(\boldsymbol{\omega}) |dS^{d-1}|(\boldsymbol{\omega})}{k_j(\boldsymbol{\omega})} - \frac{t_i^j(\boldsymbol{\omega}) |dS^{d-1}|(\boldsymbol{\omega})}{k_j(\boldsymbol{\omega})} \frac{df(\boldsymbol{\omega})}{d\omega_j} \right) m_i(\mathbf{x}, t^n)\ d\omega_j \right\} d\boldsymbol{\omega}(j) \\
&+ \sum_{j=1}^{d-1} \int_{\boldsymbol{\alpha}^l(j)}^{\boldsymbol{\beta}^l(j)} \left[\frac{\mathbf{t}^j(\boldsymbol{\omega}) \cdot \mathbf{m}(\mathbf{x}, t^n) f(\boldsymbol{\omega}) |dS^{d-1}|(\boldsymbol{\omega})}{k_j(\boldsymbol{\omega})} \right]_{\omega_j = \alpha_j}^{\omega_j = \beta_j} d\boldsymbol{\omega}(j) \\
&\overset{3.1.4}{=} \int_{O_l} \left[(d-1) f(\boldsymbol{\omega}) \mathbf{m}(\mathbf{x}, t^n) \cdot \mathbf{n}(\boldsymbol{\omega}) - \sum_{j=1}^{d-1} \frac{\mathbf{t}^j(\boldsymbol{\omega}) \cdot \mathbf{m}(\mathbf{x}, t^n)}{k_j(\boldsymbol{\omega})} \frac{df(\boldsymbol{\omega})}{d\omega_j} \right] |dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega} \qquad (C) \\
&+ \sum_{j=1}^{d-1} \int_{\boldsymbol{\alpha}^l(j)}^{\boldsymbol{\beta}^l(j)} \left[\frac{\mathbf{t}^j(\boldsymbol{\omega}) \cdot \mathbf{m}(\mathbf{x}, t^n) f(\boldsymbol{\omega}) |dS^{d-1}|(\boldsymbol{\omega})}{k_j(\boldsymbol{\omega})} \right]_{\omega_j = \alpha_j}^{\omega_j = \beta_j} d\boldsymbol{\omega}(j).
\end{aligned}
$$

In the approximate evolution operator (3.61) the function $f(\boldsymbol{\omega})$ is $n_i(\boldsymbol{\omega})$ for the m_i-component and $f(\boldsymbol{\omega}) = 1$ for z-component. Thus, the sum $\sum_{j=1}^{d-1} \frac{\mathbf{t}^j(\boldsymbol{\omega}) \cdot \mathbf{m}(\mathbf{x}, t^n)}{k_j(\boldsymbol{\omega})} \frac{df(\boldsymbol{\omega})}{d\omega_j}$ in (C) of

(3.70) is zero for z, because $\frac{df}{d\omega_j} = 0$. For m_i we obtain

$$\begin{aligned}
&(d-1)f(\boldsymbol{\omega})\mathbf{m}(\mathbf{x},t^n)\cdot\mathbf{n}(\boldsymbol{\omega}) - \sum_{j=1}^{d-1}\frac{\mathbf{t}^j(\boldsymbol{\omega})\cdot\mathbf{m}(\mathbf{x},t^n)}{k_j(\boldsymbol{\omega})}\frac{df(\boldsymbol{\omega})}{d\omega_j} \qquad (3.71)\\
&= (d-1)n_i(\boldsymbol{\omega})\mathbf{m}(\mathbf{x},t^n)\cdot\mathbf{n}(\boldsymbol{\omega}) - \sum_{j=1}^{d-1}\frac{\mathbf{t}^j(\boldsymbol{\omega})\cdot\mathbf{m}(\mathbf{x},t^n)}{k_j(\boldsymbol{\omega})}k_j(\boldsymbol{\omega})t_i^j(\boldsymbol{\omega})\\
&= (d-1)n_i(\boldsymbol{\omega})\mathbf{m}(\mathbf{x},t^n)\cdot\mathbf{n}(\boldsymbol{\omega}) - (m_i(\mathbf{x},t^n) - n_i\mathbf{n}(\boldsymbol{\omega})\cdot\mathbf{m}(\boldsymbol{\omega}))\\
&= d\ n_i(\boldsymbol{\omega})\mathbf{m}(\mathbf{x},t^n)\cdot\mathbf{n}(\boldsymbol{\omega}) - m_i(\mathbf{x},t^n).
\end{aligned}$$

Putting together (3.61), (3.70) and (3.71) we obtain the following approximate evolution operator

$$(c_b z)(P) \approx \frac{1}{|S^{d-1}|}\sum_{l=1}^{N}\int_{O_l}\left[c_b(\mathbf{x}_P)z - d\ \mathbf{m}\cdot\mathbf{n}(\boldsymbol{\omega})\right](\mathbf{x},t^n)|dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega} \qquad (3.72a)$$

$$-\frac{1}{|S^{d-1}|}\sum_{l=1}^{N}\sum_{j=1}^{d-1}\int_{\boldsymbol{\alpha}^l(j)}^{\beta^l(j)}\left[\frac{\mathbf{t}^j(\boldsymbol{\omega})\cdot\mathbf{m}(\mathbf{x},t^n)|dS^{d-1}|(\boldsymbol{\omega})}{k_j(\boldsymbol{\omega})}\right]_{\omega_j=\alpha_j^l}^{\omega_j=\beta_j^l} d\boldsymbol{\omega}(j)$$

$$m_i(P) \approx -\frac{1}{N_d^2}\sum_{l=1}^{N}\int_{O_l}\left[c_b z - (d+1)\mathbf{m}\cdot\mathbf{n}(\boldsymbol{\omega}) + \frac{m_i}{n_i(\boldsymbol{\omega})}\right](\mathbf{x},t^n)n_i(\boldsymbol{\omega})|dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega} \qquad (3.72b)$$

$$+\frac{1}{N_d^2}\sum_{l=1}^{N}\sum_{j=1}^{d-1}\int_{\boldsymbol{\alpha}^l(j)}^{\beta^l(j)}\left[\frac{\mathbf{t}^j(\boldsymbol{\omega})\cdot\mathbf{m}(\mathbf{x},t^n)n_i(\boldsymbol{\omega})|dS^{d-1}|(\boldsymbol{\omega})}{k_j(\boldsymbol{\omega})}\right]_{\omega_j=\alpha_j^l}^{\omega_j=\beta_j^l} d\boldsymbol{\omega}(j).$$

Particularly, we have for $d = 2$ and $\mathbf{n}(\omega) = (\cos(\omega), \sin(\omega))^T$

$$(c_b z)(P) \approx \frac{1}{2\pi}\int_0^{2\pi}\left[c_b(\mathbf{x}_P)z - 2\ m_1\cos(\omega) - 2\ m_2\sin(\omega)\right](\mathbf{x},t^n)\ d\omega \qquad (3.73a)$$

$$-\frac{1}{2\pi}\sum_{l=1}^{N}\left[-\sin(\omega)m_1(\mathbf{x},t^n) + \cos(\omega)m_2(\mathbf{x},t^n)\right]_{\omega=\alpha^l}^{\omega=\beta^l}$$

$$m_1(P) \approx -\frac{1}{\pi}\int_0^{2\pi}\left[c_b z + (1-3\cos^2(\omega))m_1 - 3\cos(\omega)\sin(\omega)m_2)\right](\mathbf{x},t^n)\ d\omega \qquad (3.73b)$$

$$+\frac{1}{\pi}\sum_{l=1}^{N}\left[-\sin(\omega)\cos(\omega)m_1(\mathbf{x},t^n) + \cos^2(\omega)m_2(\mathbf{x},t^n)\right]_{\omega=\alpha^l}^{\omega=\beta^l}$$

$$m_2(P) \approx -\frac{1}{\pi}\int_0^{2\pi}\left[c_b z - 3\cos(\omega)\sin(\omega))m_1 + (1-3\sin^2(\omega))m_2\right](\mathbf{x},t^n)\ d\omega \qquad (3.73c)$$

$$+\frac{1}{\pi}\sum_{l=1}^{N}\left[-\sin^2(\omega)m_1(\mathbf{x},t^n) + \cos(\omega)\sin(\omega)m_2(\mathbf{x},t^n)\right]_{\omega=\alpha^l}^{\omega=\beta^l}.$$

Remark 3.4.2 *If we would have used locally frozen coefficients, i.e. locally frozen bottom topography, to derive the evolution operator we would also obtain (3.72).*

3.5. Local evolution operator for the linear part of the alternative SWE

In the previous section we derived the explicit evolution operator (3.72). However, we want to use the evolution operator in the implicit scheme, as we did in [15]. To this end we consider the asymptotic behaviour of the evolution operator (3.72) for $\Delta t \to 0$.

This coincides with the derivation of the so-called local evolution operator in [107]. In this paper the authors derived an evolution operator for the Euler equations of gas dynamics. Letting the time increment Δt tend to zero in the operator provides an approximate solver for the generalised Riemann problem. Thus, an explicit scheme was obtained using the explicit rectangle rule as the time integrator. In [120] the same approach has been applied to the relativistic hydrodynamics equations.

Following the procedure used in [15, 107, 120] we point out that the sum (3.70) vanishes as the time increment Δt approaches zero. This can be obtained by considering the first equality of (3.70), resulted using Lemma 3.4.1. If Δt approaches zero, $m_i(\mathbf{x}, t^n)$ approaches the cell interface value $m_l|_{O_l}(\mathbf{x}_P)$ given by the smooth representation in the cell that contains the arch $\{\mathbf{x}(t^n; \boldsymbol{\omega}) : \boldsymbol{\omega} \in O_l\}$. Through m_l becoming the constant constant $m_l|_{O_l}(\mathbf{x}_P)$ we can apply the fundamental theorem of calculus for the term (A) and obtain (B) with the opposite sign in (3.70). Hence, the sum of (A) and (B) is zero in the limit $\Delta t \to 0$. Finally we obtain the local evolution operator for the linear part of the alternative SWE (2.30)

$$(c_b z)(P) \approx \frac{1}{|S^{d-1}|} \sum_{l=1}^{N} \int_{O_l} \left[c_b(\mathbf{x}_P) z - \mathbf{m} \cdot \mathbf{n}(\boldsymbol{\omega})\right](\mathbf{x}, t^n) |dS^{d-1}|(\boldsymbol{\omega})\, d\boldsymbol{\omega} \tag{3.74a}$$

$$m_i(P) \approx -\frac{1}{N_d^2} \sum_{l=1}^{N} \int_{O_l} \left[c_b(\mathbf{x}_P) z - \mathbf{m} \cdot \mathbf{n}(\boldsymbol{\omega})\right](\mathbf{x}, t^n) n_i(\boldsymbol{\omega}) |dS^{d-1}|(\boldsymbol{\omega})\, d\boldsymbol{\omega}, \tag{3.74b}$$

where

$$\mathbf{x} = \mathbf{x}(\tau; \boldsymbol{\omega}) = \mathbf{x}_P + \tau c_b(\mathbf{x}_P) \mathbf{n}(\omega), \quad \tau \geq 0. \tag{3.74c}$$

Here τ is zero or a positive small real number. In the case of $\tau = 0$ the approximate operator (3.72) corresponds to the local evolution operators used in [107, 120]. We think that the scheme is more stable, if we use a positive τ. In our numerical experiments we choose it using the CFL-condition

$$\max\left(\frac{|u_1| + c}{\Delta x}, \frac{|u_2| + c}{\Delta y}\right) \tau = CFL_{grav} \tag{3.75}$$

with the CFL-number $CFL_{grav} = 0.01$. For low Froude numbers this is basically the local time step τ from [15].

For $d = 2$ and the normal vector $\mathbf{n}(\omega) = (\cos(\omega), \sin(\omega))^T$ the local evolution operator

(3.74) reads

$$(c_b z)(P) \approx \frac{1}{2\pi} \int_0^{2\pi} \left[c_b(\mathbf{x}_P)z - \cos(\omega)m_1 - \sin(\omega)m_2\right](\mathbf{x}, t^n)\ d\omega \tag{3.76a}$$

$$m_1(P) \approx -\frac{1}{\pi} \int_0^{2\pi} \left[c_b(\mathbf{x}_P)z - \cos^2(\omega)m_1 - \cos(\omega)\sin(\omega)m_2\right](\mathbf{x}, t^n)\ d\omega, \tag{3.76b}$$

$$m_2(P) \approx -\frac{1}{\pi} \int_0^{2\pi} \left[c_b(\mathbf{x}_P)z - \cos(\omega)\sin(\omega)m_1 - \sin^2(\omega)m_2\right](\mathbf{x}, t^n)\ d\omega \tag{3.76c}$$

with $\mathbf{x}$ from (3.74c).

4. Time discretisation

In Chapter 1 we have pointed out that standard explicit schemes for the weakly compressible regime, i.e. let us say the Froude number $\varepsilon < 0.2$, are slow with regard to computational time. This is due to the strong time step restriction by the CFL-condition (1.3). Moreover, lower convergence rates or even breakdown of numerical solution may occur. In order to circumvent this problems we have introduced a splitting of the alternative SWE (2.30) into a stiff, linear subsystem governing the fast waves and a non-linear one taking care of the remaining slow waves in Chapter 2. Particularly, the corresponding systems read

$$\mathbf{w}_t + \nabla \cdot \mathcal{F}_L(\mathbf{w}) = K(\mathbf{w}), \quad \mathbf{w}_t + \nabla \cdot \mathcal{F}_{NL}(\mathbf{w}) = 0, \tag{2.37a}$$

$$\mathcal{F}_L(\mathbf{w}) = \begin{bmatrix} \mathbf{m} \\ -\dfrac{bz}{\varepsilon^2}\mathbb{1} \end{bmatrix}, \quad \mathcal{F}_{NL}(\mathbf{w}) = \begin{bmatrix} 0 \\ \dfrac{\mathbf{m}\otimes\mathbf{m}^T}{z-b} + \dfrac{z^2}{2\varepsilon^2}\mathbb{1} \end{bmatrix}, \tag{2.37b}$$

$$K(\mathbf{w}) = \begin{bmatrix} 0 \\ -\dfrac{z\nabla b}{\varepsilon^2} \end{bmatrix}, \tag{2.37c}$$

where $\mathcal{F}_L$ is the stiff, linear and $\mathcal{F}_{NL}$ the non-stiff, non-linear part of the flux. We present semi-implicit, also called implicit-explicit (IMEX), schemes of one- and multi-step type, cf. [5, 6, 17, 18, 19]. Here the stiff, linear system is treated implicitly, whereas the non-linear one explicitly. Thus, the corresponding CFL-condition for the semi-implicit, or implicit-explicit (IMEX), scheme becomes independent of the Froude number ε. Of course the presented schemes in this chapter can be applied to any ODE system with similar properties

$$y'(t) = L(y) + N(y). \tag{4.1}$$

Here the right-hand side is splitted in the two parts $L(y), N(y)$. There are several publications, where IMEX schemes for hyperbolic balance laws are used. In [38, 44, 45, 92, 108] the Euler, the isentropic Euler, the Navier-Stokes and the shallow water equations are considered. To this end, implicit and explicit backward difference formulas (BDF) or the so-called RK2CN scheme are used. Consequently, the stiff system is solved in time with the Crank-Nicholson method and the explicit one with the midpoint-rule.

In Section 4.1 we introduce the general formulation of IMEX R-K schemes in the context of the so-called partitioned schemes, see [18, 52]. Further, for a given IMEX Runge-Kutta (R-K) scheme we introduce two approaches to apply the scheme to the alternative SWE (2.30): the straight and elliptic approach. Particularly, in Section 4.2 we present a first order IMEX R-K scheme - IMEX Euler scheme - and in Section 4.3 two second order IMEX R-K schemes: ARS(2,2,2) and RK2CN scheme. In Section 4.4 we introduce the IMEX multi-step schemes. We also introduce the IMEX multi-step

straight and elliptic approach for the alternative SWE (2.30). Further we present the second order SBDF IMEX two-step scheme in Section 4.5. Finally we discuss the well-balanced property of the introduced IMEX time discretisations in Section 4.6.

4.1. IMEX Runge-Kutta schemes

An IMEX R-K scheme consists of one implicit and one explicit R-K scheme, i.e. there are two R-K tables

$$\begin{array}{c|c} \tilde{\mathbf{c}} & \tilde{A} \\ \hline & \tilde{\mathbf{b}}^T \end{array}, \quad \begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{b}^T \end{array}, \tag{4.2}$$

where the tables $(\tilde{A}, \tilde{\mathbf{b}}, \tilde{\mathbf{c}})$ represents the implicit and $(A, \mathbf{b}, \mathbf{c})$ the explicit scheme [1]. Further, we assume that

$$\tilde{\mathbf{c}} = \tilde{A}\mathbf{1}, \quad \mathbf{c} = A\mathbf{1}, \quad \mathbf{1} = (1, ..., 1)^T, \tag{4.3}$$

and that $\tilde{A}$ is a lower triangular matrix, i.e. the implicit R-K scheme is a diagonally implicit R-K (DIRK) scheme. The latter assumption assures that the explicit flux $\mathcal{F}_{NL}$ is always treated explicitly. We consider three types of IMEX R-K schemes, depending on the properties of the matrix $\tilde{A}$.

Definition 4.1.1 *We call an IMEX R-K scheme of type :*

- *A (see [94]) if the matrix $\tilde{A}$ is invertible.*
- *CK (after Carpenter and Kennedy [60]) if the matrix $\tilde{A} \in \mathbb{R}^{s\times s}$, $s \geq 2$, can be written as*

$$\tilde{A} = \begin{pmatrix} 0 & 0 \\ \boldsymbol{\alpha} & \tilde{A}_{s-1} \end{pmatrix}, \tag{4.4}$$

  where $\boldsymbol{\alpha} \in \mathbb{R}^{s-1}$ and $\tilde{A}_{s-1} \in \mathbb{R}^{s-1\times s-1}$ is invertible. If additionally $\boldsymbol{\alpha} = 0$ the scheme is called of type ARS (after Asher, Ruuth and Spiteri [5]).

Definition 4.1.2 *We call an s-stage IMEX R-K scheme consistent, if the condition*

$$\sum_{j=1}^{s} b_j = \sum_{j=1}^{s} \tilde{b}_j = 1 \tag{4.5}$$

is satisfied.

Particularly, IMEX R-K schemes belong to the class of the so-called *partitioned R-K methods*, cf. [52]. Consider the ODE system

$$z_t = f(z, m), \tag{4.6a}$$
$$m_t = g(z, m), \tag{4.6b}$$

[1] Typically the R-K scheme $(\tilde{A}, \tilde{\mathbf{b}}, \tilde{\mathbf{c}})$ is the explicit one and $(A, \mathbf{b}, \mathbf{c})$ is implicit. However, we want to follow the notation of [92], where $\tilde{\mathcal{F}}$ denotes the stiff flux.

where z, m may be vector-valued functions of different size. Due to different characters of (4.6a), (4.6b) it might be reasonable to treat these equations with different R-K schemes. For example, let the system (4.6) be stiff, where the stiffness is associated with the variable z. Then, we can apply an L-stable R-K scheme for (4.6a) and an explicit one for (4.6b). Thus, the corresponding partitioned R-K scheme reads

$$k_i = f(z^n + \Delta t \sum_{j=1}^{s} \tilde{a}_{ij} k_j, m^n + \Delta t \sum_{j=1}^{s} a_{ij} l_j), \tag{4.7a}$$

$$l_i = g(z^n + \Delta t \sum_{j=1}^{s} \tilde{a}_{ij} k_j, m^n + \Delta t \sum_{j=1}^{s} a_{ij} l_j), \tag{4.7b}$$

$$z^{n+1} = z^n + \Delta t \sum_{j=1}^{s} \tilde{b}_j k_j, \tag{4.7c}$$

$$m^{n+1} = m^n + \Delta t \sum_{j=1}^{s} b_j l_j, \tag{4.7d}$$

where the coefficients $\tilde{a}_{ij}, a_{ij}, \tilde{b}_j, b_j$ represent two R-K schemes, cf. [52]. Clearly, our splitting

$$\mathbf{w}_t = -\nabla \cdot \mathcal{F}_{NL} + [K - \nabla \cdot \mathcal{F}_L] \tag{4.8}$$

has not the same structure as (4.6). However, we can rewrite (4.8) in the following way

$$\boldsymbol{\phi}_t = -\nabla \cdot \mathcal{F}_{NL}(\mathbf{w}), \quad \boldsymbol{\psi}_t = K - \nabla \cdot \mathcal{F}_L(\mathbf{w}), \quad \mathbf{w} = \boldsymbol{\phi} + \boldsymbol{\psi}, \tag{4.9}$$

to apply the partitioned method, which results in

$$\mathbf{w}^{n+1} = \mathbf{w}^n - \Delta t \nabla \cdot \sum_{k=1}^{s} \left[\tilde{b}_k \mathcal{F}_L(\mathbf{w}_k) + b_k \mathcal{F}_{NL}(\mathbf{w}_k)\right] + \Delta t \sum_{k=1}^{s} \tilde{b}_k K(\mathbf{w}_k), \tag{4.10a}$$

$$\mathbf{w}_k = \mathbf{w}^n - \Delta t \nabla \cdot \sum_{j=1}^{k} [\tilde{a}_{kj} \mathcal{F}_L(\mathbf{w}_j) + a_{kj} \mathcal{F}_{NL}(\mathbf{w}_j)] + \Delta t \sum_{j=1}^{s} \tilde{a}_{kj} K(\mathbf{w}_j). \tag{4.10b}$$

The theory of partitioned methods provides the order conditions for numerical methods of order p, cf. [52]. We recall the conditions for $p = 1, 2$ in the following theorem.

Theorem 4.1.3 *1. A partitioned R-K scheme (4.7) is of order* 1*, if and only if both R-K schemes are of order* 1*.*

2. A partitioned R-K scheme (4.7) is of order 2*, if and only if both R-K schemes are of order* 2 *and the coupling conditions*

$$\mathbf{b} \cdot \tilde{\mathbf{c}} = \tilde{\mathbf{b}} \cdot \mathbf{c} = \frac{1}{2} \tag{4.11}$$

are satisfied.

If a space discretisation is given, time evolution of the splitted SWE (2.30) is obtained by solving the linear system corresponding to (4.10). Since it is a natural approach, we

call it the *straight approach*.
Another time marching strategy for splitted isentropic Euler equations, Euler equations and SWE, often used in literature, is based on solving an underlying elliptic equation for the pressure/ water depth. From the pressure/ water depth values at the new time level the momentum/ velocities can be updated explicitly, cf. [4, 38, 92, 108]. By using basically the same approach, we can derive an elliptic equation for the free surface elevation z. To this end, let us rewrite the R-K stages (4.10b) in the following way:

$$\hat{\mathbf{w}}_k = \mathbf{w}^n - \Delta t \nabla \cdot \sum_{j=1}^{k-1} [\tilde{a}_{kj} \mathcal{F}_L(\mathbf{w}_j) + a_{kj} \mathcal{F}_{NL}(\mathbf{w}_j)] + \Delta t \sum_{j=1}^{k-1} \tilde{a}_{kj} K(\mathbf{w}_j), \tag{4.12a}$$

$$z_k = \hat{z}_k - \Delta t \tilde{a}_{kk} \nabla \cdot \mathbf{m}_k, \tag{4.12b}$$

$$\mathbf{m}_k = \hat{\mathbf{m}}_k + \Delta t \tilde{a}_{kk} \frac{b \nabla z_k}{\varepsilon^2}, \tag{4.12c}$$

where (4.12a) is the explicit update of the stiff and non-stiff parts. Plugging (4.12c) in (4.12b), we get the desired elliptic equation

$$z_k + \left(\frac{\Delta t \tilde{a}_{kk}}{\varepsilon} \right)^2 \nabla \cdot (b \nabla z_k) = \hat{z}_k - \Delta t \tilde{a}_{kk} \nabla \cdot \hat{\mathbf{m}}_k \tag{4.13}$$

for the free surface elevation internal R-K stage z_k, $k = 1, ..., s$. Then, we can compute the internal R-K stages for the alternative SWE (2.30) by the following three steps:

1. compute the explicit update (4.12a)
2. obtain surface elevation at a new time step by solving the elliptic equation (4.13)
3. obtain new momentum at a new time step by explicit computation of (4.12c).

The time evolution using this method will be called the *elliptic approach*. Note that the straight and elliptic approach give the same results, if the solution is unique. But if we use a fully discrete scheme the two approaches may differ, cf. Chapter 6. The advantage of the elliptic approach is the following: the linear systems to be solved in every time step are two/three times smaller (for 1D/2D SWE) than solving the whole implicit part.

Remark 4.1.4 *In Chapter 5 we prove that IMEX R-K schemes of type A and CK are asymptotic preserving, if the last internal R-K stage is the solution at a new time level, i.e.* $\mathbf{w}_s = \mathbf{w}^{n+1}$. *Therefore we only use schemes with this property. Consequently we do not need to calculate (4.10a).*

4.2. First order IMEX Euler scheme

The simplest way to obtain an IMEX scheme is to apply the Euler schemes, i.e. we use implicit Euler scheme for the linear part and explicit Euler for the non-linear one. Thus, we obtain

$$\mathbf{w}^{n+1} = \mathbf{w}^n - \Delta t \nabla \cdot \left[\mathcal{F}_{NL}(\mathbf{w}^n) + \mathcal{F}_L(\mathbf{w}^{n+1}) \right] + \Delta t K(\mathbf{w}^{n+1}). \tag{4.14}$$

The corresponding R-K tables read

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}, \quad \begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}. \tag{4.15}$$

Thus, the IMEX Euler scheme is of type A.

If a space discretisation is given, the straight approach for the IMEX Euler scheme is to solve the linear system corresponding to (4.14) in every time step. The elliptic approach consists of the following three steps in every time step

$$\hat{\mathbf{w}} = \mathbf{w}^n - \Delta t \nabla \cdot \mathcal{F}_{NL}(\mathbf{w}^n), \tag{4.16a}$$

$$z^{n+1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 \nabla \cdot (b \nabla z^{n+1}) = \hat{z} - \Delta t \nabla \cdot \hat{\mathbf{m}}, \tag{4.16b}$$

$$\mathbf{m}^{n+1} = \hat{\mathbf{m}} + \Delta t \frac{b \nabla z^{n+1}}{\varepsilon^2}. \tag{4.16c}$$

4.3. Second order time discretisations

In this section we present two second order IMEX R-K schemes in order to approximate the alternative SWE (2.30).

4.3.1. ARS(2,2,2) scheme

The ARS(2,2,2) scheme is given by the R-K tables

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & \gamma & 0 & 0 \\ 1 & \delta & 1-\delta & 0 \\ \hline & \delta & 1-\delta & 0 \end{array}, \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & 0 & \gamma & 0 \\ 1 & 0 & 1-\gamma & \gamma \\ \hline & 0 & 1-\gamma & 0 \end{array}, \tag{4.17}$$

where $\gamma = 1 - 1/\sqrt{2}$ and $\delta = 1 - 1/(2\gamma)$, cf. [5]. It is of the ARS type, as indicated by its name. It consists of an explicit and implicit second order R-K scheme that satisfy the second order coupling condition (4.11). The three R-K stages applied to the splitting (2.37a) read

$$\mathbf{w}_1 = \mathbf{w}^n, \tag{4.18a}$$

$$\mathbf{w}_2 = \mathbf{w}^n - \Delta t \gamma \left[\nabla \cdot \mathcal{F}_{NL}(\mathbf{w}_1) + \nabla \cdot \mathcal{F}_L(\mathbf{w}_2) - K(\mathbf{w}_2)\right], \tag{4.18b}$$

$$\begin{aligned} \mathbf{w}_3 = \mathbf{w}^n &- \Delta t \left[\delta \nabla \cdot \mathcal{F}_{NL}(\mathbf{w}_1) + (1-\delta) \nabla \cdot \mathcal{F}_{NL}(\mathbf{w}_2)\right] \\ &- \Delta t \left[(1-\gamma)(\nabla \cdot \mathcal{F}_L(\mathbf{w}_2) - K(\mathbf{w}_2)) + \gamma(\nabla \cdot \mathcal{F}_L(\mathbf{w}_3) - K(\mathbf{w}_3))\right]. \end{aligned} \tag{4.18c}$$

The first R-K stage only constitutes that $\mathbf{w}_1 = \mathbf{w}^n$. Hence, the straight approach consists of two steps. In the first step we solve the linear system associated with the second R-K stage

$$\mathbf{w}_2 + \Delta t \gamma \left[\nabla \cdot \mathcal{F}_L(\mathbf{w}_2) - K(\mathbf{w}_2)\right] = \mathbf{w}^n - \Delta t \gamma \nabla \cdot \mathcal{F}_{NL}(\mathbf{w}^n), \tag{4.19}$$

that corresponds to the IMEX Euler scheme with the time step $\gamma\Delta t$. In the second step we solve the linear system associated with the third R-K stage

$$\begin{aligned}&\mathbf{w}^{n+1} + \Delta t\gamma(\nabla\cdot\mathcal{F}_L(\mathbf{w}^{n+1}) - K(\mathbf{w}^{n+1}))\\ &= \mathbf{w}^n - \Delta t\left[\delta\nabla\cdot\mathcal{F}_{NL}(\mathbf{w}^n) + (1-\delta)\nabla\cdot\mathcal{F}_{NL}(\mathbf{w}_2)\right]\\ &- \Delta t\left[(1-\gamma)(\nabla\cdot\mathcal{F}_L(\mathbf{w}_2) - K(\mathbf{w}_2))\right].\end{aligned} \tag{4.20}$$

Let us note that two steps are also needed for the elliptic approach. In the first step the IMEX Euler scheme with time step $\gamma\Delta t$ is applied

$$\hat{\mathbf{w}} = \mathbf{w}^n - \gamma\Delta t\nabla\cdot\mathcal{F}_{NL}(\mathbf{w}^n), \tag{4.21a}$$

$$z_2 + \left(\frac{\gamma\Delta t}{\varepsilon}\right)^2\nabla\cdot(b\nabla z_2) = \hat{z} - \gamma\Delta t\nabla\cdot\hat{\mathbf{m}}, \tag{4.21b}$$

$$\mathbf{m}_2 = \hat{\mathbf{m}} + \gamma\Delta t\frac{b\nabla z_2}{\varepsilon^2}. \tag{4.21c}$$

The second step reads

$$\begin{aligned}\hat{\mathbf{w}} = \mathbf{w}^n &- \Delta t\left[\delta\nabla\cdot\mathcal{F}_{NL}(\mathbf{w}^n) + (1-\delta)\nabla\cdot\mathcal{F}_{NL}(\mathbf{w}_2)\right]\\ &- \Delta t(1-\gamma)(\nabla\cdot\mathcal{F}_L(\mathbf{w}_2) - K(\mathbf{w}_2)),\end{aligned} \tag{4.22a}$$

$$z^{n+1} + \left(\frac{\gamma\Delta t}{\varepsilon}\right)^2\nabla\cdot(b\nabla z^{n+1}) = \hat{z} - \gamma\Delta t\nabla\cdot\hat{\mathbf{m}}, \tag{4.22b}$$

$$\mathbf{m}^{n+1} = \hat{\mathbf{m}} + \gamma\Delta t\frac{b\nabla z^{n+1}}{\varepsilon^2}. \tag{4.22c}$$

4.3.2. RK2CN scheme

The RK2CN scheme is a frequently used second order IMEX time discretisation, see e.g. [4, 92, 95, 108]. Here, the stiff subsystem is treated implicitly by means of the Crank-Nicholson scheme, whereas the second order Runge-Kutta scheme is applied for the non-linear part. More precisely, we apply the midpoint rule. Therefore, the first step of the scheme is to compute the values of $\mathbf{w}$ at the intermediate time step $t^{n+\frac{1}{2}}$. This is achieved by an IMEX Euler step. Thus, the straight approach for the RK2CN scheme reads

$$\mathbf{w}^{n+\frac{1}{2}} = \mathbf{w}^n - \frac{\Delta t}{2}\nabla\cdot\left[\mathcal{F}_{NL}(\mathbf{w}^n) + \mathcal{F}_L(\mathbf{w}^{n+\frac{1}{2}})\right] + \frac{\Delta t}{2}K(\mathbf{w}^{n+\frac{1}{2}}), \tag{4.23a}$$

$$\begin{aligned}\mathbf{w}^{n+1} = \mathbf{w}^n &- \Delta t\nabla\cdot\mathcal{F}_{NL}(\mathbf{w}^{n+\frac{1}{2}})\\ &- \frac{\Delta t}{2}\nabla\cdot\left[\mathcal{F}_L(\mathbf{w}^n) + \mathcal{F}_L(\mathbf{w}^{n+1})\right] + \frac{\Delta t}{2}\left[K(\mathbf{w}^n) + K(\mathbf{w}^{n+1})\right].\end{aligned} \tag{4.23b}$$

The RK2CN method can be described by the following R-K tables

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0\\ \frac{1}{2} & \frac{1}{2} & 0 & 0\\ 1 & 0 & 1 & 0\\ \hline & 0 & 1 & 0 \end{array}, \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0\\ \frac{1}{2} & 0 & \frac{1}{2} & 0\\ 1 & \frac{1}{2} & 0 & \frac{1}{2}\\ \hline & \frac{1}{2} & 0 & \frac{1}{2} \end{array}, \tag{4.24}$$

and is therefore of type CK. Let us note that the method satisfies the standard and coupling second order conditions (4.11).

For the elliptic approach using the RK2CN scheme we first apply the IMEX Euler step

$$\hat{\mathbf{w}} = \mathbf{w}^n - \frac{\Delta t}{2}\nabla \cdot \mathcal{F}_{NL}(\mathbf{w}^n), \tag{4.25a}$$

$$z^{n+\frac{1}{2}} + \left(\frac{\Delta t}{2\varepsilon}\right)^2 \nabla \cdot (b\nabla z_2) = \hat{z} - \frac{\Delta t}{2}\nabla \cdot \hat{\mathbf{m}}, \tag{4.25b}$$

$$\mathbf{m}^{n+\frac{1}{2}} = \hat{\mathbf{m}} + \frac{\Delta t}{2}\frac{b\nabla z^{n+\frac{1}{2}}}{\varepsilon^2}, \tag{4.25c}$$

with half time stepping and then the second step

$$\hat{\mathbf{w}} = \mathbf{w}^n - \Delta t\nabla \cdot \mathcal{F}_{NL}(\mathbf{w}^{n+\frac{1}{2}}) - \frac{\Delta t}{2}\left[\nabla \cdot \mathcal{F}_L(\mathbf{w}^n) - K(\mathbf{w}^n)\right], \tag{4.25d}$$

$$z^{n+1} + \left(\frac{\Delta t}{2\varepsilon}\right)^2 \nabla \cdot (b\nabla z^{n+1}) = \hat{z} - \frac{\Delta t}{2}\nabla \cdot \hat{\mathbf{m}}, \tag{4.25e}$$

$$\mathbf{m}^{n+1} = \hat{\mathbf{m}} + \frac{\Delta t}{2}\frac{b\nabla z^{n+1}}{\varepsilon^2}. \tag{4.25f}$$

4.4. IMEX multi-step schemes

An IMEX k-step scheme for the alternative SWE (2.30) reads

$$\begin{aligned}\mathbf{w}^{n+1} - \sum_{j=0}^{k-1}\alpha_j\mathbf{w}^{n-j} = \tilde{\beta}\left[\nabla \cdot \mathcal{F}_L(\mathbf{w}^{n+1}) - K(\mathbf{w}^{n+1})\right] \\ + \sum_{j=0}^{k-1}\left\{\tilde{\beta}_j\left[\nabla \cdot \mathcal{F}_L(\mathbf{w}^{n-j}) - K(\mathbf{w}^{n+1})\right] + \beta_j\nabla \cdot \mathcal{F}_{NL}(\mathbf{w}^{n-j})\right\}.\end{aligned} \tag{4.26}$$

It consists of the explicit k-step scheme

$$\mathbf{w}^{n+1} - \sum_{j=0}^{k-1}\alpha_j\mathbf{w}^{n-j} = \sum_{j=0}^{k-1}\beta_j\nabla \cdot \mathcal{F}_{NL}(\mathbf{w}^{n-j}), \tag{4.27a}$$

and the implicit k-step scheme

$$\begin{aligned}\mathbf{w}^{n+1} - \sum_{j=0}^{k-1}\alpha_j\mathbf{w}^{n-j} = \tilde{\beta}\left[\nabla \cdot \mathcal{F}_L(\mathbf{w}^{n+1}) - K(\mathbf{w}^{n+1})\right] \\ + \sum_{j=0}^{k-1}\tilde{\beta}_j\left[\nabla \cdot \mathcal{F}_L(\mathbf{w}^{n-j}) - K(\mathbf{w}^{n-j})\right],\end{aligned} \tag{4.27b}$$

for the ODEs

$$\mathbf{w}_t = -\nabla \cdot \mathcal{F}_{NL}(\mathbf{w}), \quad \mathbf{w}_t = K(\mathbf{w}) - \nabla \cdot \mathcal{F}_L(\mathbf{w}). \tag{4.28}$$

Definition 4.4.1 *We call the k-step schemes (4.26), (4.27) consistent, if*

$$\sum_{j=0}^{k-1} \alpha_j = 1 \tag{4.29}$$

is satisfied.

The order conditions of the IMEX multi-step schemes for constant time increment Δt can be found in [6]. Using Taylor expansion they can be easily generalised to varying time increments, e.g. for an IMEX two-step scheme, (4.26) with $k = 2$, the second order conditions read

$$\alpha_0 + \alpha_1 = 1, \tag{4.30a}$$

$$a + \alpha_1 c = \tilde{\beta} + \tilde{\beta}_1 + \tilde{\beta}_0 = \beta_0 + \beta_1, \tag{4.30b}$$

$$\frac{a^2 - \alpha_1 c^2}{2} = \tilde{\beta} a - \tilde{\beta}_1 c = -\beta_1 c. \tag{4.30c}$$

Here, the time increments are $a = t^{n+1} - t^n$, $c = t^n - t^{n-1}$. Note that the second order conditions (4.30) have two degrees of freedom: we can choose one of the α- and $\tilde{\beta}$-coefficients.

As in the case of IMEX R-K schemes we introduce the straight and elliptic approach. For the sake of completeness we repeat: If a space discretisation is given, time evolution of the SWE (2.30) can be obtained by solving the linear system corresponding to (4.26). This is referred to as the straight approach.

Similar to the IMEX R-K schemes, we can also derive an underlying elliptic equation for the free surface elevation z by rewriting (4.26) in the following way

$$\hat{\mathbf{w}} = \sum_{j=0}^{k-1} \left\{ \alpha_j \mathbf{w}^{n-j} + \beta_j \nabla \cdot \mathcal{F}_{NL}(\mathbf{w}^{n-j}) + \tilde{\beta}_j \left[\nabla \cdot \mathcal{F}_L(\mathbf{w}^{n-j}) - K(\mathbf{w}^{n-j}) \right] \right\}, \tag{4.31a}$$

$$z^{n+1} = \hat{z} + \tilde{\beta} \nabla \cdot \mathbf{m}^{n+1}, \tag{4.31b}$$

$$\mathbf{m}^{n+1} = \hat{\mathbf{m}} - \tilde{\beta} \frac{b \nabla z^{n+1}}{\varepsilon^2}. \tag{4.31c}$$

Plugging (4.31c) in (4.31b) we obtain the desired elliptic equation

$$z^{n+1} + \left(\frac{\beta}{\varepsilon} \right)^2 \nabla \cdot (b \nabla z^{n+1}) = \hat{z} + \beta \nabla \cdot \hat{\mathbf{m}}. \tag{4.32}$$

Then, we can compute time evolution for the alternative SWE (2.30) by the following three steps

1. compute the explicit update (4.31a)
2. obtain free surface elevation at a new time step by solving the elliptic equation (4.32)
3. obtain new momentum at a new time step by explicit computation of (4.31c).

The time evolution using this method is called elliptic approach. The advantage of the elliptic approach is the following: the linear systems to be solved in every time step are two/three times smaller (for 1D/2D SWE) than solving the whole implicit part.

α_0	α_1	$\tilde{\beta}$	β_0	β_1
$\frac{(a+c)^2}{c(2a+c)}$	$-\frac{a^2}{c(2a+c)}$	$-\frac{a(a+c)}{2a+c}$	$-\frac{a(a+c)^2}{c(2a+c)}$	$\frac{a^2(a+c)}{c(2a+c)}$

Table 4.1.: Coefficients of the backward difference method for non-constant times steps $a = t^{n+1} - t^n, c = t^n - t^{n-1}$.

4.5. Second order SBDF time discretisation

The BDF schemes can also be used in a semi-implicit way. Here, we can approximate the stiff / non-stiff part by implicit/ explicit BDF schemes. In [44, 59, 98] the so-called SBDF is used with constant time steps, whereas in [15, 95] the SBDF scheme is applied for adaptive time steps. To this end, the coefficients have to be chosen adaptively depending on the current and previous time steps.
Let us approximate time derivative at new time level by the adaptive BDF2 scheme, i.e.

$$\frac{\mathbf{w}^{n+1} - \alpha_0 \mathbf{w}^n - \alpha^1 \mathbf{w}^{n-1}}{-\beta} \approx \mathbf{w}^{n+1} = -\nabla \cdot \left[\mathcal{F}_L(\mathbf{w}^{n+1}) + \mathcal{F}_{NL}(\mathbf{w}^{n+1})\right] + K(\mathbf{w}^{n+1}). \quad (4.33)$$

This yields the fully implicit scheme

$$\mathbf{w}^{n+1} = \alpha_0 \mathbf{w}^n + \alpha_1 \mathbf{w}^{n-1} + \tilde{\beta}\left[\nabla \cdot \mathcal{F}_{NL}(\mathbf{w}^{n+1}) + \nabla \cdot \mathcal{F}_L(\mathbf{w}^{n+1}) - K(\mathbf{w}^{n+1})\right] \quad (4.34)$$

with $\alpha_0, \alpha_1, \beta$ coefficients from the corresponding Taylor expansion, see Table 4.1. Since we want to approximate the non-linear flux explicitly, we further approximate $\mathcal{F}_{NL}(\mathbf{w}^{n+1})$ by a linear interpolation. This yields the straight approach

$$\begin{aligned} \mathbf{w}^{n+1} &= \alpha_0 \mathbf{w}^n + \alpha_1 \mathbf{w}^{n-1} \\ &\quad + \nabla \cdot \left\{\tilde{\beta}\mathcal{F}_L(\mathbf{w}^{n+1}) + \beta_0 \mathcal{F}_{NL}(\mathbf{w}^n) + \beta_1 \mathcal{F}_{NL}(\mathbf{w}^{n-1})\right\} \\ &\quad - \tilde{\beta} K(\mathbf{w}^{n+1}). \end{aligned} \quad (4.35)$$

The corresponding elliptic approach reads

$$\hat{\mathbf{w}} = \alpha_0 \mathbf{w}^n + \alpha_{n-1} \mathbf{w}^{n-1} + \nabla \cdot \left[\beta_0 \mathcal{F}_{NL}(\mathbf{w}^n) + \beta_1 \mathcal{F}_{NL}(\mathbf{w}^{n-1})\right], \quad (4.36a)$$

$$z^{n+1} + \left(\frac{\tilde{\beta}}{\varepsilon}\right)^2 \nabla \cdot (b \nabla z^{n+1}) = \hat{z} + \tilde{\beta} \nabla \cdot \hat{\mathbf{m}}, \quad (4.36b)$$

$$\mathbf{m}^{n+1} = \hat{\mathbf{m}} - \tilde{\beta} \frac{b \nabla z^{n+1}}{\varepsilon^2}. \quad (4.36c)$$

Remark 4.5.1 *In [15] we additionally proposed to evaluate the non-linear flux $\mathcal{F}_{NL}$ at intermediate time steps $t^{n+\frac{1}{2}}$. Thus we can not imply from Section 4.4 that it is a second order time discretisation. Indeed the obtained numerical results are worse then for the time-discretisation (4.35). Therefore we do not consider this scheme anymore.*

4.6. Well-balanced property

In the introduction, we pointed out that a numerical scheme for the SWE has to be well-balanced. Let us recall:

Definition 4.6.1 *We call a numerical scheme for the alternative SWE (2.30) well-balanced, if it preserves the lake at rest equilibrium state,i.e.* $z = const., \mathbf{m} = 0$*, exactly.*

In [15], we already proved the well-balanced property for the IMEX Euler, RK2CN, and SBDF schemes, if suitable assumptions are satisfied. The aim of this section is to generalise this result for IMEX R-K and multi-step time discretisations. More precisely we prove the following theorem:

Theorem 4.6.2 *Let* $\Omega \subset \mathbb{R}^d$ *be a bounded Lipshitz-continuous domain and the bottom topography* $b \in L^\infty(\Omega)$*. Then a lake at rest state,* $z = const.$*,* $\mathbf{u} = 0$*, is preserved exactly for all times, if an IMEX R-K or an IMEX multi-step time discretisation is used and*

$$\int_{\partial\Omega} bz\nabla z \cdot \mathbf{n} \, ds \geq 0 \tag{4.37}$$

holds.

Remark 4.6.3 *Condition (4.37) is satisfied for many practical situations; for example, when the homogeneous Dirichlet or Neumann boundary conditions are assumed for the perturbation* z *or in the case of periodic boundary conditions.*

Remark 4.6.4 *Let us recall that the straight and elliptic approach time discretisations are equivalent. Thus, it suffices to show Theorem 4.6.2 for the straight approach.*

Let $\tilde{\mathbf{w}} = (Z, 0)$ be a lake at rest solution at a time step t^0 and in the case of a k-step scheme also at times $t^{-1}, ..., t^{-(k-1)}$. We prove Theorem 4.6.2 for multi-step schemes in two steps. In the first step, we show that the lake at rest state $\tilde{\mathbf{w}}$ is a solution of the discrete alternative SWE (2.30). In the second step, we show that the solution is unique. For IMEX R-K schemes we repeat these two steps for every R-K stage. In order to execute steps one and two, we need the following lemmas.

Lemma 4.6.5 *Let* $\tilde{\mathbf{w}} = (Z, 0)$ *be a lake at rest state. Then, we have*

$$\nabla \cdot \mathcal{F}_{NL}(\tilde{\mathbf{w}}) = 0, \quad \Phi(\tilde{\mathbf{w}}) := (\nabla \cdot \mathcal{F}_L - K)(\tilde{\mathbf{w}}) = 0. \tag{4.38}$$

Lemma 4.6.6 *Let the assumptions from Theorem 4.6.2 hold and* $\tilde{\mathbf{w}} = (Z, 0)$ *be a lake at rest state. Then,* $\mathbf{w} = \tilde{\mathbf{w}}$ *is the unique solution of the following equation in* $H^1(\Omega)^{d+1}$

$$\mathbf{w} = \tilde{\mathbf{w}} - \delta\Phi(\mathbf{w}), \tag{4.39}$$

where δ *is a constant.*

Proof: Due to Lemma 4.6.5, $\mathbf{w} = \tilde{\mathbf{w}} \in H^1(\Omega)^{d+1}$ is obviously a solution. It is unique, if and only if the kernel of $Id + \delta\Phi(\cdot)$ is trivial, which we show in the following. A function $\mathbf{w} \in H^1(\Omega)^{d+1}$ is an element of the kernel, if and only if

$$z + \delta \nabla \cdot \mathbf{m} = 0, \tag{4.40a}$$

$$\mathbf{m} = \delta \frac{b \nabla z}{\varepsilon^2}. \tag{4.40b}$$

Thus $b\nabla z \in H^1(\Omega)^d$ due to (4.40b). Hence we can plug (4.40b) in (4.40a) and obtain the following elliptic eigenvalue problem

$$-\nabla \cdot (b \nabla z) = \lambda z, \quad \lambda = \frac{\varepsilon^2}{\delta^2}. \tag{4.41}$$

Multiplying (4.41) with z and integrating over the domain Ω we obtain

$$0 \leq \lambda \|z\|_{L^2(\Omega)} = -\langle z, \nabla \cdot (b \nabla z) \rangle_{L^2(\Omega)} = \int_\Omega b \nabla z \cdot \nabla z \, d\mathbf{x} - \int_{\partial\Omega} bz \nabla z \cdot \mathbf{n} \, ds \leq 0, \tag{4.42}$$

where the last inequality is due to (4.37) and $b < 0$. Thus, the kernel of $Id + \delta\Phi(\cdot)$ is trivial. □

Corollary 4.6.7 *Let the assumptions from Theorem 4.6.2 hold, $\mathbf{w}^0 = \tilde{\mathbf{w}} = (Z, 0)$ be a lake at rest state at time t^0 and the alternative SWE (2.30) be discretised by an IMEX R-K scheme according to Theorem 4.6.2. Then all internal R-K stages and the solution at the following time step t^1 satisfy $\mathbf{w}_1 = ... = \mathbf{w}_s = \mathbf{w}^1 = \tilde{\mathbf{w}} \in H^1(\Omega)^{d+1}$.*

Proof: We prove that the internal R-K stages $\mathbf{w}_k = \tilde{\mathbf{w}}$, $k = 1, ..., s$, by induction. Then Lemma 4.6.5 and the R-K update (4.10a) imply that $\mathbf{w}^1 = \mathbf{w}^0 = \tilde{\mathbf{w}}$. For $k = 1$ we have

$$\mathbf{w}_1 = \tilde{\mathbf{w}} - \Delta t \tilde{a}_{11} \left[\nabla \cdot \mathcal{F}_L(\mathbf{w}_1) - K(\mathbf{w}_1) \right] \tag{4.43}$$

due to Lemma 4.6.5. From (4.43), we obtain $\mathbf{w}_1 = \tilde{\mathbf{w}}$ due to Lemma 4.6.6. The inductive step follows analogously. □

Corollary 4.6.8 *Let the assumptions from Theorem 4.6.2 hold, $\mathbf{w}^0 = \mathbf{w}^{-1} = ... = \mathbf{w}^{-(k-1)} = \tilde{\mathbf{w}} = (Z, 0)$ be a lake at rest state at time t^0 and the alternative SWE (2.30) be discretised by an k-step IMEX scheme according to Theorem 4.6.2. Then $\mathbf{w}^1 = \tilde{\mathbf{w}}$.*

Proof: Due to Lemma 4.6.5, (4.26) and the consistency condition (4.29), the solution at the next time step is given by

$$\mathbf{w}^1 = \tilde{\mathbf{w}} + \tilde{\beta} \left[\nabla \cdot \mathcal{F}_L(\mathbf{w}^1) - K(\mathbf{w}^1) \right]. \tag{4.44}$$

Again, Lemma 4.6.6 implies that $\mathbf{w}^1 = \tilde{\mathbf{w}}$. □

5. Asymptotic preserving property of time discretisation

In Chapter 1, we pointed out that standard solvers for compressible flow fail in the weakly compressible and incompressible regime, i.e. when the Froude or Mach number is small, $\varepsilon \ll 1$ or even approaches zero, numerical solution may break down or converge slower. However, one wants to use schemes that are capable of approximating solutions in all regimes in a stable and accurate way. One expects that the alternative SWE (2.30) converge to the alternative zero Froude number SWE (2.34). Therefore our scheme has to satisfy the so-called *asymptotic preserving property* introduced by Jin [57]:

- the scheme has to inherit desirable features of a compressible solver when $\varepsilon = \mathcal{O}(1)$, like non-oscillatory solution profiles and good resolution of shock type discontinuities
- the scheme has to yield a consistent approximation of the limit equations as $\varepsilon \to 0$, where a possible stability constraint must be independent of ε.

Typically, the asymptotic preserving property for the SWE (2.18) is shown for the time discretisation in the following way: We assume that the initial data are *well-prepared*, i.e. they can be written in the asymptotic expansion

$$z^l(\mathbf{x}) = z^{(0)} + \varepsilon z^{(1)} + \varepsilon^2 z^{l,(2)}(\mathbf{x}) + \mathcal{O}(\varepsilon^3), \tag{5.1a}$$

$$\mathbf{m}^l(\mathbf{x}) = \mathbf{m}^{l,(0)}(\mathbf{x}) + \varepsilon \mathbf{m}^{l,(1)}(\mathbf{x}) + \varepsilon^2 \mathbf{m}^{l,(2)}(\mathbf{x}) + \mathcal{O}(\varepsilon^3), \tag{5.1b}$$

$$\nabla \cdot \mathbf{m}^l(\mathbf{x}) = \varepsilon^2 \mathbf{m}^{l,(2)}(\mathbf{x}) + \mathcal{O}(\varepsilon^3), \tag{5.1c}$$

where the functions $z^{(i)}, \mathbf{m}^{(i)}$, $i = 0, 1, 2$, are independent of the Froude number. Moreover, we have $l = n$ for IMEX R-K schemes and $l = n - k + 1, ..., n$ for IMEX k-step schemes.

The expansion (5.1) is plugged into the numerical scheme. The aim is to show that the limit of the numerical scheme for $\varepsilon \to 0$ is a consistent approximation of the zero Froude number SWE. We refer to [14, 15, 27, 38, 51, 58, 92, 108] and the references therein for examples of asymptotic preserving schemes and relevant asymptotic techniques. Also recent publications by Boscarino et al. [17, 18, 19] are noteworthy. Here the authors study the IMEX schemes for the following ordinary differential equation

$$\mathbf{w}_t = F(\mathbf{w}) + \frac{1}{\varepsilon} G(\mathbf{w}) \tag{5.2}$$

with $\varepsilon > 0$ the stiffness parameter. Systems of such form arise from the discretisation of partial differential equations, like convection-diffusion problems or hyperbolic balance laws with a relaxation term.

In this chapter we study the asymptotic behaviour of the IMEX R-K and multi-step schemes introduced in Chapter 4. Combining the approaches used in [18] and the standard asymptotic preserving proof procedure, cf. [15], we show in Section 5.1 that an IMEX R-K scheme of type A and CK, in particular ARS, is asymptotic preserving, if it is *globally stiffly accurate*, see Definition 5.1.4. In Section 5.2, we show that any consistent IMEX multi-step scheme, cf. Section 4.4, is asymptotic preserving. Consequently, the IMEX Euler, RK2CN, ARS(2,2,2) and SBDF schemes are asymptotic preserving.

5.1. IMEX Runge-Kutta time discretisation

Following [27] we define the asymptotic preserving property in the following way:

Definition 5.1.1 *Let $\mathcal{P}^\varepsilon$ be a singular perturbation problem of a problem $\mathcal{P}^0$, $\varepsilon \to 0$. Then, a consistent scheme $\mathcal{P}^{\varepsilon,h}$ for $\mathcal{P}^\varepsilon$ with a discretisation parameter h is asymptotic preserving, if the following conditions are satisfied:*

- *the limit of the scheme $\mathcal{P}^{\varepsilon,h}$ as $\varepsilon \to 0$, $\mathcal{P}^{0,h}$, is a consistent approximation of $\mathcal{P}^0$*
- *$\mathcal{P}^{\varepsilon,h}$ is stable independently on the parameter ε; in particular a stability condition does not depend on ε.*

We have studied the asymptotic preserving property of the IMEX Euler and RK2CN scheme for the alternative SWE (2.30) in our recent publication [15]. Following the analysis presented in [18], we can prove the asymptotic preserving property of IMEX R-K time discretisations of type A and CK for the alternative SWE (2.30), cf. Chapter 4. Therefore we rewrite the IMEX time discretisation of the alternative SWE (2.30) in the following way

$$z^{n+1} = z^n - \Delta t \sum_{k=1}^{s} \tilde{b}_k g_1(\mathbf{m}_k), \tag{5.3a}$$

$$\varepsilon^2 \mathbf{m}^{n+1} = \varepsilon^2 \mathbf{m}^n - \Delta t \sum_{k=1}^{s} \tilde{b}_k g_2(z_k) - \Delta t \sum_{k=1}^{s} b_k \left[f_1(z_k) + \varepsilon^2 f_2(z_k, \mathbf{m}_k) \right], \tag{5.3b}$$

$$g_1(\mathbf{m}) = \nabla \cdot \mathbf{m}, \quad g_2(z) = -b \nabla z, \quad f_1(z) = z \nabla z, \quad f_2(z, m) = \nabla \cdot \frac{\mathbf{m} \otimes \mathbf{m}}{z - b}. \tag{5.3c}$$

Here, the internal R-K stages $z_k, \mathbf{m}_k$ are given by

$$z_k = z^n - \Delta t \sum_{j=1}^{k} \tilde{a}_{kj} g_1(\mathbf{m}_j), \tag{5.3d}$$

$$\varepsilon^2 \mathbf{m}_k = \varepsilon^2 \mathbf{m}^n - \Delta t \sum_{j=1}^{k} \tilde{a}_{kj} g_2(z_j) - \Delta t \sum_{j=1}^{k-1} a_{kj} \left[f_1(z_j) + \varepsilon^2 f_2(z_j, \mathbf{m}_j) \right], \tag{5.3e}$$

where the coefficients $\tilde{a}_{kj}, \tilde{b}_k$ correspond to a DIRK method and a_{kj}, b_k to an explicit R-K scheme.

5.1.1. IMEX Runge-Kutta schemes of type A

We show in this section that an IMEX R-K scheme (5.3) of type A is asymptotic preserving, if it is *globally stiffly accurate*, see Definition 5.1.4. To this end we first show an important property of the internal R-K stages.

Lemma 5.1.2 *Assume well-prepared initial data and periodic boundary conditions for the alternative SWE (2.30). If we choose the IMEX R-K time discretisation (5.3) of type A, then the R-K stages* $(z_k, \mathbf{m}_k)$, $1 \leq k \leq s$, *satisfy*

$$z_k = z^n + \mathcal{O}(\varepsilon^2), \quad \nabla \cdot \mathbf{m}_k = \mathcal{O}(\varepsilon^2) \tag{5.4}$$

under the assumption that the internal R-K stages remain bounded with respect to ε.

Proof: Introducing the notation

$$\bar{z} = \begin{bmatrix} z_1 \\ \vdots \\ z_s \end{bmatrix}, \quad \bar{\mathbf{m}} = \begin{bmatrix} \mathbf{m}_1 \\ \vdots \\ \mathbf{m}_s \end{bmatrix}, \tag{5.5}$$

we set

$$f(\bar{z}, \bar{\mathbf{m}}) = \begin{bmatrix} f(z_1, \mathbf{m}_1) \\ \vdots \\ f(z_s, \mathbf{m}_s) \end{bmatrix}. \tag{5.6}$$

Then, we can rewrite time discretisation (5.3) of the SWE in the following way

$$z^{n+1} = z^n - \Delta t \tilde{b}^T g_1(\bar{\mathbf{m}}), \tag{5.7a}$$

$$\varepsilon^2 \mathbf{m}^{n+1} = \varepsilon^2 \mathbf{m}^n - \Delta t \tilde{b}^T g_2(\bar{z}) - \Delta t b^T \left[f_1(\bar{z}) + \varepsilon^2 f_2(\bar{z}, \bar{\mathbf{m}}) \right], \tag{5.7b}$$

$$\tilde{A}^{-1}(\bar{z} - \mathbf{1} \otimes z^n) = -\Delta t g_1(\bar{\mathbf{m}}), \tag{5.7c}$$

$$\varepsilon^2 \bar{\mathbf{m}} = \varepsilon^2 \mathbf{1} \otimes \mathbf{m}^n - \Delta t \tilde{A} \otimes \mathbb{1}_d \left[g_2(\bar{z}) + (\tilde{A}^{-1} A) \otimes \mathbb{1}_d f_1(\bar{z}) \right] - \Delta t \varepsilon^2 A \otimes \mathbb{1}_d f_2(\bar{z}, \bar{\mathbf{m}}). \tag{5.7d}$$

Multiplying (5.7d) with $\tilde{A}^{-1} \otimes \mathbb{1}_d$ we obtain

$$g_2(\bar{z}) + (\tilde{A}^{-1} A) \otimes \mathbb{1}_d f_1(\bar{z}) = \mathcal{O}(\varepsilon^2). \tag{5.8}$$

Plugging in the definitions of g_2 and f_1 in (5.8) gives us

$$-b \nabla z_k + \sum_{j=1}^{k} \sum_{l=1}^{j-1} \omega_{kj} a_{jl} z_l \nabla z_l = \mathcal{O}(\varepsilon^2), \quad k = 1, ..., s, \tag{5.9}$$

where $\tilde{A}^{-1} = (\omega_{kj})$. Here, we have used the properties of the matrices $A, \tilde{A}$ to obtain the given summation indices:

- A is a lower triangular matrix with zero diagonal

- $\tilde{A}$ is a regular lower tridiagonal matrix, and therefore $\tilde{A}^{-1}$, too.

From (5.9) it follows easily by induction that $\nabla z_k = \mathcal{O}(\varepsilon^2)$ for $k = 1, ..., s$: For $k = 1$ the sum in (5.9) is zero since the index set is empty. Thus $\nabla z_1 = \mathcal{O}(\varepsilon^2)$. Let $\nabla z_k = \mathcal{O}(\varepsilon)^2$ for $1 \leq k \leq n < s$. Then $\nabla z_{l+1} = \mathcal{O}(\varepsilon^2)$ since the sum in (5.9) is a linear combination of $z_i \nabla z_i$ for $i = 1, ..., l$, and thus it is $\mathcal{O}(\varepsilon^2)$ due to the induction hypothesis.

Let us now integrate (5.7c) over the whole domain. Using periodic boundary conditions and the Gauss divergence rule we get

$$0 = \Delta t \int_{\partial\Omega} \mathbf{m}_i \cdot \mathbf{n} \, ds = \int_{\Omega} \sum_{k=1}^{i} \omega_{ik}(z^n - z_k) \, dx = |\Omega| \sum_{k=1}^{i} \omega_{ik}(z^n - z_k) + \mathcal{O}(\varepsilon^2), \tag{5.10}$$

where the last equality holds due to $\nabla z_k = \mathcal{O}(\varepsilon^2)$, $k = 1, ..., s$. Since $\tilde{A}^{-1}$ is regular, we obtain $z_k = z^n + \mathcal{O}(\varepsilon^2)$ and consequently $\nabla \cdot \mathbf{m}_k = \mathcal{O}(\varepsilon^2)$ for $k = 1, ..., s$. □

Remark 5.1.3 *Lemma 5.1.2 remains valid for other boundary conditions as long as the momentum flux vanishes sufficiently fast for decreasing ε, i.e. $\int_{\partial\Omega} \mathbf{m}_k \cdot \mathbf{n} \, ds = \mathcal{O}(\varepsilon^2)$ for $k = 1, ..., s$.*

Definition 5.1.4 *We call an s-stage IMEX R-K scheme globally stiffly accurate, if*

$$b^T = e_s^T A, \quad \tilde{b}^T = e_s^T \tilde{A}, \quad c_s = \tilde{c}_s = 1. \tag{5.11}$$

Then, the numerical solution obtained by the IMEX R-K scheme is the last internal R-K stage $z_s, \mathbf{m}_s$.

Corollary 5.1.5 *With the assumptions from Lemma 5.1.2 any IMEX scheme of type A guarantees that z^{n+1} approaches z^n as ε approaches zero. However, it is not necessary that $\nabla \cdot \mathbf{m}^{n+1} = 0$ in the low Froude number limit. We can enforce the momentum divergence-free by using globally stiffly accurate IMEX R-K schemes of type A. Then $z^{n+1} = z_s$ and $\mathbf{m}^{n+1} = \mathbf{m}_s$ and the constraints of the zero Froude number SWE are automatically satisfied. Moreover, the scheme is asymptotic preserving.*

Proof: We plug in well-prepared initial data (5.3) into the IMEX scheme of type A, cf. Section 4.1, and consider the low Froude number limit

$$z^{n+1,(0)} = z^{n,(0)}, \tag{5.12a}$$

$$\begin{aligned} \mathbf{m}^{n+1,(0)} = \mathbf{m}^{n,(0)} &+ \Delta t \sum_{j=1}^{s} \tilde{b}_j b \nabla z_j^{n,(2)} \\ &- \Delta t \sum_{j=1}^{s} b_j \left[z^{n,(0)} \nabla z_j^{n,(2)} + \nabla \cdot \frac{\mathbf{m}_j^{n,(0)} \otimes \mathbf{m}_j^{n,(0)}}{z^{n,(0)} - b} \right], \end{aligned} \tag{5.12b}$$

$$\begin{aligned} \mathbf{m}_k^{n,(0)} = \mathbf{m}^{n,(0)} &+ \Delta t \sum_{j=1}^{k} \tilde{a}_{kj} b \nabla z_j^{n,(2)} \\ &- \Delta t \sum_{j=1}^{k-1} a_{kj} \left[z^{n,(0)} \nabla z_j^{n,(2)} + \nabla \cdot \frac{\mathbf{m}_j^{n,(0)} \otimes \mathbf{m}_j^{n,(0)}}{z^{n,(0)} - b} \right]. \end{aligned} \tag{5.12c}$$

Since $\mathbf{m}^{n+1,(0)}$ is divergence-free, (5.12) is a consistent approximation of the zero Froude number SWE (2.34) of the same order as the IMEX R-K scheme. □

5.1.2. IMEX Runge-Kutta schemes of type CK

Similarly to the previous section, we demonstrate the asymptotic preserving property for CK type schemes.

Lemma 5.1.6 *Assume well-prepared initial data and periodic boundary conditions for the alternative SWE (2.30). If we choose the IMEX R-K time discretisation (5.3) of type CK, then the R-K stages* $(z_k, \mathbf{m}_k)$, $1 \leq k \leq s$, *satisfy*

$$z_k = z^n + \mathcal{O}(\varepsilon^2), \quad \nabla \cdot \mathbf{m}_k = \mathcal{O}(\varepsilon^2) \tag{5.13}$$

under the assumption that the internal R-K stages remain bounded with respect to ε.

Proof: The proof is analogous to the proof of Lemma 5.1.2. Let us use the notation from the proof of Lemma 5.1.2. Moreover, we decompose a vector $\mathbf{v} \in \mathbb{R}^s$ in the following way

$$\mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_s \end{bmatrix} = \begin{bmatrix} v_1 \\ v' \end{bmatrix}. \tag{5.14}$$

The first internal stage of the R-K scheme of type CK is always the old time step, i.e. $z_1 = z^n, \mathbf{m}_1 = \mathbf{m}^n$. Let us use the notation

$$\bar{z}' = \begin{bmatrix} z_2 \\ \vdots \\ z_s \end{bmatrix}, \quad \bar{\mathbf{m}}' = \begin{bmatrix} \mathbf{m}_2 \\ \vdots \\ \mathbf{m}_s \end{bmatrix}, \tag{5.15}$$

and define for a function $f = f(z, \mathbf{m})$

$$f(\bar{z}', \bar{\mathbf{m}}') = \begin{bmatrix} f(z_2, \mathbf{m}_2) \\ \vdots \\ f(z_s, \mathbf{m}_s) \end{bmatrix}. \tag{5.16}$$

Then, we can rewrite time discretisation (5.3) of the SWE in the following way

$$\begin{aligned}
z^{n+1} &= z^n - \Delta t \tilde{b}_1 g_1(\mathbf{m}^n) - \Delta t (\tilde{b}')^T g_1(\bar{\mathbf{m}}'), && \text{(5.17a)}\\
\varepsilon^2 \mathbf{m}^{n+1} &= \varepsilon^2 \mathbf{m}^n - \Delta t \tilde{b}_1 g_2(z^n) - \Delta t b_1 \left[f_1(z^n) + \varepsilon^2 f_2(z^n, \mathbf{m}^n) \right] && \text{(5.17b)}\\
&\quad - \Delta t (\tilde{b}')^T g_2(\bar{z}') - \Delta t (b')^T \left[f_1(\bar{z}') + \varepsilon^2 f_2(\bar{z}', \bar{\mathbf{m}}') \right], \\
\bar{z}' &= \mathbf{1} z^n - \Delta t \boldsymbol{\alpha} g_1(\mathbf{m}^n) - \Delta t \hat{A} g_1(\bar{\mathbf{m}}') && \text{(5.17c)}\\
\varepsilon^2 \bar{\mathbf{m}}' &= \varepsilon^2 \mathbf{1} \otimes \mathbf{m}^n - \Delta t \boldsymbol{\alpha} \otimes \mathbf{1} g_2(z^n) && \text{(5.17d)}\\
&\quad - \Delta t \hat{A} \otimes \mathbb{1}_d \left[g_2(\bar{z}') + \hat{A}^{-1} \otimes \mathbb{1}_d (A \otimes \mathbb{1}_d f_1(\bar{z}))' \right] - \Delta t \varepsilon^2 (A \otimes \mathbb{1}_d f_2(\bar{z}, \bar{\mathbf{m}}))'.
\end{aligned}$$

Then, multiplying (5.17d) with $\hat{A}^{-1} \otimes \mathbb{1}_d$ gives us

$$g_2(\bar{z}') + (\hat{A}^{-1}\boldsymbol{\alpha}) \otimes \mathbb{1}_d g_2(z^n) + \hat{A}^{-1} \otimes \mathbb{1}_d (A \otimes \mathbb{1}_d f_1(\bar{z}))' = \mathcal{O}(\varepsilon^2). \tag{5.18}$$

The rest is completely analogous to the proof of Lemma 5.1.2. □

Let us note that also for the CK type time discretisation we have the same properties above. For the completeness we recall them in what follows.

Remark 5.1.7 *Lemma 5.1.6 remains valid for other boundary conditions as long as the momentum flux vanishes sufficiently fast for decreasing ε, i.e. $\int_{\partial\Omega} \mathbf{m}_k \cdot \mathbf{n}\, ds = \mathcal{O}(\varepsilon^2)$ for $k = 1, ..., s$.*

Corollary 5.1.8 *With the assumptions from Lemma 5.1.6 any IMEX R-K scheme of type CK guarantees that $z^{n+1} = z^n$ for $\varepsilon \to 0$. However, it is not necessary that $\nabla \cdot \mathbf{m}^{n+1} = 0$ in the low Froude number limit. We can enforce the momentum divergence-free by using globally stiffly accurate IMEX R-K schemes of type CK. Then $z^{n+1} = z_s$ and $\mathbf{m}^{n+1} = \mathbf{m}_s$ and the constraints of the zero Froude number SWE are automatically satisfied. Moreover, the scheme is asymptotic preserving.*

5.2. IMEX multi-step schemes

We show in this section that a consistent IMEX multi-step scheme, cf Section 4.4, is also asymptotic preserving.

Lemma 5.2.1 *Assume periodic boundary conditions and well-prepared initial data $\mathbf{w}^{n-j}$ for $j = 0, ..., k-1$. Then an IMEX k-step scheme (4.26) applied to the alternative SWE (2.30) satisfies*

$$z^{n+1} = z^n + \mathcal{O}(\varepsilon^2), \quad \nabla \cdot \mathbf{m}^{n+1} = \mathcal{O}(\varepsilon^2), \tag{5.19}$$

if the consistency condition

$$1 = \sum_{j=0}^{k-1} \alpha_j \tag{5.20}$$

is satisfied and $\mathbf{w}^{n+1}$ is bounded with respect to ε. Moreover, then the scheme is asymptotic preserving.

Proof: A k-step IMEX scheme applied to the SWE (2.30) reads

$$z^{n+1} = \tilde{\beta} g_1(\mathbf{m}^{n+1}) + \sum_{j=0}^{k-1} \left[\alpha_j z^{n-j} + \tilde{\beta}_j g_1(\mathbf{m}^{n-j})\right], \tag{5.21a}$$

$$\varepsilon^2 \mathbf{m}^{n+1} = \tilde{\beta} g_2(z^{n+1}) + \sum_{j=0}^{k-1} \left[\varepsilon^2 \alpha_j \mathbf{m}^{n-j} + \tilde{\beta}_j g_2(z^{n-j}) + \beta_j (f_1(z^{n-j}) + \varepsilon^2 f_2(z^{n-j}, \mathbf{m}^{n-j}))\right], \tag{5.21b}$$

where g_1, g_2, f_1, f_2 are from (5.3c). Since the explicit data is well-prepared and $\tilde{\beta} \neq 0$ we obtain

$$\nabla z^{n+1} = \mathcal{O}(\varepsilon^2) \tag{5.22}$$

from (5.21b). Integrating (5.21a) over the whole domain Ω, we obtain

$$\int_\Omega z^{n+1}\, dx = \int_\Omega \sum_{j=0}^{k-1} \alpha_j z^{n-j}\, dx, \tag{5.23}$$

since the integrals over $g_1(\mathbf{m}^{n-j})$ for $j = -1, 0, ..., k-1$ vanish due to the Gauss divergence theorem and periodic boundary conditions. The integrands on the left- and right-hand side are constant up to $\mathcal{O}(\varepsilon^2)$-terms and $\alpha_0 + ... + \alpha_{k-1} = 1$. Thus, we have

$$z^{n+1} = z^n + \mathcal{O}(\varepsilon^2), \tag{5.24}$$

and (5.21a) implies

$$\nabla \cdot \mathbf{m}^{n+1} = \mathcal{O}(\varepsilon^2) \tag{5.25}$$

due to well-prepared momentum initial data and (5.24). □

6. Finite volume IMEX schemes

The aim of this chapter is to develop fully discrete IMEX R-K and IMEX multi-step finite volume schemes for the alternative SWE (2.30). To this end we give a short introduction to the finite volume method in the first section. Here we approximate the *flux integrals* using *numerical fluxes* as well as the approximate evolution operators, cf. Chapter 3. In the second section we derive well-balanced source term discretisations for the straight approach IMEX Euler scheme, cf. Section 4.2. In the third section, we combine the developed well-balanced source term discretisations from Section 6.2 and the finite volume space discretisation with the IMEX time discretisations from Chapter 4. This leads us to general straight approach IMEX finite volume schemes. Then we introduce and discuss the elliptic approach IMEX R-K and multi-step schemes. To this end we derive the elliptic approach IMEX Euler scheme in Section 6.4 and generalise it to elliptic approach IMEX R-K and multi-step schemes in Section 6.5. The elliptic approach IMEX finite volume schemes in Section 6.5 are derived by Gauss elimination of straight approach ones. In Section 6.6 we present second order elliptic approach IMEX finite volume schemes by discretising the ODE system (4.12) in another way. In Section 6.8 we consider the resulting linear systems of IMEX finite volume schemes. Particularly, the non-singularity is of great interest. To this end the theory of circulant block matrices is used, which is presented briefly in Section 6.7.

6.1. Finite volume method

The numerical approximation of hyperbolic balance laws

$$\mathbf{w}_t + \nabla \cdot \mathcal{F}(\mathbf{w}) = K(\mathbf{w}), \quad \mathbf{w}(\mathbf{x}, 0) = \mathbf{w}_0(\mathbf{x}), \tag{6.1}$$

is of great interest and therefore various classes of schemes have been proposed in literature so far. The most popular classes are the finite difference (FD), finite element (FE) and finite volume (FV) schemes. Though, during the last two decades the discontinuous Galerkin (DG) schemes have been developed and are used frequently nowadays. Due to a possible development of discontinuities - even in the case of well-posed problems with smooth initial data - the finite volume and discontinuous Galerkin schemes are favoured, since representation and numerical treatment of discontinuities is straightforward and additional effort is unnecessary. In addition, conserved physical quantities, like mass, momentum or energy, remain conserved, too. Detailed introduction on these schemes can be found in [41, 75, 77] and the references therein. We refer a reader to [26] for the literature to DG schemes.

Let us consider the system (6.1) on the domain $\Omega \subset \mathbb{R}^d$ and suitable boundary conditions.

Definition 6.1.1 *A set of disjoint subdomains $\Omega_i \subset \Omega$, $i = 1, ..., N$, is called finite volume mesh, if*

$$\bar{\Omega} = \bigcup_{i=1}^{N} \bar{\Omega}_i. \tag{6.2}$$

Then, a subdomain Ω_i, $i \in \{1, ..., N\}$, is called (finite volume) cell.

Typically, finite volume cells are polytopes, i.e. line segments, polygons, polyhedrons in 1D, 2D, 3D. Thus, we consider the cells to be polytopes in the following.

Definition 6.1.2 *Let $\Gamma_{ij} := \Omega_i \cap \Omega_j$, $i, j \in \{1, ..., N\}$, be a non-empty intersection of two different cells. Then, Ω_i, Ω_j are neighbours, if the intersection Γ_{ij} contains an $(d-1)$-dimensional manifold. Obviously, Γ_{ij} is the union of β_{ij} $(d-1)$-dimensional manifolds $\Gamma_{ij}^{\alpha} = \Gamma_{ji}^{\alpha}$*

$$\Gamma_{ij} = \bigcup_{\alpha=1}^{\beta_{ij}} \Gamma_{ij}^{\alpha}. \tag{6.3}$$

Further, we can define the index set of neighbours of Ω_i as

$$N(i) := \{j \in \{1, ..., N\} : \Omega_i, \Omega_j \text{ are neighbours}\}. \tag{6.4}$$

Let us integrate (6.1) from t^n to t^{n+1} in time and average over a finite volume cell Ω_i

$$\mathbf{w}_i^{n+1} - \mathbf{w}_i^n + \frac{1}{|\Omega_i|} \int_{t^n}^{t^{n+1}} \int_{\Omega_i} \nabla \cdot \mathcal{F}(\mathbf{w}) \, d\mathbf{x}dt = \frac{1}{|\Omega_i|} \int_{t^n}^{t^{n+1}} \int_{\Omega_i} K \, d\mathbf{x}dt, \tag{6.5}$$

where

$$\mathbf{w}_i^k := \frac{1}{|\Omega_i|} \int_{\Omega_i} \mathbf{w}(\mathbf{x}, t^k) \, d\mathbf{x} \tag{6.6}$$

denotes the integral averages. Further, we apply the Gauss divergence theorem to obtain

$$\mathbf{w}_i^{n+1} = \mathbf{w}_i^n - \frac{1}{|\Omega_i|} \int_{t^n}^{t^{n+1}} \int_{\partial\Omega_i} \mathcal{F}(\mathbf{w}) \cdot \mathbf{n} \, dsdt + \frac{1}{|\Omega_i|} \int_{t^n}^{t^{n+1}} \int_{\Omega_i} K \, d\mathbf{x}dt. \tag{6.7}$$

For the approximation of $\int_{t^n}^{t^{n+1}} \int_{\Omega_i} K \, d\mathbf{x}dt$ we can use some suitable quadrature rules. We will discuss this part in more detail in Section 6.2. Here we just concentrate on the approximation of the flux integrals,

$$\int_{t^n}^{t^{n+1}} \int_{\partial\Omega_i} \mathcal{F}(\mathbf{w}) \cdot \mathbf{n} \, dsdt = \sum_{j \in N(i)} \sum_{\alpha=1}^{\beta_{ij}} \int_{t^n}^{t^{n+1}} \int_{\partial\Gamma_{ij}^{\alpha}} \mathcal{F}(\mathbf{w}) \cdot \mathbf{n}_{ij}^{\alpha} \, dsdt, \tag{6.8}$$

to evaluate the cell averages $\mathbf{w}_i$ in time. Note that in 1D the FV cells Ω_i are open intervals (a_i, a_{i+1}) with $a_i < a_{i+1}$ for $i = 1, ..., N$. The surface integral (6.8) is simply the difference

$$\int_{t^n}^{t^{n+1}} [\mathcal{F}(\mathbf{w})]_{x=a_i}^{x=a_{i+1}} \; dt. \tag{6.9}$$

However the computation of (6.8) is a quite delicate problem: we have approximated the solution $\mathbf{w}$ of (6.1) by the piecewise constant function

$$\mathbf{w}(\mathbf{x}, t^k) \approx \mathbf{w}^k(\mathbf{x}) := \mathbf{w}_i^k, \quad \mathbf{x} \in \Omega_i. \tag{6.10}$$

Therefore, there is no well-defined value for $\mathbf{w}^k$ on the cell boundaries. Note that this is not a problem of an unsuitable ansatz space for the numerical solution, since the solutions of hyperbolic partial differential equations may be discontinuous.

The surface integral in (6.8) can be approximated by the so-called *numerical flux H* via

$$\sum_{j \in N(i)} \sum_{\alpha=1}^{\beta_{ij}} \int_{\Gamma_{ij}^\alpha} \mathcal{F}(\mathbf{w}(\mathbf{x}, t^k)) \; ds \approx \sum_{j \in N(i)} \sum_{\alpha=1}^{\beta_{ij}} |\Gamma_{ij}^\alpha| H(\mathbf{w}_i^k, \mathbf{w}_j^k, \mathbf{n}_{ij}^\alpha) =: \mathcal{H}_i(\mathbf{w}^k) \tag{6.11}$$

for multi-dimensional and

$$\Big[\mathcal{F}(\mathbf{w}(\mathbf{x}, t^k))\Big]_{x=a_i}^{x=a_{i+1}} \approx H(\mathbf{w}_i^k, \mathbf{w}_{i+1}^k, 1) + H(\mathbf{w}_i^k, \mathbf{w}_{i-1}^k, -1) =: \mathcal{H}_i(\mathbf{w}^k) \tag{6.12}$$

for 1D problems. Then, we can evaluate the numerical solution in time via the *finite volume update*

$$\mathbf{w}_i^{n+1} = \mathbf{w}_i^n - \frac{\Delta t}{|\Omega_i|} \Big[\theta \mathcal{H}_i(\mathbf{w}^n) + (1-\theta)\mathcal{H}_i(\mathbf{w}^{n+1})\Big] + \frac{1}{|\Omega_i|} \int_{t^n}^{t^{n+1}} \int_{\Omega_i} K \; d\mathbf{x}dt, \tag{6.13}$$

where $\theta \in [0, 1]$.

Definition 6.1.3 *We call a function*

$$H : \mathbb{R}^p \times \mathbb{R}^p \times S^{d-1} \to \mathbb{R}^p \tag{6.14}$$

a numerical flux. A numerical flux H is called

- *continuous, if the mapping H is continuous.*
- *consistent, if it satisfies*

$$H(\mathbf{u}, \mathbf{u}, \mathbf{n}) = \mathcal{F}(\mathbf{u}) \cdot \mathbf{n} \tag{6.15}$$

for all $\mathbf{u} \in \mathbb{R}^p$.

- *conservative, if it satisfies*

$$H(\mathbf{u},\mathbf{v},\mathbf{n}) = -H(\mathbf{v},\mathbf{u},-\mathbf{n}) \tag{6.16}$$

for all $\mathbf{u},\mathbf{v}\in\mathbb{R}^p, \mathbf{n}\in S^{d-1}$.

Remark 6.1.4

1. *If the solution* $\mathbf{w}(\cdot,t^k)=\mathbf{a}$ *is constant in space at time* t^k*, then the value of* $\mathbf{w}$ *on cell interfaces* Γ_{ij} *is well-defined. Moreover, we have* $\mathcal{F}(\mathbf{w})=\mathcal{F}(\mathbf{a})$ *and* $\mathbf{w}_i^k=\mathbf{a}$*,* $i=1,...,N$*. In this simple scenario a consistent flux provides the exact integration.*

2. *Finite volume schemes are often used to approximate solution of hyperbolic conservation laws*

$$\mathbf{w}_t + \nabla\cdot\mathcal{F}(\mathbf{w}) = 0, \tag{6.17}$$

where $\mathbf{w}$ *consists of conserved physical quantities, e.g. mass, momentum, energy. The surface integral*

$$\int_{t^n}^{t^{n+1}}\int_{\Gamma_{ij}^\alpha} \mathcal{F}(\mathbf{w})\mathbf{n}\; ds\; dt \tag{6.18}$$

is the magnitude of the conserved physical quantities $\mathbf{w}$ *travelling from* Ω_i *to* Ω_j *across the surface* Γ_{ij}^α *in the time interval* $[t^n,t^{n+1}]$*. Thus, the same magnitude travels from* Ω_j *to* Ω_i*, but with the opposite sign. This property is satisfied in a discrete way by conservative numerical fluxes. Thus the quantities* $\mathbf{w}$ *are conserved for all times, if no additional quantities flow in through the boundary. Hence, the corresponding finite volume scheme is indeed conservative.*

In Chapter 8 we present our 1D and 2D numerical experiments for the alternative SWE (2.30), where finite volume schemes are applied on a mesh having regular rectangle cells for 2D and regular line segment cells for 1D. More precisely, the 1D finite volume mesh for $\Omega=[0,1]$ is given by

$$\Omega_i = ((i-1)\Delta x, i\Delta x), \quad i=1,...,N, \tag{6.19}$$

and the 2D mesh for $\Omega=[0,1]\times[0,1]$ by

$$\Omega_{ij} = ((i-1)\Delta x, i\Delta x)\times((j-1)\Delta y, j\Delta y), \quad i=1,...,N, \quad j=1,...,M, \tag{6.20}$$

where $\Delta x = 1/N$, $\Delta y = 1/M$. Therefore the intersection of two neighbouring cells is a point in 1D and a line segment in 2D. Particularly, we have

$$\begin{aligned}\int_{\partial\Omega_{ij}} \mathcal{F}(\mathbf{w}(\cdot,t^k))\mathbf{n}\; ds &= \int_{(j-1)\Delta y}^{j\Delta y} \Big[\mathcal{F}(\mathbf{w}(\cdot,t^k))\mathbf{e}_1\Big]_{x=(i-1)\Delta x}^{x=i\Delta x}\; dy \\ &+ \int_{(i-1)\Delta x}^{i\Delta x} \Big[\mathcal{F}(\mathbf{w}(\cdot,t^k))\mathbf{e}_2\Big]_{y=(j-1)\Delta y}^{y=j\Delta y}\; dx \\ &\approx \sum_{l\in\{-1,1\}} \Big[\Delta y H(\mathbf{w}_{ij}^k, \mathbf{w}_{i+l,j}^k, l\mathbf{e}_1) + \Delta x H(\mathbf{w}_{ij}^k, \mathbf{w}_{i,j+l}^k, l\mathbf{e}_2)\Big] = \mathcal{H}_{ij}(\mathbf{w}^k).\end{aligned} \tag{6.21}$$

In the following we restrict ourselves to regular meshes, cf. (6.19) or (6.20). We use the following numerical fluxes to approximate the flux integral (6.21):

Example 6.1.5

- *Central finite difference flux (CFD): we approximate the values of the flux $\mathcal{F}(\mathbf{w})\mathbf{n}$ on the surface between to neighbouring cells with values $\mathbf{u}, \mathbf{v}$ by a simple average, i.e.*

$$H_{cfd}(\mathbf{u}, \mathbf{v}, \mathbf{n}) = \frac{\mathcal{F}(\mathbf{u}) + \mathcal{F}(\mathbf{v})}{2}\mathbf{n}. \tag{6.22}$$

  If we apply the central finite difference flux to a 1D problem, the explicit finite volume update (6.13) reads

$$\mathbf{w}_i^{n+1} = \mathbf{w}_i^n - \frac{\Delta t}{2\Delta x}\left[\mathcal{F}(\mathbf{w}_{i+1}^n) - \mathcal{F}(\mathbf{w}_{i-1}^n)\right]. \tag{6.23}$$

  Clearly the flux divergence $\nabla \cdot \mathcal{F}(\mathbf{w}) = \mathcal{F}(\mathbf{w})_x$ is approximated by central finite differences. The corresponding explicit scheme (6.13) is well-known to be unconditionally unstable according to von Neumann-stability analysis. However, the same analysis demonstrates, that it is stable, if the central finite difference flux is used implicitly.

- *Lax-Friedrichs and Rusanov flux: the explicit finite volume scheme with central finite difference flux can be stabilised by introducing diffusive terms. These can be done by modifying the CFD flux in the following way*

$$H(\mathbf{u}, \mathbf{v}, \mathbf{n}) = H_{cfd}(\mathbf{u}, \mathbf{v}, \mathbf{n}) - \frac{1}{2\lambda}(\mathbf{v} - \mathbf{u}). \tag{6.24}$$

  For $\lambda = d\ \Delta t/\Delta x$ we obtain the Lax-Friedrichs flux H_{LF}. If $\lambda = \max\{|\lambda_{\mathbf{u}}|, |\lambda_{\mathbf{v}}|\}$, we obtain the Rusanov flux H_{Rus}, where $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}}$ are the spectral radius of the matrix pencils $\mathbb{P}(\mathbf{u}, \mathbf{n}), \mathbb{P}(\mathbf{v}, \mathbf{n})$, cf. Section 2.1.

- *HLLC flux: The HLLC (Harten-Lax-van Leer-contact) scheme is based on an approximate Riemann solver, cf. [109, 111]. It is an extension of the HLL scheme to enable contact discontinuities, which was proposed by Toro et al. [110]. If we have $\mathbf{n} = \pm\mathbf{e}_k$, $k = 1, ..., d$, the corresponding flux function reads*

$$H_{HLLC}(\mathbf{w}_L, \mathbf{w}_R, \mathbf{n}) = \begin{cases} \mathcal{F}(\mathbf{w}_L)\mathbf{n} & \text{if } 0 \le S_L \\ \mathcal{F}(\mathbf{w}_L)\mathbf{n} + S_L(\mathbf{w}_L^* - \mathbf{w}_L) & \text{if } S_L \le 0 \le S^* \\ \mathcal{F}(\mathbf{w}_R)\mathbf{n} + S_R(\mathbf{w}_R^* - \mathbf{w}_R) & \text{if } S^* \le 0 \le S_R \\ \mathcal{F}(\mathbf{w}_R)\mathbf{n} & \text{if } S_R \le 0 \end{cases}, \tag{6.25a}$$

  where S_L, S_R are the smallest and fastest signal velocities in the normal direction of the solution and S^* is an intermediate velocity. Let us assume that $\mathbf{n} = \mathbf{e}_1$ and

$d = 2$. *Then, following the textbooks [109, 111] we obtain*

$$z^* = z_L^* = z_R^* = \frac{(m_1)_L - (m_1)_R - S_L z_L + S_R z_R}{S_R - S_L}, \tag{6.25b}$$

$$m_1^* = (m_1)_L^* = (m_1)_R^* = \frac{S_R(m_1)_L - S_L(m_1)_R + S_L S_R(z_R - z_L)}{S_R - S_L}, \tag{6.25c}$$

$$S^* = \frac{m_1^*}{z^* - b^*}, \quad (m_2)_L^* = (m_2)_L, \quad (m_2)_R^* = (m_2)_R \tag{6.25d}$$

for the alternative SWE (2.30). We use the estimates

$$S_L = \min\left\{\frac{(m_1)_L}{z_L - b_L} - c_L, \frac{(m_2)_R}{z_r - b_R} - c_R\right\}, \tag{6.25e}$$

$$S_R = \max\left\{\frac{(m_1)_L}{z_L - b_L} + c_L, \frac{(m_2)_R}{z_r - b_R} + c_R\right\}, \tag{6.25f}$$

and approximate b^* by

$$b^* = \frac{b_L + b_R}{2}. \tag{6.25g}$$

The extension for other directions of $\mathbf{n}$ and spatial dimensions $d > 2$ are straightforward by separating velocities into normal and tangential directions.

- *Van Leer flux: Let $\mathbb{P}$ be a diagonalisable matrix pencil, cf. Definition 2.1.1. Then the Van Leer flux reads as follows, cf. [41],*

$$H_{VL}(\mathbf{u}, \mathbf{v}, \mathbf{n}) = H_{cfd}(\mathbf{u}, \mathbf{v}, \mathbf{n}) - \left|\mathbb{P}\left(\frac{\mathbf{u} + \mathbf{v}}{2}, \mathbf{n}\right)\right| \frac{\mathbf{v} - \mathbf{u}}{2}. \tag{6.26}$$

Here, the matrix $\mathbb{P}$ is defined in the following way: Let $\Lambda = R^{-1}\mathbb{P}R$ be a Jordan normal form of $\mathbb{P}$ with diagonal entries $\lambda_1, ..., \lambda_p$. Further let Λ^+, Λ^- be diagonal matrices with diagonal entries $\lambda_i^+ = \max\{0, \lambda_i\}$ and $\lambda_i^- = \min\{0, \lambda_i\}$ for $i = 1, ..., p$. Then we have $\Lambda = \Lambda^+ + \Lambda^-$. Thus we define

$$|\mathbb{P}| = R(\Lambda^+ - \Lambda^-)R^{-1}. \tag{6.27}$$

- *EG flux: in Chapter 3 we have derived the approximate evolution operators (3.72) and (3.74). Instead of using a standard one-dimensional numerical flux, we can predict values on cell interfaces Γ_{ij} using a multidimensional evolution operator*

$$EG_{\Delta t} : \mathbb{R}^{3MN} \times \mathbb{R}^d \to \mathbb{R}^p, \quad (\mathbf{w}^k, \mathbf{x}_P) \mapsto EG_{\Delta t}(\mathbf{w}^k, \mathbf{x}_P) \approx \mathbf{w}(\mathbf{x}_P, t^k + \Delta t), \tag{6.28}$$

and approximate the flux integral (6.21) by means of a numerical quadrature. This

leads us to the following explicit and implicit numerical flux

$$\mathcal{H}_{ij}(\mathbf{w}^{k+1}) = \sum_{l=1}^{m} \gamma_l \left\{ \Delta y \left[\mathcal{F}(EG_{\Delta t}(\mathbf{w}^k, (x, y_l))) \right]_{x=(i-1)\Delta x}^{x=i\Delta x} \right\} \mathbf{e}_1 \tag{6.29a}$$
$$+ \sum_{l=1}^{m} \gamma_l \left\{ \Delta x \left[\mathcal{F}(EG_{\Delta t}(\mathbf{w}^k, (x_l, y))) \right]_{y=(j-1)\Delta y}^{y=j\Delta y} \right\} \mathbf{e}_2,$$
$$\mathcal{H}_{ij}(\mathbf{w}^{k}) = \sum_{l=1}^{m} \gamma_l \left\{ \Delta y \left[\mathcal{F}(EG_{0}(\mathbf{w}^k, (x, y_l))) \right]_{x=(i-1)\Delta x}^{x=i\Delta x} \right\} \mathbf{e}_1 \tag{6.29b}$$
$$+ \sum_{l=1}^{m} \gamma_l \left\{ \Delta x \left[\mathcal{F}(EG_{0}(\mathbf{w}^k, (x_l, y))) \right]_{y=(j-1)\Delta y}^{y=j\Delta y} \right\} \mathbf{e}_2,$$

where γ_l, x_l, y_l, $l = 1, ..., m$, are the weights and nodes of the corresponding quadrature. In our numerical experiments we use the Simpson quadrature with the following weights and nodes

$$\gamma_1 = \gamma_3 = \frac{1}{6}, \quad \gamma_2 = \frac{4}{6}, \quad x_l = \left(i - 1 + \frac{l-1}{2} \right) \Delta x, \quad y_l = \left(j - 1 + \frac{l-1}{2} \right) \Delta y, \tag{6.30}$$

where $l = 1, 2, 3$. This is reasonable due to the following reasons:

- *It is a fifth order numerical quadrature, cf. [53], thus the error introduced by the Simpson rule should be negligible for second order schemes.*
- *The corners of a cell Ω_{ij} are nodes. Thus, all bordering cells, i.e. the intersection with Ω_{ij} is non-empty, are used by the evolution operator, which should lead to a better approximation of multi-dimensional wave phenomena.*

Note that the trapezoidal rule could be also used for the reasons mentioned above. But Lukacova et al. [87] pointed out that the trapezoidal rule leads to an unconditionally unstable scheme, whereas the Simpson rule leads conditionally stable ones.

Remark 6.1.6 *Note that the numerical fluxes introduced in Example 6.1.5 are continuous, consistent and conservative.*

So far, we have considered only first order schemes, since the ansatz space consists of piecewise constant functions and we have used just a first order time integrator - explicit Euler scheme. Higher order time integrators for the balance law (6.1) can be easily obtained by applying a high-order ODE solver to

$$\mathbf{w}_t = K(\mathbf{w}) - \nabla \cdot \mathcal{F}(\mathbf{w}). \tag{6.31}$$

However not every ODE solver will produce a stable scheme. To this end *strong stability preserving (SSP)* ODE solver can be used, see [48].

Typically, a high-order space approximation is achieved in the finite volume framework by the so-called *reconstruction step*. Here, derivatives of $\mathbf{w}$ inside each cell are obtained by using data from nearby cells. Assuming $\mathbf{w}_{ij}^k$ to be the value at the centre $\mathbf{x}_{ij}^M := (x_i^M, y_j^M)$ of the cell Ω_{ij}, this leads to a piecewise polynomial representation $R\mathbf{w}^k$ of $\mathbf{w}^k$. Here R denotes the reconstruction. We use this new representation in the finite volume update

(6.13) in the following way: instead of plugging in the numerical flux the values at the centre of the cells, $\mathbf{w}_{ij}^k$, we use the limiting value from inside a cell on a corresponding boundary. More precisely, we use the following flux integral approximation

$$\begin{aligned}\mathcal{H}_{ij}(R\mathbf{w}^k) := \sum_{l\in\{-1,1\}} &\left[\Delta y H\left((R\mathbf{w}^k)_{ij}\left(l\frac{\Delta x}{2},0\right),(R\mathbf{w}^k)_{i+l,j}\left(-l\frac{\Delta x}{2},0\right),l\mathbf{e}_1\right)\right] \\ &+\left[\Delta x H\left((R\mathbf{w}^k)_{ij}\left(0,l\frac{\Delta y}{2}\right),(R\mathbf{w}^k)_{i,j+l}\left(0,-l\frac{\Delta y}{2}\right),l\mathbf{e}_2\right)\right],\end{aligned} \tag{6.32}$$

where

$$(R\mathbf{w}^k)_{ij}(x,y) := R\mathbf{w}^k\big|_{\Omega_{ij}}(x_i^M+x,y_j^M+y). \tag{6.33}$$

This is also known as the *MUSCL (monotonic upstream-centred scheme for conservation laws)* approach, see the text books [41, 77] or the original papers by van Leer [113, 114, 115, 116, 117].

If we use the evolution operator (3.72) or (3.74) we obtain high order accuracy in space by applying the reconstructed numerical solution to predict the cell interface values, e.g. the implicit numerical flux integral approximation reads

$$\begin{aligned}\mathcal{H}_{ij}(\mathbf{w}^k) = &\sum_{l=1}^m \gamma_l\left\{\Delta x\left[\mathcal{F}(EG_0(R\mathbf{w}^k,(x,y_l)))\right]_{x=(i-1)\Delta x}^{x=i\Delta x}\right\}\mathbf{e}_1 \\ &+\sum_{l=1}^m \gamma_l\left\{\Delta y\left[\mathcal{F}(EG_0(R\mathbf{w}^k,(x_l,y)))\right]_{y=(j-1)\Delta y}^{y=j\Delta y}\right\}\mathbf{e}_2,\end{aligned} \tag{6.34}$$

where γ_l, x_l, y_l, $l=1,...,m$, are the weights and nodes of the corresponding quadrature.

In the following example we introduce some basic reconstruction methods for the 2D case. The corresponding 1D reconstruction is straightforward.

Example 6.1.7

1. *The simplest reconstruction is the piecewise constant reconstruction, i.e.*

$$(\mathbf{w}_x)_{ij} := (\mathbf{w}_x)_{ij} := 0. \tag{6.35}$$

   Using this piecewise constant reconstruction in (6.32) gives (6.21).

2. *The simplest way two obtain a second order reconstruction is to use central finite difference approximations for the derivatives of $\mathbf{w}_{ij}$:*

$$(\mathbf{w}_x)_{ij} := \frac{\mathbf{w}_{i+1,j}-\mathbf{w}_{i-1,j}}{2\Delta x},\quad (\mathbf{w}_x)_{ij} := \frac{\mathbf{w}_{i,j+1}-\mathbf{w}_{i,j-1}}{2\Delta y}. \tag{6.36}$$

   Thus we obtain a linear polynomial representation in each cell. We refer to this method as the linear reconstruction. For smooth solutions without large gradient fluctuations, the linear reconstruction works fine. But in the vicinity of discontinuities or large gradients it may cause under- and overshoots that result in oscillations.

3. *We can extend the linear reconstruction in a multi-dimensional way to the so-called bilinear reconstruction. Here, we compute the first order derivatives in a similar way*

$$(\mathbf{w}_x)_{ij} := \frac{\mathbf{w}_{i+1,j} + 0.5(\mathbf{w}_{i+1,j+1} + \mathbf{w}_{i+1,j-1}) - \mathbf{w}_{i-1,j} - 0.5(\mathbf{w}_{i-1,j+1} + \mathbf{w}_{i-1,j-1})}{4\Delta x}, \tag{6.37a}$$

$$(\mathbf{w}_y)_{ij} := \frac{\mathbf{w}_{i,j+1} + 0.5(\mathbf{w}_{i+1,j+1} + \mathbf{w}_{i-1,j+1}) - \mathbf{w}_{i,j-1} - 0.5(\mathbf{w}_{i+1,j-1} + \mathbf{w}_{i-1,j-1})}{4\Delta y},$$

and additionally the mixed second order derivative

$$(\mathbf{w}_{xy})_{ij} := \frac{\mathbf{w}_{i+1,j+1} + \mathbf{w}_{i-1,j-1} - \mathbf{w}_{i+1,j-1} - \mathbf{w}_{i-1,j+1}}{4\Delta x \Delta y}. \tag{6.37b}$$

As for the linear reconstruction, the bilinear reconstruction may produce oscillations near discontinuities or strong gradient fluctuations.

4. *To prevent oscillations due to under- or overshoots we can use the minmod reconstruction, cf. [41, 75, 77]:*

$$(\mathbf{w}_x)_{ij} := \begin{cases} 0 & \text{if } \operatorname{sgn}(s_L) \neq \operatorname{sgn}(s_R) \\ \min\{s_L, s_R\} & \text{if } s_L \geq 0 \\ \max\{s_L, s_R\} & \text{if } s_L < 0 \end{cases} \tag{6.38a}$$

with

$$s_R = \frac{\mathbf{w}_{i+1,j} - \mathbf{w}_{ij}}{\Delta x}, \quad s_L = \frac{\mathbf{w}_{ij} - \mathbf{w}_{i-1,j}}{\Delta x}. \tag{6.38b}$$

Note that the relations in (6.38) are componentwise. The y-derivative can be obtained analogously.

Remark 6.1.8 *The analysis of convergence and convergence speed has been addressed in several papers. However, only the order of convergence $p = 1/2$ or $p = 1/4$ was obtained. We refer the reader for convergence results and error estimates to [68, 69, 70] and the references therein.*

Remark 6.1.9 *In this section we have introduced the finite volume method for the hyperbolic balance law (6.1), where the flux function $\mathcal{F}$ depends only on the solution $\mathbf{w}$. In the case of the alternative SWE (2.30) the flux function depends additionally on the bottom topography. Therefore we generalise the introduced finite volume method to space-dependent flux functions by evaluating the bottom topography in the same manner as the corresponding free surface elevation. For example, consider the CFD numerical flux for the linear part of the alternative SWE (2.30) with a reconstruction R*

$$\begin{aligned} &H_{cfd}((R\mathbf{w}^k)_{ij}, (R\mathbf{w}^k)_{i+1,j}, \mathbf{e}_1) \\ &= \begin{bmatrix} (Rm_1)^k_{ij}(\frac{\Delta x}{2}, 0) + (Rm_1)^k_{i+1,j}(-\frac{\Delta x}{2}, 0) \\ -\dfrac{(Rb_{ij})(\frac{\Delta x}{2}, 0)(Rz_{ij})(\frac{\Delta x}{2}, 0) + (Rb_{i+1,j})(-\frac{\Delta x}{2}, 0)(Rz_{i+1j})(-\frac{\Delta x}{2}, 0)}{\varepsilon^2} \\ 0 \end{bmatrix}. \end{aligned} \tag{6.39}$$

Further, we call a numerical flux consistent, if it is consistent for every constant bottom topography. The definition of a conservative numerical flux remains.

6.2. Straight approach: well-balanced IMEX Euler scheme

In the previous section we have introduced the finite volume method and presented some standard numerical fluxes, e.g. Rusanov, Van Leer numerical fluxes, as well as the EG numerical flux based on the multidimensional approximate evolution operator, see Example 6.1.5. Approximating the linear and nonlinear flux integrals by means of numerical fluxes in combination with the straight approach of the IMEX Euler time discretisation (4.14) we obtain the IMEX Euler finite volume scheme

$$\mathbf{w}_{ij}^{n+1} = \mathbf{w}_{ij}^{n} - \frac{\Delta t}{\Delta x \Delta y}\left[\mathcal{H}_{ij}^{L}(R\mathbf{w}^{n+1}) + \mathcal{H}_{ij}^{NL}(R\mathbf{w}^{n}) - \mathcal{K}_{ij}(R\mathbf{w}^{n+1})\right]. \tag{6.40}$$

Here $\mathcal{H}^{L}, \mathcal{H}^{NL}$ denote the approximations of the flux integral (6.21) of the linear and nonlinear fluxes $\mathcal{F}_{L}, \mathcal{F}_{NL}$ and

$$\mathcal{K}_{ij}(R\mathbf{w}^{n+1}) \approx \int\limits_{\Omega_{ij}} K(\mathbf{w}(\mathbf{x}, t^{n+1}))\, d\mathbf{x} = \int\limits_{\Omega_{ij}} \begin{bmatrix} 0 \\ -\dfrac{z(t^{n+1})\nabla b}{\varepsilon^2} \end{bmatrix} d\mathbf{x} \tag{6.41}$$

is a source term approximation. In the introduction, cf. Chapter 1, we have pointed out that a numerical scheme has to be well-balanced to provide reliable numerical results for the hyperbolic balance law (2.1a). Therefore the source term approximation is crucial. The aim of this section is to obtain suitable source term approximations and thus to get the fully discrete first order well-balanced IMEX Euler finite volume scheme (6.40).

Let us remind that we call scheme (6.40) well-balanced, if it preserves every lake at rest state, cf. Definition 4.6.1. Let k be an arbitrary time step, e.g. $k = n$ or $k = n + 1$. Then, if scheme (6.40) is well-balanced, it satisfies

$$\mathcal{K}_{ij}(R\tilde{\mathbf{w}}^{k}) = \mathcal{H}_{ij}^{L}(R\tilde{\mathbf{w}}^{k}) + \mathcal{H}_{ij}^{NL}(R\tilde{\mathbf{w}}^{k}) \tag{6.42}$$

for every discrete lake at rest state $\tilde{\mathbf{w}}^{k}$ such that $z = Z = \text{const.}, \mathbf{m} = 0$. Note that the lake at rest state is constant, thus we have $R\tilde{\mathbf{w}}^{k} = \tilde{\mathbf{w}}^{k}$ for suitable boundary conditions, e.g. periodic, extrapolated or wall boundary conditions. Further, the approximated integral over the nonlinear part is zero for every lake at rest state, if the corresponding numerical flux is consistent. Thus, the well-balanced condition (6.42) simplifies to

$$\mathcal{K}_{ij}(\tilde{\mathbf{w}}^{k}) = \mathcal{H}_{ij}^{L}(\tilde{\mathbf{w}}^{k}) \tag{6.43}$$

for every lake at rest state $\tilde{\mathbf{w}}^{k}$. In particular, the source term discretisation depends on the numerical flux used for the linear part of the alternative SWE (2.30).

Example 6.2.1 *The necessary condition (6.43) for the scheme (6.40) to be well-balanced reads*

$$\mathcal{K}_{ij}(\tilde{\mathbf{w}}^k) = -\frac{Z}{2\varepsilon^2}\begin{bmatrix} 0 \\ \Delta y(b_{i+1,j} - b_{i-1,j}) \\ \Delta x(b_{i,j+1} - b_{i,j-1}) \end{bmatrix} = \mathcal{H}^L_{ij}(\tilde{\mathbf{w}}^k) \tag{6.44}$$

for the CFD flux and

$$\mathcal{K}_{ij}(\tilde{\mathbf{w}}^k) = -\frac{Z}{\varepsilon^2}\sum_{l=1}^{3}\gamma_l \begin{bmatrix} 0 \\ \Delta y\left[b(x_3, y_l) - b(x_1, y_l)\right] \\ \Delta x\left[b(x_l, y_3) - b(x_l, y_1)\right] \end{bmatrix} = \mathcal{H}^L_{ij}(\tilde{\mathbf{w}}^k) \tag{6.45}$$

for the EG flux, cf. Example 6.1.5. Here γ_i, x_i, y_i, $i = 1, 2, 3$, are the Simpson rules weights and nodes from (6.30). Note that the evolution operators (3.72), (3.74) are well-balanced, i.e. $EG_{\Delta t}(\tilde{\mathbf{w}}^k, (x, y)) = z^* = Z$, which can be easily verified.

In order to preserve the lake at rest equilibrium state, we apply the approach used in [15, 85]. Let us demonstrate this approach on the first non-zero source component. We rewrite the volume integral

$$-\frac{1}{\varepsilon^2}\int\limits_{\Omega_{ij}} z(\cdot, t^k) b_x \, d\mathbf{x} \approx -\frac{1}{\varepsilon^2}\int\limits_{(i-1)\Delta x}^{i\Delta x}\int\limits_{(j-1)\Delta y}^{j\Delta y} \frac{z((i-1)\Delta x, y, t^k) + z(i\Delta x, y, t^k)}{2} b_x(x, y) \, d\mathbf{x} \tag{6.46a}$$

$$= -\frac{1}{\varepsilon^2}\int\limits_{(j-1)\Delta y}^{j\Delta y} \frac{z((i-1)\Delta x, y, t^k) + z(i\Delta x, y, t^k)}{2}(b(i\Delta x, y) - b((i-1)\Delta x, y)) \, dy.$$

in a surface integral. The volume integral for the other non-zero source component can be rewritten analogously in a surface integral by averaging z along the y-axis

$$-\frac{1}{\varepsilon^2}\int\limits_{\Omega_{ij}} z(\cdot, t^k) b_y \, d\mathbf{x} \tag{6.46b}$$

$$\approx -\frac{1}{\varepsilon^2}\int\limits_{(i-1)\Delta x}^{i\Delta x} \frac{z(x, (j-1)\Delta y, t^k) + z(x, j\Delta y, t^k)}{2}(b(x, j\Delta y) - b(x, (j-1)\Delta y)) \, dx.$$

Then, we apply the same technique to predict cell interface values as in the case of the linear part flux integral.

Example 6.2.2

1. *CFD numerical flux: Here, we predict point values on cell interfaces by averaging over corresponding neighbouring cells, e.g. for $(x, y) \in \Omega_{ij}$ we have*

$$\mathbf{w}(i\Delta x, y, t^k) \approx \frac{(R\mathbf{w}^k)_{ij}(\frac{\Delta x}{2}, 0) + (R\mathbf{w}^k)_{i+1,j}(-\frac{\Delta x}{2}, 0)}{2} =: \mathbf{w}^k_{i+\frac{1}{2},j}, \tag{6.47a}$$

$$\mathbf{w}(x, j\Delta y, t^k) \approx \frac{(R\mathbf{w}^k)_{ij}(0, \frac{\Delta y}{2}) + (R\mathbf{w}^k)_{i,j+1}(0, -\frac{\Delta y}{2})}{2} =: \mathbf{w}^k_{i,j+\frac{1}{2}}, \tag{6.47b}$$

and similar for the bottom topography. Consequently, the corresponding source term approximation reads

$$\mathcal{K}_{ij}(\mathbf{w}^k) = -\frac{\Delta x \Delta y}{\varepsilon^2} \begin{bmatrix} 0 \\ \mu_x[(Rz)_{ij}] D_x[(Rb)_{ij}] \\ \mu_y[(Rz)_{ij}] D_y[(Rb)_{ij}] \end{bmatrix}, \tag{6.48a}$$

$$\mu_x[(Rz)_{ij}] = \frac{z^k_{i+\frac{1}{2},j} + z^k_{i-\frac{1}{2},j}}{2}, \quad \mu_y[(Rz)_{ij}] = \frac{z^k_{i,j+\frac{1}{2}} + z^k_{i,j-\frac{1}{2}}}{2}, \tag{6.48b}$$

$$D_x[(Rz)_{ij}] = \frac{z^k_{i+\frac{1}{2},j} - z^k_{i-\frac{1}{2},j}}{\Delta x}, \quad D_y[(Rz)_{ij}] = \frac{z^k_{i,j+\frac{1}{2}} - z^k_{i,j-\frac{1}{2}}}{\Delta y} \cdot \mu^y_{ij}(Rz) \tag{6.48c}$$

2. *EG numerical flux: Here, the point values are predicted by an approximated evolution operator, cf. (3.72), (3.74), and the surface integral is computed by means of the Simpson rule. Thus, the corresponding source term approximation reads*

$$\mathcal{K}_{ij}(\mathbf{w}^k) = -\frac{1}{2\varepsilon^2} \sum_{l=1}^{3} \gamma_l \begin{bmatrix} 0 \\ \Delta y \left[z^*(x_3, y_l) + z^*(x_1, y_l)\right] \left[b(x_3, y_l) - b(x_1, y_l)\right] \\ \Delta x \left[z^*(x_l, y_3) + z^*(x_l, y_1)\right] \left[b(x_l, y_3) - b(x_l, y_1)\right] \end{bmatrix}, \tag{6.49}$$

where γ_i, x_i, y_i, $i = 1, 2, 3$, are the Simpson rules weights and nodes from (6.30) and $(z^*, (\mathbf{m}^*)^T)^T(x_i, y_j) = \mathbf{w}^*(x_i, y_j) = EG(R\mathbf{w}^k, (x_i, y_j))$, $i, j = 1, 2, 3$, are the cell interface values predicted by an approximated evolution operator.

Indeed, the source term discretisations from Example 6.2.2 satisfy condition (6.44). Thus, if the linear system corresponding to the numerical scheme (6.40) posses a unique solution, the scheme is well-balanced - and we have the following theorem.

Theorem 6.2.3 *Assume that the numerical solution is unique. Further, we apply a consistent numerical flux to approximate the nonlinear part. Then, the straight approach IMEX Euler finite volume scheme (6.40) is well-balanced, if the CFD or the EG numerical flux and the corresponding source term discretisations (6.48), (6.49) are used.*

Remark 6.2.4 *The source term discretisations (6.48), (6.49) satisfy the well-balanced condition (6.44) due to the averages of the free surface elevation z in front of the bottom topography difference. In fact any other average of the free surface elevation z instead leads to a (possibly inconsistent) source term approximation that satisfy condition (6.44). Particularly, the average*

$$\mu_x[(Rz)_{ij}] = \frac{(Rz)_{i-1,j}(\frac{\Delta x}{2}, 0) + (Rz)_{i+1,j}(-\frac{\Delta x}{2}, 0)}{2} \tag{6.50}$$

and analogously the average in y-direction μ_y are of interest, cf. Section 6.8 and Chapter 7.

Remark 6.2.5 *If we use piecewise constant reconstruction, the difference operator (6.48c) is the central difference operator (6.36).*

6.3. Straight approach: well-balanced IMEX Runge-Kutta and multi-step schemes

In the previous section we have derived fully discrete IMEX Euler schemes that are well-balanced, if the numerical solution is unique. The aim of this section is to generalise this result to the IMEX R-K schemes of the type A, CK and to the IMEX multi-step schemes, cf. Chapter 4.

Let us recall the application of a straight approach IMEX R-K scheme of type A or CK to the alternative SWE (2.30)

$$\mathbf{w}^{n+1} = \mathbf{w}^n - \Delta t \nabla \cdot \sum_{k=1}^{s} \left[\tilde{b}_k \mathcal{F}_L(\mathbf{w}_k) + b_k \mathcal{F}_{NL}(\mathbf{w}_k) \right] + \Delta t \sum_{k=1}^{s} \tilde{b}_k K(\mathbf{w}_k), \tag{4.10a}$$

$$\mathbf{w}_k = \mathbf{w}^n - \Delta t \nabla \cdot \sum_{l=1}^{k} \left[\tilde{a}_{kl} \mathcal{F}_L(\mathbf{w}_l) + a_{kl} \mathcal{F}_{NL}(\mathbf{w}_l) \right] + \Delta t \sum_{l=1}^{k} \tilde{a}_{kl} K(\mathbf{w}_l). \tag{4.10b}$$

We remind that the nonlinear part of the flux $\mathcal{F}_{NL}$ is only evaluated explicitly, since $a_{ii} = 0$ for $i = 1, ..., s$. Averaging (4.10) over the cell Ω_{ij} and applying numerical fluxes to approximate the flux integrals we obtain the straight approach IMEX R-K finite volume scheme

$$\mathbf{w}_{ij}^{n+1} = \mathbf{w}_{ij}^n - \frac{\Delta t}{\Delta x \Delta y} \sum_{k=1}^{s} \left[\tilde{b}_k \mathcal{H}_{ij}^L(R\mathbf{w}_k) + b_k \mathcal{H}_{ij}^{NL}(R\mathbf{w}_k) \right] + \frac{\Delta t}{\Delta x \Delta y} \sum_{k=1}^{s} \tilde{b}_k \mathcal{K}_{ij}(R\mathbf{w}_k), \tag{6.52a}$$

$$(\mathbf{w}_k)_{ij} = \mathbf{w}_{ij}^n - \frac{\Delta t}{\Delta x \Delta y} \sum_{l=1}^{k} \left[\tilde{a}_{kl} \mathcal{H}_{ij}^L(R\mathbf{w}_l) + a_{kl} \mathcal{H}_{ij}^{NL}(R\mathbf{w}_l) \right] + \frac{\Delta t}{\Delta x \Delta y} \sum_{l=1}^{k} \tilde{a}_{kl} \mathcal{K}_{ij}(R\mathbf{w}_l). \tag{6.52b}$$

Analogously we obtain from the straight approach k-step time discretisation

$$\begin{aligned} \mathbf{w}^{n+1} - \sum_{l=0}^{k-1} \alpha_l \mathbf{w}^{n-l} &= \tilde{\beta} \left[\nabla \cdot \mathcal{F}_L(\mathbf{w}^{n+1}) - K(\mathbf{w}^{n+1}) \right] \\ &\quad + \sum_{l=1}^{k-1} \left\{ \tilde{\beta}_l \left[\nabla \cdot \mathcal{F}_L(\mathbf{w}^{n-l}) - K(\mathbf{w}^{n-l}) \right] + \beta_l \mathcal{F}_{NL}(\mathbf{w}^{n-l}) \right\} \end{aligned} \tag{4.26}$$

the straight approach IMEX k-step finite volume scheme

$$\begin{aligned} \mathbf{w}_{ij}^{n+1} - \sum_{l=0}^{k-1} \alpha_l \mathbf{w}_{ij}^{n-l} &= \frac{\tilde{\beta}}{\Delta x \Delta y} \left[\mathcal{H}_{ij}^L(R\mathbf{w}^{n+1}) - \mathcal{K}_{ij}(R\mathbf{w}^{n+1}) \right] \\ &\quad + \frac{1}{\Delta x \Delta y} \sum_{l=1}^{k-1} \left\{ \tilde{\beta}_l \left[\mathcal{H}_{ij}^L(R\mathbf{w}^{n-l}) - \mathcal{K}_{ij}(R\mathbf{w}^{n-l}) \right] + \beta_l \mathcal{H}_{ij}^{NL}(R\mathbf{w}^{n-l}) \right\}. \end{aligned} \tag{6.53}$$

In Section 4.6 we proved that straight approach IMEX R-K and multi-step time discretisations are well-balanced under mild conditions, cf. Theorem 4.6.2. The main ingredients of the corresponding proof are Lemmas 4.6.5, 4.6.6. The fully discrete version of Lemma 4.6.5 shows, that the numerical flux differences

$$\mathcal{H}_{ij}^{NL}(\tilde{\mathbf{w}}^k) = \mathcal{H}_{ij}^L(\tilde{\mathbf{w}}^k) - \mathcal{K}_{ij}(\tilde{\mathbf{w}}^k) = 0 \tag{6.54}$$

for all discrete lake at rest states $\tilde{\mathbf{w}}^k$.

In the previous section we have pointed out that the nonlinear flux difference $\mathcal{H}^{NL}$ vanishes if an old time solution is a lake at rest state and the underlying numerical flux is consistent. Also we choose the source term discretisation $\mathcal{K}$ so that the discrete stiff part $\mathcal{H}^{L} - \mathcal{K}$ is zero for any lake at rest state. Lemma 4.6.6 can be replaced by Theorem 6.2.3. Thus, we can repeat the proof completely for the fully discrete schemes (6.52), (6.53) where the CFD or the EG numerical flux and the corresponding source term approximation are used. Consequently, we obtain the following theorem.

Theorem 6.3.1 *Assume that the numerical solution is unique. Further, we apply a consistent numerical flux to approximate the nonlinear part. Then, the finite volume IMEX R-K and multi-step scheme (6.52), (6.53) are well-balanced, if the CFD or the EG numerical flux, cf. Example 6.1.5, and the corresponding source term discretisations (6.48), (6.49) are used.*

Remark 6.3.2 *In general we have to use an IMEX time discretisation of p-th order and a piecewise polynomial of degree $(p-1)$ reconstruction to to obtain an IMEX scheme of p-th order. Thus, using for example ENO or WENO reconstructions, see [41, 77], we can obtain straight approach IMEX finite volume schemes of arbitrary high order.*

6.4. Elliptic approach: well-balanced IMEX Euler finite volume scheme

In Section 4.1 we have proposed to use an underlying elliptic equation for the free surface elevation to propagate the numerical solution in time. The aim of this section is to derive suitable spatial discretisations for this reformulation.

To this end we use the elliptic approach IMEX Euler time discretisation, that consists of the following three steps

$$\hat{\mathbf{w}} = \mathbf{w}^n - \Delta t \nabla \cdot \mathcal{F}_{NL}(\mathbf{w}^n), \tag{4.16a}$$

$$z^{n+1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 \nabla \cdot (b \nabla z^{n+1}) = \hat{z} - \Delta t \nabla \cdot \hat{\mathbf{m}}, \tag{4.16b}$$

$$\mathbf{m}^{n+1} = \hat{\mathbf{m}} + \Delta t \frac{b \nabla z^{n+1}}{\varepsilon^2}. \tag{4.16c}$$

Note that in the scheme (4.16) there is no interplay between a flux and a source term as for the corresponding straight approach (4.14). Thus it is "almost automatically" well-balanced.

The first step, (4.16a), is obviously the explicit Euler time discretisation for the hyperbolic conservation law $\mathbf{w}_t + \nabla \cdot \mathcal{F}_{NL}(\mathbf{w}) = 0$. Hence we apply the finite volume approximation

$$\hat{\mathbf{w}}_{ij} = \mathbf{w}_{ij}^n - \Delta t \mathcal{H}_{ij}^{NL}(R\mathbf{w}^n), \tag{6.56a}$$

where $\mathcal{H}^{NL}$ is a flux integral approximation by means of consistent numerical fluxes and R a reconstruction from Example 6.1.7. Averaging (4.16b) and (4.16c) over a cell Ω_{ij} we

obtain

$$z_{ij}^{n+1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 (\nabla \cdot (b \nabla z^{n+1}))_{ij} = \hat{z}_{ij} - \Delta t (\nabla \cdot \hat{\mathbf{m}})_{ij}, \tag{6.56b}$$

$$\mathbf{m}_{ij}^{n+1} = \hat{\mathbf{m}}_{ij} + \Delta t \frac{(b \nabla z)_{ij}}{\varepsilon^2}. \tag{6.56c}$$

We can use the standard central difference approximations in (6.56b), (6.56c)

$$((bz_x)_x)_{ij} = \frac{z_{i+1,j}(b_{i+1,j} + b_{ij}) - z_{ij}(b_{i+1,j} + 2b_{ij} + b_{i-1,j}) + z_{i-1,j}(b_{ij} + b_{i-1,j})}{2\Delta x^2}, \tag{6.57a}$$

$$((\hat{m}_1)_x)_{ij} = D_x[(\hat{m}_1)_{ij}], \tag{6.57b}$$

$$(bz_x)_{ij} = b_{ij} D_x[z_{ij}] \tag{6.57c}$$

and similar for the y-derivatives. Here D_x is the central difference operator

$$D_x[f_{ij}] = \frac{f_{i+1,j} - f_{i-1,j}}{2\Delta x}. \tag{6.57d}$$

Another way to obtain an elliptic approach IMEX Euler scheme is based on rewriting the straight approach IMEX Euler scheme (6.40). Using the CFD numerical flux (6.22) and the corresponding source term approximation (6.48) the straight approach IMEX finite volume scheme reads

$$z_{ij}^{n+1} = \hat{z}_{ij} - \Delta t \left\{ D_x[(Rm_1^{n+1})_{ij}] + D_y[(Rm_2^{n+1})_{ij}] \right\}, \tag{6.58a}$$

$$\mathbf{m}_{ij}^{n+1} = \hat{\mathbf{m}}_{ij} + \frac{\Delta t}{\varepsilon^2} \left\{ \begin{bmatrix} D_x[(Rb)_{ij}(Rz^{n+1})_{ij}] \\ D_y[(Rb)_{ij}(Rz^{n+1})_{ij}] \end{bmatrix} - \begin{bmatrix} \mu_x[(Rz^{n+1})_{ij}] D_x[(Rb)_{ij}] \\ \mu_y[(Rz^{n+1})_{ij}] D_y[(Rb)_{ij}] \end{bmatrix} \right\}, \tag{6.58b}$$

where μ_x, μ_y are average the operators (6.48b) or (6.50) and D_x, D_y the difference operators (6.48c). Here $\hat{\mathbf{w}} = (\hat{z}, \hat{\mathbf{m}}^T)^T$ is the explicit update of the nonlinear part $\mathcal{F}_{NL}$ (6.56a). Note that the terms in the curved brackets in (6.58) give the approximations

$$(\nabla \cdot (\hat{\mathbf{m}}))_{ij} = D_x[(Rm_1^{n+1})_{ij}] + D_y[(Rm_2^{n+1})_{ij}], \tag{6.59a}$$

$$(b \nabla z)_{ij} = \begin{bmatrix} D_x[(Rb)_{ij}(Rz)_{ij}] \\ D_y[(Rb_{ij})(Rz)_{ij}] \end{bmatrix} - \begin{bmatrix} \mu_x[(Rz)_{ij}] D_x[(Rb)_{ij}] \\ \mu_y[(Rz)_{ij}] D_y[(Rb)_{ij}] \end{bmatrix}. \tag{6.59b}$$

Further, plugging (6.58b) in (6.58a) we obtain

$$z_{ij}^{n+1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 (\nabla \cdot (b \nabla z^{n+1}))_{ij} = \hat{z}_{ij} - \Delta t \left\{ D_x[(R\hat{m}_1)_{ij}] + D_y[(R\hat{m}_2)_{ij}] \right\} \tag{6.60}$$

with

$$((bz_x)_x)_{ij} = D_x[R \left\{ D_x[(Rb)_{ij}(Rz)_{ij}] - \mu_x[(Rz)_{ij}] D_x[(Rb)_{ij}] \right\}], \tag{6.61}$$

and analogously for $((bz_y)_y)_{ij}$.

Example 6.4.1 *The approximations (6.59b), (6.61) read as follows, if we we choose the piecewise constant or the linear reconstruction, cf. Example 6.1.7, and use the average operator (6.48b)*

$$(b\nabla z)_{ij} = \begin{bmatrix} \dfrac{b_{i+1,j}+b_{i-1,j}}{4}\dfrac{z_{i+1,j}-z_{i-1,j}}{2\Delta x} + b_{i+1,j}\dfrac{z_{i+1,j}-z_{i,j}}{4\Delta x} + b_{i-1,j}\dfrac{z_{ij}-z_{i-1,j}}{4\Delta x} \\ \dfrac{b_{i,j+1}+b_{i,j-1}}{4}\dfrac{z_{i,j+1}-z_{i,j-1}}{2\Delta y} + b_{i,j+1}\dfrac{z_{i,j+1}-z_{i,j}}{4\Delta y} + b_{i,j-1}\dfrac{z_{ij}-z_{i,j-1}}{4\Delta y} \end{bmatrix}, \tag{6.62a}$$

$$\begin{aligned}((bz_x)_x)_{ij} = \frac{1}{(4\Delta x)^2}\Big\{ & z_{i+2,j}(3b_{i+2,j}+b_{ij}) + z_{i+1,j}2(b_{ij}-b_{i+2,j}) \\ & - z_{ij}(b_{i+2,j}+6b_{ij}+b_{i-2,j}) + z_{i-1,j}2(b_{ij}-b_{i-2,j}) + z_{i-2,j}(3b_{i-2,j}+b_{ij})\Big\},\end{aligned} \tag{6.62b}$$

and

$$\begin{aligned}(bz_x)_{ij} = \frac{1}{128\Delta x}\{ & z_{i+2}\left[3b_{i+2}-10b_{i+1}-4b_i-6b_{i-1}-b_{i-2}\right] \\ & -z_{i-2}\left[3b_{i-2}-10b_{i-1}-4b_i-6b_{i+1}-b_{i+2}\right] \\ & +4z_{i+1}\left[-3b_{i+2}+10b_{i+1}+12b_i+6b_{i-1}-b_{i-2}\right] \\ & -4z_{i-1}\left[-3b_{i-2}+10b_{i-1}+12b_i+6b_{i+1}-b_{i+2}\right] \\ & +6z_i\left[b_{i+2}-b_{i-2}-2(b_{i+1}-b_{i-1})\right]\},\end{aligned} \tag{6.63a}$$

$$\begin{aligned}((bz_x)_x)_{ij} = \{ & z_{i+4}\left[-3b_{i+4}+10b_{i+3}+4b_{i+2}+6b_{i+1}-b_i\right] \\ & + z_{i-4}\left[-3b_{i-4}+10b_{i-3}+4b_{i-2}+6b_{i-1}-b_i\right] \\ & - 2z_{i+3}\left[-6b_{i+4}+11b_{i+3}+54b_{i+2}+24b_{i+1}+16b_i-3b_{i-1}\right] \\ & - 2z_{i-3}\left[-6b_{i-4}+11b_{i-3}+54b_{i-2}+24b_{i-1}+16b_i-3b_{i+1}\right] \\ & + 6z_{i+2}\left[-b_{i+4}-10b_{i+3}+40b_{i+2}+46b_{i+1}+25b_i-4b_{i-1}\right] \\ & + 6z_{i-2}\left[-b_{i-4}-10b_{i-3}+40b_{i-2}+46b_{i-1}+25b_i-4b_{i+1}\right] \\ & + z_{i+1}\left[-4b_{i+4}+60b_{i+3}-24b_{i+2}+22b_{i+1}+120b_i-12b_{i-1}+36b_{i-2}-6b_{i-3}\right] \\ & + z_{i-1}\left[-4b_{i-4}+60b_{i-3}-24b_{i-2}+22b_{i-1}+120b_i-12b_{i+1}+36b_{i+2}-6b_{i+3}\right] \\ & - z_i\left[-b_{i+4}-b_{i-4}-18(b_{i+3}+b_{i-3})+148(b_{i+2}+b_{i-2})+226(b_{i+1}+b_{i-1})+474b_i\right] \\ & \}\frac{1}{(32\Delta x)^2}.\end{aligned} \tag{6.63b}$$

If we use the average operators (6.50) instead, they read

$$(b\nabla z)_{ij} = \frac{1}{4\Delta x}\begin{bmatrix} (b_{i+1,j}+b_{i-1,j})(z_{i+1,j}-z_{i-1,j}) \\ (b_{i,j+1}+b_{i,j-1})(z_{i,j+1}-z_{i,j-1}) \end{bmatrix}, \tag{6.64a}$$

$$((bz_x)_x)_{ij} = \frac{1}{8\Delta x^2}\Big\{z_{i+2,j}(b_{i+2,j}+b_{ij}) - z_{ij}(b_{i+2,j}+2b_{ij}+b_{i-2,j}) + z_{i-2,j}(b_{i-2,j}+b_{ij})\Big\}, \tag{6.64b}$$

and

$$
\begin{aligned}
(bz_x)_{ij} = \frac{1}{64\Delta x} \{ & z_{i+2} \left[b_{i+2} - 2b_{i+1} - 2b_i - 6b_{i-1} + b_{i-2} \right] \\
& - z_{i-2} \left[b_{i-2} - 2b_{i-1} - 2b_i - 6b_{i+1} + b_{i+2} \right] \\
& + 4z_{i+1} \left[-b_{i+2} + 2b_{i+1} + 6b_i + 6b_{i-1} - b_{i-2} \right] \\
& - 4z_{i-1} \left[-b_{i-2} + 2b_{i-1} + 6b_i + 6b_{i+1} - b_{i+2} \right] \\
& + 12 z_i \left[b_{i+1} - b_{i-1} \right] \} ,
\end{aligned} \tag{6.65a}
$$

$$
\begin{aligned}
((bz_x)_x)_{ij} = \{ & z_{i+4} \left[-b_{i+4} + 2b_{i+3} + 2b_{i+2} + 6b_{i+1} - b_i \right] \\
& + z_{i-4} \left[-b_{i-4} + 2b_{i-3} + 2b_{i-2} + 6b_{i-1} - b_i \right] \\
& - 2z_{i+3} \left[-2b_{i+4} + b_{i+3} + 18b_{i+2} + 18b_{i+1} + 16b_i - 3b_{i-1} \right] \\
& - 2z_{i-3} \left[-2b_{i-4} + b_{i-3} + 18b_{i-2} + 18b_{i-1} + 16b_i - 3b_{i+1} \right] \\
& + 12z_{i+2} \left[-3b_{i+3} + 4b_{i+2} + 13b_{i+1} + 12b_i - 2b_{i-1} \right] \\
& + 12z_{i-2} \left[-3b_{i-3} + 4b_{i-2} + 13b_{i-1} + 12b_i - 2b_{i+1} \right] \\
& - 2z_{i+1} \left[2b_{i+4} - 12b_{i+3} - 48b_{i+2} - b_{i+1} + 32b_i - 6b_{i-1} - 18b_{i-2} + 3b_{i-3} \right] \\
& - 2z_{i-1} \left[2b_{i-4} - 12b_{i-3} - 48b_{i-2} - b_{i-1} + 32b_i - 6b_{i+1} - 18b_{i+2} + 3b_{i+3} \right] \\
& + z_i \left[b_{i+4} + b_{i-4} + 18(b_{i+3} + b_{i-3}) - 146(b_{i+2} + b_{i-2}) - 122(b_{i+1} + b_{i-1}) - 94b_i \right] \\
& \} \frac{1}{512\Delta x^2} .
\end{aligned} \tag{6.65b}
$$

Note that only the discretisation (6.64) leads to a symmetric discretisation matrix, cf. Section 6.8.1.

Consequently we obtain two finite volume IMEX Euler schemes (6.56) for the elliptic approach. Each consists of three steps:

1. compute the explicit update for the nonlinear flux via (6.56a)
2. solve the elliptic equation (6.56b) for the free surface elevation z at a new time level using either (6.57a), (6.57b) or (6.61), (6.57b)
3. compute the explicit momentum update at a new time level (6.56c) using (6.57c) or (6.59b).

Remark 6.4.2 *For constant bottom topography and piecewise constant reconstruction, cf. Example 6.1.7, i.e. $b = b_{ij} = const.$ for all $i = 1, ..., N, j = 1, ..., M$, the approximations (6.57a), (6.61) read*

$$
((bz_x)_x)_{ij} = b \frac{z_{i+1,j} - 2z_{ij} + z_{i-1,j}}{\Delta x^2}, \tag{6.66a}
$$

$$
((bz_x)_x)_{ij} = b \frac{z_{i+2,j} - 2z_{ij} + z_{i-2,j}}{(2\Delta x)^2}, \tag{6.66b}
$$

and (6.57c) and (6.59b) are equal. Thus we have

$$
(bz_x)_{ij} = b \frac{z_{i+1,j} - z_{i-1,j}}{2\Delta x}. \tag{6.66c}
$$

Thus, the size of the stencil for the discrete Laplacian is here the only difference. This result was obtained analogously by Degond et al. in [38] for the isentropic Euler equations.

Since the two elliptic approach IMEX Euler finite volume schemes differ in their size of the stencil we refer to the scheme with the wider stencil (6.56), (6.59), (6.61) as ELLW scheme - *elliptic wide* - and to the other scheme as ELL scheme - *elliptic* - (6.56),(6.57).

Theorem 6.4.3 *The ELL and ELLW schemes are well-balanced, if the numerical solution is unique and a consistent numerical flux is used to approximate the explicit parts (6.56a).*

Proof: Let the initial conditions be a lake at rest state $\tilde{\mathbf{w}}$ with constant free surface elevation $z = Z$ and zero velocity, thus $\mathbf{m} = 0$. In Section 6.3 we pointed out that $\mathcal{H}_{ij}^{NL}(R\tilde{\mathbf{w}}) = 0$. Thus the ELL, ELLW schemes read

$$\hat{\mathbf{w}}_{ij} = \tilde{\mathbf{w}}_{ij}, \tag{6.67a}$$

$$z_{ij}^{n+1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 (\nabla \cdot (b \nabla z^{n+1}))_{ij} = Z, \tag{6.67b}$$

$$\mathbf{m}_{ij}^{n+1} = \Delta t \frac{(b_{ij} \nabla z^{n+1})_{ij}}{\varepsilon^2}. \tag{6.67c}$$

The linear system corresponding to (6.67b) admits the solution $z_{ij}^{n+1} = Z$ for all i, j, since the discrete elliptic term vanishes for constant data. Moreover the numerical solution is unique, thus $z_{ij}^{n+1} = Z$ for all i, j. Then the momentum remains zero, i.e. $\mathbf{m}_{ij}^{n+1} = 0$. □

Note that we derived the ELLW scheme by Gauss elimination of the linear system corresponding to an IMEX Euler finite volume scheme. We can see a similarity with the well-known Schur-complement. Thus the following theorem holds:

Theorem 6.4.4 *Assume that*

1. *the CFD numerical flux (6.22) and the corresponding source term approximation (6.48) are used*
2. *the numerical flux $\mathcal{H}^{NL}$ to approximate the nonlinear flux part in the straight approach IMEX Euler scheme is the same as for the ELLW scheme.*

Then the ELLW scheme (6.56), (6.57a), (6.57c), (6.61) is algebraically equivalent to the straight approach IMEX Euler finite volume scheme described in Section 6.2. Here we mean by algebraically equivalent, that for given initial data both schemes provide the same results.

6.5. Elliptic approach: well-balanced IMEX Runge-Kutta and multi-step schemes

The aim of this section is to generalise the results from the previous section for the elliptic approach IMEX Euler finite volume schemes to general IMEX R-K and IMEX

multi-step time discretisations. To this end we provide fully discrete IMEX R-K and multi-step finite volume schemes and study their properties.

Let us start with the elliptic approach IMEX R-K scheme (4.10a), (4.12), (4.13)

$$\mathbf{w}^{n+1} = \mathbf{w}^n - \Delta t \nabla \cdot \sum_{k=1}^{s} \left[\tilde{b}_k \mathcal{F}_L(\mathbf{w}_k) + b_k \mathcal{F}_{NL}(\mathbf{w}_k) \right] + \Delta t \sum_{k=1}^{s} \tilde{b}_k K(\mathbf{w}_k), \tag{4.10a}$$

$$\hat{\mathbf{w}}_k = \mathbf{w}^n - \Delta t \nabla \cdot \sum_{j=1}^{k-1} \left[\tilde{a}_{kj} \mathcal{F}_L(\mathbf{w}_j) + a_{kj} \mathcal{F}_{NL}(\mathbf{w}_j) \right] + \Delta t \sum_{j=1}^{k-1} \tilde{a}_{kj} K(\mathbf{w}_j), \tag{4.12a}$$

$$z_k + \left(\frac{\Delta t \tilde{a}_{kk}}{\varepsilon} \right)^2 \nabla \cdot (b \nabla z_k) = \hat{z}_k - \Delta t \tilde{a}_{kk} \nabla \cdot \hat{\mathbf{m}}_k, \tag{4.13}$$

$$\mathbf{m}_k = \hat{\mathbf{m}}_k + \Delta t \tilde{a}_{kk} \frac{b \nabla z_k}{\varepsilon^2}. \tag{4.12c}$$

As in the previous section, we apply the finite volume space discretisation to (4.10a), (4.12a). Here we approximate the flux integrals over linear and nonlinear parts of the flux $\mathcal{F}_L, \mathcal{F}_{NL}$ by numerical fluxes $\mathcal{H}^L, \mathcal{H}^{NL}$

$$\mathbf{w}_{ij}^{n+1} = \mathbf{w}_{ij}^n - \frac{\Delta t}{\Delta x \Delta y} \sum_{k=1}^{s} \left[\tilde{b}_k \mathcal{H}_{ij}^L(R\mathbf{w}_k) + b_k \mathcal{H}_{ij}^{NL}(R\mathbf{w}_k) \right] + \frac{\Delta t}{\Delta x \Delta y} \sum_{k=1}^{s} \tilde{b}_k \mathcal{K}_{ij}(R\mathbf{w}_k), \tag{6.68a}$$

$$(\hat{\mathbf{w}}_k)_{ij} = \mathbf{w}_{ij}^n - \frac{\Delta t}{\Delta x \Delta y} \sum_{j=1}^{k-1} \left[\tilde{a}_{kj} \mathcal{H}_{ij}^L(R\mathbf{w}_j) + a_{kj} \mathcal{H}_{ij}^{NL}(R\mathbf{w}_j) \right] + \frac{\Delta t}{\Delta x \Delta y} \sum_{j=1}^{k-1} \tilde{a}_{kj} \mathcal{K}_{ij}(R\mathbf{w}_j), \tag{6.68b}$$

$$(z_k)_{ij} + \left(\frac{\Delta t \tilde{a}_{kk}}{\varepsilon} \right)^2 (\nabla \cdot (b \nabla z_k))_{ij} = (\hat{z}_k)_{ij} - \Delta t \tilde{a}_{kk} (\nabla \cdot \hat{\mathbf{m}}_k)_{ij}, \tag{6.68c}$$

$$(\mathbf{m}_k)_{ij} = (\hat{\mathbf{m}}_k)_{ij} + \Delta t \tilde{a}_{kk} \frac{(b \nabla z_k)_{ij}}{\varepsilon^2}. \tag{6.68d}$$

Here R denotes the reconstruction, cf. Example 6.1.7 and $((bz_x)_x)_{ij}, ((\hat{m}_1)_x)_{ij}, (bz_x)_{ij}$ the approximations (6.61), (6.59a), (6.59b). Hence, we obtain the fully discrete elliptic approach IMEX R-K finite volume scheme (6.68), (6.59), (6.61).

Analogously we can obtain the elliptic approach IMEX multi-step finite volume scheme

$$(\hat{\mathbf{w}})_{ij} = \sum_{l=0}^{k-1} \left\{ \alpha_l \mathbf{w}_{ij}^{n-l} + \frac{\tilde{\beta}_l}{\Delta x \Delta y} \left[\mathcal{H}_{ij}^L(R\mathbf{w}^{n-l}) - \mathcal{K}_{ij}(R\mathbf{w}^{n-l}) \right] + \frac{\beta_l}{\Delta x \Delta y} \mathcal{H}_{ij}^{NL}(\mathbf{w}^{n-l}) \right\}, \tag{6.69a}$$

$$z_{ij}^{n+1} + \left(\frac{\tilde{\beta}}{\varepsilon} \right)^2 (\nabla \cdot (b \nabla z^{n+1}))_{ij} = \hat{z}_{ij} + \tilde{\beta} (\nabla \cdot \hat{\mathbf{m}}_k)_{ij}, \tag{6.69b}$$

$$\mathbf{m}_{ij}^{n+1} = \hat{\mathbf{m}}_{ij} - \tilde{\beta} \frac{(b \nabla z^{n+1})_{ij}}{\varepsilon^2}. \tag{6.69c}$$

Due to the derivation of the above IMEX schemes a more general version of Theorem 6.4.4 holds:

Theorem 6.5.1 *The elliptic approach IMEX R-K or IMEX multi-step finite volume schemes (6.68), (6.69) with the approximations (6.59), (6.61) are algebraically equivalent to the corresponding straight approach IMEX R-K or IMEX multi-step scheme described in Section 6.3, respectively, if*

1. *the CFD numerical flux (6.22) and the corresponding source term approximation (6.48) are used*

2. *the used numerical flux $\mathcal{H}^{NL}$ to approximate the nonlinear flux is the same for the straight and elliptic approach scheme.*

As already mentioned in the Theorem 6.4.4 algebraically equivalent means, that for given initial data both schemes provide the same result.

Since the elliptic approach IMEX finite volume schemes (6.68), (6.69) with the approximations (6.59), (6.61) can be rewritten as the straight approach IMEX finite volume schemes, the well-balanced property is inherited:

Corollary 6.5.2 *The elliptic approach IMEX R-K and IMEX multi-step finite volume schemes (6.68), (6.69) with the approximations (6.59), (6.61) are well-balanced, if the numerical solution is unique and a consistent numerical flux is used to approximate the explicit parts.*

Further, we can obtain elliptic approach IMEX finite volume schemes of arbitrary order with high order reconstructions and IMEX time discretisations, cf. Remark 6.3.2.

6.6. Second order IMEX Runge-Kutta and multi-step schemes

In the Sections 6.2-6.5 we have derived straight and elliptic approach IMEX finite volume schemes of arbitrary order. However high order elliptic approach IMEX finite volume schemes introduced in the previous section are algebraically equivalent to the straight approach IMEX finite volume schemes. The aim of this section is to introduce alternative second order elliptic approach IMEX finite volumes schemes.

To this end, let us first point out, that the approximations (6.59), (6.61) are of second order if the linear reconstruction R from the Example 6.1.7: Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a smooth function with point values f_{ij} on an equidistant grid with increments $\Delta x, \Delta y$. Then we have

$$D_x[(Rf)_{ij}] = \frac{-f_{i+2,j} + 6f_{i+1,j} - 6f_{i-1,j} + f_{i-2,j}}{8\Delta x}. \tag{6.70}$$

Using the Taylor expansion we obtain that (6.70) is a second order approximation of $(f_x)_{ij}$. Indeed the approximation order depends on the degree of the piecewise polynomial reconstruction. Thus the order of the spatial discretisation of the linear parts is not increased using linear or bilinear reconstruction, cf. Example 6.1.7, when the CFD numerical flux (6.22) with the source term approximation (6.48) is used. Thus we can use

piecewise constant reconstruction maintaining the order of spatial discretisation. Consequently, we can use the approximations $(\nabla \cdot \hat{\mathbf{m}})_{ij}$, $(b\nabla z)_{ij}$ and $(\nabla \cdot (b\nabla z))_{ij}$ from Section 6.4.

Theorem 6.6.1 *The elliptic approach IMEX R-K and IMEX multi-step finite volume schemes (6.68), (6.69) with the approximations (6.57a), (6.57b), (6.57c) or (6.61), (6.57b), (6.59b) are well-balanced, if the numerical solution is unique and a consistent numerical flux is applied to approximate the nonlinear terms.*

Proof: Let us consider a consistent IMEX k-step time discretisation and let $\tilde{\mathbf{w}} = \mathbf{w}^n = \mathbf{w}^{n-1} = ... = \mathbf{w}^{n-k+1}$ be a lake at rest state. In Section 6.2 we have pointed out that the flux differences $\mathcal{H}_{ij}^{NL}(\tilde{\mathbf{w}}) = \mathcal{H}_{ij}^{L}(\tilde{\mathbf{w}}) - \mathcal{K}_{ij}(\tilde{\mathbf{w}}) = 0$. Then, the explicit update (6.69a) reads $(\hat{\mathbf{w}})_{ij} = \tilde{\mathbf{w}}$. The implicit part of an elliptic approach IMEX k-step scheme is the same as of the elliptic approach IMEX Euler scheme. Thus, due to Theorem 6.4.3 $\mathbf{w}^{n+1} = \tilde{\mathbf{w}}$ and the k-step scheme is well-balanced.

For an elliptic approach multi-step IMEX R-K scheme we use induction over the internal R-K stages with the same argument. □

6.7. Circulant block matrices

In the following Section 6.8, 7.1, 7.2 we will study the non-singularity of the arising linear systems and the asymptotic preserving property of the derived IMEX finite volume schemes. If we use periodic boundary conditions matrices of a special structure, so-called *circulant block matrices*, arise. Due to this special structure we can simplify calculations. Thus we are able to obtain more results.

The aim of this section is to introduce circulant block matrices and discuss their properties. To this end, we first present the circulant matrices in Section 6.7.1. Then we extend the theory of circulant matrices to circulant block matrices in Section 6.7.2. For more details on circulant matrices we refer a reader to the monograph [30] or to the paper [67] and references therein.

6.7.1. Circulant matrices

Let us consider the $n \times n$-matrix

$$U := \begin{bmatrix} 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & & \ddots & \ddots & 1 \\ 1 & 0 & ... & ... & 0 \end{bmatrix}. \tag{6.71}$$

It is easily checked that the l-th power of U is given by the relation

$$(U^l)_{ij} = \begin{cases} 1 & \text{if } j - i = l \ mod \ n \\ 0 & \text{else} \end{cases}. \tag{6.72}$$

Consequently, U generates a cyclic group of order n with matrix multiplication as group operation. Particularly $U^n = \mathbb{1}$ is the identity.

Definition 6.7.1 *The set of circulant $n \times n$-matrices is*

$$Circ_n := \left\{ \sum_{l=0}^{n-1} a_l U^l \mid a_l \in \mathbb{C} \right\}. \tag{6.73}$$

Example 6.7.2 *A general 4×4 circulant matrix reads*

$$C = \begin{pmatrix} z_0 & z_1 & z_2 & z_3 \\ z_3 & z_0 & z_1 & z_2 \\ z_2 & z_3 & z_0 & z_1 \\ z_1 & z_2 & z_3 & z_0 \end{pmatrix}, \tag{6.74}$$

where $z_i \in \mathbb{C}$ for $i = 0, ..., 3$.

Theorem 6.7.3

1. *Matrix multiplication is commutative in $Circ_n$.*

2. *The mapping*

$$\mathcal{C} : \mathbb{C}^n \mapsto Circ_n, \quad (z_0, ..., z_{n-1}) \mapsto \sum_{l=0}^{n-1} z_l U^l \tag{6.75}$$

is an isomorphism. Moreover $(z_0, ..., z_{n-1})$ is the first row of the matrix $\mathcal{C}(z_0, ..., z_{n-1})$.

3. *The Jordan normal form of the circulant matrix $\mathcal{C}(z_0, ..., z_{n-1})$ is*

$$\begin{bmatrix} \lambda_0 & 0 & ... & 0 \\ 0 & \lambda_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & ... & 0 & \lambda_{n-1} \end{bmatrix}, \quad \lambda_k = \sum_{j=0}^{n-1} z_j \omega^{jk}, \quad k = 0, ..., n-1, \tag{6.76}$$

where $\omega = e^{i\frac{2\pi}{n}}$. The corresponding eigenvector to the k-th eigenvalue is

$$\mathbf{r}_k = \frac{1}{\sqrt{n}} (1, \omega^k, \omega^{2k}, ..., \omega^{(n-1)k})^T, \quad k = 0, ...n-1. \tag{6.77}$$

Moreover, the eigenvectors $\mathbf{r}_k, k = 0, ..., n-1$, are independent of the matrix $\mathcal{C}(z_0, ..., z_{n-1})$.

Proof: The first two statements follow directly from the definition. For the third one, it suffices to compute the eigenvalues and eigenvectors of the matrix U.
Laplace expansion of the determinant along the first row or column leads to the characteristic polynomial of $U \in Circ_n$

$$\chi_U(x) = x^n - 1. \tag{6.78}$$

Thus the eigenvalues of U are the n-th roots of unity, i.e. $1, \omega, \omega^2, ..., \omega^{n-1}$.
The calculation

$$U\mathbf{r}_k = \frac{1}{\sqrt{n}} \begin{bmatrix} \omega^k \\ \omega^{2k} \\ \vdots \\ \omega^{(n-1)k} \\ 1 = \omega^{nk} \end{bmatrix} = \omega^k \frac{1}{\sqrt{n}} \begin{bmatrix} 1 \\ \omega^k \\ \omega^{2k} \\ \vdots \\ \omega^{(n-1)k} \end{bmatrix} = \omega^k \mathbf{r}_k \tag{6.79}$$

proves that $\mathbf{r}_k$ is the k-th eigenvector. □

The equation for the eigenvalue λ_k (6.76) leads to the linear system

$$\begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_{n-1} \end{bmatrix} = \sqrt{n} F \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_{n-1} \end{bmatrix}, \quad F = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & ... & ... & 1 \\ 1 & \omega & \omega^2 & ... & \omega^{n-1} \\ \vdots & \omega^2 & \omega^4 & ... & \omega^{2(n-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \omega^{n-1} & \omega^{2(n-1)} & ... & \omega^{(n-1)(n-1)} \end{bmatrix}, \tag{6.80}$$

that describes the dependence of the eigenvalues $\lambda_k, k = 0, ..., n-1$, from the coefficients $z_k, k = 0, ..., n-1$, and vice versa via the Fourier matrix F. Let us recall that $F^{-1} = \bar{F} = F^*$. This, leads us to the following representation theorem for circulant matrices.

Corollary 6.7.4 *Let*

$$diag : \mathbb{R}^{n \times n} \to \mathbb{R}^n, \quad A = (a_{ij}) \mapsto (a_{11}, a_{22}, ..., a_{nn}) \tag{6.81}$$

be the mapping that maps a matrix onto its diagonal and $D^n(\mathbb{C})$ the set of diagonal $(n \times n)$-matrices over $\mathbb{C}$. Then:

1. *The diagram*

$$\begin{array}{ccc} \mathbb{C}^n & \to & \mathbb{C}^n \\ \boldsymbol{\lambda} = diag(\Lambda) & \mapsto & \mathbf{z} = \dfrac{F^* \boldsymbol{\lambda}}{\sqrt{n}} \\ \uparrow & & \downarrow \\ D^n(\mathbb{C}) & \leftarrow & Circ_n \\ \Lambda = F^* A F & \leftarrow & A = \mathcal{C}(\mathbf{z}) \end{array} \tag{6.82}$$

   commutes and every mapping is an isomorphism.

2. *A matrix A is circulant, if and only if F^*AF is a diagonal matrix.*

3. *The inverse of a non-singular circulant matrix is circulant.*

Proof:

1. This is clear from the above discussion.

2. This follows from the first statement.

3. Let A be a non-singular, cyclic matrix with inverse A^{-1}. Then

$$F^* A^{-1} F = (F^* A F)^{-1} \tag{6.83}$$

is a diagonal matrix and thus, due to the second statement, A^{-1} is circulant.

□

We will not use the following lemma, however it is easy to proof and is of general interest.

Lemma 6.7.5 *Let $\mathcal{C}(z_0, ..., z_{n-1}) = A \in Circ_n$ be a circulant matrix with a dominant entry z_j, i.e.*

$$|z_j| > \sum_{i \neq j} |z_i|. \tag{6.84}$$

Then the matrix A is non-singular.

Proof: Let us assume that the matrix A is singular, i.e. there exists a vector $v = (v_0, ..., v_{n-1})^T \neq 0$ in the kernel of A. Without loss of generality, z_0 is the dominant entry and v_0 is the entry of v of maximum absolute value, i.e. $|v_0| = \|v\|_\infty$. Then, one finds

$$0 = (Av)_1 = \sum_{k=0}^{n-1} z_k v_k, \tag{6.85}$$

that is a contradiction since

$$0 = \left| \sum_{k=0}^{n-1} z_k v_k \right| \geq |z_0 v_0| - \left| \sum_{k=1}^{n-1} z_k v_k \right| \geq |z_0 v_0| - \sum_{k=1}^{n-1} |z_k v_k| \geq |z_0 v_0| - \sum_{k=1}^{n-1} |z_k v_0| > 0. \tag{6.86}$$

Thus, A is non-singular. □

6.7.2. Circulant block matrices of circulant matrices

Let us extend the statements of the previous section to circulant block matrices of circulant matrices. Such a matrix is obtained by taking a circulant matrix $\mathcal{C}(z_0, ..., z_{n-1})$ and replace the entrys z_i by circulant matrices Z_i, e.g.: Let $Z_0, Z_1, Z_2 \in Circ_n$. Then

$$A = \begin{bmatrix} Z_0 & Z_1 & Z_2 \\ Z_2 & Z_0 & Z_1 \\ Z_1 & Z_2 & Z_0 \end{bmatrix} \in \mathbb{C}^{3n \times 3n} \tag{6.87}$$

is a circulant block matrix of circulant matrices. For the sake of brevity we will refer to these matrices in the following as circulant block matrices.

Definition 6.7.6 *Let $\otimes$ denote the Kronecker product and $U \in Circ_m$ the matrix (6.71). Then, the set of circulant m-block n-matrices is*

$$Circ_{m \times n} = \left\{ \sum_{j=0}^{m-1} U^j \otimes A_j : A_j \in Circ_n \right\}. \tag{6.88}$$

Theorem 6.7.7

1. *The mapping*

$$\mathcal{C} : Circ_n^m \mapsto Circ_{m\times n}, \quad (A_0, ..., A_{m-1}) \mapsto \sum_{l=0}^{m-1} U^l \otimes A_l \tag{6.89}$$

is an isomorphism.

2. *Let $A, B \in Circ_{m\times n}$. Then $AB = BA \in Circ_{m\times n}$.*

3. *The Jordan normal form of the circulant block matrix $\mathcal{C}(A_0, ..., A_{m-1})$, $A_0, ..., A_{m-1} \in Circ_n$ is*

$$\begin{bmatrix} \Lambda_0 & 0 & ... & 0 \\ 0 & \Lambda_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & ... & 0 & \Lambda_{m-1} \end{bmatrix}, \quad \Lambda_k = \begin{bmatrix} \lambda_{k0} & 0 & ... & 0 \\ 0 & \lambda_{k1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & ... & 0 & \lambda_{k,n-1} \end{bmatrix}, \tag{6.90a}$$

$$\lambda_{kl} := \sum_{j=0}^{m-1} \omega^{jk} \lambda_l(A_j), \quad 0 \le k \le m-1, \quad 0 \le l \le n-1, \tag{6.90b}$$

where $\lambda_l(A_j)$ is the l-th eigenvalue of A_j and $\omega = e^{i\frac{2\pi}{m}}$. The matrix containing the corresponding eigenvectors in the columns is $F^{mn} := F_m \otimes F_n$, where F_m, F_n denote the m-, n-dimensional Fourier matrix (6.80).

Proof: The first statement follows from the definition.
The second one is an implication of the Kronecker product rule combined with the commutative matrix multiplication in $Circ_n$, i.e.

$$(U^j \otimes A)(U^k \otimes B) = (U^j U^k) \otimes (AB) = (U^k U^j) \otimes (BA) = (U^k \otimes B)(U^j \otimes A) \tag{6.91}$$

for $U \in Circ_m, A, B \in Circ_n$.
The third statement is obtained as a consequence of Corollary 6.7.4 and

$$F_m^* \otimes F_n^* \cdot U^j \otimes A_j \cdot F_m \otimes F_n = (F_m^* U^j F_m) \otimes (F_n^* A_j F_n). \tag{6.92}$$

□

Let $\mathbf{e}_j \in \mathbb{R}^m$ be the j-th standard basis vector and $\mathbf{z}^{(l)} \in \mathbb{R}^n$ the vector containing the first line of the matrix A_l, i.e. $\mathcal{C}(\mathbf{z}^{(l)}) = A_l$. Then due to Theorems 6.7.7, 6.7.3 we get the following correspondence of eigenvalues λ_{kl}, $0 \le k \le m-1, 0 \le l \le n-1$ to the vectors $\mathbf{z}^{(l)}, l = 0, ..., m-1$,

$$\begin{aligned} \boldsymbol{\lambda} &= \sum_{k=0}^{m-1} (\mathbf{e}_{k+1} \otimes diag(\Lambda_k)) = \sum_{k=0}^{m-1} \left(\mathbf{e}_{k+1} \otimes \left(\sum_{j=0}^{m-1} \omega^{jk} \boldsymbol{\lambda}^{(j)} \right) \right) = \sum_{j=0}^{m-1} \left(\left(\sum_{k=0}^{m-1} \omega^{jk} \mathbf{e}_{k+1} \right) \otimes \boldsymbol{\lambda}^{(j)} \right) \\ &= \sum_{j=0}^{m-1} (\sqrt{m} F_m \mathbf{e}_{j+1}) \otimes (\sqrt{n} F_n \mathbf{z}^{(j)}) = \sqrt{mn} (F_m \otimes F_n) \sum_{j=0}^{m-1} (\mathbf{e}_{j+1} \otimes \mathbf{z}^{(j)}) \\ &= \sqrt{mn} (F_m \otimes F_n) \mathbf{z}, \end{aligned} \tag{6.93}$$

where $\boldsymbol{\lambda} = ((\boldsymbol{\lambda}^{(0)})^T, ..., (\boldsymbol{\lambda}^{(m-1)})^T)^T$, $\mathbf{z} = (\mathbf{z}^{(0)T}, ..., \mathbf{z}^{(n-1)T})^T$ and $\boldsymbol{\lambda}^{(j)} = (\lambda_0(A_j), ..., \lambda_{n-1}(A_j))^T = \sqrt{n} F \mathbf{z}^{(j)}$. This correspondence implies the following corollary.

Corollary 6.7.8

1. *The diagram*

$$\begin{array}{rcl} \mathbb{C}^{mn} & \to & \mathbb{C}^{mn} \\ \boldsymbol{\lambda} = diag(\Lambda) & \mapsto & \mathbf{z} = \dfrac{(F_m^* \otimes F_n^*)\boldsymbol{\lambda}}{\sqrt{mn}} \\ \uparrow & & \downarrow \\ D^{mn\times mn}(\mathbb{C}) & \leftarrow & Circ_{m\times n} \\ \Lambda = (F_m^* \otimes F_n^*)A(F_m \otimes F_n) & \leftarrow & A = \mathcal{C}(\mathcal{C}(\mathbf{z}^{(0)}), ..., \mathcal{C}(\mathbf{z}^{(l-1)})) \end{array} \tag{6.94}$$

   commutes and every mapping is an isomorphism.

2. *A matrix A is circulant block matrix, if and only if $(F_m^* \otimes F_n^*)A(F_m \otimes F_n)$ is a diagonal matrix.*

3. *The inverse of a non-singular circulant bloc matrix is a circulant block matrix.*

The proofs are analogous to those of Corollary 6.7.4.

6.8. Resulting linear systems

In the previous sections we have derived fully discrete IMEX finite volume schemes for the SWE (2.30). Due to the implicit treatment of the stiff part $\nabla \cdot \mathcal{F}_L - K$ (2.37) at least one linear system has to be solved per time step. The aim of this section is to discuss the corresponding linear systems and their properties. Particularly the non-singularity of the linear system is of great interest.

To this end let us consider the straight approach IMEX R-K scheme

$$\mathbf{w}_{ij}^{n+1} = \mathbf{w}_{ij}^{n} - \frac{\Delta t}{\Delta x \Delta y} \sum_{k=1}^{s} \left[\tilde{b}_k \mathcal{H}_{ij}^{L}(R\mathbf{w}_k) + b_k \mathcal{H}_{ij}^{NL}(R\mathbf{w}_k) \right] + \frac{\Delta t}{\Delta x \Delta y} \sum_{k=1}^{s} \tilde{b}_k \mathcal{K}_{ij}(R\mathbf{w}_k), \tag{6.52a}$$

$$(\mathbf{w}_k)_{ij} = \mathbf{w}_{ij}^{n} - \frac{\Delta t}{\Delta x \Delta y} \sum_{l=1}^{k} \left[\tilde{a}_{kl} \mathcal{H}_{ij}^{L}(R\mathbf{w}_l) + a_{kl} \mathcal{H}_{ij}^{NL}(R\mathbf{w}_l) \right] + \frac{\Delta t}{\Delta x \Delta y} \sum_{l=1}^{k} \tilde{a}_{kl} \mathcal{K}_{ij}(R\mathbf{w}_l), \tag{6.52b}$$

and the IMEX multi-step scheme

$$\begin{aligned} \mathbf{w}_{ij}^{n+1} - \sum_{l=0}^{k-1} \alpha_l \mathbf{w}_{ij}^{n-l} = {} & \frac{\tilde{\beta}}{\Delta x \Delta y} \left[\mathcal{H}_{ij}^{L}(R\mathbf{w}^{n+1}) - \mathcal{K}_{ij}(R\mathbf{w}^{n+1}) \right] \\ & + \frac{1}{\Delta x \Delta y} \sum_{l=1}^{k-1} \left\{ \tilde{\beta}_l \left[\mathcal{H}_{ij}^{L}(R\mathbf{w}^{n-l}) - \mathcal{K}_{ij}(R\mathbf{w}^{n-l}) \right] + \beta_l \mathcal{H}_{ij}^{NL}(R\mathbf{w}^{n-l}) \right\}. \end{aligned} \tag{6.53}$$

We write the implicit terms on the left-hand side and the explicit ones on the right-hand side. Then the internal IMEX R-K stage equation (6.52b) or the IMEX multi-step

update (6.53) can be rewritten in the following way

$$\mathbf{w}_{ij}^{new} + \frac{\delta}{\Delta x \Delta y}\left[\mathcal{H}^L(R\mathbf{w}_{ij}^{new}) - \mathcal{K}(\mathbf{w}_{ij}^{new})\right] = \hat{\mathbf{w}}_{ij}. \tag{6.95}$$

Here R is a reconstruction, $\delta > 0$ depends on the time increment(s) and $\mathbf{w}^{new}$ is either the numerical solution at a new time step (IMEX multi-step scheme) or an internal R-K stage (IMEX R-K scheme). Let us introduce the *state vector*

$$\mathbf{W} := \begin{bmatrix} \mathbf{Z} \\ \mathbf{M} \end{bmatrix} \in \mathbb{R}^{3NM}, \quad \mathbf{M} = \begin{bmatrix} \mathbf{M}_1 \\ \mathbf{M}_2 \end{bmatrix} \in \mathbb{R}^{2NM}, \tag{6.96a}$$

where

$$\mathbf{Z}^T = \left[z_{11}, ..., z_{N1}, z_{12}, ..., z_{N2}, ..., z_{1,M}, ..., z_{NM}\right], \tag{6.96b}$$

similar notation is used for $\mathbf{M}_1$, $\mathbf{M}_2$. Recall that N, M is the number of cells in the x- and y-direction. Further let us consider the mapping

$$A : \mathbb{R}^{3MN} \to \mathbb{R}^{3MN}, \quad \mathbf{W}^k = (\mathbf{w}_{ij}^k) \mapsto A\mathbf{W}^k = \left(\frac{\mathcal{H}_{ij}^L(R\mathbf{w}^k) - \mathcal{K}_{ij}(R\mathbf{w}^k)}{\Delta x \Delta y}\right). \tag{6.97}$$

If $\mathcal{H}^L$, $\mathcal{K}$ and R are linear functions, then A is a $(3MN \times 3MN)$-matrix and the linear system corresponding to (6.95) reads

$$(\mathbb{1} + \delta A)\,\mathbf{W}^{new} = \hat{\mathbf{W}}. \tag{6.98}$$

Note that the linear system (6.98) corresponds to the space-continuous equation $\mathbf{w} + \delta\Phi(\mathbf{w}) = \hat{\mathbf{w}}$, cf. Section 4.6. For example the CFD or the EG numerical flux (6.22), (6.29b) with the corresponding source terms (6.48), (6.49) can be used with the piecewise constant, linear or bilinear reconstruction, cf. Example 6.1.7. Due to consistency reasons and the structure of the implicit part $\nabla \cdot \mathcal{F}_L - K$ we can write the matrix A in the following way

$$A = F_L + F_D, \quad F_L = \begin{bmatrix} 0 & \mathfrak{D}_x & \mathfrak{D}_y \\ -\frac{1}{\varepsilon^2}\overline{\mathfrak{B}\mathfrak{D}}_x & 0 & 0 \\ -\frac{1}{\varepsilon^2}\overline{\mathfrak{B}\mathfrak{D}}_y & 0 & 0 \end{bmatrix}. \tag{6.99}$$

Here $\mathfrak{D}_x, \mathfrak{D}_y, \overline{\mathfrak{B}\mathfrak{D}}_x, \overline{\mathfrak{B}\mathfrak{D}}_y \in \mathbb{R}^{MN \times MN}$ approximate the differential operators $\partial/\partial x$, $\partial/\partial y$, $b\partial/\partial x$, $b\partial/\partial y$, where $\overline{\mathfrak{B}\mathfrak{D}}_x, \overline{\mathfrak{B}\mathfrak{D}}_y$ is only notation for a matrix and in general not the product of two matrices. Hence, F_L approximates $\nabla \cdot \mathcal{F}_L - K$ and F_D contains diffusive terms. In the following we consider the CFD and EG numerical fluxes, cf. Example 6.1.5, to approximate the stiff parts in Section 6.8.1 and 6.8.2.

6.8.1. Central difference numerical flux

The aim of this section is to study the non-singularity of the arising linear systems when the CFD numerical flux is used. In the semi-discrete case we have proved uniqueness of the solution in the following way: we considered the underlying elliptic equation (4.41)

and used the fact that its eigenvalues are negative. We want to adopt this method to the fully discrete schemes. To this end we derive a linear system that corresponds to the elliptic equation (4.41) by Gauss elimination. In the case of the CFD numerical flux the matrix $F_D = 0$. Thus we obtain from the Gauss elimination

$$\left[\begin{array}{ccc|c} \mathbb{1} & \delta\mathfrak{D}_x & \delta\mathfrak{D}_y & \hat{\mathbf{Z}} \\ -\frac{\delta}{\varepsilon^2}\overline{\mathfrak{B}\mathfrak{D}}_x & \mathbb{1} & 0 & \hat{\mathbf{M}}_1 \\ -\frac{\delta}{\varepsilon^2}\overline{\mathfrak{B}\mathfrak{D}}_y & 0 & \mathbb{1} & \hat{\mathbf{M}}_2 \end{array}\right] \quad I - \delta\mathfrak{D}_x II - \delta\mathfrak{D}_y III \tag{6.100}$$

$$\left[\begin{array}{ccc|c} \mathbb{1} + \left(\frac{\delta}{\varepsilon}\right)^2\left[\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x + \mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y\right] & 0 & 0 & \hat{\mathbf{Z}} - \delta\left[\mathfrak{D}_x\hat{\mathbf{M}}_1 + \mathfrak{D}_y\hat{\mathbf{M}}_2\right] \\ -\frac{\delta}{\varepsilon^2}\overline{\mathfrak{B}\mathfrak{D}}_x & \mathbb{1} & 0 & \hat{\mathbf{M}}_1 \\ -\frac{\delta}{\varepsilon^2}\overline{\mathfrak{B}\mathfrak{D}}_y & 0 & \mathbb{1} & \hat{\mathbf{M}}_2 \end{array}\right].$$

This is indeed the method that we used in Section 6.5 to derive elliptic approach finite volume schemes from the straight approach. Thus it suffices to consider only the elliptic approach IMEX finite volume schemes. Further, the matrix $\mathbb{1} + \delta A$ is non-singular, due to (6.100), for all δ and $\varepsilon > 0$, if and only if $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x + \mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y$ has no negative eigenvalue. In the following we study the definiteness of $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x + \mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y$, since this is a sufficient criterion. To this end one can use the following lemma.

Lemma 6.8.1 *If $\overline{\mathfrak{B}\mathfrak{D}}_x = \mathfrak{B}\mathfrak{D}_x, \overline{\mathfrak{B}\mathfrak{D}}_y = \mathfrak{B}\mathfrak{D}_y$ with a negative definite matrix $\mathfrak{B} \in \mathbb{R}^{NM\times NM}$ and $\mathfrak{D}_x, \mathfrak{D}_y$ are anti-symmetric, then the matrix $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x + \mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y$ is positive semi-definite.*

Proof: We compute

$$\mathbf{W}^T\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x\mathbf{W} = \mathbf{W}^T\mathfrak{D}_x^T(-\mathfrak{B})\mathfrak{D}_x\mathbf{W} \geq 0. \tag{6.101}$$

Doing the analogous calculation for $\mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y$, we obtain the statement of the lemma. □

The condition $\overline{\mathfrak{B}\mathfrak{D}}_x = \mathfrak{B}\mathfrak{D}_x, \overline{\mathfrak{B}\mathfrak{D}}_y = \mathfrak{B}\mathfrak{D}_y$ seems not to be satisfied for the presented elliptic approach IMEX finite volume IMEX schemes. However if the bottom topography is constant, $b = b_{ij} = const.$, we have

$$((bz_x)_x)_{ij} = bD_x[RD_x[(Rz)_{ij}]], \tag{6.102}$$

which corresponds to $b\mathfrak{D}_x^2\mathbf{Z}$. It is also easy to prove that the used discrete derivatives lead to anti-symmetric matrices $\mathfrak{D}_x, \mathfrak{D}_y$, if the piecewise constant, linear or bilinear reconstruction, cf. Example 6.1.7, and periodic boundary conditions are used. Consequently we obtain:

Corollary 6.8.2 *Let the bottom topography be constant and the boundary conditions be periodic. Then the straight approach IMEX finite volume schemes and the corresponding elliptic approach IMEX finite volume schemes introduced in this chapter have a unique numerical solution, if the CFD numerical flux (6.22) is used.*

Let us now consider non-constant bottom topography. Instead of Lemma 6.8.1 we can use the the Gershgorin circle theorem [43] to estimate the eigenvalues:

Lemma 6.8.3 *A diagonally dominant matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$, i.e. $\sum_{j=1}^{n} |a_{ij}| \leq 2|a_{ii}|$, with positive a_{ii}, $i = 1, ..., n$, is positive semi-definite.*

Corollary 6.8.4 *Let the boundary conditions be periodic, fixed wall boundary or extrapolated. Further we use the CFD numerical flux (6.22) and the piecewise constant reconstruction (6.22) for stiff, linear parts. Then, the derived elliptic approach IMEX R-K and multi-step schemes have a unique numerical solution, if we use the approximations (6.57) or (6.59), (6.61) with the average operator (6.50). Due to Theorem 6.5.1 the corresponding straight approach IMEX finite volume schemes have also a unique numerical solution.*

Proof: Obviously the matrices $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x + \mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y$ corresponding to the discretisations (6.57) or (6.59), (6.61) are diagonally dominant, cf. (6.64b). Let us recall that our bottom topography $b < 0$. Thus the diagonal entries are positive. Due to Lemma 6.8.3 the corresponding matrices $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x + \mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y$ are therefore positive semi-definite. □

Remark 6.8.5 *If the average operator (6.48b) or linear or bilinear reconstruction, cf. Example 6.1.7, are used, the matrix $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x + \mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y$ is not diagonal dominant. Thus we can not use Lemma 6.8.3 to prove the positive semi-definite property. Indeed we show in the following example, that the matrix $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x + \mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y$ is not positive semi-definite, if the average operator (6.48b) is used. However here the matrix $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x + \mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y$ is not symmetric, thus it remains unknown if the corresponding linear system is non-singular.*

Example 6.8.6 *For the CFD numerical flux the matrix A is given by*

$$\frac{\mathcal{H}^L_{ij}(R\mathbf{w}^k) - \mathcal{K}_{ij}(R\mathbf{w}^k)}{\Delta x\Delta y} = \begin{bmatrix} D_x[(Rm_1)_{ij}] + D_y[(Rm_2)_{ij}] \\ -\frac{1}{\varepsilon^2}\left[D_x[(Rb)_{ij}(Rz)_{ij}] - \mu_x[(Rz)_{ij}]D_x[(Rb)_{ij}]\right] \\ -\frac{1}{\varepsilon^2}\left[D_y[(Rb)_{ij}(Rz)_{ij}] - \mu_x[(Rz)_{ij}]D_y[(Rb)_{ij}]\right] \end{bmatrix}, \quad (6.103)$$

where D_x, D_y are the approximations of the first order derivatives $\partial/\partial x, \partial/\partial y$ (6.48c), and μ_x, μ_y the average operators (6.48b) or (6.50).

1. *We calculate the matrix F_L (6.99) using the piecewise constant reconstruction (6.22), the average operator (6.48b) and periodic boundary conditions:*

$$\mathfrak{D}_x = \frac{1}{2\Delta x}\mathbb{1}_M \otimes \begin{bmatrix} 0 & 1 & 0 & ... & 0 & -1 \\ -1 & 0 & 1 & 0 & & 0 \\ 0 & -1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 1 & 0 \\ 0 & & 0 & -1 & 0 & 1 \\ 1 & 0 & ... & 0 & -1 & 0 \end{bmatrix} =: \frac{1}{2\Delta x}\mathbb{1}_M \otimes D_N, \quad (6.104a)$$

$$\mathfrak{D}_y = \frac{1}{2\Delta y}D_M \otimes \mathbb{1}_N, \quad (6.104b)$$

$$\overline{\mathfrak{B}\mathfrak{D}}_x = \frac{1}{2\Delta x}\begin{bmatrix} \overline{BD}^{(1)}_x & 0 & ... & .0 \\ 0 & \overline{BD}^{(2)}_x & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & ... & 0 & \overline{BD}^{(N)}_x \end{bmatrix}, \quad (6.104c)$$

$$\overline{BD}_x^{(j)} = \begin{bmatrix} -\frac{b_{2j}-b_{Nj}}{2} & \frac{3b_{2j}+b_{Nj}}{4} & 0 & ... & 0 & -\frac{b_{2j}+3b_{Nj}}{4} \\ -\frac{b_{3j}+3b_{1j}}{4} & -\frac{b_{3j}-b_{1j}}{2} & \frac{3b_{3j}+b_{1j}}{4} & 0 & ... & 0 \\ 0 & -\frac{b_{4j}+3b_{2j}}{4} & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \frac{3b_{N-1,j}+b_{N-3,j}}{4} & 0 \\ 0 & ... & 0 & -\frac{b_{Nj}+3b_{N-2,j}}{4} & -\frac{b_{Nj}-b_{N-2,j}}{2} & \frac{3b_{Nj}+b_{N-2,j}}{4} \\ \frac{3b_{1j}+b_{N-1,j}}{4} & 0 & ... & 0 & -\frac{b_{1j}+3b_{N-1,j}}{4} & -\frac{b_{1j}-b_{N-1,j}}{2} \end{bmatrix}, \tag{6.104d}$$

$$\overline{\mathfrak{BD}}_y = \frac{1}{2\Delta y} \begin{bmatrix} \overline{BD}_y^{(c,1)} & \overline{BD}_y^{(u,1)} & 0 & ... & 0 & \overline{BD}_y^{(d,1)} \\ \overline{BD}_y^{(d,2)} & \overline{BD}_y^{(c,2)} & \overline{BD}_y^{(u,2)} & 0 & ... & 0 \\ 0 & \overline{BD}_y^{(d,3)} & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \overline{BD}_y^{(u,M-2)} & 0 \\ 0 & & 0 & \overline{BD}_y^{(d,M-1)} & \overline{BD}_y^{(c,M-1)} & \overline{BD}_y^{(u,M-1)} \\ \overline{BD}_y^{(u,M)} & 0 & ... & 0 & \overline{BD}_y^{(d,M)} & \overline{BD}_y^{(c,M)} \end{bmatrix}, \tag{6.104e}$$

$$\overline{BD}_y^{(u,i)} = \begin{bmatrix} \frac{3b_{i2}+b_{iM}}{4} & 0 & ... & .0 \\ 0 & \frac{3b_{i3}+b_{i1}}{4} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & ... & 0 & \frac{3b_{i1}+b_{i,M-1}}{4} \end{bmatrix}, \tag{6.104f}$$

$$\overline{BD}_y^{(c,i)} = \begin{bmatrix} -\frac{b_{i2}-b_{iM}}{2} & 0 & ... & .0 \\ 0 & -\frac{b_{i3}-b_{i1}}{2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & ... & 0 & -\frac{b_{i1}-b_{i,M-1}}{2} \end{bmatrix}, \tag{6.104g}$$

$$\overline{BD}_y^{(d,i)} = \begin{bmatrix} -\frac{3b_{i2}+b_{iM}}{4} & 0 & ... & .0 \\ 0 & -\frac{3b_{i3}+b_{i1}}{4} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & ... & 0 & -\frac{3b_{i1}+b_{i,M-1}}{4} \end{bmatrix}. \tag{6.104h}$$

It seems pretty awkward to show that $\mathfrak{D}_x\overline{\mathfrak{BD}}_x + \mathfrak{D}_y\overline{\mathfrak{BD}}_y$ is positive semi-definite for all negative bottom topographies. However we can break the proof down to the 1D case: Therefore, note that if we take the bottom topography b and the free surface state vector $\mathbf{Z}$ constant in y-direction it holds $\mathbf{Z}^T(\mathfrak{D}_x\overline{\mathfrak{BD}}_x + \mathfrak{D}_y\overline{\mathfrak{BD}}_y)\mathbf{Z} = \mathbf{Z}^T\mathfrak{D}_x\overline{\mathfrak{BD}}_x\mathbf{Z}$. Furthermore, one finds $\overline{BD}_x := \overline{BD}_x^{(1)} = ... = \overline{BD}_x^{(N)}$. Thus, $D\overline{BD}_x$ is positive semi-definite for every negative bottom topography, if $\mathfrak{D}_x\overline{\mathfrak{BD}}_x + \mathfrak{D}_y\overline{\mathfrak{BD}}_y$ is positive semi-definite for every negative bottom topography. Consequently, we obtain that if $\mathfrak{D}_x\overline{\mathfrak{BD}}_x + \mathfrak{D}_y\overline{\mathfrak{BD}}_y$ is positive semi-definite for every negative bottom topography b, then the corresponding 1D discretisation is also positive semi-definite for every negative bottom topography. Vice versa, if the 1D discretisation leads positive semi-definite matrices, i.e. $D\overline{BD}_x^{(i)}$ is positive semi-definite for $i = 1, ..., N$,

then $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x$ is positive semi-definite. We show that $\mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y$ is positive semi-definite, if and only if $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x$ is positive semi-definite. Then, $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x + \mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y$ is positive semi-definite for every negative bottom topography b, if and only if the corresponding 1D discretisation is also positive semi-definite for every negative bottom topography. To this end we introduce the permutation matrix $P \in \mathbb{R}^{MN\times MN}$ that is given uniquely by the following property

$$P\mathbf{Z} = (z_{11}, z_{12}, z_{13}, ..., z_{1M}, z_{21}, z_{22}, ..., z_{2M}, ..., z_{NM})^T, \tag{6.105}$$

i.e. the vector $P\mathbf{Z}$ describes the same discrete state as $\mathbf{Z}$, if the x-axis is switched with the y-axis and Δx is switched with Δy. Thus $P^2 = \mathbb{1}$ and $P\mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_yP$ is $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x$ with switched M and N. Consequently, if $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x$ is positive semi-definite, then $P\mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_yP$ is positive semi-definite and $\mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y$ too, since $P = P^{-1}$.

We showed that it suffices to consider the 1D case, i.e. $\mathfrak{D}_x\overline{\mathfrak{B}\mathfrak{D}}_x+\mathfrak{D}_y\overline{\mathfrak{B}\mathfrak{D}}_y$ is positive semi-definite for all negative bottom topographies, if and only if $D\overline{BD}_x$ is positive semi-definite for all negative bottom topographies. Unfortunately $D\overline{BD}_x$ is not positive semi-definite. This can be obtained in the following way: Let us assume that $D\overline{BD}_x$ is positive semi-definite. Then $\mathbb{1}\delta + D\overline{BD}_x$ is positive definite for all $\delta > 0$. A matrix A is positive/negative (semi-)definite, if and only if its symmetric part $(A+A^T)/2$ is positive/negative (semi-)definite. Thus $2\delta\mathbb{1}+D\overline{BD}_x+(D\overline{BD}_x)^T$ is positive definite. Due to the Hurwitz-/Sylvester criterion the principal minors of $2\delta\mathbb{1} + D\overline{BD}_x + (D\overline{BD}_x)^T$ are larger then zero. It can be calculated that

$$D\overline{BD}_x + (D\overline{BD}_x)^T = \begin{bmatrix} \alpha_1 & \beta_1 & \gamma_1 & 0 & ... & 0 & \hat{\gamma}_1 & \hat{\beta}_1 \\ \hat{\beta}_2 & \alpha_2 & \beta_2 & \gamma_2 & 0 & ... & 0 & \hat{\gamma}_2 \\ \hat{\gamma}_3 & \hat{\beta}_3 & \alpha_3 & \beta_3 & \gamma_3 & 0 & ... & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & ... & 0 & \hat{\gamma}_{N-2} & \hat{\beta}_{N-2} & \alpha_{N-2} & \beta_{N-2} & \gamma_{N-2} \\ \gamma_{N-1} & 0 & ... & 0 & \hat{\gamma}_{N-1} & \hat{\beta}_{N-1} & \alpha_{N-1} & \beta_{N-1} \\ \beta_N & \gamma_N & 0 & ... & 0 & \hat{\gamma}_N & \hat{\beta}_N & \alpha_N \end{bmatrix}, \tag{6.106}$$

$$\alpha_i = -\frac{b_{i-2} + 6b_i + b_{i+2}}{3}, \quad \beta_i = \frac{b_{i+1} - b_{i-1} - b_{i+2} + b_i}{2},$$
$$\hat{\beta}_i = \frac{b_i - b_{i-2} - b_{i+1} + b_{i-1}}{2}, \quad \gamma_i = b_i + b_{i+2}, \quad \hat{\gamma}_i = b_{i-2} + b_i.$$

If we set $b_N = b_1 = b_2 = b_4 = -1$ and $b_3 = -20$ we obtain that the second principal minor is negative for sufficient small δ. Thus the matrix $D\overline{BD}_x$ is not positive semi-definite. However the chosen bottom topography is quite singular. Recall that the positive semi-definite property is just a sufficient condition for the non-singularity. Thus we do not know, if the described matrix $\mathbb{1}\delta + F_L$ is non-singular or not.

2. *We calculate the matrix F_L (6.99) using piecewise constant reconstruction, see Example 6.1.7, the average operator (6.50) and periodic boundary conditions: Using*

average (6.50) instead of (6.48c) changes only the matrices $\overline{BD}_x^{(j)}$, $j = 1, ..., M$, $\overline{BD}_y^{(u,i)}, \overline{BD}_y^{(c,i)}, \overline{BD}_y^{(d,i)}$, $j = 1, ..., N$

$$\overline{BD}_x^{(j)} = \begin{bmatrix} 0 & \frac{b_{2j}+b_{Nj}}{2} & 0 & ... & 0 & -\frac{b_{2j}+b_{Nj}}{2} \\ -\frac{b_{3j}+b_{1j}}{2} & 0 & \frac{b_{3j}+b_{1j}}{2} & 0 & ... & 0 \\ 0 & -\frac{b_{4j}+b_{2j}}{2} & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \frac{b_{N-1,j}+b_{N-3,j}}{2} & 0 \\ 0 & ... & 0 & -\frac{b_{Nj}+b_{N-2,j}}{2} & & \frac{b_{Nj}+b_{N-2,j}}{2} \\ \frac{b_{1j}+b_{N-1,j}}{2} & 0 & ... & 0 & -\frac{b_{1j}+b_{N-1,j}}{2} & 0 \end{bmatrix}, \tag{6.107a}$$

$$\overline{BD}_y^{(u,i)} = \begin{bmatrix} \frac{b_{i2}+b_{iM}}{2} & 0 & ... & .0 \\ 0 & \frac{b_{i3}+b_{i1}}{2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & ... & 0 & \frac{b_{i1}+b_{i,M-1}}{2} \end{bmatrix}, \tag{6.107b}$$

$$\overline{BD}_y^{(c,i)} = 0, \quad \overline{BD}_y^{(d,i)} = -\overline{BD}_y^{(u,i)}. \tag{6.107c}$$

With the argumentation from the Example 6.8.6.1 it suffices to show that $D\overline{BD}_x^{(1)}$ *is positive semi-definite for any negative bottom topography. A straightforward calculation yields*

$$\mathbf{x}^T D\overline{BD}_x^{(1)}\mathbf{x} = \sum_{i=1}^{N} -\frac{b_{i1}}{2}\left[(x_i - x_{i+1})^2 + (x_i - x_{i-1})^2\right] \geq 0, \tag{6.108}$$

where $x_0 := x_N$ *and* $x_{N+1} := x_1$. *This coincides with the statement of Corollary 6.8.4.*

6.8.2. EG numerical flux

The aim of this section is to prove that the linear system (6.98) posses a unique solution for all Δt and $\varepsilon > 0$, if the bottom topography is constant, $M = N$, the boundary conditions are periodic and the EG numerical flux (6.29b) is used to approximate the stiff, linear parts. Then the matrix $A \in \mathbb{R}^{3N^2 \times 3N^2}$, given by (6.97), consists of the nine circulant block matrices $C_1, ..., C_9 \in Circ_{N\times N}$,

$$A = \begin{bmatrix} c_b C_1 & C_2 & C_3 \\ c_b^2 C_4 & c_b C_5 & c_b C_6 \\ c_b^2 C_7 & c_b C_8 & c_b C_9 \end{bmatrix} \in \mathbb{R}^{3N^2 \times 3N^2}, \quad c_b = \frac{\sqrt{-b}}{\varepsilon}. \tag{6.109}$$

We can describe the matrices $C_1, ..., C_9$ using the notation

$$\mathcal{C}^i(a_0, a_1, ..., a_{2i+1}) := \mathcal{C}(a_0, a_1, ..., a_i, 0, ..., 0, a_{i+1}, ..., a_{2i+1}), \tag{6.110}$$

where $\mathcal{C}$ denotes the isomorphisms (6.75) or (6.89). We have

$$C_1 = \frac{8\alpha}{\pi}\mathcal{C}^1(4\mathcal{C}^1(5,-1,-1),\mathcal{C}^1(-4,-1,-1),\mathcal{C}^1(-4,-1,-1)) \hat{=} -\frac{2\Delta x}{\pi}\Delta + \mathcal{O}(\Delta x^2) \tag{6.111a}$$

$$C_2 = \alpha\mathcal{C}^1(\mathcal{C}^1(0,\gamma,-\gamma),\mathcal{C}^1(0,\beta,-\beta),\mathcal{C}^1(0,\beta,-\beta)) \hat{=} D_x + \mathcal{O}(\Delta x^2) \tag{6.111b}$$

$$C_3 = \alpha\mathcal{C}^1(\mathcal{C}^1(0,0,0),\mathcal{C}^1(\gamma,\beta,\beta),\mathcal{C}^1(-\gamma,-\beta,-\beta)) \hat{=} D_y + \mathcal{O}(\Delta x^2) \tag{6.111c}$$

$$C_4 = \alpha\mathcal{C}^1(10\mathcal{C}^1(0,1,-1),\mathcal{C}^1(0,1,-1),\mathcal{C}^1(0,1,-1)) \hat{=} D_x + \mathcal{O}(\Delta x^2) \tag{6.111d}$$

$$C_5 = \frac{2\alpha}{\pi}\mathcal{C}^1(10\mathcal{C}^1(2,-1,-1),\mathcal{C}^1(2,-1,-1),\mathcal{C}^1(2,-1,-1)) \hat{=} -\frac{\Delta x}{\pi}D_{xx} + \mathcal{O}(\Delta x^2) \tag{6.111e}$$

$$C_6 = \frac{2\alpha}{\pi}\mathcal{C}^1(\mathcal{C}^1(0,0,0),\mathcal{C}^1(0,-1,1),\mathcal{C}^1(0,1,-1)) \hat{=} -\frac{\Delta x}{3\pi}D_{xy} + \mathcal{O}(\Delta x^2) \tag{6.111f}$$

$$C_7 = \alpha\mathcal{C}^1(\mathcal{C}^1(0,0,0),\mathcal{C}^1(10,1,1),\mathcal{C}^1(-10,-1,-1)) \hat{=} D_y + \mathcal{O}(\Delta x^2), \tag{6.111g}$$

$$C_8 = C_6 \hat{=} -\frac{\Delta x}{3\pi}D_{xy} + \mathcal{O}(\Delta x^2) \tag{6.111h}$$

$$C_9 = \frac{2\alpha}{\pi}\mathcal{C}^1(\mathcal{C}^1(20,2,2),\mathcal{C}^1(-10,-1,-1),\mathcal{C}^1(-10,-1,-1)) \hat{=} -\frac{\Delta x}{\pi}D_{yy} + \mathcal{O}(\Delta x^2), \tag{6.111i}$$

$$\alpha = \frac{1}{24\Delta x}, \quad \beta = 1 + \frac{2}{\pi}, \quad \gamma = 10 - \frac{4}{\pi}. \tag{6.111j}$$

Using the Taylor expansion it is easy to show that the matrices $C_1, ..., C_9$ are approximations of differential operators, see right-hand side of (6.111). Further, due to (6.98) we have

$$\begin{aligned}
&(\mathbb{1}_3 \otimes (F^{NN})^*)(\mathbb{1} + \Delta t A)(\mathbb{1}_3 \otimes F^{NN})(\mathbb{1}_3 \otimes (F^{NN})^*)\mathbf{W} \qquad (6.112)\\
&= \left[\mathbb{1} + \Delta t \begin{bmatrix} J(C_1) & J(C_2) & J(C_3) \\ J(C_4) & J(C_5) & J(C_6) \\ J(C_7) & J(C_8) & J(C_9) \end{bmatrix}\right] \tilde{\mathbf{W}} = (\mathbb{1}_3 \otimes F^*_{NN})\hat{\mathbf{W}},\\
&\tilde{\mathbf{W}} = (\mathbb{1}_3 \otimes (F^{NN})^*)\mathbf{W}, \quad F^{NN} = F_N \otimes F_N, \quad J(C_i) = (F^{NN})^* C_i F_{NN}, \quad i = 1, ..., 9,
\end{aligned}$$

where F_N is the $(N \times N)$-Fourier matrix (6.80). Due to Theorems 6.7.3, 6.7.7 $J(C_i)$ is the Jordan normal form of C_i for $i = 1, ..., 9$. Moreover $J(C_i)$ is a diagonal matrix with the diagonal

$$diag(J(C_i)) = (\lambda^i_{11}, ..., \lambda^i_{1N}, \lambda^i_{21}, ..., \lambda^i_{2N}, ..., , \lambda^i_{N1}, ..., \lambda^i_{NN}), \tag{6.113}$$

for , $i = 1, ..., 9$.

We can calculate the eigenvalues of the matrices $C_1, ..., C_9$ using Theorems 6.7.3, 6.7.7:

$$\begin{aligned}
\sigma(C_1) &= \left\{ \frac{8\alpha}{\pi}[20 - 8\cos\left(\frac{2\pi}{N}l\right) - 8\cos\left(\frac{2\pi}{N}k\right) - 4\cos\left(\frac{2\pi}{N}l\right)] : k,l = 0,...,N-1 \right\} \\
\sigma(C_2) &= \left\{ i2\alpha \sin\left(\frac{2\pi}{N}l\right)\left[\gamma + \beta 2\cos\left(\frac{2\pi}{N}k\right)\right] : k,l = 0,...,N-1 \right\} \\
\sigma(C_3) &= \left\{ i2\alpha \sin\left(\frac{2\pi}{N}k\right)\left[\gamma + \beta 2\cos\left(\frac{2\pi}{N}l\right)\right] : k,l = 0,...,N-1 \right\} \\
\sigma(C_4) &= \left\{ i4\alpha \sin\left(\frac{2\pi}{N}l\right)\left[5 + \cos\left(\frac{2\pi}{N}k\right)\right] : k,l = 0,...,N-1 \right\} \\
\sigma(C_5) &= \left\{ \frac{8\alpha}{\pi}\left[1 - \cos\left(\frac{2\pi}{N}l\right)\right]\left[5 + \cos\left(\frac{2\pi}{N}k\right)\right] : k,l = 0,...,N-1 \right\} \\
\sigma(C_6) &= \left\{ \frac{8\alpha}{\pi}\sin\left(\frac{2\pi}{N}k\right)\sin\left(\frac{2\pi}{N}l\right) : k,l = 0,...,N-1 \right\} \\
\sigma(C_7) &= \left\{ i4\alpha \sin\left(\frac{2\pi}{N}k\right)\left[5 + \cos\left(\frac{2\pi}{N}l\right)\right] : k,l = 0,...,N-1 \right\} \\
\sigma(C_8) &= \sigma(C_1^{(6)}) \\
\sigma(C_9) &= \left\{ \frac{8\alpha}{\pi}\left[1 - \cos\left(\frac{2\pi}{N}k\right)\right]\left[5 + \cos\left(\frac{2\pi}{N}l\right)\right] : k,l = 0,...,N-1 \right\}, \\
\alpha &= \frac{1}{24\Delta x}, \quad \beta = 1 + \frac{2}{\pi}, \quad \gamma = 10 - \frac{4}{\pi}.
\end{aligned} \tag{6.114}$$

Since the matrices $C_1, ..., C_9$ commute, the matrix $\mathbb{1} + \Delta t A$ is non-singular, if and only if the matrix

$$\begin{aligned}
cdet(\mathbb{1} + \Delta t A) := (\mathbb{1} + \Delta t c_b C_1)(\mathbb{1} + \Delta t c_b C_2)(\mathbb{1} + \Delta t c_b C_3) + (\Delta t c_b)^3(C_3C_6C_7 + C_3C_4C_8) \\
- (\Delta t c_b)^2(C_2C_4(\mathbb{1} + \Delta t c_b C_9) + (\mathbb{1} + \Delta t c_b C_1)C_6C_8) \in Circ_{N\times N}
\end{aligned} \tag{6.115}$$

is non-singular. Note that the matrix $cdet(\mathbb{1} + \Delta t A)$ is obtained by computing the determinant of a 3×3-matrix following Sarrus scheme. Using Theorems 6.7.3, 6.7.7 we obtain the eigenvalues of $cdet(\mathbb{1} + \Delta t A)$

$$\begin{aligned}
\sigma(cdet(\mathbb{1} + \Delta t A)) = \{\lambda_{kl} = 1 + (\Delta t c_b)(\lambda_{kl}^1 + \lambda_{kl}^5 + \lambda_{kl}^9) + (\Delta t c_b)^2(\lambda_{kl}^1\lambda_{kl}^5 + \lambda_{kl}^1\lambda_{kl}^9 + \lambda_{kl}^5\lambda_{kl}^9) \\
+ (\Delta t c_b)^3 \det \begin{bmatrix} \lambda_{kl}^1 & \lambda_{kl}^2 & \lambda_{kl}^3 \\ \lambda_{kl}^4 & \lambda_{kl}^5 & \lambda_{kl}^6 \\ \lambda_{kl}^7 & \lambda_{kl}^8 & \lambda_{kl}^9 \end{bmatrix} : 0 \le k,l \le N-1\},
\end{aligned} \tag{6.116}$$

where λ_{kl}^i is the corresponding eigenvalue of C_i in (6.114), $i = 1, ..., 9$.

Lemma 6.8.7 *The spectrum of $cdet(\mathbb{1} + \Delta t A)$ is real and bounded from below by* 1.

Proof: Let us consider an eigenvalue λ_{kl} of $cdet(\mathbb{1} + \Delta t A)$ as a polynomial in $\Delta t c_b$ of degree three, cf. (6.116). We show that all coefficients are larger then zero. Thus $\lambda_{kl} \ge 1$, since $\Delta t c_b > 0$.

- Obviously the eigenvalues of C_1, C_5, C_9 are larger or equal to zero, thus the coefficient of $\Delta t c_b$ is larger or equal to zero.

- It is easily seen, that every term in the coefficient of $(\Delta t c_b)^2$ is positive but $-\lambda_{kl}^6\lambda_{kl}^8 = -(\lambda_{kl}^6)^2$. However

$$\begin{aligned}&\lambda_{kl}^5\lambda_{kl}^9 - (\lambda_{kl}^6)^2 \\ &= \left(\frac{8\alpha}{\pi}\right)^2\left[\left(5+\cos\left(\frac{2\pi k}{N}\right)\right)\left(1-\cos\left(\frac{2\pi k}{N}\right)\right)\left(5+\cos\left(\frac{2\pi l}{N}\right)\right)\left(1-\cos\left(\frac{2\pi l}{N}\right)\right)\right. \\ &\left.-\sin^2\left(\frac{2\pi k}{N}\right)\sin^2\left(\frac{2\pi l}{N}\right)\right] \geq 0\end{aligned} \tag{6.117}$$

  follows from the following inequality

$$(5+\cos\phi)(1-\cos\phi) = 5 - 4\cos\phi - \cos^2\phi \geq 1 - \cos^2\phi = \sin^2\phi. \tag{6.118}$$

  Thus the coefficient is positive.

- The coefficient of $(\Delta t c_b)^3$ can be written as

$$\lambda_{kl}^1(\lambda_{kl}^5\lambda_{kl}^9 - (\lambda_{kl}^6)^2) + \lambda_{kl}^2(\lambda_{kl}^6\lambda_{kl}^7 - \lambda_{kl}^4\lambda_{kl}^9) + \lambda_{kl}^3(\lambda_{kl}^4\lambda_{kl}^8 - \lambda_{kl}^5\lambda_{kl}^7). \tag{6.119}$$

  We already proved that the first summand in (6.119) is positive. The second one is positive due to

$$\begin{aligned}&\lambda_{kl}^2(\lambda_{kl}^6\lambda_{kl}^7 - \lambda_{kl}^4\lambda_{kl}^9) \\ &= (2\alpha)^3\sin^2\left(\frac{2\pi l}{N}\right)\left(10 - \frac{4}{\pi} + 2\left(1+\frac{2}{\pi}\cos\left(\frac{2\pi k}{N}\right)\right)\right)16\left(1-\cos\left(\frac{2\pi k}{N}\right)\right) \geq 0.\end{aligned} \tag{6.120}$$

  Note that the third summand is obtained by changing k with l in (6.120). Hence it is also larger then zero. This concludes the proof.

□

Remark 6.8.8 *Note that*

- λ_{kl}^1 *is zero, if and only if* $k = l = 0$. λ_{kl}^5 *is zero, if and only if* $l = 0$. λ_{kl}^9 *is zero, if and only if* $k = 0$. *Thus* $\lambda_{kl}^1(\lambda_{kl}^5\lambda_{kl}^9 - (\lambda_{kl}^6)^2) = 0$, *if and only if* $k = 0$ *or* $l = 0$.
- $\lambda_{kl}^2(\lambda_{kl}^6\lambda_{kl}^7 - \lambda_{kl}^4\lambda_{kl}^9) = 0$, *if* $k = 0$ *or* $l = 0$, *cf. (6.120).*
- $\lambda_{kl}^3(\lambda_{kl}^4\lambda_{kl}^8 - \lambda_{kl}^5\lambda_{kl}^7)$ *is obtained by changing* k *with* l *in* $\lambda_{kl}^2(\lambda_{kl}^6\lambda_{kl}^7 - \lambda_{kl}^4\lambda_{kl}^9) = 0$. *Thus it is zero, if* $k = 0$ *or* $l = 0$.

Recall that $c_b = \sqrt{-b}/\varepsilon$. *Hence an eigenvalue* λ_{kl} *of the of* $cdet(\mathbb{1} + \Delta t A)$, *cf. (6.116), satisfies*

$$\lambda_{kl} = \begin{cases}\mathcal{O}(\varepsilon^{-3}) & , \text{ if } k,l \neq 0 \\ \mathcal{O}(\varepsilon^{-2}) & , \text{ if either } k = 0 \text{ or } l = 0 \\ 1 & , \text{ if } k = l = 0.\end{cases} \tag{6.121}$$

7. Asymptotic preserving property of fully discrete schemes

In Chapter 5 we have discussed the asymptotic preserving property of the time discretisations by means of IMEX R-K schemes of type A, CK and IMEX multi-step schemes, cf. Chapter 4. We have proved, in particular, that the IMEX Euler, RK2CN, ARS(2,2,2) and SBDF time discretisations, cf. Chapter 4, are asymptotic preserving, if suitable assumptions are satisfied. To this end, we have used asymptotic analysis arguments to obtain that ∇z^{n+1} and $\nabla \cdot \mathbf{m}^{n+1}$ are zero in the low Froude number limit. For a semi-discrete or fully discrete numerical scheme, it is not clear, that the corresponding numerical solutions have a limit as the Froude number approaches zero. However, let us assume that there is a limit of the numerical solutions as ε approaches zero. Then we would like to know, if the fully discrete solution is an approximation to the zero Froude number equations and if it is so, how good the approximation is.

In [38, 51] fully discrete schemes for the isentropic Euler and Navier-Stokes equations were shown to be asymptotic preserving. However, the authors assumed convergence of the numerical solutions as the Froude number ε approaches zero or used formal arguments.

The aim of this chapter is to study the asymptotic preserving property of the IMEX finite volume schemes. More precisely, let $\mathbf{W}^{n+1} = ((\mathbf{Z}^{n+1})^T, (\mathbf{M}^{n+1})^T)^T$, cf. (6.96), be the numerical solution at time t^{n+1} of the alternative SWE (2.30), obtained by some IMEX finite volume scheme, cf. Chapters 4 and 6, with discrete well-prepared initial condition $\mathbf{W}^n$ for some IMEX R-K time discretisation and $\mathbf{W}^n, \mathbf{W}^{n-1}, ..., \mathbf{W}^{n-k+1}$ for some IMEX k-step scheme, cf. Chapter 4. Thus

$$\begin{aligned}
\mathbf{Z}^{n-i+1} &= \mathbf{Z}^{n-i+1,(0)} + \varepsilon \mathbf{Z}^{n-i+1,(1)} + \varepsilon^2 \mathbf{Z}^{n-i+1,(2)} + \mathcal{O}(\varepsilon^3) && \text{(7.1a)}\\
&= (Z + \varepsilon Z')\mathbf{1} + \varepsilon^2 \mathbf{Z}^{n-i+1,(2)} + \mathcal{O}(\varepsilon^3),\\
\mathbf{M}^{n-i+1} &= \mathbf{M}^{n-i+1,(0)} + \varepsilon \mathbf{M}^{n-i+1,(1)} + \varepsilon^2 \mathbf{M}^{n-i+1,(2)} + \mathcal{O}(\varepsilon^3), && \text{(7.1b)}\\
\nabla_h \cdot \mathbf{M}^{n-i+1,(0)} &= \mathcal{O}(\max\{\Delta x^p, \Delta y^p\}), \quad p \in \mathbb{N}, \quad i = 1, ..., k, && \text{(7.1c)}
\end{aligned}$$

where $\nabla_h \cdot$ is a discrete divergence operator of order p, e.g. (7.4a). We have proved that numerical solutions of the IMEX finite volume schemes exist and are unique under certain conditions independent on the Froude number $\varepsilon > 0$, cf. Section 6. Thus there is a numerical solution $\mathbf{W}^{n+1} = \mathbf{W}_\varepsilon^{n+1}$ for each $\varepsilon > 0$. In Section 7.1 we consider the IMEX Euler time discretisation and provide conditions on the spatial discretisation that imply:

1. The sequence of numerical solutions $(\mathbf{W}_\varepsilon^{n+1})_\varepsilon$ converges to $\mathbf{W}_0^{n+1}$ as the Froude number ε approaches zero.

2. The limit $\mathbf{W}_0^{n+1}$ is a consistent approximation of the zero Froude number SWE (2.34). Moreover, the scheme is asymptotic preserving, cf. Definition 5.1.1.

Further, we demonstrate that some of our IMEX finite volume schemes presented before satisfy these conditions. In Section 7.2 we generalise the results from Section 7.1 to general IMEX R-K and IMEX multi-step schemes.

7.1. IMEX Euler time discretisation

The aim of this section is to study the asymptotic preserving property of the IMEX Euler finite volume schemes. To this end we provide conditions on the spatial discretisation so that the numerical solution $\mathbf{W}^{n+1} = \mathbf{W}_\varepsilon^{n+1}$ at a new time step t^{n+1} converges as the Froude number ε approaches zero. Then we point out that certain spatial discretisations from Chapter 6 satisfy these conditions. First we consider the CFD numerical flux and then the EG numerical flux, cf. Example 6.1.7.

Let A be the matrix given by (6.97). Then the straight approach IMEX Euler finite volume scheme reads

$$(\mathbb{1} + \Delta t A)\mathbf{W}^{n+1} = \hat{\mathbf{W}}, \tag{7.2}$$

cf. Section 6.2, where $\hat{\mathbf{W}}$ is the result of explicit approximation. Here we use $\mathbf{W}^{n+1}$ instead of more precise $\mathbf{W}_\varepsilon^{n+1}$ for simplicity of presentation. Analogously to (6.97) we can define the matrix

$$E : \mathbb{R}^{MN\times MN}, \quad \mathbf{Z} = (z_{ij}) \mapsto E\mathbf{Z} = ((\nabla\cdot(b\nabla z))_{ij}) \tag{7.3}$$

and the vectors

$$\nabla_h\cdot\hat{\mathbf{M}} := ((\nabla\cdot\hat{\mathbf{m}})_{11},...,(\nabla\cdot\hat{\mathbf{m}})_{N1},(\nabla\cdot\hat{\mathbf{m}})_{12},...,(\nabla\cdot\hat{\mathbf{m}})_{N2},...,(\nabla\cdot\hat{\mathbf{m}})_{NM})^T, \tag{7.4a}$$

$$\overline{\mathbf{B}\nabla_h}\mathbf{Z} := ((bz_x)_{11},...,(bz_x)_{N1},(bz_x)_{12},...,(bz_x)_{NM},(bz_y)_{11},...,(bz_y)_{NM})^T. \tag{7.4b}$$

Here $(\nabla\cdot(b\nabla z))_{ij}, (\nabla\cdot\hat{\mathbf{m}})_{ij}, (b\nabla z)_{ij}$ stand for the approximations used in Sections 6.4, 6.5. Then the elliptic approach IMEX Euler scheme reads

$$\left(\mathbb{1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 E\right)\mathbf{Z}^{n+1} = \hat{\mathbf{Z}} - \Delta t\nabla_h\cdot\hat{\mathbf{M}}, \tag{7.5a}$$

$$\mathbf{M}^{n+1} = \hat{\mathbf{M}} + \Delta t\frac{\overline{\mathbf{B}\nabla_h}\mathbf{Z}^{n+1}}{\varepsilon^2}. \tag{7.5b}$$

7.1.1. Central finite difference numerical flux

In order to prove the asymptotic preserving property for the straight approach IMEX Euler finite volume scheme (7.2), we could use the following theorem.

Theorem 7.1.1 *Consider a straight approach IMEX finite volume scheme (7.2), where the CFD numerical flux (6.22) and a corresponding source term approximation (6.48a) are used to approximate the implicit parts. The Lax-Friedrichs or Rusanov numerical fluxes are used for the explicit, non-linear terms. Further let the following conditions be satisfied:*

1. $\mathbb{1} + \Delta t A$ *is non-singular for all* $\varepsilon > 0$.
2. *The limit*

$$\lim_{\varepsilon \to 0} (\mathbb{1} + \Delta t A)^{-1} \tag{7.6}$$

   exists.

3. *The kernel intersection of* $\mathcal{N}(\overline{\mathfrak{B}\mathfrak{D}}_x), \mathcal{N}(\overline{\mathfrak{B}\mathfrak{D}}_y)$ *is*

$$\mathcal{N}(\overline{\mathfrak{B}\mathfrak{D}}_x) \cap \mathcal{N}(\overline{\mathfrak{B}\mathfrak{D}}_y) = \langle \mathbf{1} \rangle := \{ c\mathbf{1} | c \in \mathbb{R} \}. \tag{7.7}$$

Then

$$\overline{\mathfrak{B}\mathfrak{D}}_x \mathbf{Z}^{n+1} = \overline{\mathfrak{B}\mathfrak{D}}_y \mathbf{Z}^{n+1} = \mathcal{O}(\varepsilon^2), \tag{7.8a}$$
$$\mathbf{Z}^{n+1} = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2), \tag{7.8b}$$
$$\mathfrak{D}_x \mathbf{M}_1 + \mathfrak{D}_y \mathbf{M}_2 = \mathcal{O}(\varepsilon^2). \tag{7.8c}$$

Hence the scheme is asymptotic preserving.

Proof: Since the limit (7.6) exists, there is a low Froude number limit of the numerical solution

$$\mathbf{W}^{n+1,(0)} := \lim_{\varepsilon \to 0} \mathbf{W}^{n+1} = \lim_{\varepsilon \to 0} (\mathbb{1} + \Delta t A)^{-1} \hat{\mathbf{W}}. \tag{7.9}$$

Particularly $\mathbf{W}^{n+1}$ is uniformly bounded with respect to $\varepsilon \in (0, 1)$. Vice versa it holds

$$\lim_{\varepsilon \to 0} (\mathbb{1} + \Delta t A) \mathbf{W}^{n+1} = \lim_{\varepsilon \to 0} \hat{\mathbf{W}}. \tag{7.10}$$

Let us recall that the matrix A has the following structure, cf. (6.99),

$$A = \begin{bmatrix} 0 & \mathfrak{D}_x & \mathfrak{D}_y \\ -\frac{1}{\varepsilon^2} \overline{\mathfrak{B}\mathfrak{D}}_x & 0 & 0 \\ -\frac{1}{\varepsilon^2} \overline{\mathfrak{B}\mathfrak{D}}_y & 0 & 0 \end{bmatrix}. \tag{7.11}$$

Since the limit on the left-hand side of (7.10) exists and $\mathbf{W}^{n+1}$ is bounded, $\mathbf{Z}^{n+1,(0)} = \tilde{\mathbf{Z}} + \mathcal{O}(\varepsilon^2)$, where $\tilde{\mathbf{Z}}$ is in the kernel intersection $\mathcal{N}(\overline{\mathfrak{B}\mathfrak{D}}_x) \cap \mathcal{N}(\overline{\mathfrak{B}\mathfrak{D}}_y) = \langle \mathbf{1} \rangle$. Hence we can write $\tilde{\mathbf{Z}} = \tilde{Z}\mathbf{1}$. The free surface elevation is a conservative quantity and the scheme is conservative. Thus we have

$$\tilde{Z} + \mathcal{O}(\varepsilon^2) = Z + \varepsilon Z' + \mathcal{O}(\varepsilon^2), \tag{7.12}$$

which gives us $\mathbf{Z}^{n+1} = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2)$.

Note that if we use Lax-Friedrichs or Rusanov numerical flux, then $\hat{\mathbf{Z}} = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2)$. Thus we obtain

$$\begin{aligned}\mathbf{Z}^n + \mathcal{O}(\varepsilon^2) + \Delta t(\mathfrak{D}_x \mathbf{M}_1^{n+1} + \mathfrak{D}_y \mathbf{M}_2^{n+1}) &= \mathbf{Z}^{n+1} + \Delta t(\mathfrak{D}_x \mathbf{M}_1^{n+1} + \mathfrak{D}_y \mathbf{M}_2^{n+1}) \\ &= \hat{\mathbf{Z}} = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2)\end{aligned} \tag{7.13}$$

from the discrete continuity equation. Consequently we have $\mathfrak{D}_x \mathbf{M}_1^{n+1} + \mathfrak{D}_y \mathbf{M}_2^{n+1} = \mathcal{O}(\varepsilon^2)$ and provides a discrete divergence constraint for the momentum in the low Froude number limit. Moreover, the straight approach IMEX finite volume scheme (7.2) reads

$$\mathbf{Z}^{n+1,(0)} = \mathbf{Z}^{n,(0)}, \quad \mathbf{M}^{n+1,(0)} = \hat{\mathbf{M}} + \Delta t \overline{B \nabla}_h \mathbf{Z}^{n+1,(2)} \tag{7.14}$$

in the low Froude number limit $\varepsilon \to 0$, which is a consistent approximation of the alternative zero Froude number SWE. □

The analogous theorem for the elliptic approach reads

Theorem 7.1.2 *Consider an elliptic approach IMEX finite volume scheme (7.5), where the free surface elevation is conserved and the Lax-Friedrichs or Rusanov numerical flux is used for the explicit, non-linear terms. Further let the following conditions be satisfied:*

1. $\mathbb{1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 E$ *is non-singular for all* $\varepsilon > 0$.
2. *The limit*

$$\lim_{\varepsilon \to 0} \left(\mathbb{1} + \left(\frac{\Delta t}{\varepsilon} \right)^2 E \right)^{-1} \tag{7.15}$$

   exists.

3. *The kernel of* $E, \overline{\mathbf{B} \nabla_h}$ *satisfy*

$$\mathcal{N}(E) = \langle \mathbf{1} \rangle \subset \mathcal{N}(\overline{\mathbf{B} \nabla_h}). \tag{7.16}$$

Then

$$\frac{\Delta t}{\varepsilon^2} E \mathbf{Z}^{n+1} + \nabla_h \cdot \hat{\mathbf{M}} = \mathcal{O}(\varepsilon^2), \tag{7.17a}$$

$$\mathbf{Z}^{n+1} = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2), \tag{7.17b}$$

$$\lim_{\varepsilon \to 0} \mathbf{M}^{n+1} \text{ exists.} \tag{7.17c}$$

Furthermore, if

$$E = \nabla_h \cdot \overline{\mathbf{B} \nabla_h}, \tag{7.18}$$

then

$$\nabla_h \cdot \mathbf{M}^{n+1} = \nabla_h \cdot \hat{\mathbf{M}} + \frac{\Delta t}{\varepsilon^2} E \mathbf{Z}^{n+1} = \mathcal{O}(\varepsilon^2). \tag{7.19}$$

Hence the scheme is asymptotic preserving.

Remark 7.1.3 *The elliptic approach IMEX Euler finite volume scheme (7.5) can be written as straight approach IMEX Euler finite volume scheme, if and only if condition (7.18) is satisfied. Thus, it seems that an elliptic approach IMEX Euler scheme, that can not be represented using the straight approach, does not satisfy a divergence constraint for the momentum in the low Froude number limit.*

Proof: Since the limit (7.15) exists, there is a low Froude number limit of the numerical solution

$$\mathbf{Z}^{n+1,(0)} := \lim_{\varepsilon\to 0} \mathbf{Z}^{n+1} = \lim_{\varepsilon\to 0} \left(\mathbb{1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 E \right)^{-1} (\hat{\mathbf{Z}} - \nabla \cdot \hat{\mathbf{M}}). \tag{7.20}$$

As in the the proof of Theorem 7.1.1 we obtain (7.17a), (7.17b). Since the initial data are well-prepared, we have $\mathbf{Z}^{n+1} = (Z + \varepsilon Z')\mathbf{1} + \mathcal{O}(\varepsilon^2)$. Hence

$$\mathbf{M}^{n+1,(0)} := \lim_{\varepsilon\to 0} \mathbf{M}^{n+1} = \lim_{\varepsilon\to 0} \left(\hat{\mathbf{M}} + \frac{\Delta t}{\varepsilon^2} \overline{\mathbf{B}\nabla_h} \mathbf{Z}^{n+1} \right). \tag{7.21}$$

exists.

If additionally (7.18) holds, then multiplying the momentum update (7.5b) with the discrete divergence operator $\nabla_h \cdot$ and using (7.17a) leads (7.19). Consequently, we obtain the discrete limit equations (7.14). □

Theorems 7.1.1, 7.1.2 guarantee the IMEX Euler finite volume scheme to be asymptotic preserving, if the above mentioned conditions $1, 2, 3$ are satisfied. In the following, we study which of the spatial discretisations introduced in Chapter 6 satisfy these conditions and are therefore asymptotic preserving. Indeed, all conditions are typically not satisfied, but sometimes we can circumvent this drawback and still obtain the asymptotic preserving property of an IMEX scheme by following the guideline of the proofs of Theorems 7.1.1, 7.1.2.

In Section 6.8 we provided two different conditions, which imply that the matrix $\mathbb{1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 E$ is non-singular for all Δt and $\varepsilon > 0$, cf. Corollary 6.8.2 and 6.8.4. In both cases the matrix E is symmetric and positive semi-definite. Thus we can apply the following lemma.

Lemma 7.1.4 *Let $E \in \mathbb{R}^{n\times n}$ be a symmetric and positive semi-definite matrix independent of a parameter ε. Then $\left(\mathbb{1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 E \right)$ is non-singular for all Δt and $\varepsilon > 0$ and we have*

$$\left(\mathbb{1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 E \right)^{-1} \mathbf{v} = \pi_{\mathcal{N}(E)}(\mathbf{v}) + \mathcal{O}(\varepsilon^2), \tag{7.22}$$

where $\mathbf{v} \in \mathbb{R}^n$ is a vector independent of ε and $\pi_{\mathcal{N}(E)}$ denotes the orthogonal projection onto the kernel of the matrix E.

Proof: The matrix E is diagonalisable with real eigenvalues $\mu_1, ..., \mu_n$. Thus the matrix $\left(\mathbb{1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 E\right)^{-1}$ is also diagonalisable with eigenvalues

$$\lambda_i = \frac{\varepsilon^2}{\varepsilon^2 + \Delta t^2 \mu_i} = \begin{cases} \mathcal{O}(\varepsilon^2) & \text{if } \mu_i \neq 0 \\ 1 & \text{if } \mu_i = 0 \end{cases}, \quad i = 1, ..., n. \tag{7.23}$$

Moreover, the basis of eigenvectors of E and $\left(\mathbb{1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 E\right)^{-1}$ is the same. Let $\mathbf{r}_i, i \in I$, be the eigenvectors to the eigenvalue zero and $\mathbf{v} \in \mathbb{R}^n$ a vector independent of the parameter ε. Then, we have

$$\left(\mathbb{1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 E\right)^{-1} \mathbf{v} = \sum_{i \in I} (\mathbf{r}_i \cdot \mathbf{v}) \mathbf{r}_i + \mathcal{O}(\varepsilon^2) \tag{7.24}$$

due to (7.23). □

Remark 7.1.5 *Lemma 7.1.4 can be extended in the following way: Let $E \in \mathbb{R}^{n \times n}$ be a matrix independent of a parameter ε. Then, there exists a $\delta > 0$ so that $(\mathbb{1} + \Delta t^2 E/\varepsilon^2)$ is non-singular for all $\varepsilon \in (0, \delta)$. If E is positive semi-definite we can choose $\delta = \infty$. Using similar arguments as in the proof of Lemma 7.1.4 on the Jordan normal form of E, it follows easily that the limit of $(\mathbb{1} + \Delta t^2 E/\varepsilon^2)^{-1}$ exists as ε approaches zero, if and only if the algebraic and geometric multiplicities of the eigenvalue $\lambda = 0$ coincide. Moreover, this limit is the projection along $\mathcal{R}(E)$ onto $\mathcal{N}(E)$.*

Hence, conditions 1 and 2 of Theorems 7.1.2 and 7.1.3 are satisfied under the conditions of Corollary 6.8.2 or 6.8.4 . However, the third condition may not be satisfied. Indeed, the kernel of the corresponding matrix E may contain the so-called *checkerboard modes*, cf. Example 7.1.6; for the piecewise constant, linear or bilinear reconstruction, cf. Example 6.1.7; if the the number of cells in x- or y-direction is even. We show this fact for piecewise constant reconstruction and periodic boundary conditions in the following example.

Example 7.1.6 *Let us consider the elliptic approach IMEX Euler finite volume scheme (6.40) with piecewise constant reconstruction and constant bottom topography $b = const < 0$. Using the notation*

$$\mathcal{C}^i(a_0, a_1, ..., a_{2i+1}) = \mathcal{C}(a_0, a_1, ..., a_i, 0, ..., 0, a_{i+1}, ..., a_{2i+1}), \tag{6.110}$$

where $\mathcal{C}$ denotes the isomorphisms (6.75) or (6.89), we can write the matrix E in the following way

$$E = \frac{b}{4} \mathcal{C}^2 \left(\mathcal{C}^2 \left(\frac{-2}{\Delta x^2} + \frac{-2}{\Delta y^2}, 0, \frac{1}{\Delta x^2}, \frac{1}{\Delta x^2}, 0 \right), 0, \mathcal{C}^0 \left(\frac{1}{\Delta y^2} \right), \mathcal{C}^0 \left(\frac{1}{\Delta y^2} \right), 0 \right). \tag{7.25}$$

Thus, due to Theorems 6.7.3, 6.7.7 the eigenvalues of E read

$$\lambda_{kl} = b \left[\frac{-2 + \omega_N^{2l} + \bar{\omega}_N^{2l}}{4\Delta x^2} + \frac{-2 + \omega_M^{2k} + \bar{\omega}_M^{2k}}{4\Delta y^2} \right] = b \left[\frac{-1 + \cos\left(\frac{4\pi l}{N}\right)}{2\Delta x^2} + \frac{-1 + \cos\left(\frac{4\pi k}{M}\right)}{2\Delta y^2} \right], \tag{7.26}$$

where $\omega_N = e^{i2\pi/N}, \omega_M = e^{i2\pi/M}$ *and* $k = 0, ..., M-1, l = 0, ..., N-1$. *Hence,* $\lambda_{kl} = 0$, *if and only if* $k \in \{0, M/2\}$ *and* $l \in \{0, N/2\}$. *The corresponding eigenvectors* $\mathbf{r}_{kl}$ *read*

$$\mathbf{r}_{00} = \mathbf{1} \otimes \mathbf{1}, \quad \mathbf{r}_{0,N/2} = \mathbf{1} \otimes \tilde{\mathbf{1}}, \quad \mathbf{r}_{M/2,0} = \tilde{\mathbf{1}} \otimes \mathbf{1}, \quad \mathbf{r}_{M/2,N/2} = \tilde{\mathbf{1}} \otimes \tilde{\mathbf{1}}, \tag{7.27}$$

where $\tilde{\mathbf{1}} = (1, -1, 1, -1, 1, -1, ..., 1, -1)^T$. *Due to (7.5a) the leading order terms* $\mathbf{Z}^{n+1,(0)}$, $\mathbf{Z}^{n+1,(1)}$ *are in the kernel of* E. *The eigenvectors* $\mathbf{r}_{0,N/2}, \mathbf{r}_{M/2,0}, \mathbf{r}_{M/2,N/2}$ *- the checkerboard modes - may lead to the checkerboard effect for the free surface elevation, cf. [42]. If* E *is positive semi-definite, which is our case, we can remove the checkerboard modes by adding implicit diffusion to the continuity equation, i.e. we use the matrix* $\mathbb{1} + \left(\frac{\Delta t}{\varepsilon}\right)^2 [E - D]$, *where* $-D$ *is a positive semi-definite matrix with* $\mathcal{N} = \langle \mathbf{1} \rangle$. *Then* $\mathcal{N}(E - D) = \langle \mathbf{1} \rangle$. *Further* D *has to vanish as* $\Delta x, \Delta y \to 0$, *so that the scheme remains consistent. Here we can choose* D *in the following way*

$$D = \mathcal{C}^1 \left(\mathcal{C}^1 \left(-\frac{2}{\Delta x} - \frac{2}{\Delta y}, \frac{1}{\Delta x}, \frac{1}{\Delta x} \right), \mathcal{C}^0 \left(\frac{1}{\Delta y} \right), \mathcal{C}^0 \left(\frac{1}{\Delta y} \right) \right). \tag{7.28}$$

Then the scheme also remains conservative. However, we will not consider this approach any more throughout the thesis. Note that similar type of diffusion is automatically included in the continuity equation, when the EG numerical flux (6.29b) is used.

Though the kernel of the matrix E contain checkerboard modes, we can prove that $\mathbf{Z}^{n+1} = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2)$ in the setting of Example 7.1.6. To this end we calculate the inverse in multi-dimensional Fourier modes, cf. Section 6.7. Hence, we obtain the asymptotic preserving property following the proof of Theorem 7.1.1 or 7.1.2.

Lemma 7.1.7 *The IMEX Euler finite volume schemes*

- *(6.40) with the CFD numerical flux (6.22) for the stiff, implicit parts and the Lax-Friedrichs or Rusanov numerical flux for the non-stiff, explicit terms*
- *(6.56) with the approximations (6.59), (6.61) and the Lax-Friedrichs or Rusanov numerical flux for the non-stiff, explicit terms*

are asymptotic preserving, if the bottom topography is constant, the boundary conditions are periodic and the reconstruction is piecewise constant, linear or bilinear, cf. Example 6.1.7.

Proof: It suffices to prove that $\mathbf{Z}^{n+1} = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2)$ for the elliptic approach IMEX finite volume scheme, since it provides the same results as the straight approach, cf. Theorem 6.4.4. To this end we use the theory of cyclic block matrices, cf. Section 6.7. Note that the matrix E and the discrete divergence operator $\nabla_h \cdot$ can be described in the following way

$$E = b(\mathfrak{D}_x^2 + \mathfrak{D}_y^2), \tag{7.29a}$$

$$\nabla_h \cdot \hat{\mathbf{M}} = \mathfrak{D}_x \hat{\mathbf{M}}_1 + \mathfrak{D}_y \hat{\mathbf{M}}_2, \tag{7.29b}$$

where $b < 0$ is the constant bottom topography and $\mathfrak{D}_x, \mathfrak{D}_y$ are anti-symmetric matrices approximating the first order derivatives in the x- and y-direction, cf. Section 6.8.

Let us point out that

$$\left(\mathbb{1}+\left(\frac{\Delta t}{\varepsilon}\right)^2 E\right)^{-1}\mathfrak{D}_x,\left(\mathbb{1}+\left(\frac{\Delta t}{\varepsilon}\right)^2 E\right)^{-1}\mathfrak{D}_y=\mathcal{O}(\varepsilon^2). \tag{7.30}$$

To this end we consider the eigenvalues: Let $\lambda^x_{kl},\lambda^y_{kl}$ be eigenvalues of $\mathfrak{D}_x,\mathfrak{D}_y$, μ_{kl} eigenvalues of $\left(\mathbb{1}+\left(\frac{\Delta t}{\varepsilon}\right)^2 E\right)^{-1}$ and η^x_{kl} eigenvalues of $\left(\mathbb{1}+\left(\frac{\Delta t}{\varepsilon}\right)^2 E\right)^{-1}\mathfrak{D}_x$, $k=0,...,N-1$, $l=0,...,M-1$. Then it holds

$$\mu_{kl}=\frac{\varepsilon^2}{\varepsilon^2+\Delta t^2 b\left[\left(\lambda^x_{kl}\right)^2+\left(\lambda^y_{kl}\right)^2\right]}, \tag{7.31a}$$

$$\eta^x_{kl}=\frac{\varepsilon^2\lambda^x_{kl}}{\varepsilon^2+\Delta t^2 b\left[\left(\lambda^x_{kl}\right)^2+\left(\lambda^y_{kl}\right)^2\right]}=\mathcal{O}(\varepsilon^2), \tag{7.31b}$$

$$\left[\left(\lambda^x_{kl}\right)^2+\left(\lambda^y_{kl}\right)^2\right]=\left[\frac{-1+\cos\left(\frac{4\pi l}{N}\right)}{2\Delta x^2}+\frac{-1+\cos\left(\frac{4\pi k}{M}\right)}{2\Delta y^2}\right], \tag{7.31c}$$

and analogously for the y-direction.

Since the matrix E is symmetric and positive semi-definite, cf. (7.29a), (7.31c), we can apply Lemma 7.1.4 and (7.30). Further the diffusive terms from the explicit update are $\mathcal{O}(\varepsilon^2)$ due to well-prepared initial data. Thus we obtain

$$\mathbf{Z}^{n+1}=\pi_{\mathcal{N}(E)}(\hat{\mathbf{Z}})+\mathcal{O}(\varepsilon^2)=\pi_{\mathcal{N}(E)}(\mathbf{Z}^n)+\mathcal{O}(\varepsilon^2)=\mathbf{Z}^n+\mathcal{O}(\varepsilon^2). \tag{7.32}$$

□

Remark 7.1.8 *The proof of Lemma 7.1.7 is based on showing $\mathcal{R}(\nabla_h\cdot)\perp\mathcal{N}(E)$. If the matrix E is symmetric, the multiplication of $\hat{\mathbf{Z}}+\nabla_h\cdot\hat{\mathbf{M}}$ with the matrix $(\mathbb{1}+\Delta t^2E/\varepsilon^2)^{-1}$ from the left is the orthogonal projection of $\hat{\mathbf{Z}}$ onto the kernel of E up to $\mathcal{O}(\varepsilon^2)$ terms, cf. Lemma 7.1.4. This yields $\mathbf{Z}^{n+1}=\mathbf{Z}^n+\mathcal{O}(\varepsilon^2)$.*

In the case of a matrix E, which satisfies the conditions of Remark 7.1.5, the matrix $(\mathbb{1}+\Delta t^2E/\varepsilon^2)^{-1}$ is the projection onto $\mathcal{N}(E)$ along $\mathcal{R}(E)$. Hence, we have $\mathbf{Z}^{n+1}=\mathbf{Z}^n+\mathcal{O}(\varepsilon^2)$, if $\mathcal{R}(\nabla_h\cdot)=\mathcal{R}(E)$. Thus, in the case of a non-constant bottom topography an elliptic approach IMEX finite volume scheme, which conserves the free surface elevation, would be asymptotic preserving, if the following properties would be satisfied:

1. *$E=\nabla_h\cdot\overline{B\nabla}_h$*
2. *$\langle\mathbf{1}\rangle\subset\mathcal{N}(E),\mathcal{N}(\overline{B\nabla}_h)$*
3. *the blocks corresponding to the eigenvalue $\lambda=0$ of the Jordan normal form of E are diagonal*
4. *$\mathcal{R}(\nabla_h\cdot)=\mathcal{R}(E)$.*

The properties 1,2 depend on the chosen scheme and are easy to control. However, we are not able to decide, if the properties 3,4 are satisfied (in the case of a non-constant bottom topography).

Lemma 7.1.9 *If we use the elliptic approach IMEX Euler finite volume scheme (6.56), (6.57) with a constant bottom topography and periodic boundary conditions, then* $\mathcal{N}(E) = \langle\ \mathbf{1}\ \rangle$. *Thus* $\mathbf{Z}^{n+1} = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2)$. *However, we can not prove a divergence constraint.*

Proof: Using the notation (6.110) we can write the matrix E in the following way

$$E = b\mathcal{C}^1\left(\mathcal{C}^1\left(-\frac{2}{\Delta x^2} - \frac{2}{\Delta y^2}, \frac{1}{\Delta x^2}, \frac{1}{\Delta x^2}\right), \mathcal{C}^0\left(\frac{1}{\Delta y^2}\right), \mathcal{C}^0\left(\frac{1}{\Delta y^2}\right)\right). \tag{7.33}$$

Due to Theorems 6.7.3 and 6.7.7 the eigenvalues λ_{kl} of the matrix E read

$$\lambda_{kl} = 2b\left[\frac{-1 + \cos\left(\frac{2\pi l}{N}\right)}{\Delta x^2} + \frac{-1 + \cos\left(\frac{2\pi k}{M}\right)}{\Delta y^2}\right]. \tag{7.34}$$

Thus $\lambda_{kl} = 0$, if and only if $k = l = 0$. Further, the corresponding eigenvector to λ_{00} is $\mathbf{1}$. □

Remark 7.1.10 *Let us consider one of the IMEX finite volume schemes from Lemma 7.1.7, where the initial data* $\mathbf{Z}^n, \mathbf{M}^n$ *are perturbed by an error* $\mathcal{O}(\delta)$. *Then we obtain*

$$\mathbf{Z}^{n+1} = \pi_{\mathcal{N}(E)}(\hat{\mathbf{Z}}) + \mathcal{O}(\varepsilon^2) = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\delta), \tag{7.35a}$$

$$\mathfrak{D}_x\mathbf{M}_1^{n+1} + \mathfrak{D}_y\mathbf{M}_2^{n+1} = \mathcal{O}(\delta) + \mathcal{O}(\varepsilon^2). \tag{7.35b}$$

Here $\mathbf{Z}^{n+1}$ *is in* $\mathcal{N}(E)$ *up to* $\mathcal{O}(\varepsilon^2)$*-terms. If* $\mathcal{N}(E) = \langle\mathbf{1}\rangle$, *the free surface elevation will become constant in the low Froude number limit after one time step. But if the kernel of* E *contains checkerboard modes, we should observe the checkerboard effect in the low Froude number limit of the free surface elevation. Thus it seems more robust to use schemes, where the kernel of* E *is* $\langle\mathbf{1}\rangle$.

Since $\mathbf{Z}^{n+1}$ *is in* $\mathcal{N}(E)$ *up to* $\mathcal{O}(\varepsilon^2)$*-terms,*

$$\mathbf{Z}^{n+k} = \mathbf{Z}^{n+1} + \mathcal{O}(\varepsilon^2), \quad \mathfrak{D}_x\mathbf{M}_1^{n+k} + \mathfrak{D}_y\mathbf{M}_2^{n+k} = \mathcal{O}(\varepsilon^2), \quad k \geq 2. \tag{7.36}$$

The statement for the free surface elevation in (7.36) is also true for the IMEX Euler finite volume scheme from Lemma 7.1.9.

7.1.2. EG numerical flux

The aim of this section is to study the asymptotic preserving property, cf. Definition 5.1.1, of the IMEX Euler finite volume scheme (7.2), where the EG numerical flux (6.29b) is used the for evaluation of the cell interface fluxes. Further we assume constant bottom topography. Following the procedure in Section 7.1.1 we first present conditions that guarantee the asymptotic preserving property. Then we check when these are satisfied.

Theorem 7.1.11 *Consider a straight approach IMEX finite volume scheme (7.2), where the stiff, linear parts are approximated with the EG numerical flux (6.29b) and the corresponding source term approximation (6.49). The non-stiff parts are approximated using the Lax-Friedrichs or Rusanov fluxes. Further let the matrix* A *consist of the matrices* $C_1, ..., C_9$, *cf. (6.109) and assume:*

1. $\mathbb{1} + \Delta t A$ *is non-singular for all* $\varepsilon > 0$

2. *The limit*

$$\lim_{\varepsilon \to 0} (\mathbb{1} + \Delta t A)^{-1} \tag{7.37}$$

exists.

3. *The kernel intersection of* C_1, C_4, C_7 *is*

$$\mathcal{N}(C_1) \cap \mathcal{N}(C_4) \cap \mathcal{N}(C_7) = \langle \mathbf{1} \rangle. \tag{7.38}$$

Then

$$\lim_{\varepsilon \to 0} \mathbf{Z}^{n+1} = \mathbf{Z}^{n,(0)}. \tag{7.39}$$

Moreover it holds

$$C_2 \mathbf{M}_1^{n+1} + C_3 \mathbf{M}_3^{n+1} = \mathcal{O}(\varepsilon), \tag{7.40a}$$
$$\mathbf{Z}^{n+1} = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2), \tag{7.40b}$$

if $\mathbf{Z}^{n+1} = \alpha \mathbf{1} + \mathcal{O}(\varepsilon^2)$ *where* α *may depend on* ε*. Here (7.40a) is a discrete divergence constraint on the momentum in the low Froude number limit. Hence the scheme is asymptotic preserving.*

Proof: Since the limit (7.37) exists, the sequence of solutions $\mathbf{W}^{n+1} = \mathbf{W}_\varepsilon^{n+1}$, $\varepsilon > 0$, has a limit

$$\mathbf{W}^{n+1,(0)} := \lim_{\varepsilon \to 0} (\mathbb{1} + \Delta t A)^{-1} \hat{\mathbf{W}}. \tag{7.41}$$

Obviously it holds $C_1 \mathbf{Z}^{n+1,(0)} = C_4 \mathbf{Z}^{n+1,(0)} = C_7 \mathbf{Z}^{n+1,(0)} = 0$. Thus the limit free surface elevation is constant due to (7.38). Since $\mathbf{Z}^{n,(0)} = Z\mathbf{1}$ and the scheme is conservative, we have

$$\sum_{i,j} z_{ij}^{n,(0)} = \sum_{i,j} z_{ij}^{n+1,(0)}. \tag{7.42}$$

Thus $\mathbf{Z}^{n+1,(0)} = \mathbf{Z}^{n,(0)} = Z\mathbf{1}$.

Since we use the Lax-Friedrichs or Rusanov numerical flux $\hat{\mathbf{Z}} = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2)$. Thus, if additionally $\mathbf{Z}^{n+1} = \alpha \mathbf{1} + \mathcal{O}(\varepsilon^2)$, then

$$\begin{aligned} &(\mathbb{1} + \Delta t c_b C_1)\, \mathbf{Z}^{n+1} + \Delta t (C_2 \mathbf{M}_1^{n+1} + C_3 \mathbf{M}_3^{n+1}) \\ &= \mathbf{Z}^{n+1} + \Delta t (C_2 \mathbf{M}_1^{n+1} + C_3 \mathbf{M}_3^{n+1}) + \mathcal{O}(\varepsilon) = \mathbf{Z}^n + \Delta t (C_2 \mathbf{M}_1^{n+1} + C_3 \mathbf{M}_3^{n+1}) + \mathcal{O}(\varepsilon) \\ &= \hat{\mathbf{Z}} = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2), \end{aligned} \tag{7.43}$$

since $c_b = \mathcal{O}(1/\varepsilon)$, cf. (6.109). Consequently $C_2 \mathbf{M}_1^{n+1} + C_3 \mathbf{M}_2^{n+1} = \mathcal{O}(\varepsilon)$. □

Let us now discuss the assumptions of Theorem 7.1.11. We have proved in Section 6.8.2 that the matrix $\mathbb{1} + \Delta t A$ is non-singular under certain conditions. In the following lemma we show that these conditions imply the assumptions from Theorem 7.1.11. If these conditions are not satisfied, we do not know the structure of the matrix A and therefore we are not able to verify the validity of the assumptions.

Lemma 7.1.12 *If the piecewise constant reconstruction (6.36), periodic boundary conditions, constant bottom topography are used and the number of cells in x- and y-direction is equal, i.e. $M = N$, then the assumptions of Theorem 7.1.11 are satisfied. Thus the corresponding scheme is asymptotic preserving.*

Proof:

1. The non-singularity of $(\mathbb{1} + \Delta t A)$ has been already proven in Section 6.8.2.

2. The matrices $C_1, ..., C_9$ commute, since they are circulant block matrices, cf. Section 6.7. Thus we can use a block variant of Cramer's rule, i.e.

$$(\mathbb{1} + \Delta t A)^{-1} = \mathbb{1}_3 \otimes cdet(\mathbb{1} + \Delta t A)^{-1}(\mathbb{1} + \Delta t A)^{cr}, \tag{7.44a}$$

$$(\mathbb{1} + \Delta t A)^{cr} = \begin{bmatrix} (\mathbb{1} + \Delta t A)_1^{cr} & (\mathbb{1} + \Delta t A)_2^{cr} & (\mathbb{1} + \Delta t A)_3^{cr} \\ (\mathbb{1} + \Delta t A)_4^{cr} & (\mathbb{1} + \Delta t A)_5^{cr} & (\mathbb{1} + \Delta t A)_6^{cr} \\ (\mathbb{1} + \Delta t A)_7^{cr} & (\mathbb{1} + \Delta t A)_8^{cr} & (\mathbb{1} + \Delta t A)_9^{cr} \end{bmatrix}, \tag{7.44b}$$

$$(\mathbb{1} + \Delta t A)_1^{cr} = (\mathbb{1} + \Delta t c_b C_5)(\mathbb{1} + \Delta t c_b C_9) - (\Delta t c_b)^2 C_6 C_8, \tag{7.44c}$$
$$(\mathbb{1} + \Delta t A)_2^{cr} = \Delta t(\Delta t c_b C_3 C_8 - C_2(\mathbb{1} + \Delta t c_b C_9)) \tag{7.44d}$$
$$(\mathbb{1} + \Delta t A)_3^{cr} = \Delta t(\Delta t c_b C_2 C_6 - C_3(\mathbb{1} + \Delta t c_b C_5)) \tag{7.44e}$$
$$(\mathbb{1} + \Delta t A)_4^{cr} = \Delta t c_b^2(\Delta t c_b C_6 C_7 - C_4(\mathbb{1} + \Delta t c_b C_9)) \tag{7.44f}$$
$$(\mathbb{1} + \Delta t A)_5^{cr} = (\mathbb{1} + \Delta t c_b C_1)(\mathbb{1} + \Delta t c_b C_9) - (\Delta t c_b)^2 C_3 C_7 \tag{7.44g}$$
$$(\mathbb{1} + \Delta t A)_6^{cr} = \Delta t c_b(\Delta t c_b C_3 C_4 - (\mathbb{1} + \Delta t c_b C_1) C_6) \tag{7.44h}$$
$$(\mathbb{1} + \Delta t A)_7^{cr} = \Delta t c_b^2(\Delta t c_b C_4 C_8 - (\mathbb{1} + \Delta t c_b C_5) C_7) \tag{7.44i}$$
$$(\mathbb{1} + \Delta t A)_8^{cr} = \Delta t c_b(\Delta t c_b C_2 C_7 - (\mathbb{1} + \Delta t c_b C_1) C_8) \tag{7.44j}$$
$$(\mathbb{1} + \Delta t A)_9^{cr} = (\mathbb{1} + \Delta t c_b C_1)(\mathbb{1} + \Delta t c_b C_5) - (\Delta t c_b)^2 C_2 C_4, \tag{7.44k}$$

where the matrix $cdet(A)$ is given in (6.115). Moreover we can diagonalise each of the nine blocks $(\mathbb{1} + \Delta t A)_i^{cr}$, $i = 1, ..., 9$, and $cdet(\mathbb{1} + \Delta t A)$ by a change of basis, cf. Section 6.7. Then the low Froude number limit (7.37) exists, if and only if the low Froude number limit of the diagonal entries exist. Let λ_{kl} be an eigenvalue of $cdet(\mathbb{1} + \Delta t A)$, cf. (6.116). Then λ_{kl}^{-1} is the corresponding eigenvalue of $cdet(\mathbb{1} + \Delta t A)^{-1}$. Due to Remark 6.8.8, we have

$$\lambda_{kl}^{-1} = \begin{cases} \mathcal{O}(\varepsilon^3) & , k, l \neq 0 \\ \mathcal{O}(\varepsilon^2) & , \text{ either } k = 0 \text{ or } l = 0 \\ 1 & , k = l = 0. \end{cases}. \tag{7.45}$$

An eigenvalue μ^i of the matrices $(\mathbb{1} + \Delta t A)_i^{cr}$, $i = 1, ..., 9$, satisfies

$$\mu^i = \begin{cases} \mathcal{O}(\varepsilon^{-2}) & , \text{either } k \neq 0 \text{ or } l \neq 0 \\ \mathcal{O}(1) & , k = l = 0. \end{cases}, \tag{7.46}$$

due to (6.114). Consequently the limit (7.37) exists.

3. For a matrix C_i the eigenvalue λ_{00}^i with the eigenvector $\mathbf{1}$ equals zero, $i = 1, ..., 9$. Moreover λ_{00}^1 is the only zero eigenvalue of C_1, cf. (6.114).
Thus $\mathcal{N}(C_1) = \mathcal{N}(C_1) \cap \mathcal{N}(C_2) \cap \mathcal{N}(C_3) = \langle \mathbf{1} \rangle$.

4. The eigenvalues μ_{kl} of $(\mathbb{1} + \Delta t A)_i^{cr}$, $i = 2, 3$, satisfy

$$\mu_{kl} = \begin{cases} 0 & , k = l = 0 \\ \mathcal{O}(1) & , \text{either } k = 0 \text{ or } l = 0 \\ \mathcal{O}(\varepsilon^{-1}) & , \text{else} \end{cases} \tag{7.47}$$

due to Theorems (6.7.3), (6.7.7). Thus, using (7.45), (7.46) we have

$$\begin{aligned} \mathbf{Z}^{n+1} &= cdet(\mathbb{1} + \Delta t A)^{-1} \left[(\mathbb{1} + \Delta t A)_1^{cr} \hat{\mathbf{Z}} + \sum_{i=1}^{2} (\mathbb{1} + \Delta t A)_i^{cr} \hat{\mathbf{M}}_i \right] \\ &= cdet(\mathbb{1} + \Delta t A)^{-1} (\mathbb{1} + \Delta t A)_1^{cr} \hat{\mathbf{Z}} + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{7.48}$$

From (6.114) we obtain that $\mathbf{1}$ is an eigenvector of $(\mathbb{1} + \Delta t A)^{cr}$ corresponding to the eigenvalue $\mu_{00} = 1$. Similarly, $\mathbf{1}$ is an eigenvector of $cdet(\mathbb{1} + \Delta t A)^{-1}$ corresponding to the eigenvalue $\lambda_{00} = 1$. Since $\mathbf{Z}^n$ is well-prepared we have

$$\begin{aligned} \mathbf{Z}^{n+1} &= cdet(\mathbb{1} + \Delta t A)^{-1} (\mathbb{1} + \Delta t A)_1^{cr} \hat{\mathbf{Z}} \\ &= cdet(\mathbb{1} + \Delta t A)^{-1} (\mathbb{1} + \Delta t A)_1^{cr} (\mathbf{Z}^{n,(0)} + \varepsilon \mathbf{Z}^{n,(1)}) + \mathcal{O}(\varepsilon^2) \\ &= \mathbf{Z}^n + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{7.49}$$

□

As in the previous section, we consider a perturbation of the well-prepared initial data in the following remark.

Remark 7.1.13 *Let us consider the IMEX finite volume scheme from Theorem 7.1.11, where the initial data $\mathbf{Z}^n, \mathbf{M}^n$ are perturbed by an error $err = \mathcal{O}(\delta)$. Then we obtain*

$$\mathbf{Z}^{n+1} = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\delta), \tag{7.50a}$$

$$C_2 \mathbf{M}_1^{n+1} + C_3 \mathbf{M}_2^{n+1} = \mathcal{O}(\delta) + \mathcal{O}(\varepsilon). \tag{7.50b}$$

Here $\mathbf{Z}^{n+1}$ is in $\mathcal{N}(C_1) = \langle \mathbf{1} \rangle$ up to $\mathcal{O}(\varepsilon^2)$-terms. Thus

$$\mathbf{Z}^{n+k} = \mathbf{Z}^{n+1} + \mathcal{O}(\varepsilon^2), \quad C_2 \mathbf{M}_1^{n+k} + C_3 \mathbf{M}_2^{n+k} = \mathcal{O}(\varepsilon), \quad k \geq 2. \tag{7.51a}$$

7.2. IMEX Runge-Kutta and IMEX multi-step time discretisations

In the previous section we have studied the asymptotic preserving property of the IMEX Euler finite volume schemes. Moreover we have provided conditions on the spatial discretisation that guarantee the asymptotic preserving property. We show in this section, that basically the same conditions guarantee the asymptotic preserving property for general IMEX R-K and IMEX multi-step schemes.

Lemma 7.2.1 *Every lemma, corollary and theorem from the Section 7.1 can be extended to be valid for globally stiffly accurate type A and ARS IMEX R-K and consistent IMEX multi-step time discretisations with $\tilde{\beta}_0 = 0$ instead of the IMEX Euler one, cf. Chapter 4.*

Proof: Every statement in the previous section has been proven by using the structure of the linear system and the fact that that the explicitly updated free surface elevation satisfies $\hat{\mathbf{Z}} = (Z + \varepsilon Z')\mathbf{1} + \mathcal{O}(\varepsilon^2)$. Note that if an IMEX R-K time discretisation is used, we solve several linear systems with slight different right-hand side. If an IMEX multi-step discretisation is used, we solve one linear system with a different right-hand side. However the left-hand side of the linear system is the same as in the IMEX Euler case besides the time-increment Δt. For the IMEX R-K and IMEX multi-step time discretisation $\hat{\mathbf{Z}} = (Z + \varepsilon Z')\mathbf{1} + \mathcal{O}(\varepsilon^2) + \nabla_h \cdot \tilde{\mathbf{M}}$, where $\tilde{\mathbf{M}}$ is a linear combination of momentum at already calculated time steps or internal R-K stages. In Section 7.1 we pointed out, that $(\mathbb{1} + \Delta t^2/\varepsilon^2 A)^{-1}\nabla_h \cdot = \mathcal{O}(\varepsilon^2)$ for the considered schemes. Thus, we have $\mathbf{Z}^{n+1} = \mathbf{Z}^n + \mathcal{O}(\varepsilon^2)$. Further, we get a divergence constraint due to the restrictions on time discretisation. Consequently we can just repeat the proofs, where an induction over the R-K stages is necessary for the IMEX R-K time discretisation. □

Remark 7.2.2 *Typically the momentum initial data do not satisfy a discrete divergence constraint, i.e. $\nabla_h \cdot \mathbf{m} \neq \mathcal{O}(\varepsilon)$. For some high-order time discretisations, e.g. RK2CN scheme, cf. Section 4.3.2, the discrete divergence of the initial momentum is used and thus additional numerical error may be added. In particular,*

$$\mathbf{Z}^{n+1} + \frac{\Delta t}{2}\nabla_h \cdot \mathbf{M}^{n+1} = \mathbf{Z}^n - \frac{\Delta t}{2}\nabla_h \cdot \mathbf{M}^n + \mathcal{O}(\varepsilon^2) \tag{7.52}$$

holds for the RK2CN time discretisation, cf. (4.23b), and we have $\nabla_h \cdot (\mathbf{M}^{n+1} - \mathbf{M}^n) = \mathcal{O}(\varepsilon^2)$, since $\mathbf{Z}^{n+1} - \mathbf{Z}^n = \mathcal{O}(\varepsilon^2)$. Thus the divergence constraint of the momentum is satisfied at a new time step t^{n+1} if and only if it is satisfied at an old time step. Consequently, we had to restrict the time discretisations in Lemma 7.2.1. However, if the discrete momentum initial data satisfies the discrete divergence constraint, Lemma 7.2.1 is valid for any globally stiffly accurate IMEX R-K and consistent IMEX multi-step time discretisations.

8. Numerical experiments

In Chapter 4 we have introduced the IMEX R-K and IMEX multi-step time discretisations. For the semi-discrete schemes we have proved that the schemes are well-balanced and asymptotic preserving under certain conditions, cf. Chapters 4, 5. Particularly these results hold for the IMEX Euler, RK2CN, ARS222 and SBDF time discretisations. In Chapter 6 we have introduced the combination of the IMEX time discretisations with the finite volume space discretisation. Also for these fully discrete schemes we have showed that the well-balanced and asymptotic preserving property are satisfied under certain assumptions, cf. Chapters 6, 7. The aim of this chapter is to study the behaviour of the previously introduced first and second order IMEX finite volume schemes on a series of numerical experiments. Particular questions we are interested in are: accuracy, stability and asymptotic behaviour as the Froude number ε approaches zero.

First, we study the shock capturing properties of IMEX finite volume scheme in Section 8.1. In Section 8.2 we study the accuracy of IMEX schemes by comparing numerical and exact solutions. Further we measure convergence rates of the IMEX schemes for various Froude numbers in the range from 10^{-8} to 0.8. Then we study the asymptotic preserving property considering the algebraic constraints of free surface elevation and the divergence of momentum for low Froude numbers. In the end we perform long time simulations to verify the stability of IMEX finite volume schemes. In Section 8.3 we examine the well-balanced property. To this end, we consider time evolution of a lake at rest state with a smooth and a discontinuous bottom topography.

In order to refer easily to the various IMEX finite volume schemes, we use the abbreviations in Table 8.1. Here, CFD, EG and Rusanov refer to the numerical fluxes (6.22), (6.29b) and (6.24). The piecewise constant and linear reconstructions are defined in Example 6.1.7. The two elliptic approaches ELL (6.56),(6.57) and ELLW (6.56), (6.59), (6.61) are explained in Sections 6.4, 6.5. Further, we compare IMEX and explicit schemes. Here we use first and second order HLLC schemes, cf. (6.25). The explicit finite volume first order scheme is obtained using the explicit Euler time discretisation combined with the HLLC flux. We abbreviate this scheme with HLLC. For the second order scheme we use the Runge-Kutta scheme in time and the MUSCL approach with the linear reconstruction, cf. Example 6.1.7. The second order scheme is referred to as RK2HLLC.

8.1. 1D Riemann problems

For 1D Riemann problems the solution of the SWE (2.30) is known: if an initial jump is small enough, than the 1D Riemann problem solution consists of shocks, contact discontinuities and rarefaction waves, see [47]. Therefore they are typical benchmarks for numerical schemes to verify their accuracy and stability for Riemann-type initial data.

time discretisation	**numerical flux for stiff parts**	**numerical flux for non-stiff parts**	**reconstruction**	**abbreviation**
IMEX Euler	CFD	Rusanov	pcw. const.	CFDRUS
IMEX Euler	EG	Rusanov	pcw. const.	EGRUS
ARS(2,2,2)	CFD	Rusanov	linear	ARSCFDRUS
ARS(2,2,2)	EG	Rusanov	linear	ARSEGRUS
RK2CN	CFD	Rusanov	linear	RK2CNCFDRUS
RK2CN	EG	Rusanov	linear	RK2CNEGRUS
SBDF	CFD	Rusanov	linear	BDFCFDRUS
SBDF	EG	Rusanov	linear	BDFEGRUS
time discretisation	**elliptic approach**	**numerical flux used for non-stiff parts**	**reconstruction**	**abbreviation**
IMEX Euler	ELLW	Rusanov	pcw. const.	RUSELLW
IMEX Euler	ELL	Rusanov	pcw. const.	RUSELL
ARS(2,2,2)	ELLW	Rusanov	linear	ARSRUSELLW
ARS(2,2,2)	ELL	Rusanov	linear	ARSRUSELL
RK2CN	ELLW	Rusanov	linear	RK2CNRUSELLW
RK2CN	ELL	Rusanov	linear	RK2CNRUSELL
SBDF	ELLW	Rusanov	linear	BDFRUSELLW
SBDF	ELL	Rusanov	linear	BDFRUSELL

Table 8.1.: Abbreviations for IMEX finite volume schemes.

Particularly, observation of non-oscillating behaviour of numerical solution in the vicinity of discontinuities is very important. Therefore, we adopt the test [38, Example 6.1] for the isentropic Euler equations with the equation of state $p(\rho) = \rho^2$ by replacing density with water depth. Let us recall that up to source terms the SWE (2.18) are identical to the isentropic Euler equations, if the equation of state is $p(\rho) = \rho^2/2$. Therefore a direct comparison with the results obtained in [38] is not possible. Nevertheless qualitative comparison is reasonable.

The initial data reads

$$h(x,0) = 1, \qquad m(x,0) = 1 - \frac{\varepsilon^2}{2}, \quad x \in [0, 0.2] \cap (0.8, 1], \tag{8.1a}$$

$$h(x,0) = 1 + \varepsilon^2, \qquad m(x,0) = 1, \quad x \in (0.2, 0.3], \tag{8.1b}$$

$$h(x,0) = 1, \qquad m(x,0) = 1 + \frac{\varepsilon^2}{2}, \quad x \in (0.3, 0.7], \tag{8.1c}$$

$$h(x,0) = 1 - \varepsilon^2, \qquad m(x,0) = 1, \quad x \in (0.7, 0.8], \tag{8.1d}$$

the boundary conditions are taken to be periodic. The solution of this example consists of shocks, contact discontinuities and rarefaction waves, cf. [38]. For $\varepsilon = 0.8$ we can see the initial data and its discontinuities in Figure 8.1. Figures 8.2-8.7 present numerical solutions of various IMEX finite volume schemes at time $T = 0.05$ and Froude number $\varepsilon \in \{0.1, 0.8\}$. The solutions were computed on a 200 cells mesh, except for the reference solution that was obtained on a 2000 cells mesh by the RK2HLLC scheme. In Figure 8.2 the results of the first order schemes EGRUS, RUSELLW, RUSELL are presented.

For Froude number $\varepsilon = 0.8$ the numerical solutions of the first order schemes seems fine, whereas for $\varepsilon = 0.1$ the solutions are damped. The results obtained by the RUSELLW and RUSELL schemes can not be distinguished and are a little closer to the reference solution, than the solutions obtained by the EGRUS scheme. The results of the second order schemes ARSEGRUS, ARSRUSELLW, RK2CNEGRUS, RK2CNRUSELLW, BDFEGRUS, BDFRUSELLW, BDFRUSELL are shown in Figures 8.3-8.7. We left out the numerical solutions obtained by the ARSRUSELL and RK2CNRUSELL schemes due to their strong oscillatory behaviour. Since no slope limiting was applied, the numerical solutions of the second order schemes show under- and overshoots on discontinuities. In order to prevent under- and overshoots we can use the minmod reconstruction (6.38) for the explicit, non-stiff terms. Then the fluctuations in the numerical solution disappear for the Froude number $\varepsilon = 0.8$, cf. Figure 8.6. However, for the Froude number $\varepsilon = 0.1$ we needed in addition implicit diffusion to the elliptic equation of the free surface elevation to reduce large under- and overshoots, cf. Figure 8.7. More precisely, we added a discrete version of $(\Delta x^2 z_{xx} + \Delta y^2 z_{yy})/4$.

Note that the limiting of the implicitly treated fluxes is not analogously to the limiting of explicit flux functions, since the fast waves associated with the implicit fluxes can travel across more than one cell. Thus the location of a discontinuity in a next time step t^{n+1} is unknown.

Though addition of arbitrary diffusion is not a satisfying method to limit the slopes, we will not deepen on other solutions to obtain oscillation-free profiles in presence of discontinuities. To the best knowledge of the author it remains an open problem how to limit efficiently the slopes in the evaluation of implicit parts in high order schemes.

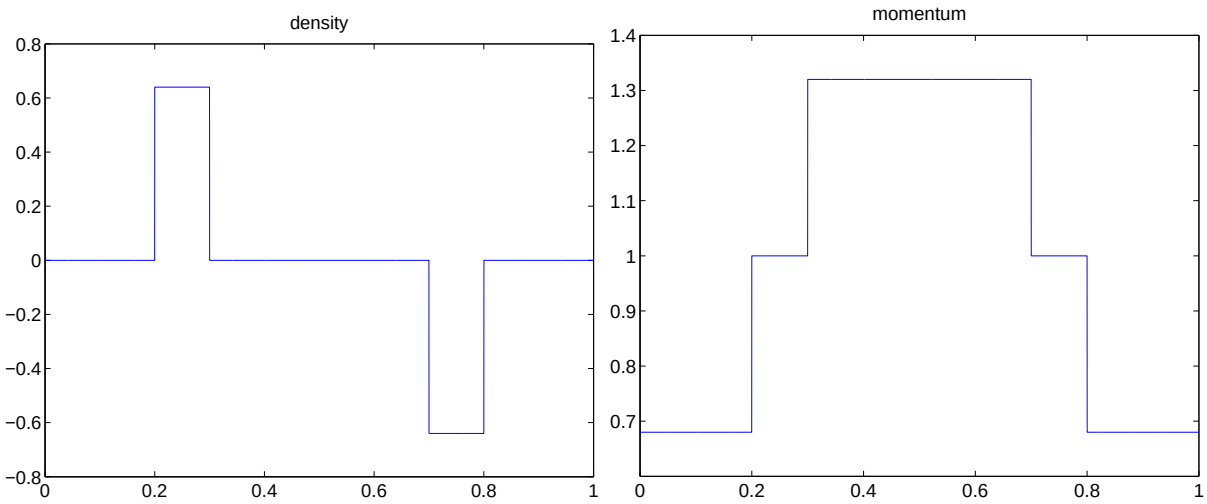

Figure 8.1.: Initial conditions of the 1D Riemann problems test with Froude number $\varepsilon = 0.8$. Left: free surface elevation. Right: momentum.

Remark 8.1.1 *In [27, 38, 51, 108] splittings of the isentropic Euler and Navier-Stokes equations are also splitted into a subsystem governing slow waves and another one governing fast waves, where a parameter controls the magnitude of stiff terms treated implicitly and explicitly. For Froude number $\varepsilon = \mathcal{O}(1)$ the numerical results in [51] provide oscillation-free profiles by applying the minmod limiter (6.38) to the non-stiff part. In [108] the numerical results of Riemann problems show oscillations.*

Remark 8.1.2 *In order to limit the stiff terms we are going to apply the minmod limiter. Consequently, the linear systems in the finite volume updates will change to nonlinear systems. For example, let A, $\tilde{A}$ be the operators from (6.97), where the linear reconstruction is used by A and the minmod reconstruction by $\tilde{A}$. Thus A is a linear mapping and $\tilde{A}$ a nonlinear one. We approximate the solution of the nonlinear finite volume update*

$$\left(\mathbb{1} + \delta\tilde{A}\right)\mathbf{W} = \hat{\mathbf{W}} \tag{8.2}$$

by applying the defect correction method, cf. [106], i.e. we use the iteration

$$(\mathbb{1} + \delta A)\,\mathbf{W}^{n+1,0} = \hat{\mathbf{W}}, \tag{8.3}$$

$$(\mathbb{1} + \delta A)\,\mathbf{W}^{n+1,i+1} = \hat{\mathbf{W}} + \delta(A - \tilde{A})\mathbf{W}^{n+1,i}, \quad i \in \mathbb{N}. \tag{8.4}$$

If a few iterations are enough to suppress the oscillations, the IMEX scheme should remain efficient.

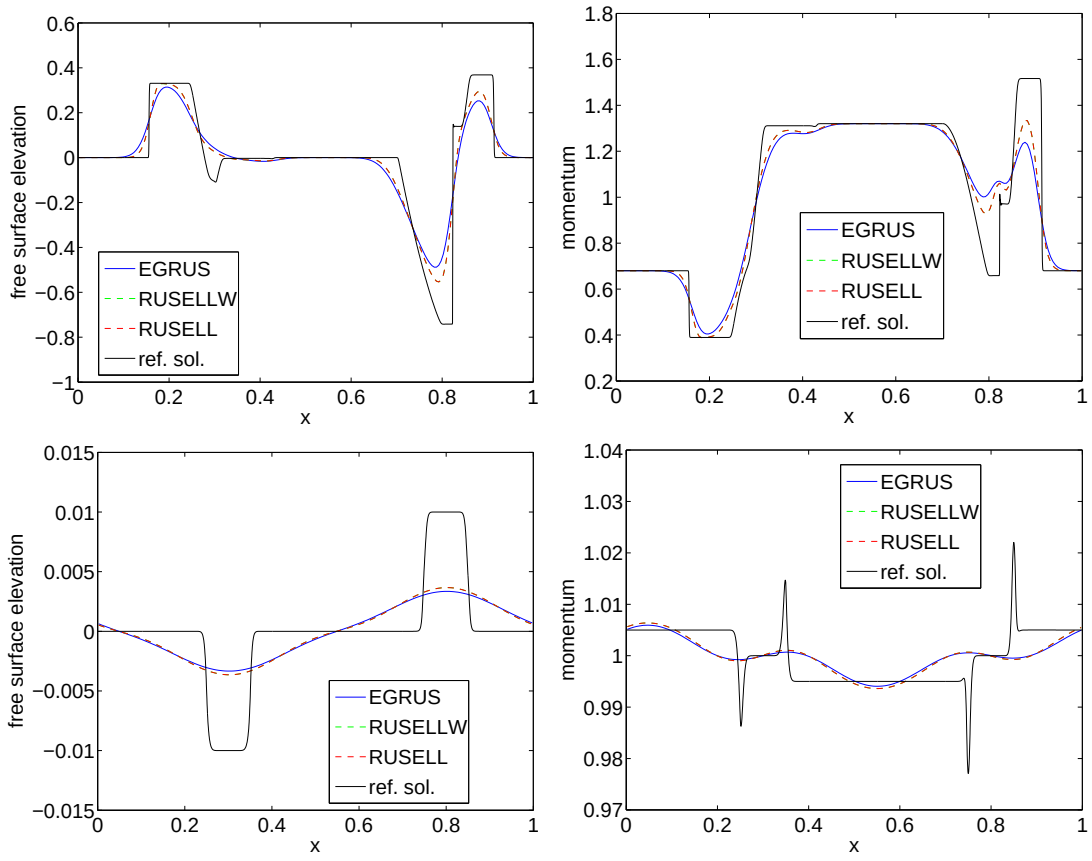

Figure 8.2.: Numerical solutions of the Riemann problems test at time $T = 0.05$ obtained by the first order EGRUS, RUSELLW, RUSELL schemes. $CFL_u = 0.45, \Delta x = 0.05$. Top: Froude number $\varepsilon = 0.8$. Bottom: Froude number $\varepsilon = 0.1$.

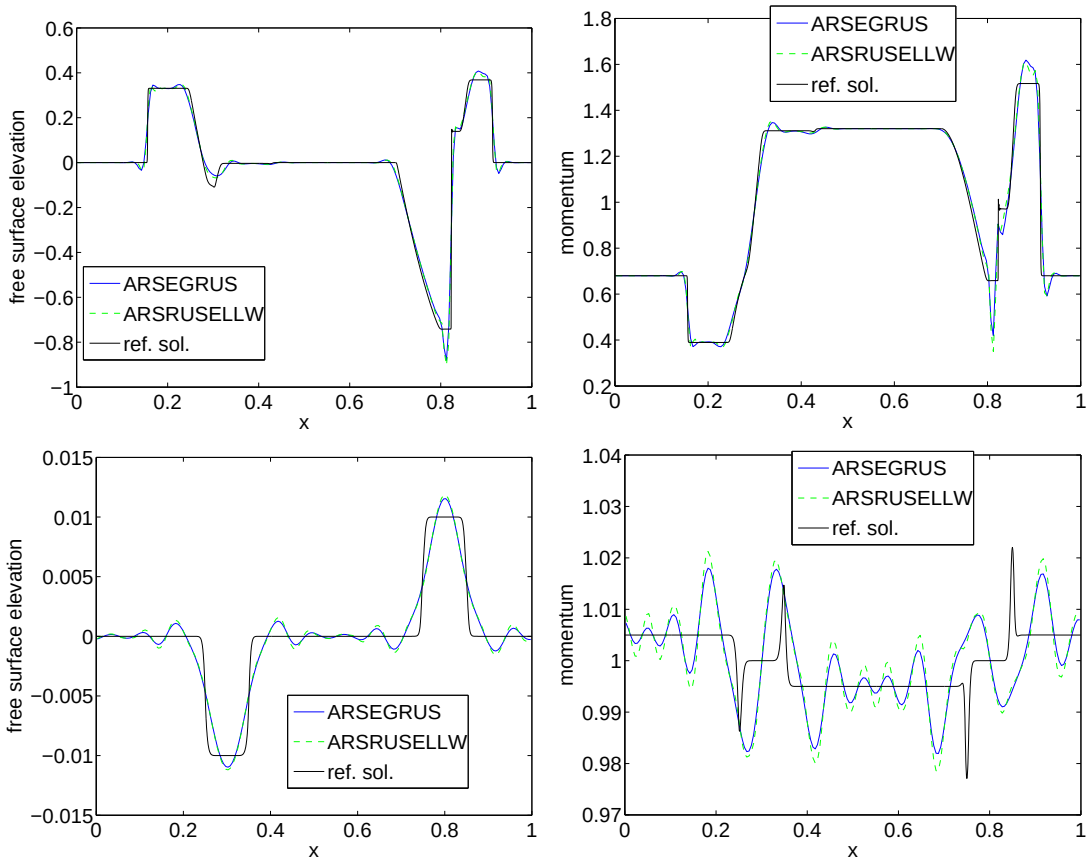

Figure 8.3.: Numerical solutions of the Riemann problems test at time $T = 0.05$ obtained by the second order ARSEGRUS, ARSRUSELLW schemes. $CFL_u = 0.45, \Delta x = 0.05$. Top: Froude number $\varepsilon = 0.8$. Bottom: Froude number $\varepsilon = 0.1$.

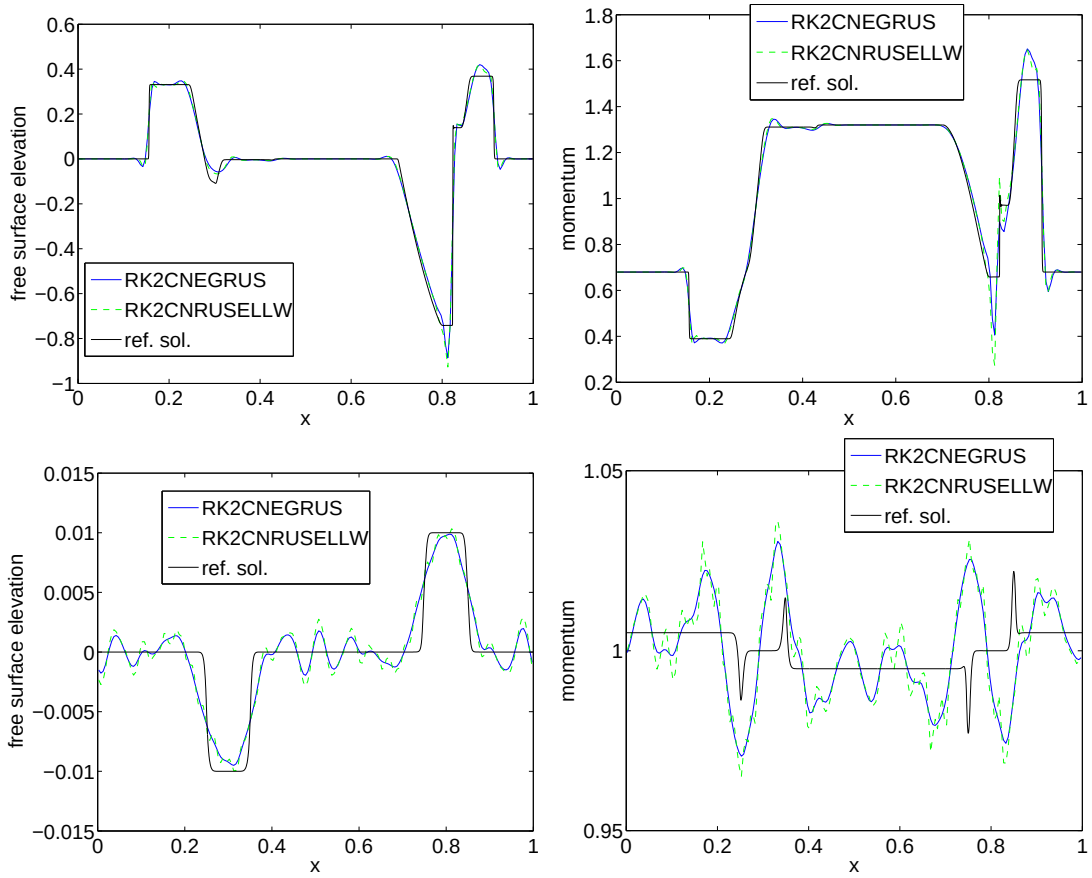

Figure 8.4.: Numerical solutions of the Riemann problems test at time $T = 0.05$ obtained by the second order RK2CNEGRUS, RK2CNRUSELLW schemes. $CFL_u = 0.45$, $\Delta x = 0.05$. Top: Froude number $\varepsilon = 0.8$. Bottom: Froude number $\varepsilon = 0.1$.

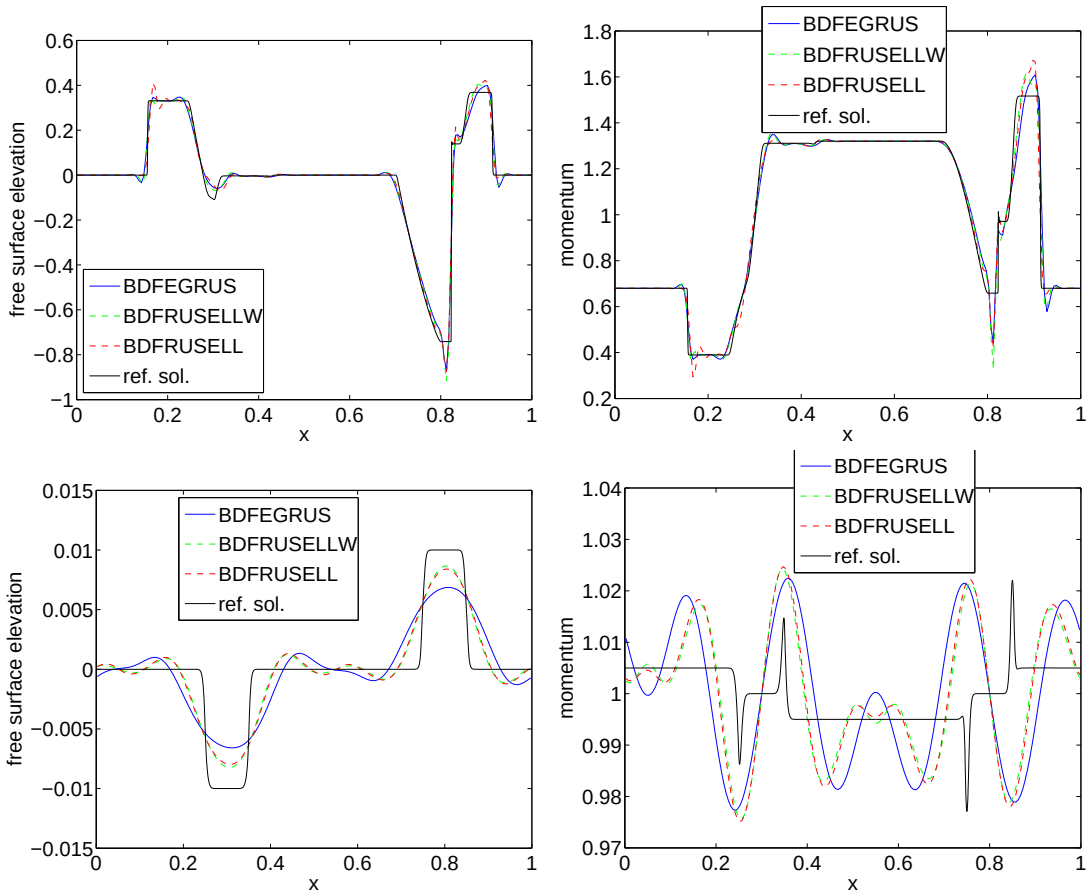

Figure 8.5.: Numerical solutions of the Riemann problems test at time $T = 0.05$ obtained by the second order BDFEGRUS, BDFEGRUSELLW, BDFRUSELL schemes. $CFL_u = 0.45, \Delta x = 0.05$. Top: Froude number $\varepsilon = 0.8$. Bottom: Froude number $\varepsilon = 0.1$.

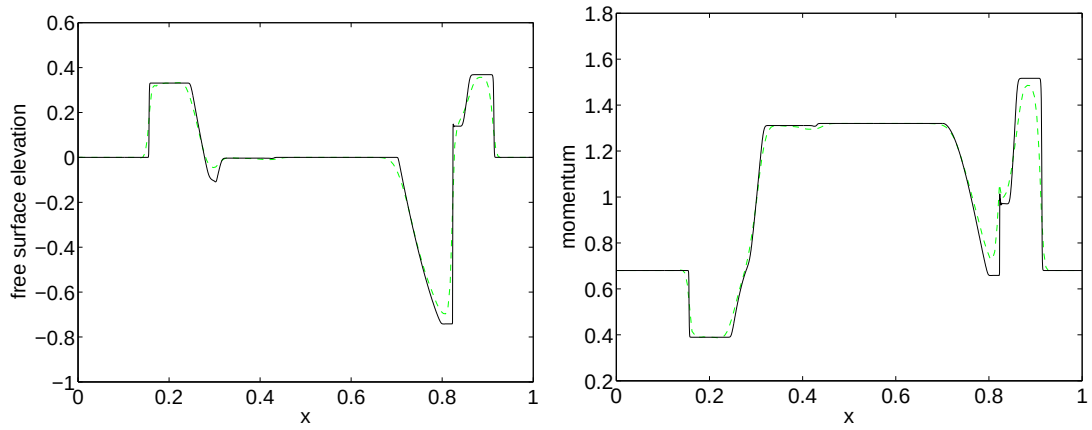

Figure 8.6.: Numerical solution of the Riemann problems test at time $T = 0.05$ obtained by second order BDFRUSELLW scheme with the minmod reconstruction for the non-stiff parts. $CFL_u \approx 0.3, \Delta t = 0.0005, \Delta x = 0.05$, Froude number $\varepsilon = 0.8$. Solid line: reference solution. Dashed line: BDFRUSELLW solution.

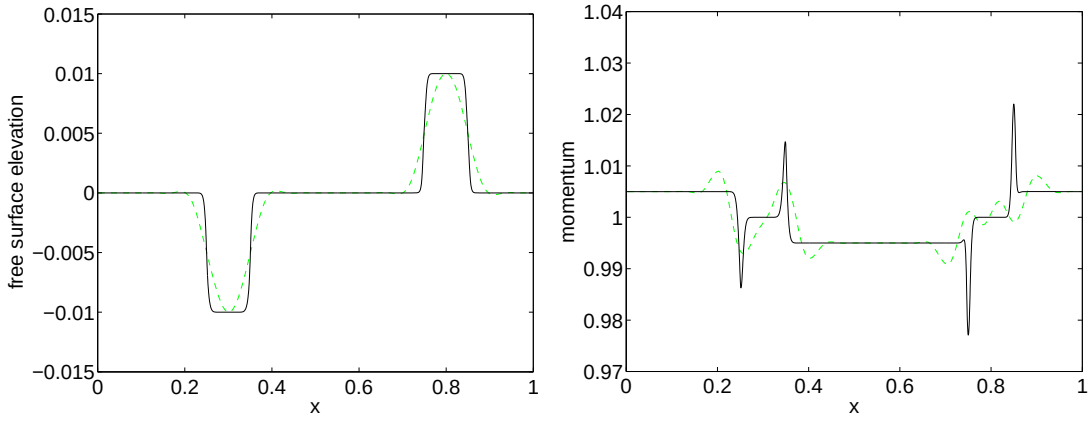

Figure 8.7.: Numerical solution of the Riemann problems test at time $T = 0.05$ obtained by second order BDFRUSELLW scheme with the minmod reconstruction for the non-stiff parts and additional implicit diffusion. $CFL_u \approx 0.1, \Delta t = 0.0005, \Delta x = 0.05$, Froude number $\varepsilon = 0.1$. Solid line: reference solution. Dashed line: BDFRUSELLW solution.

8.2. Travelling vortex

In [100] the exact solution to the 2D SWE for the so-called travelling vortex test are presented, see also [105] for further experiments. The initial data read

$$h(x,y,0) = 110 + \begin{cases} \left(\frac{\varepsilon\Gamma}{\omega}\right)^2 (k(\omega r_c) - k(\pi)) & \text{if } \omega r_c \leq \pi \\ 0 & \text{otherwise} \end{cases}, \tag{8.5a}$$

$$u(x,y,0) = 0.6 + \begin{cases} \Gamma(1+\cos(\omega r_c))(0.5-y) & \text{if } \omega r_c \leq \pi \\ 0 & \text{otherwise} \end{cases}, \tag{8.5b}$$

$$v = \begin{cases} \Gamma(1+\cos(\omega r_c))(x-0.5) & \text{if } \omega r_c \leq \pi \\ 0 & \text{otherwise} \end{cases}, \tag{8.5c}$$

$$r_c = \|\mathbf{x} - (0.5, 0.5)^T\|, \quad \Gamma = 1.5, \quad \omega = 4\pi, \tag{8.5d}$$

$$k(r) = 2\cos(r) + 2r\sin(r) + \frac{1}{8}\cos(2r) + \frac{r}{4}\sin(2r) + \frac{3}{4}r^2. \tag{8.5e}$$

where $(x,y) \in \Omega = [0,1] \times [0,1]$ and $\mathbf{u} = (u,v)^T$. Here Γ is the so-called *vortex intensity parameter*, r_c the distance from the vortex core, and ω an angular wave frequency that specifies the vortex width. Using periodic boundary conditions we are in the setting considered in the asymptotic analysis in Chapter 2, 5, 7.
The initial data describe a rotating vortex that is placed in the middle of the computational domain Ω. The exact solution gives the transport of the travelling vortex with the reference advection velocity $\mathbf{u}_{ref} = (0.6, 0)^T$. Due to the periodic boundary condition the exact solution

$$h(x,y,t) = h(x - t/T, y, 0), \tag{8.6a}$$

$$u(x,y,t) = u(x - t/T, y, 0), \tag{8.6b}$$

$$v(x,y,t) = v(x - t/T, y, 0), \tag{8.6c}$$

is periodic with the period $T = 5/3$. In Figure 8.8 the initial data for the Froude numbers $\varepsilon \in \{0.8, 0.01\}$ are shown. For small Froude numbers the absolute value of the free surface elevation decreases and the momentum does not experience much change. Further the difference between the momentum in x- and y-direction is just a constant and a rotation. This momentum symmetry holds also for the numerical results. Hence, we present only the momentum in x-direction in the following.

In Figures 8.9-8.15 numerical solutions of the first order explicit and IMEX Euler finite volume schemes for Froude numbers in the range from 10^{-8} to 0.8 using 160×160 mesh cells at time $T = 0.1$ are shown. In Figures 8.9-8.11 we compare the HLLC, EGRUS and CFDRUS scheme. For the HLLC and EGRUS scheme we observe the development of numerical artifacts around the vortex for $\varepsilon \in \{0.1, 0.01\}$. For decreasing ε these perturbations increase in size relative to the depth of the vortex and corrupt the numerical solution. If the EGRUS scheme is used, one can get rid of these artifacts by setting the matrices $C_5 = C_6 = C_8 = C_9 = 0$ in (6.109). The consequent scheme is still consistent since the matrices C_5, C_6, C_8, C_9 can be understood as numerical diffusion

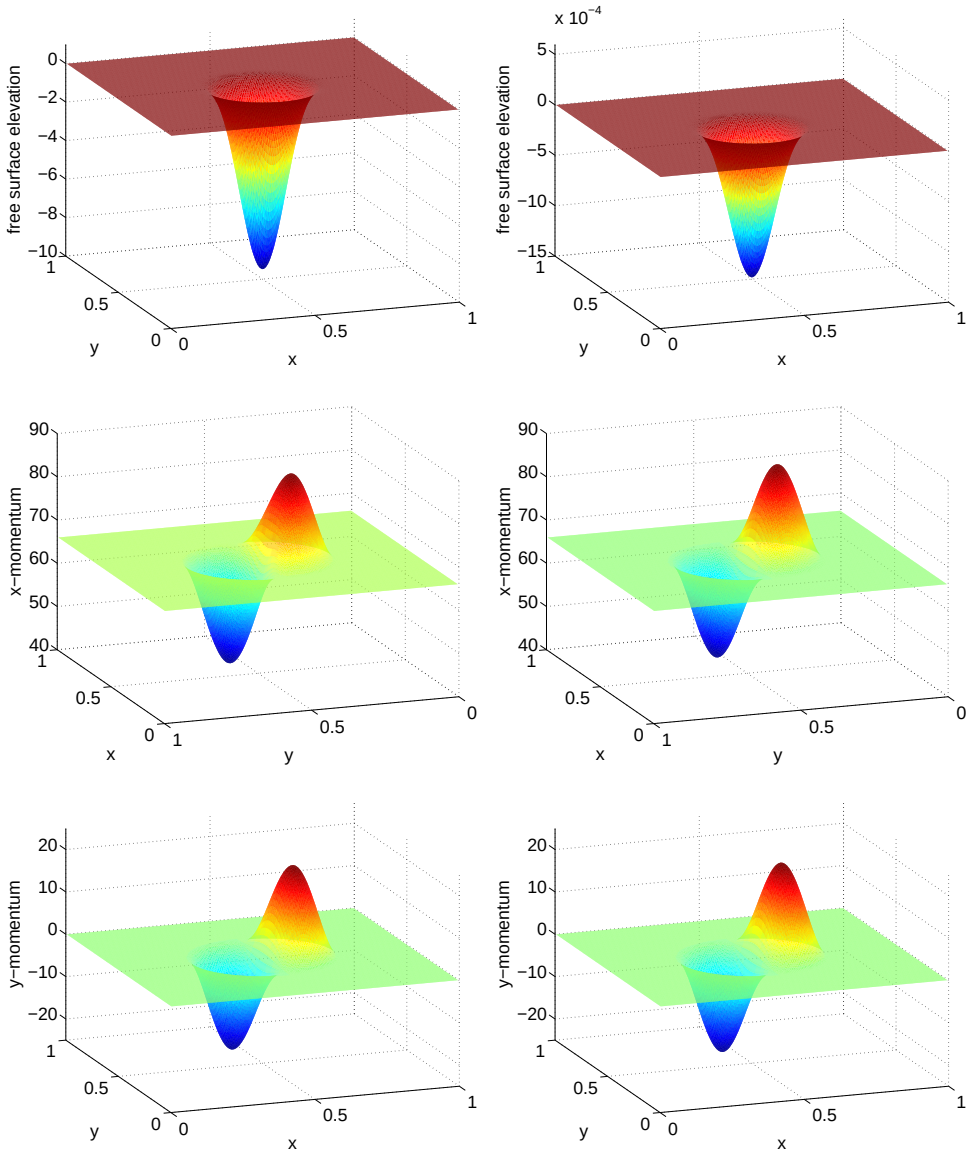

Figure 8.8.: Initial conditions of the travelling vortex test. The pictures show from top to bottom the free surface elevation and the momentum in x- and y-direction. Left: $\varepsilon = 0.8$. Right: $\varepsilon = 0.01$.

terms of the EG numerical flux. Though this seems like a good numerical scheme we will not consider it in the following.

The CFDRUS scheme seems not to allow such numerical artifacts. In Figure 8.12 the numerical solution at time $T = 0.1$ obtained by the CFDRUS scheme with $CFL_u = 0.45$ and Froude numbers $\varepsilon \in \{10^{-6}, 10^{-8}\}$ is presented on a 160×160 cells grid. Here all numerical solutions are reasonable, except the free surface elevation for $\varepsilon = 10^{-8}$, where we can recognise a checkerboard instability. In Figure 8.13 the checkerboard modes of this numerical solution are presented, i.e. the numerical solution is considered either on even or odd cells in x- and y-direction. In Example 7.1.6 we pointed out that checkerboard modes might appear, if the number of cells in x- or y-direction is even. Indeed, using a 161×161 cells grid prevents the checkerboard instability, cf. Figure 8.14. However, if we compute the solution with the RUSELL or RUSELLW scheme on a 160×160 cells grid with $CFL_u = 0.45$, $\varepsilon = 10^{-8}$, we do not observe any checkerboard instabilities. For the RUSELL scheme, this is consistent with Lemma 7.1.9, since there are no checkerboard modes in the kernel of the matrix E, cf. Chapter 7. However the RUSELLW scheme is equivalent to the CFDRUS scheme, cf. Theorem 6.4.4. Thus, it seems that the checkerboard modes are stimulated due to double precision errors in the CFDRUS scheme for low Froude numbers.

In Figures 8.16-8.26 numerical solutions of various second order finite volume schemes for $\varepsilon \in \{0.1, 10^{-3}, 10^{-5}\}$ using 160×160 mesh cells at time $T = 0.1$ are shown. To this end the linear reconstruction (6.36) is used. Firstly, the numerical solutions obtained by the the RK2HLLC, ARSEGRUS, RK2CNEGRUS, BDFEGRUS schemes are shown in Figures 8.16-8.18. As in the solutions obtained by the the HLLC and EGRUS schemes we observe perturbations of the free surface elevation near the vortex. Similarly, the perturbations increase in size relative to the vortex and corrupt the results for decreasing Froude numbers. However the perturbations are smaller than in the case of the HLLC and EGRUS schemes.

In Figures 8.19-8.22 the numerical solutions of the travelling vortex experiment obtained by the ARSRUSELLW, RK2CNRUSELLW and BDFRUSELLW schemes at the time $T = 0.1$ are presented for the Froude numbers $\varepsilon \in \{0.1, 10^{-3}, 10^{-5}\}$. We use typically the CFL_u-number 0.45. However we need to set $CFL_u = 0.3$ for the BDFRUSELLW scheme to obtain correct results for $\varepsilon \in \{10^{-3}, 10^{-5}\}$, cf. Figures 8.20-8.22. The numerical solutions of the ARSRUSELLW and the BDFRUSELLW scheme give very good approximations. In the results obtained by the RK2CNRUSELLW scheme numerical artifacts appear and corrupt the solution of the free surface elevation for $\varepsilon \in \{10^{-3}, 10^{-5}\}$. The numerical solution for $\varepsilon = 0.1$ is a good approximation. The choice of lower CFL_u-numbers does not improve the numerical solution of the RK2CNRUSELLW scheme.

In Figures 8.23-8.26 the numerical solutions of the travelling vortex experiment obtained by the ARSRUSELL, RK2CNRUSELL and BDFRUSELL schemes at the time $T = 0.1$ are presented for the Froude numbers $\varepsilon \in \{0.1, 10^{-3}, 10^{-5}\}$. The ARSRUSELL scheme provides oscillatory solutions of the free surface elevation for the low Froude numbers $\varepsilon \in \{10^{-3}, 10^{-5}\}$. The behaviour of the numerical solutions of the RK2CNRUSELL and BDFRUSELL schemes is analogously to the RK2CNRUSELLW and BDFRUSELLW schemes: for $\varepsilon \in \{10^{-3}, 10^{-5}\}$ numerical artifacts appear in the solutions obtained by the RK2CNRUSELL scheme, whereas the CFL_u-number needs to be reduced to $CFL_u = 0.3$

to obtain correct results by the BDFRUSELL scheme.

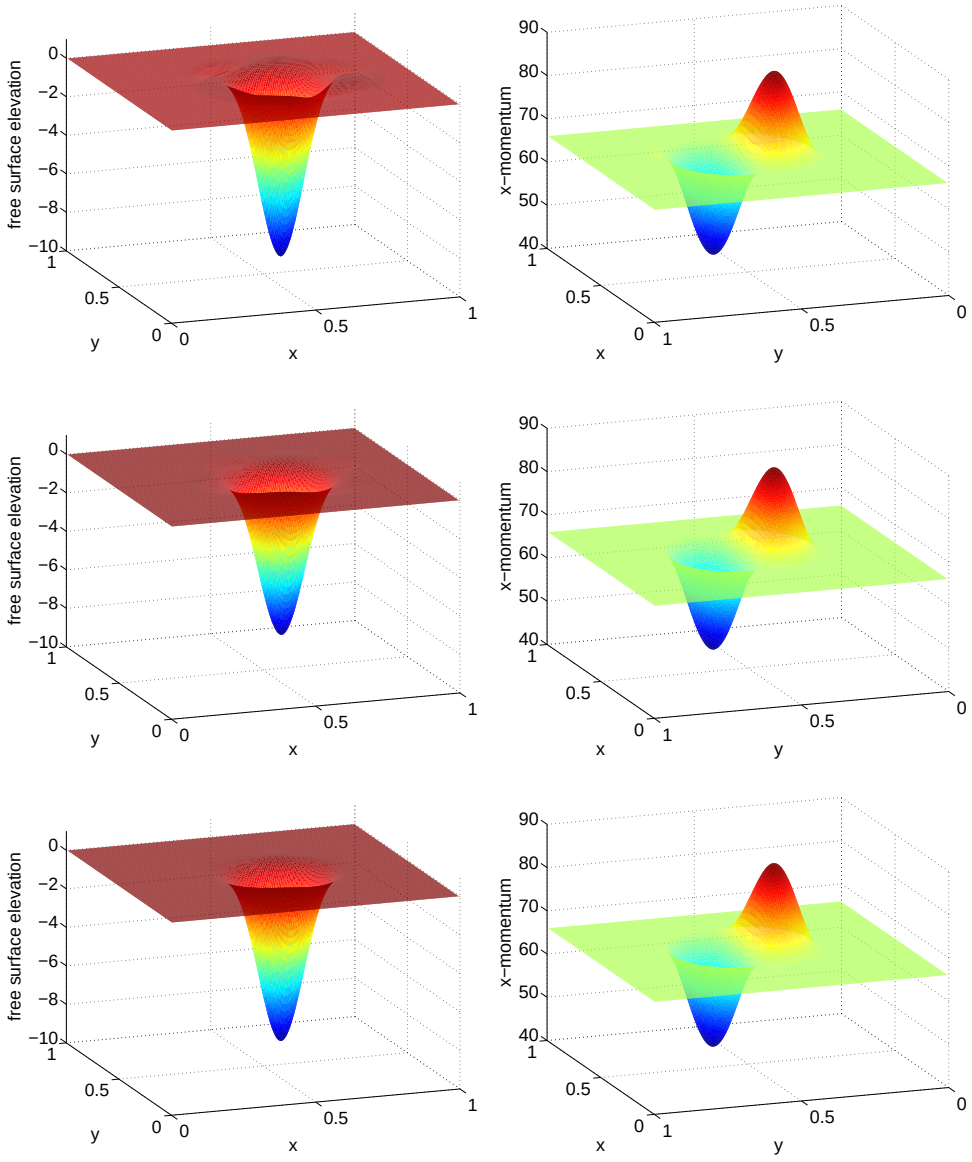

Figure 8.9.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$. The Froude number is $\varepsilon = 0.8$ and $CFL_u = CFL = 0.45$. From top to bottom the solution is obtained by the HLLC, EGRUS, CFDRUS schemes. The constant $CFL_g = 0.01$ was used in the EGRUS scheme.

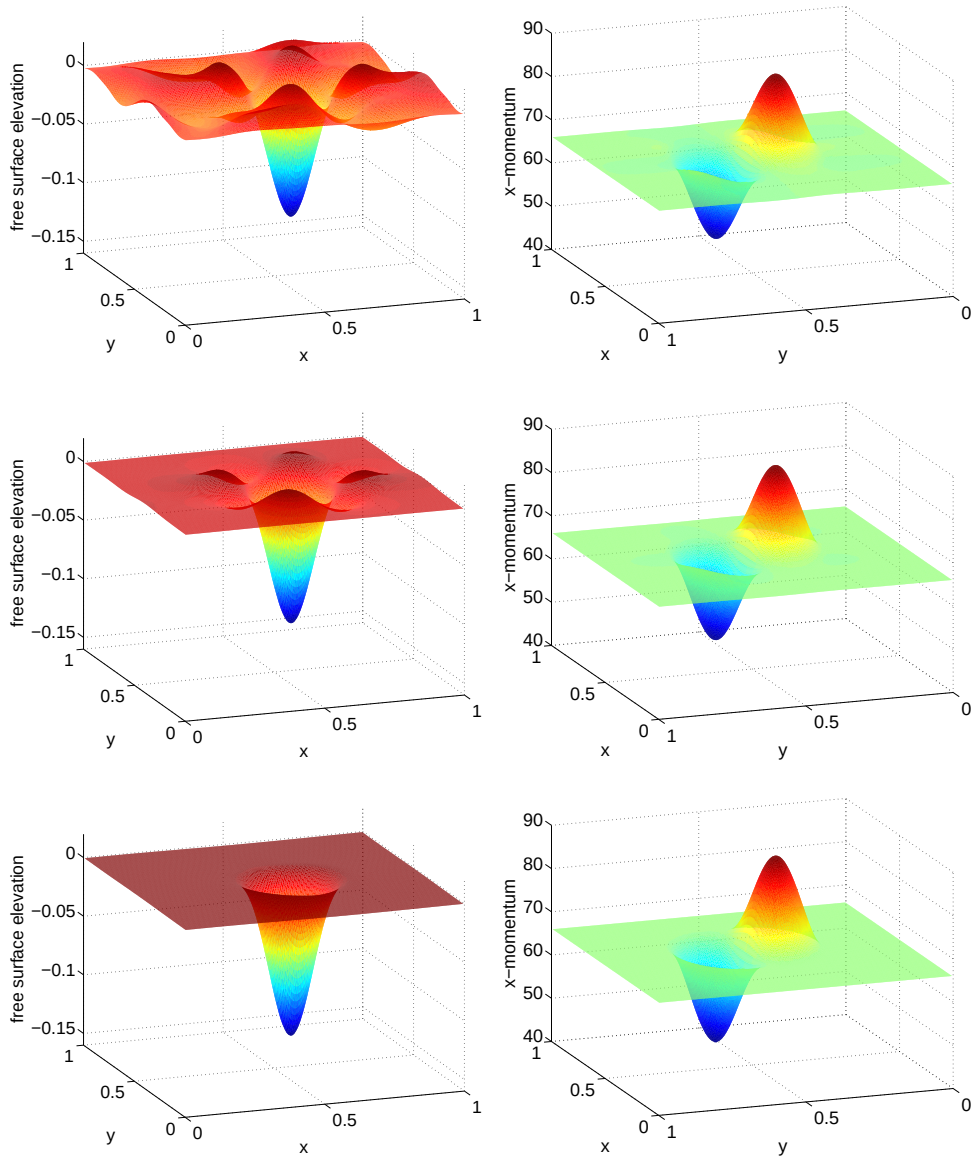

Figure 8.10.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$. The Froude number is $\varepsilon = 0.1$ and $CFL_u = CFL = 0.45$. From top to bottom the solution is obtained by the HLLC, EGRUS, CFDRUS schemes. The constant $CFL_g = 0.01$ was used in the EGRUS scheme.

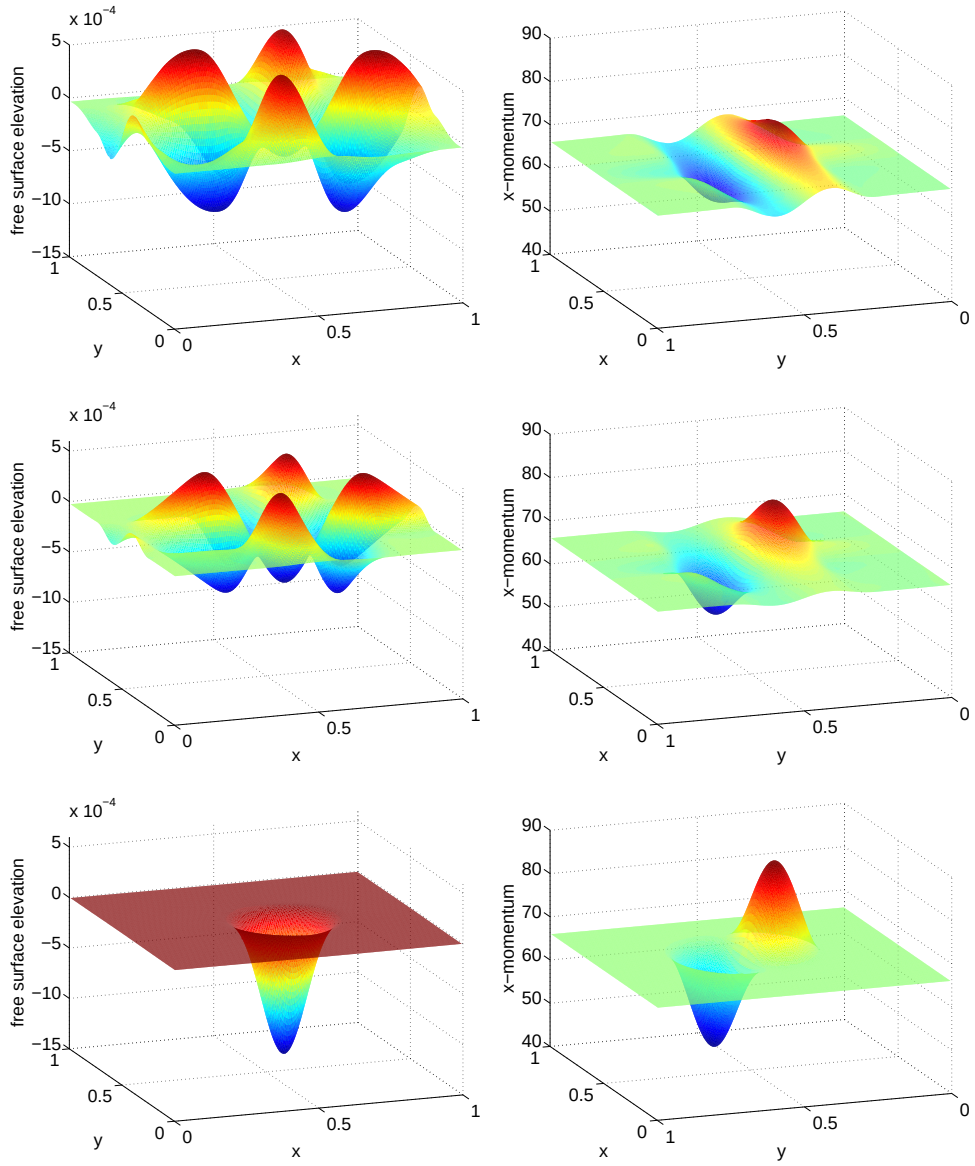

Figure 8.11.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$. The Froude number is $\varepsilon = 0.01$ and $CFL_u = CFL = 0.45$. From top to bottom the solution is obtained by the HLLC, EGRUS, CFDRUS schemes. The constant $CFL_g = 0.01$ was used in the EGRUS scheme.

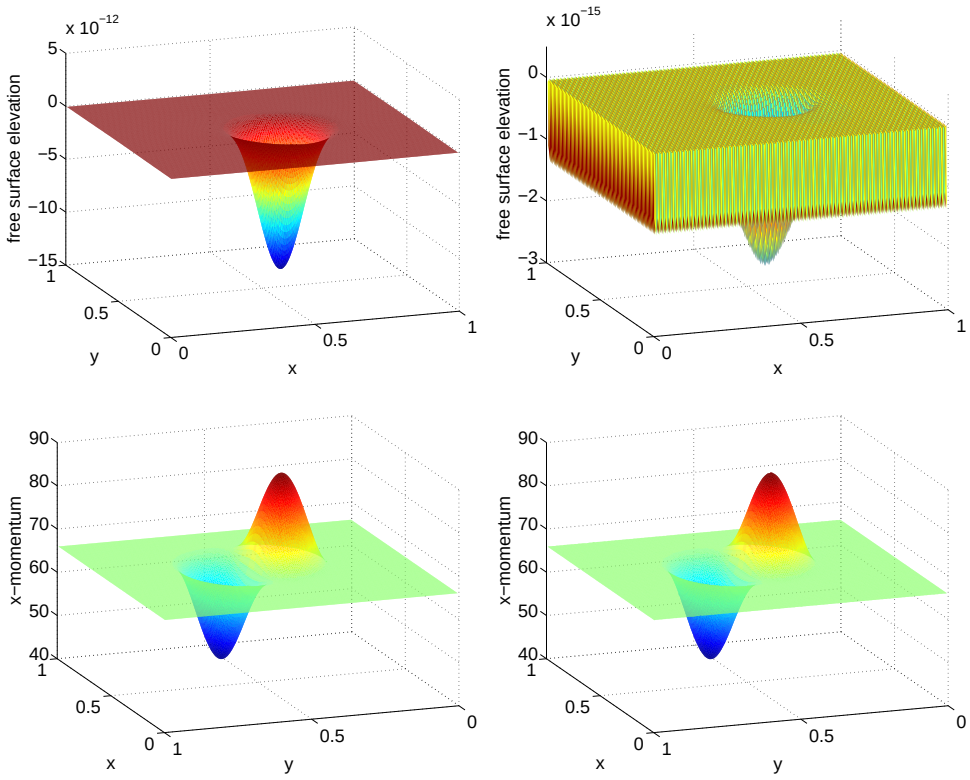

Figure 8.12.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$ computed by the CFDRUS scheme and $CFL_u = 0.45$. Left: $\varepsilon = 10^{-6}$. Right: $\varepsilon = 10^{-8}$.

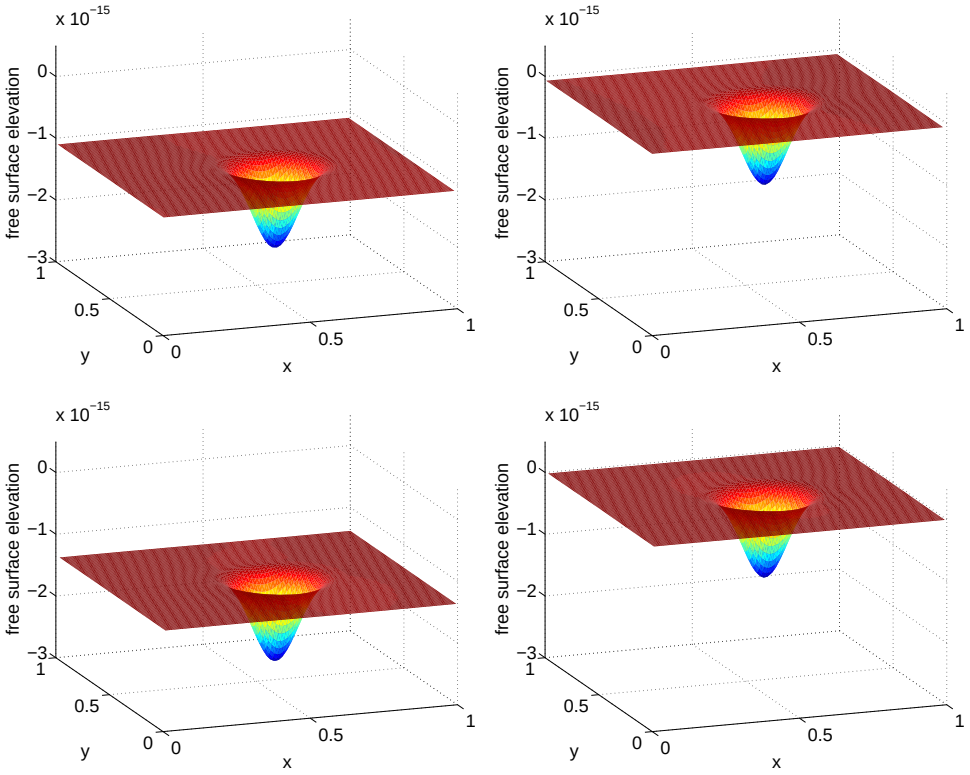

Figure 8.13.: Checkerboard modes of the numerical solution of the travelling vortex experiment at time $T = 0.1$ computed by the CFDRUS scheme, $CFL_u = 0.45$, $\varepsilon = 10^{-8}$, 160×160 mesh cells. Top left: odd cells in x- and y-direction. Top right: even cells in $x-$ and odd in y-direction. Bottom left: odd cells in x- and even cells in y-direction. Bottom right: odd cells in x- and y-direction.

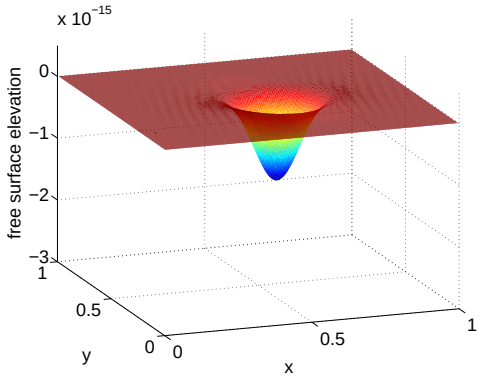

Figure 8.14.: Free surface elevation numerical solution of the travelling vortex experiment at time $T = 0.1$ computed by the CFDRUS scheme, $CFL_u = 0.45$, $\varepsilon = 10^{-8}$, 161×161 mesh cells.

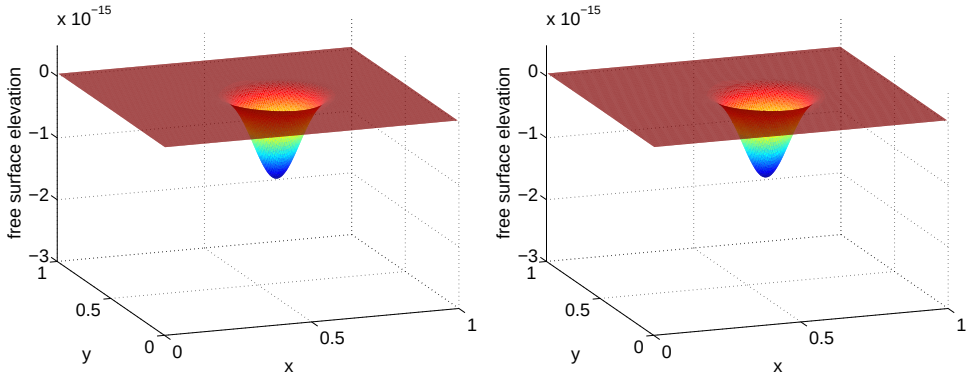

Figure 8.15.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$ computed by the elliptic approach IMEX Euler finite volume schemes, $CFL_u = 0.45$, $\varepsilon = 10^{-8}$. Left: RUSELL scheme. Right: RUSELLW scheme.

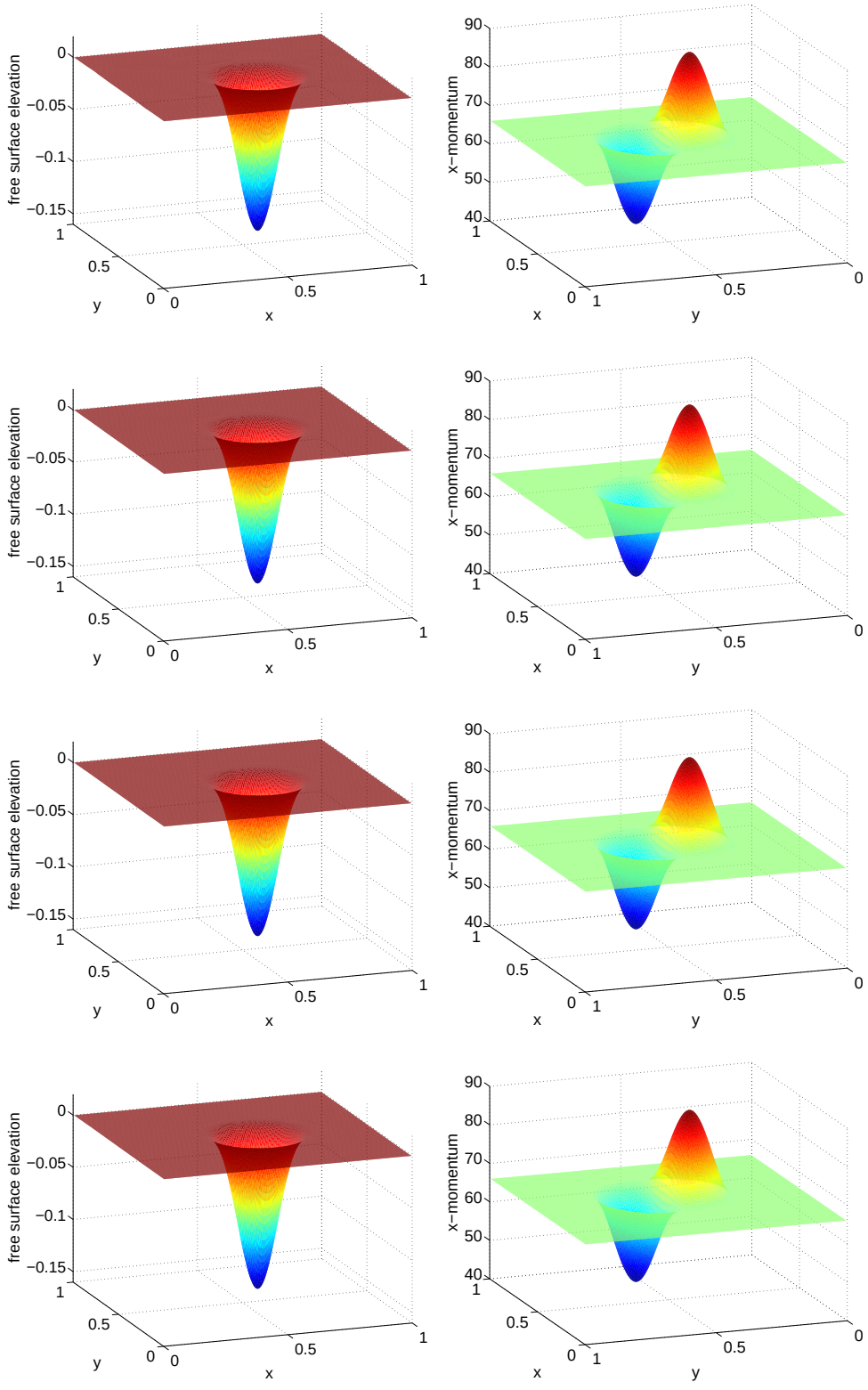

Figure 8.16.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$ computed with straight approach second order IMEX finite volume schemes, $CFL_u = 0.45$, Froude number $\varepsilon = 0.1$. From top to bottom the solution is obtained by the RK2HLLC, ARSEGRUS, RK2CNEGRUS, BDFEGRUS schemes. Left: free surface elevation. Right: momentum in x-direction.

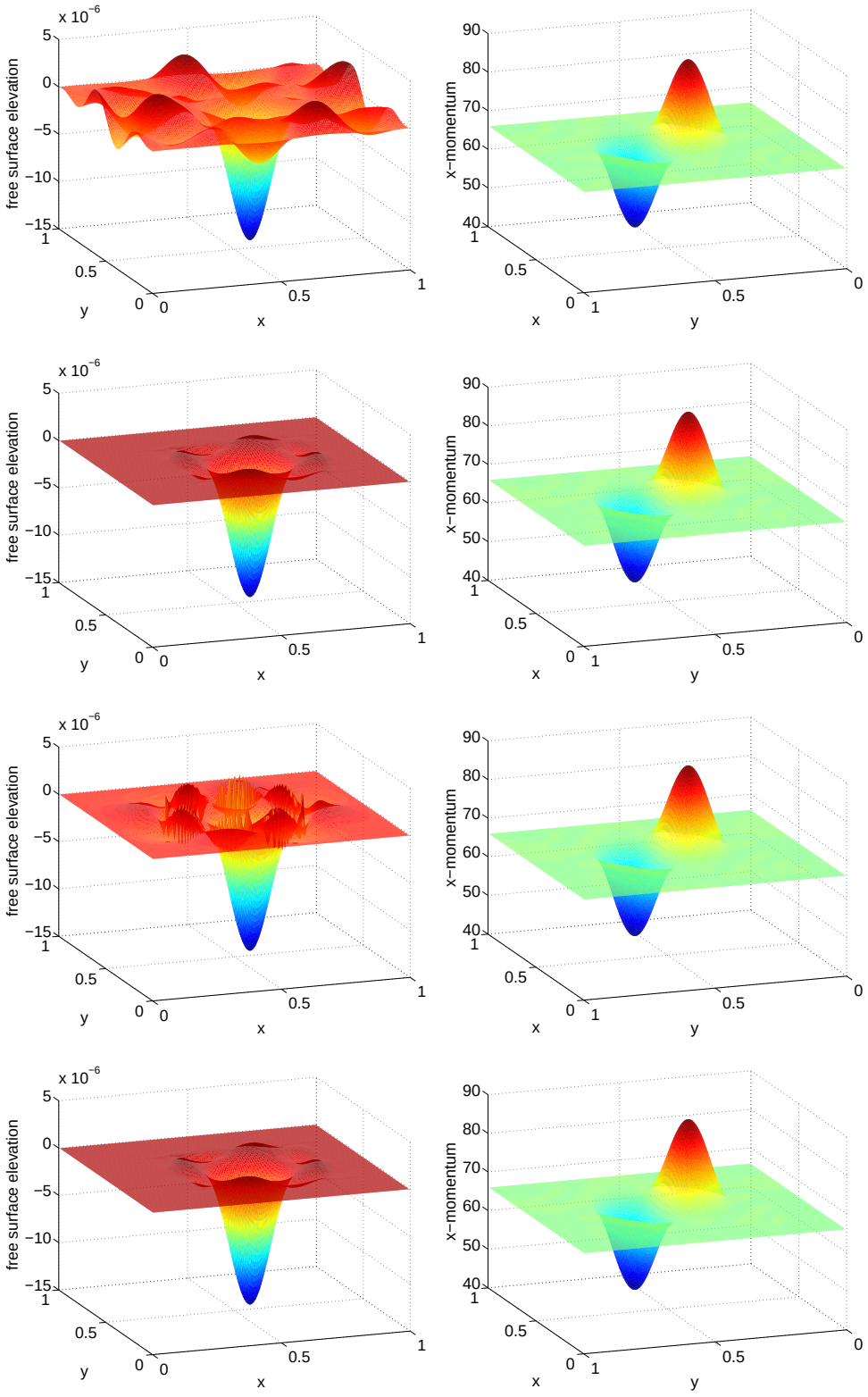

Figure 8.17.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$ computed with straight approach second order IMEX finite volume schemes, $CFL_u = 0.45$, Froude number $\varepsilon = 10^{-3}$. From top to bottom the solution is obtained by the RK2HLLC, ARSEGRUS, RK2CNEGRUS, BDFEGRUS schemes. Left: free surface elevation. Right: momentum in x-direction.

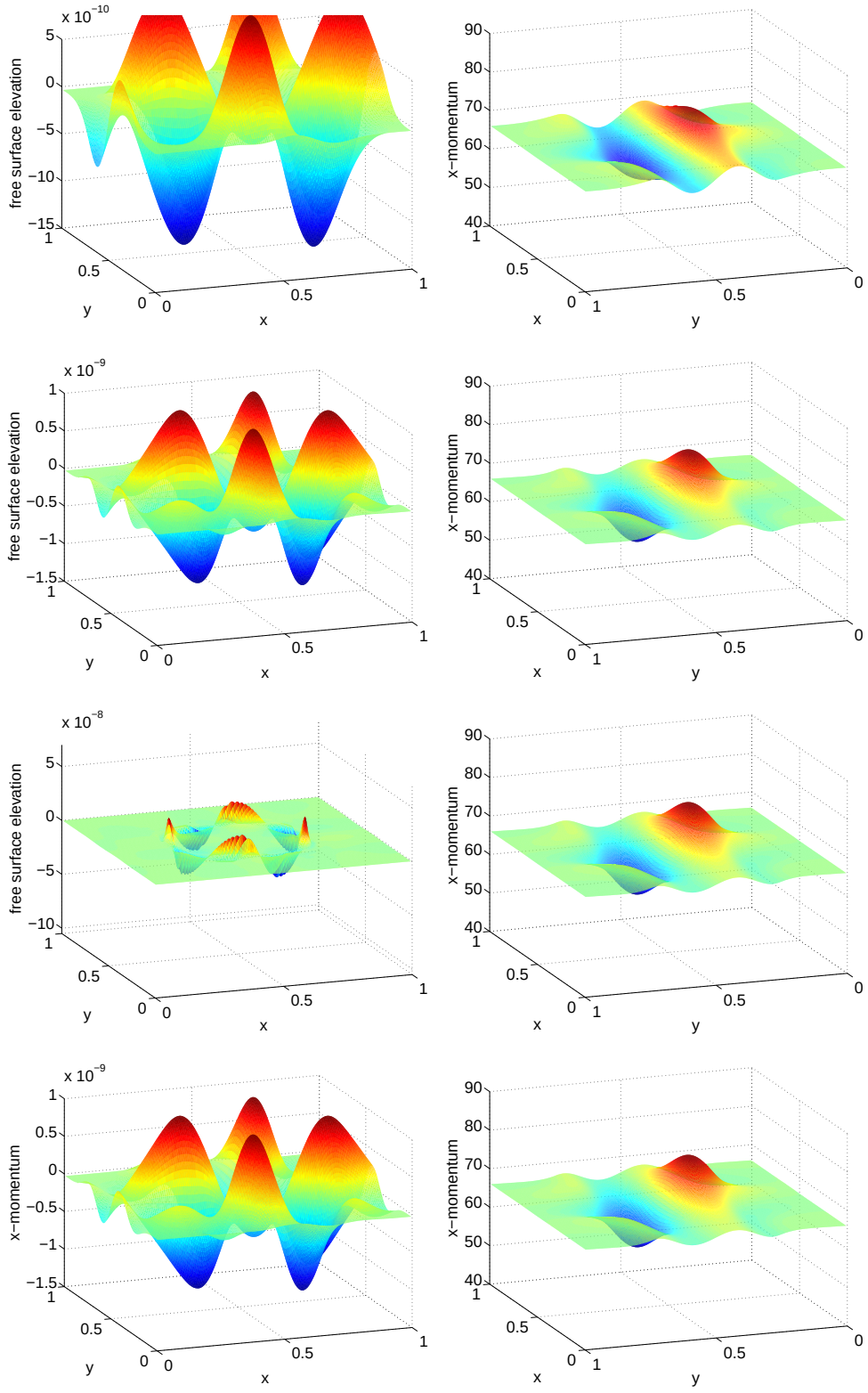

Figure 8.18.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$ computed with straight approach second order IMEX finite volume schemes, $CFL_u = 0.45$, Froude number $\varepsilon = 10^{-5}$. From top to bottom the solution is obtained by the RK2HLLC, ARSEGRUS, RK2CNEGRUS, BDFEGRUS schemes. Left: free surface elevation. Right: momentum in x-direction.

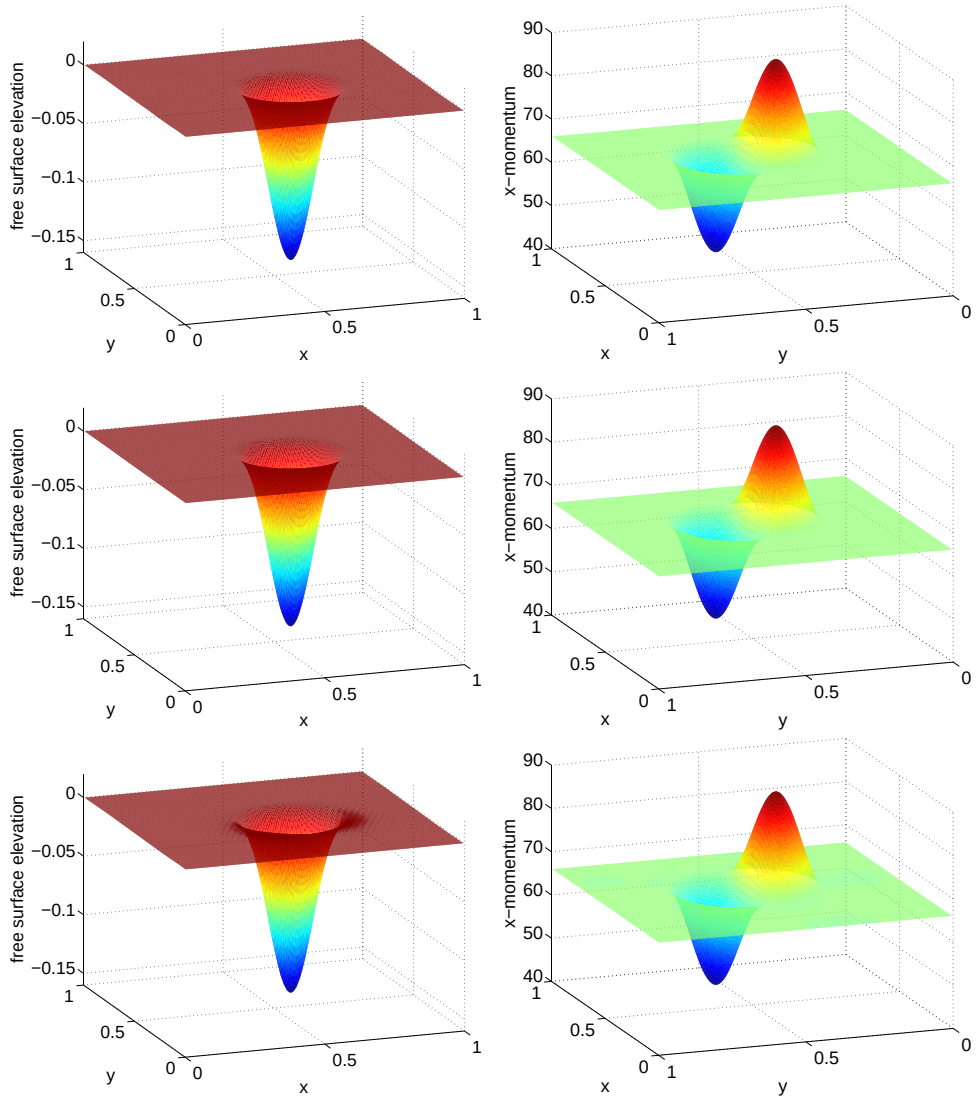

Figure 8.19.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$ computed by the elliptic approach second order IMEX finite volume schemes, $CFL_u = 0.45$, Froude number $\varepsilon = 0.1$. From top to bottom the solution is obtained by the ARSRUSELLW, RK2CNRUSELLW, BDFRUSELLW schemes. Left: free surface elevation. Right: momentum in x-direction.

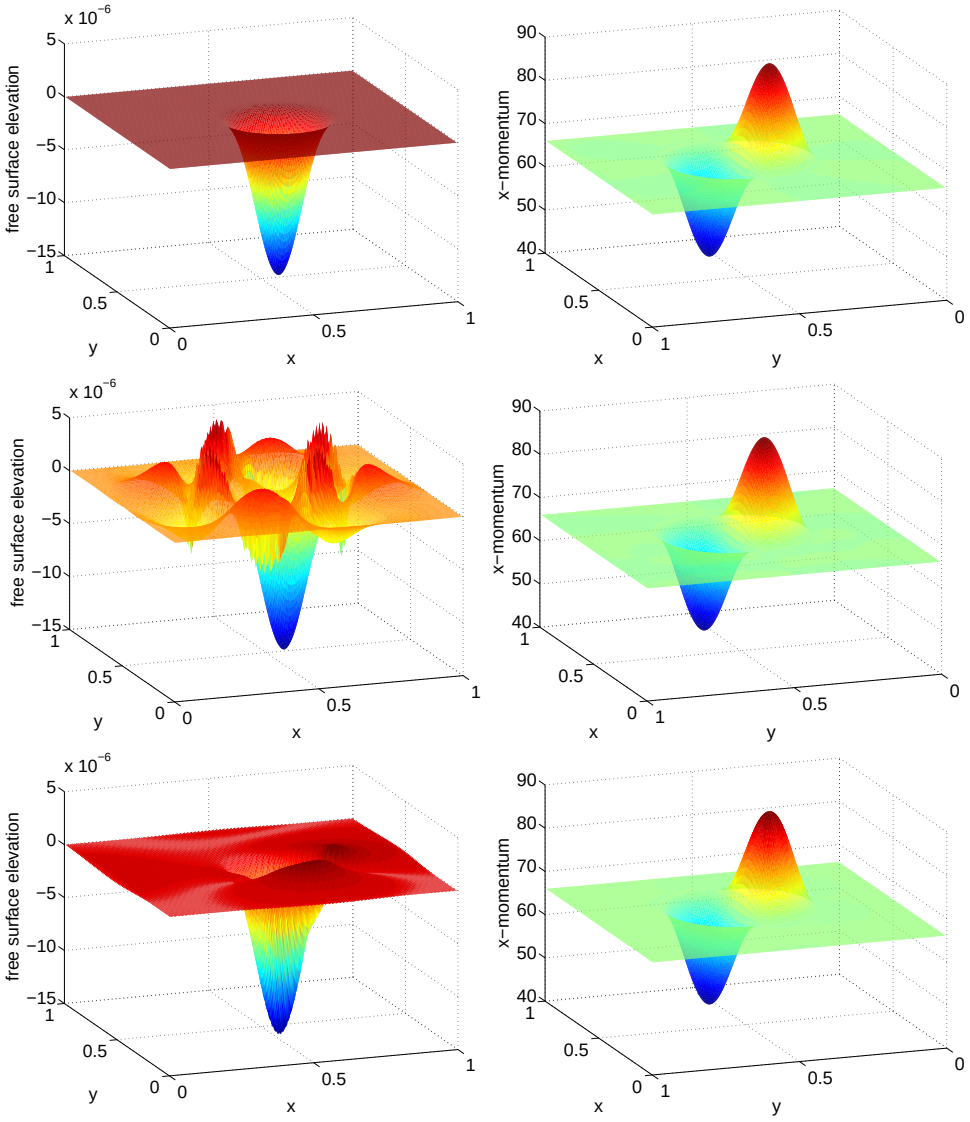

Figure 8.20.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$ computed with elliptic approach second order IMEX finite volume schemes, $CFL_u = 0.45$, Froude number $\varepsilon = 10^{-3}$. From top to bottom the solution is obtained by the ARSRUSELLW, RK2CNRUSELLW, BDFRUSELLW schemes. Left: free surface elevation. Right: momentum in x-direction.

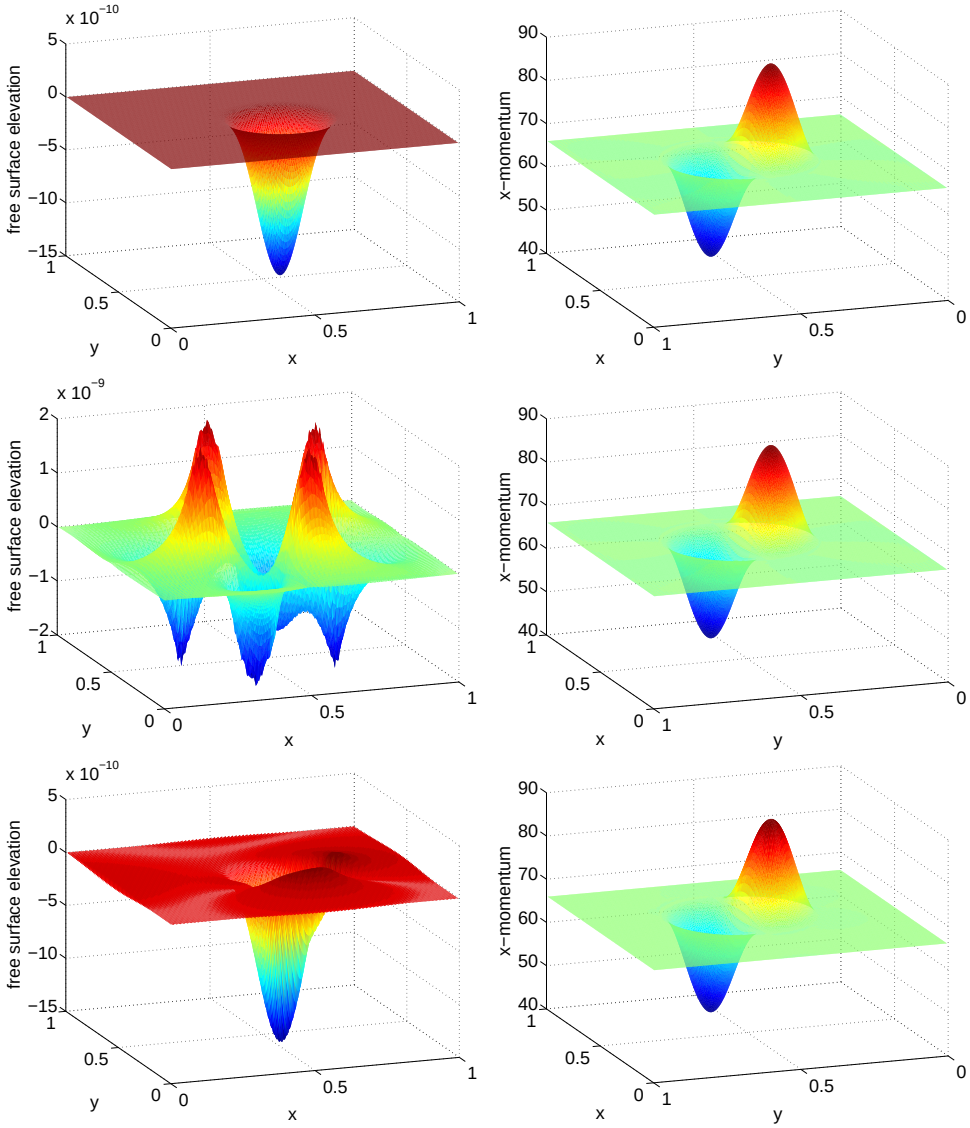

Figure 8.21.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$ computed with elliptic approach second order IMEX finite volume schemes, $CFL_u = 0.45$, Froude number $\varepsilon = 10^{-5}$. From top to bottom the solution is obtained by the ARSRUSELLW, RK2CNRUSELLW, BDFRUSELLW schemes. Left: free surface elevation. Right: momentum in x-direction.

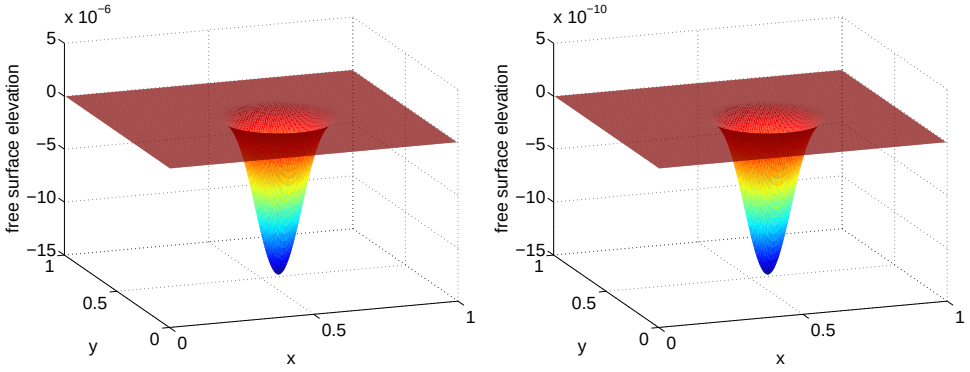

Figure 8.22.: Numerical free surface elevation solutions of the travelling vortex experiment at time $T = 0.1$ computed by the BDFRUSELLW scheme using $CFL_u = 0.3$. Left: $\varepsilon = 10^{-3}$. Right: $\varepsilon = 10^{-5}$.

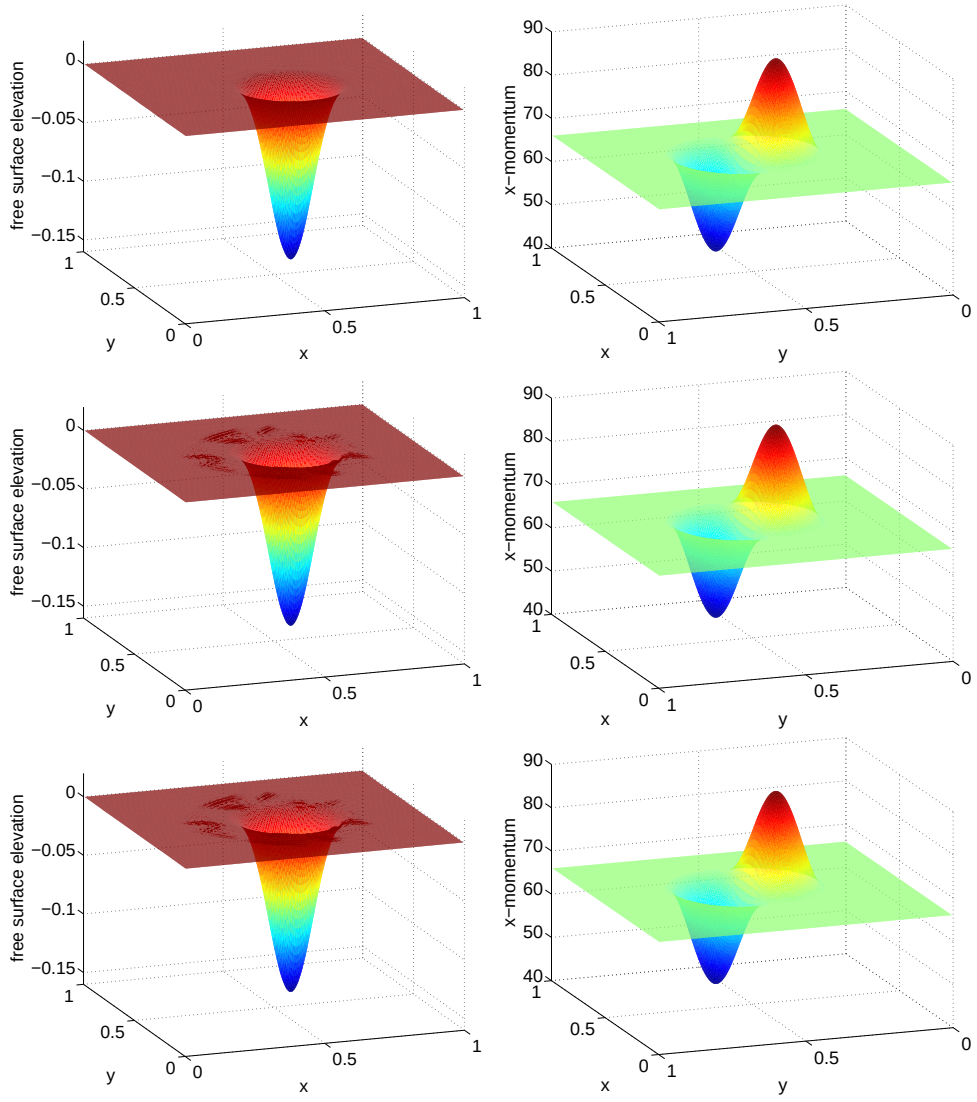

Figure 8.23.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$ computed by the elliptic approach second order IMEX finite volume schemes, $CFL_u = 0.45$, Froude number $\varepsilon = 0.1$. From top to bottom the solution is obtained by the ARSRUSELL, RK2CNRUSELL, BDFRUSELL schemes. Left: free surface elevation. Right: momentum in x-direction.

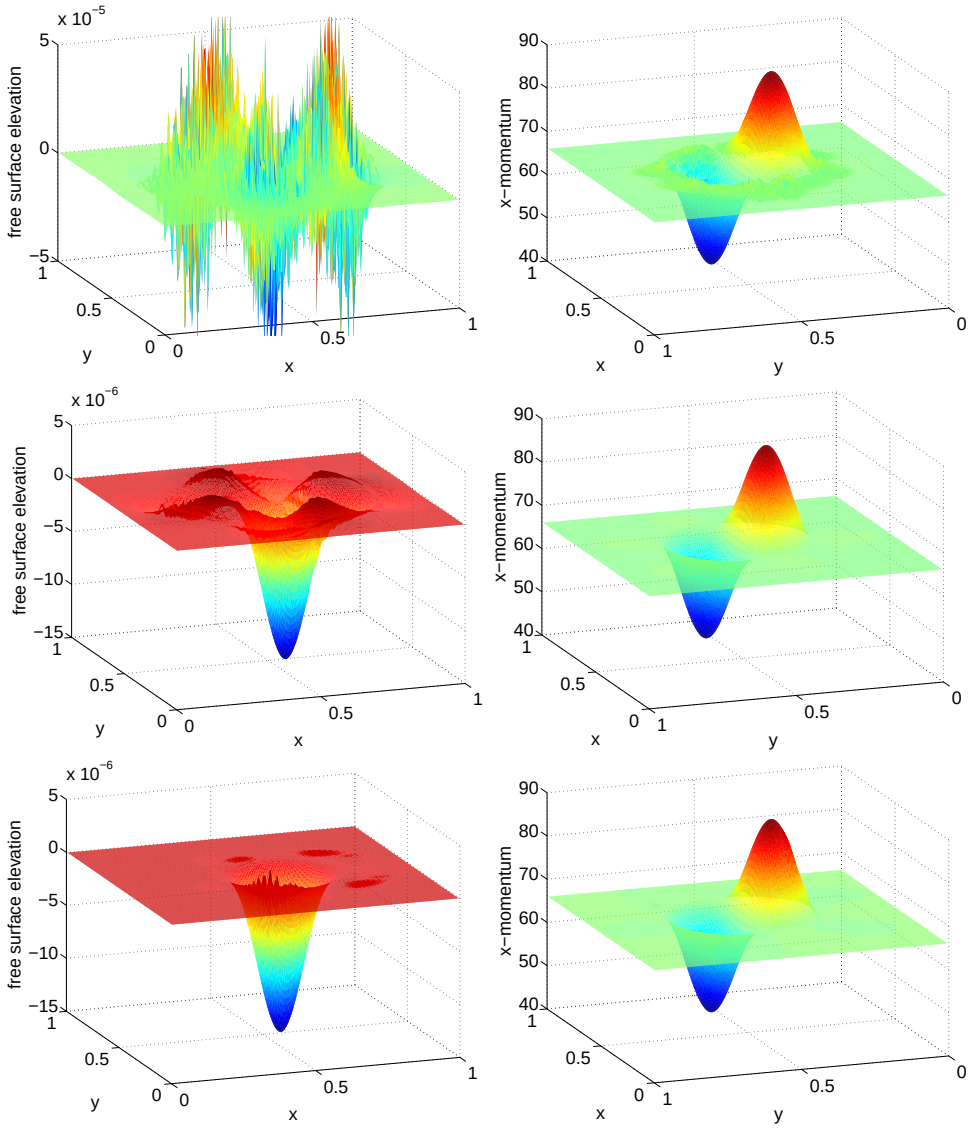

Figure 8.24.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$ computed with elliptic approach second order IMEX finite volume schemes, $CFL_u = 0.45$, Froude number $\varepsilon = 10^{-3}$. From top to bottom the solution is obtained by the ARSRUSELL, RK2CNRUSELL, BDFRUSELL schemes. Left: free surface elevation. Right: momentum in x-direction.

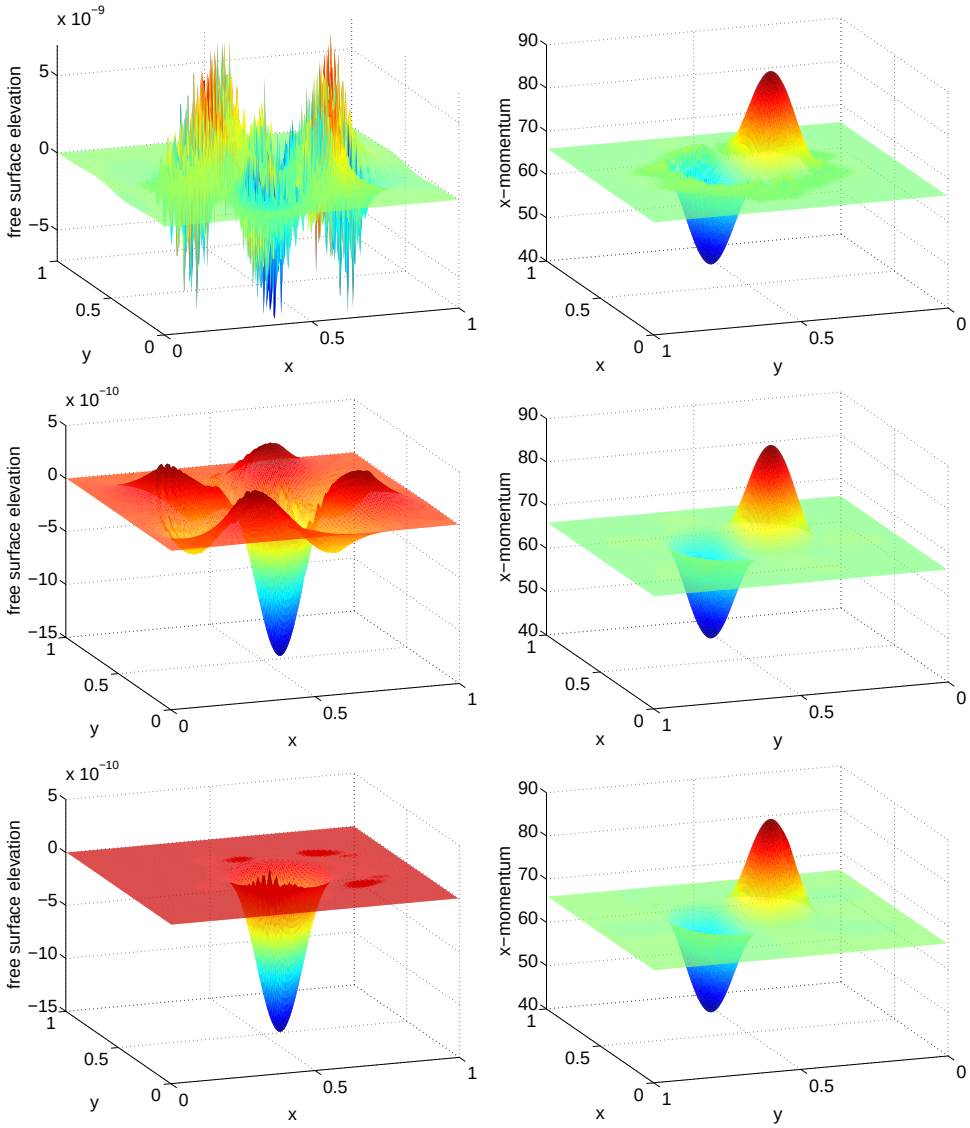

Figure 8.25.: Numerical solutions of the travelling vortex experiment at time $T = 0.1$ computed with elliptic approach second order IMEX finite volume schemes, $CFL_u = 0.45$, Froude number $\varepsilon = 10^{-5}$. From top to bottom the solution is obtained by the ARSRUSELL, RK2CNRUSELL, BDFRUSELL schemes. Left: free surface elevation. Right: momentum in x-direction.

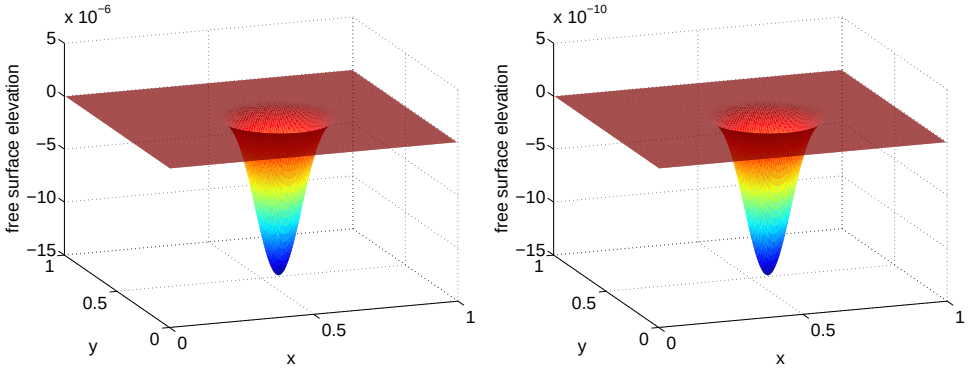

Figure 8.26.: Numerical free surface elevation solutions of the travelling vortex experiment at the time $T = 0.1$ obtained by the BDFRUSELL scheme using $CFL_u = 0.3$. Left: $\varepsilon = 10^{-3}$. Right: $\varepsilon = 10^{-5}$.

8.2.1. Experimental order of convergence

Assume that $\mathbf{w}_N$ is the numerical solution of the SWE (2.30) computed by a p-th order finite volume scheme on a $N \times N$ cells mesh, cf. Section 6.1. Further let $\mathbf{w}$ be the exact solution. Then

$$\|\mathbf{w}_N - \mathbf{w}\| = \frac{C}{N^p} \tag{8.7}$$

implies

$$p = \log_2 \left(\frac{\|\mathbf{w}_N - \mathbf{w}\|}{\|\mathbf{w}_{2N} - \mathbf{w}\|} \right) =: EOC. \tag{8.8}$$

Thus using the formula (8.8) we can measure the *experimental order of convergence* (EOC) of a scheme. We will compute the EOC of IMEX finite volume schemes that are discretised with either first or second order approximations in time and space to validate their convergence order. Particularly we consider the first order discretised EGRUS, CFDRUS, RUSELLW, RUSELL schemes and the second order discretised ARSEGRUS, RK2CNEGRUS, BDFEGRUS, ARSRUSELLW, RK2CNRUSELLW, BDFRUSELLW, ARSRUSELL, RK2CNRUSELL, BDFRUSELL schemes, cf. Table 8.1.

Tables 8.2-8.6 present the EOC of the HLLC, EGRUS, CFDRUS, RUSELLW and RUSELL schemes at time $T = 0.1$ for Froude numbers ε in the range from 10^{-16} to 0.8. The CFL-number $CFL = 0.45$ was used for the explicit HLLC scheme; for the IMEX schemes $CFL_u = 0.45$. Note that the EGRUS scheme is independent of the CFL_g-number as long as $CFL_g < 1$, since the numerical solutions are piecewise constant. Table 8.3 indicates that the EGRUS scheme is of first order for $\varepsilon \in \{0.8, 0.1\}$. Note that the convergence rates increase with increasing cell number N^2. We expect to obtain first order convergence rates for all Froude numbers, if the grid is sufficiently fine. Further, let us fix the number of cells N and consider the error in the momentum. It increases significantly with decreasing Froude number. However, the momentum in the exact solution (8.6) almost does not change for small Froude numbers. Thus the fact that the scheme is not able to resolve the free surface elevation z for low Froude numbers impacts the quality of the numerical solution of the momentum. However, note that the errors of the free surface elevation are lower than $\Delta x = \Delta y$ when the convergence rate is significantly lower than one. The HLLC scheme behaves similarly, but the errors are slightly larger due to a larger dissipation.

The EOC of the CFDRUS and the RUSELLW schemes are presented in Tables 8.4, 8.5. These demonstrate that the CFDRUS and RUSELLW schemes are equal up to double precision errors. However the RUSELLW scheme is more stable as we have already pointed out. Thus, it suffices to consider the RUSELLW scheme. It is remarkable that the first order convergence rates as well as the errors of the momentum are uniform with respect to the Froude number. Moreover the convergence rates of the free surface elevation are around one up to the Froude number $\varepsilon = 10^{-6}$. For smaller Froude numbers the convergence stops probably due to double precision errors. Even if the convergence in the z-component breaks down since the free surface elevation is of double precision order, the convergence rate of the momentum remains. We observe analogous results for the RUSELL scheme, cf. Table 8.6.

travelling vortex, $\varepsilon = 0.8, CFL = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	0.46403		0.62193		0.65557	
20	0.27183	0.7715	0.41111	0.5972	0.46938	0.4820
40	0.15361	0.8235	0.24355	0.7553	0.28181	0.7360
80	0.08269	0.8935	0.13402	0.8618	0.15591	0.8540
160	0.04306	0.9414	0.07055	0.9257	0.08240	0.9200

travelling vortex, $\varepsilon = 0.1, CFL = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.627e-002		1.56741		1.39070	
20	2.109e-002	-0.3740	1.25232	0.3238	1.25671	0.1462
40	1.785e-002	0.2409	0.91297	0.4560	0.92926	0.4355
80	1.204e-002	0.5680	0.58834	0.6339	0.60701	0.6144
160	7.117e-003	0.7583	0.34534	0.7686	0.35853	0.7596

travelling vortex, $\varepsilon = 0.01, CFL = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	7.334e-005		1.75490		1.55767	
20	7.452e-005	-0.0229	1.73044	0.0203	1.69204	-0.1194
40	7.460e-005	-0.0015	1.73704	-0.0055	1.70807	-0.0136
80	1.563e-004	-1.0674	1.65521	0.0696	1.63879	0.0597
160	2.478e-004	-0.6648	1.40345	0.2380	1.39689	0.2304

Table 8.2.: EOC of the first order HLLC scheme; travelling vortex test.

Comparing the errors and convergence rates in Tables 8.3-8.6 the results are similar for $\varepsilon = 0.8$. For lower Froude numbers, the RUSELLW and RUSELL schemes outperform EGRUS clearly, which give slightly better results than the HLLC scheme.

travelling vortex, $\varepsilon = 0.8, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	0.27275		0.52234		1.11499	
10	0.25602	0.0913	0.71106	-0.4450	0.81517	0.4519
20	0.20070	0.3513	0.46092	0.6255	0.67460	0.2731
40	0.13650	0.5562	0.27917	0.7234	0.41320	0.7072
80	0.08311	0.7158	0.15565	0.8428	0.23342	0.8239
160	0.04661	0.8344	0.08270	0.9123	0.12521	0.8986

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	7.163e-03		0.55269		1.22576	
10	1.332e-02	-0.8952	1.14562	-1.0516	1.12356	0.1256
20	1.140e-02	0.2243	0.89290	0.3596	1.01654	0.1444
40	7.926e-03	0.5249	0.61738	0.5324	0.70429	0.5294
80	4.946e-03	0.6803	0.37546	0.7175	0.43380	0.6991
160	2.925e-03	0.7579	0.21019	0.8370	0.24445	0.8275

travelling vortex, $\varepsilon = 0.01, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	8.700e-05		0.72935		1.34663	
10	1.330e-04	-0.6119	1.65234	-1.1798	1.51589	-0.1708
20	1.325e-04	0.0049	1.64049	0.0104	1.60403	-0.0815
40	1.590e-04	-0.2632	1.55681	0.0755	1.52462	0.0732
80	1.756e-04	-0.1431	1.30703	0.2523	1.30074	0.2291
160	1.465e-04	0.2617	0.96001	0.4452	0.96531	0.4303

Table 8.3.: EOC of the first order EGRUS scheme; travelling vortex test.

travelling vortex, $\varepsilon = 0.8, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	2.779e-01		0.51960		1.10574	
10	2.831e-01	-0.0268	0.63395	-0.2870	0.81631	0.4378
20	1.515e-01	0.9021	0.38808	0.7080	0.64017	0.3507
40	8.667e-02	0.8056	0.22732	0.7716	0.38544	0.7319
80	4.784e-02	0.8573	0.12468	0.8664	0.21589	0.8362
160	2.540e-02	0.9135	0.06570	0.9244	0.11496	0.9091

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	8.070e-03		0.41534		1.12315	
10	7.033e-03	0.1983	0.73944	-0.8321	0.89730	0.3239
20	3.291e-03	1.0956	0.40569	0.8660	0.66916	0.4232
40	2.205e-03	0.5775	0.22985	0.8197	0.39979	0.7431
80	1.370e-03	0.6873	0.12631	0.8637	0.22383	0.8369
160	7.731e-04	0.8249	0.06683	0.9184	0.11936	0.9071

travelling vortex, $\varepsilon = 0.01, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	8.998e-05		0.41380		1.12320	
10	5.844e-05	0.6227	0.72954	-0.8180	0.89078	0.3345
20	3.565e-05	0.7131	0.40402	0.8526	0.66915	0.4127
40	2.134e-05	0.7400	0.22980	0.8140	0.40046	0.7407
80	1.295e-05	0.7204	0.12624	0.8642	0.22430	0.8362
160	7.289e-06	0.8297	0.06672	0.9201	0.11964	0.9067

travelling vortex, $\varepsilon = 10^{-6}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	9.011e-13		0.41381		1.12319	
10	5.847e-13	0.6241	0.72932	-0.8176	0.89070	0.3346
20	3.578e-13	0.7086	0.40402	0.8521	0.66916	0.4126
40	2.135e-13	0.7446	0.22981	0.8140	0.40046	0.7407
80	1.426e-13	0.5827	0.12625	0.8642	0.22431	0.8362
160	7.958e-14	0.8412	0.06672	0.9200	0.11964	0.9067

travelling vortex, $\varepsilon = 10^{-8}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	1.127e-15		0.81442		1.28396	
10	4.865e-16	1.2116	0.89240	-0.1319	0.93102	0.4637
20	1.904e-16	1.3533	0.42158	1.0819	0.67095	0.4726
40	4.591e-16	-1.2699	0.23645	0.8342	0.40191	0.7393
80	1.036e-15	-1.1738	0.13081	0.8541	0.22492	0.8375
160	6.757e-16	0.6162	0.08141	0.6842	0.12018	0.9043

Table 8.4.: EOC of the first order CFDRUS scheme; travelling vortex test.

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	8.070e-03		0.41534		1.12315	
10	7.033e-03	0.1983	0.73944	-0.8321	0.89730	0.3239
20	3.291e-03	1.0956	0.40569	0.8660	0.66916	0.4232
40	2.205e-03	0.5775	0.22985	0.8197	0.39979	0.7431
80	1.370e-03	0.6873	0.12631	0.8637	0.22383	0.8369
160	7.731e-04	0.8249	0.06683	0.9184	0.11936	0.9071

travelling vortex, $\varepsilon = 10^{-6}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	9.010e-13		0.41381		1.12319	
10	5.847e-13	0.6238	0.72932	-0.8176	0.89070	0.3346
20	3.580e-13	0.7076	0.40402	0.8521	0.66916	0.4126
40	2.134e-13	0.7462	0.22981	0.8140	0.40046	0.7407
80	1.421e-13	0.5871	0.12625	0.8642	0.22431	0.8362
160	7.618e-14	0.8993	0.06672	0.9200	0.11964	0.9067

travelling vortex, $\varepsilon = 10^{-8}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	4.705e-17		0.41381		1.12319	
10	4.774e-17	-0.0209	0.72932	-0.8176	0.89070	0.3346
20	3.766e-17	0.3422	0.40402	0.8521	0.66916	0.4126
40	5.430e-17	-0.5281	0.22981	0.8140	0.40046	0.7407
80	6.001e-17	-0.1442	0.12625	0.8642	0.22431	0.8362
160	7.565e-17	-0.3341	0.06672	0.9200	0.11964	0.9067

travelling vortex, $\varepsilon = 10^{-16}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	0.000e+00		0.41381		1.12319	
10	0.000e+00	NaN	0.72932	-0.8176	0.89070	0.3346
20	0.000e+00	NaN	0.40402	0.8521	0.66916	0.4126
40	0.000e+00	NaN	0.22981	0.8140	0.40046	0.7407
80	0.000e+00	NaN	0.12625	0.8642	0.22431	0.8362
160	0.000e+00	NaN	0.06672	0.9200	0.11964	0.9067

Table 8.5.: EOC of the first order RUSELLW scheme; travelling vortex test.

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	7.592e-03		0.39525		1.11277	
10	6.277e-03	0.2743	0.67555	-0.7733	0.89671	0.3114
20	3.127e-03	1.0052	0.40353	0.7434	0.66925	0.4221
40	2.197e-03	0.5097	0.22987	0.8118	0.39983	0.7432
80	1.369e-03	0.6823	0.12632	0.8638	0.22384	0.8369
160	7.731e-04	0.8243	0.06683	0.9185	0.11936	0.9071

travelling vortex, $\varepsilon = 10^{-6}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	7.917e-13		0.39163		1.11182	
10	6.572e-13	0.2687	0.67540	-0.7862	0.89569	0.3119
20	3.430e-13	0.9379	0.40336	0.7437	0.66915	0.4207
40	2.126e-13	0.6904	0.22993	0.8109	0.40050	0.7405
80	1.854e-13	0.1970	0.12627	0.8647	0.22432	0.8362
160	7.371e-14	1.3310	0.06673	0.9202	0.11965	0.9068

travelling vortex, $\varepsilon = 10^{-8}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	1.178e-17		0.39163		1.11182	
10	6.751e-17	-2.5193	0.67540	-0.7862	0.89569	0.3119
20	4.269e-17	0.6612	0.40336	0.7437	0.66915	0.4207
40	4.538e-17	-0.0882	0.22993	0.8109	0.40050	0.7405
80	7.414e-17	-0.7081	0.12627	0.8647	0.22432	0.8362
160	7.294e-17	0.0236	0.06673	0.9202	0.11965	0.9068

travelling vortex, $\varepsilon = 10^{-16}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	0.000e+00		0.39163		1.11182	
10	0.000e+00	NaN	0.67540	-0.7862	0.89569	0.3119
20	0.000e+00	NaN	0.40336	0.7437	0.66915	0.4207
40	0.000e+00	NaN	0.22993	0.8109	0.40050	0.7405
80	0.000e+00	NaN	0.12627	0.8647	0.22432	0.8362
160	0.000e+00	NaN	0.06673	0.9202	0.11965	0.9068

Table 8.6.: EOC of the first order RUSELL scheme; travelling vortex test.

Tables 8.7-8.21 present the EOC of the schemes with second order time and space discretisation, i.e. RK2HLLC, ARSEGRUS, RK2CNEGRUS, BDFEGRUS, ARSRUSELLW, RK2CNRUSELLW, BDFRUSELLW, ARSRUSELL, RK2CNRUSELL, BDFRUSELL schemes. First, the EOC of the RK2HLLC, ARSEGRUS, RK2CNEGRUS and BDFEGRUS schemes for the Froude numbers $\varepsilon \in \{0.8, 0.1, 10^{-2}, 10^{-3}, 10^{-5}\}$ at time $T = 0.1$ are presented in Tables 8.7-8.10. We observe mostly second and sometimes third order convergence rates for the Froude numbers $\varepsilon \in \{0.8, 0.1, 10^{-2}, 10^{-3}\}$, cf. Tables 8.8-8.10. As for the first order EGRUS scheme, the error of the momentum increases for a fixed grid and decreasing Froude number. Also the error of the free surface elevation is below $\Delta x^2 = \Delta y^2$, when its convergence rates breaks down. Comparing the EG numerical flux-based IMEX schemes, one should use either the ARSEGRUS or the BDFEGRUS schemes since the errors of the free surface elevation are lower due to the lack of oscillations in the solutions obtained by the RK2CNEGRUS scheme, cf. Figures 8.16-8.18. We can not decide whether the results of the explicit RK2HLLC scheme are better or worse than the results obtained by the IMEX schemes in general. For $\varepsilon = 0.8$ the results obtained by the RK2HLLC scheme are more accurate than those obtained by the IMEX schemes; however for $\varepsilon \in \{10^{-2}, 10^{-3}\}$ they are less accurate.

If we use the CFD numerical flux for the stiff linear part, we can rewrite the straight approach ARSCFDRUS, RK2CNCFDRUS and BDFCFDRUS schemes into the elliptic approach ARSRUSELLW, RK2CNRUSELLW and BDFRUSELLW schemes. The corresponding EOC are presented in the Tables 8.11-8.15. Using $CFL_u = 0.45$ for the ARSRUSELLW and RK2CNRUSELLW schemes and $CFL_u = 0.3$ for BDFRUSELLW we obtain second order convergence rates for the momentum. Moreover the errors and the convergence rates of the momentum are almost independent of the Froude number ε. For $\varepsilon = \{0.8, 0.1, 0.01\}$ we observe second order convergence rates. For $\varepsilon \in \{10^{-3}, 10^{-5}\}$ these three schemes show different behaviour. For the ARSRUSELLW scheme we observe second order convergence rates until a certain error threshold that is significantly below $\Delta x^2 = \Delta y^2$. Then the convergence seems to break down. The RK2CNRUSELLW scheme seems to develop second order convergence rates for $\varepsilon = 10^{-3}$ and $N \geq 160$. However,for $\varepsilon = 10^{-5}$ there is no second order convergence rate at all. The BDFRUSELLW scheme shows uniform convergence rates for all considered Froude numbers. Thus the BDFRUSELLW scheme gives better results than the ARSRUSELLW and RK2CNRUSELLW schemes. Also the RK2CNRUSELLW scheme provides worse results than the ARSRUSELLW scheme due to the numerical artifacts in the free surface elevation and the associated larger errors. Let us point out that we use a lower CFL_u-number for the BDFRUSELLW scheme to maintain stability. Lowering the CFL_u-number for the ARSRUSELLW and RK2CNRUSELLW schemes does not improve the convergence rates. Also it does not remove the perturbations of the free surface elevation obtained by the RK2CNRUSELLW scheme.

travelling vortex, $\varepsilon = 0.8, CFL = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	2.876e-001		0.33035		0.42825	
20	6.658e-002	2.1111	0.09891	1.7398	0.12724	1.7509
40	1.437e-002	2.2122	0.02510	1.9782	0.03673	1.7925
80	3.318e-003	2.1147	0.00657	1.9333	0.01033	1.8299
160	8.124e-004	2.0299	0.00172	1.9367	0.00280	1.8826

travelling vortex, $\varepsilon = 0.1, CFL = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	2.615e-002		0.91674		0.81096	
20	6.743e-003	1.9553	0.29332	1.6440	0.32008	1.3412
40	1.457e-003	2.2098	0.06259	2.2284	0.07545	2.0848
80	4.104e-004	1.8285	0.01251	2.3233	0.01666	2.1792
160	9.709e-005	2.0796	0.00288	2.1198	0.00405	2.0401

travelling vortex, $\varepsilon = 0.01, CFL = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	9.003e-005		1.79961		1.54227	
20	2.696e-004	-1.5822	1.22506	0.5548	1.20910	0.3511
40	1.362e-004	0.9848	0.39574	1.6302	0.39775	1.6040
80	3.191e-005	2.0939	0.07951	2.3154	0.08089	2.2978
160	9.081e-006	1.8131	0.01453	2.4522	0.01497	2.4344

travelling vortex, $\varepsilon = 10^{-3}, CFL = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	7.283e-007		1.75773		1.59368	
20	7.214e-007	0.0137	1.78600	-0.0230	1.76907	-0.1506
40	3.740e-006	-2.3741	1.39022	0.3614	1.38240	0.3558
80	1.884e-006	0.9896	0.48647	1.5149	0.48658	1.5064
160	6.958e-007	1.4367	0.09820	2.3085	0.09832	2.3072

travelling vortex, $\varepsilon = 10^{-5}, CFL = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	7.283e-11		1.75818		1.59367	
20	7.459e-11	-0.0343	1.73237	0.0213	1.73852	-0.1255
40	7.501e-11	-0.0081	1.74215	-0.0081	1.73615	0.0020
80	7.386e-11	0.0223	1.76908	-0.0221	1.76848	-0.0266
160	3.004e-11	1.2977	0.72428	1.2884	1.16964	0.5964

Table 8.7.: EOC of the second order RK2HLLC scheme; travelling vortex test.

travelling vortex, $\varepsilon = 0.8, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	1.947e-01		0.50154		1.11200	
10	1.848e-01	0.0755	0.40423	0.3112	0.63672	0.8044
20	6.271e-02	1.5587	0.11984	1.7540	0.21824	1.5447
40	1.333e-02	2.2340	0.02866	2.0639	0.05380	2.0201
80	2.785e-03	2.2592	0.00703	2.0267	0.01336	2.0100
160	6.945e-04	2.0036	0.00178	1.9807	0.00344	1.9566

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	2.164e-02		0.55106		1.15913	
10	1.266e-02	0.7742	0.63363	-0.2014	0.79673	0.5409
20	4.702e-03	1.4285	0.18732	1.7582	0.29123	1.4519
40	8.265e-04	2.5081	0.04003	2.2262	0.06912	2.0750
80	1.692e-04	2.2885	0.00895	2.1614	0.01631	2.0831
160	4.307e-05	1.9737	0.00221	2.0208	0.00408	2.0002

travelling vortex, $\varepsilon = 0.01, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	2.245e-04		0.90435		1.42922	
10	2.326e-04	-0.0508	1.66005	-0.8763	1.41262	0.0169
20	2.010e-04	0.2107	0.81981	1.0179	0.84674	0.7384
40	4.279e-05	2.2317	0.20862	1.9744	0.22225	1.9298
80	6.043e-06	2.8238	0.03921	2.4115	0.04366	2.3476
160	8.490e-07	2.8314	0.00711	2.4638	0.00837	2.3830

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	7.102e-06		0.84762		1.42607	
10	6.578e-06	0.1107	1.75797	-1.0524	1.58747	-0.1547
20	1.717e-06	1.9377	1.77070	-0.0104	1.74213	-0.1341
40	2.496e-06	-0.5396	0.99094	0.8375	0.98578	0.8215
80	5.529e-07	2.1745	0.25784	1.9423	0.25918	1.9273
160	7.334e-08	2.9143	0.04853	2.4096	0.04895	2.4045

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	8.266e-10		0.77207		1.37399	
10	8.867e-11	3.2208	1.77541	-1.2014	1.61877	-0.2365
20	7.447e-11	0.2517	1.73522	0.0330	1.74597	-0.1091
40	7.503e-11	-0.0107	1.74235	-0.0059	1.74064	0.0044
80	9.515e-11	-0.3428	1.83111	-0.0717	1.82986	-0.0721
160	2.919e-10	-1.6173	1.23092	0.5730	1.23079	0.5721

Table 8.8.: EOC of the second order ARSEGRUS scheme; travelling vortex test.

travelling vortex, $\varepsilon = 0.8, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	2.607e-01		0.47649		1.10965	
10	1.839e-01	0.5031	0.40390	0.2385	0.63687	0.8010
20	6.030e-02	1.6090	0.11953	1.7566	0.21802	1.5466
40	1.263e-02	2.2555	0.02871	2.0580	0.05371	2.0212
80	2.687e-03	2.2324	0.00704	2.0276	0.01335	2.0084
160	6.739e-04	1.9955	0.00179	1.9794	0.00344	1.9557

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	1.521e-02		0.54495		1.17860	
10	1.591e-02	-0.0644	0.64293	-0.2386	0.80072	0.5577
20	3.702e-03	2.1035	0.18776	1.7758	0.28844	1.4730
40	1.062e-03	1.8019	0.03955	2.2470	0.06893	2.0650
80	1.705e-04	2.6384	0.00892	2.1493	0.01626	2.0838
160	4.267e-05	1.9984	0.00221	2.0131	0.00407	1.9975

travelling vortex, $\varepsilon = 0.01, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	6.195e-04		1.08290		1.68528	
10	1.764e-03	-1.5093	1.68163	-0.6350	1.46433	0.2027
20	3.101e-04	2.5076	0.82018	1.0358	0.84167	0.7989
40	5.476e-05	2.5016	0.20762	1.9820	0.22125	1.9276
80	1.855e-05	1.5618	0.03932	2.4007	0.04367	2.3410
160	4.526e-06	2.0352	0.00713	2.4634	0.00840	2.3786

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	8.241e-06		1.33334		1.98019	
10	1.038e-04	-3.6550	1.51762	-0.1868	1.74937	0.1788
20	2.849e-05	1.8656	1.78786	-0.2364	1.74914	0.0002
40	5.113e-06	2.4781	0.99114	0.8511	0.98419	0.8296
80	1.088e-06	2.2329	0.25782	1.9428	0.25893	1.9264
160	3.319e-07	1.7126	0.04852	2.4097	0.04894	2.4036

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	8.531e-10		1.37485		2.03019	
10	1.517e-08	-4.1519	0.73978	0.8941	1.46632	0.4694
20	5.725e-08	-1.9166	1.30225	-0.8159	1.67391	-0.1910
40	4.051e-08	0.4991	1.91009	-0.5526	1.90078	-0.1834
80	4.914e-09	3.0432	1.83051	0.0614	1.82977	0.0549
160	7.669e-10	2.6800	1.23098	0.5724	1.23074	0.5721

Table 8.9.: EOC of the second order RK2CNEGRUS scheme; travelling vortex test.

travelling vortex, $\varepsilon = 0.8, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	2.607e-01		0.47649		1.10965	
10	1.862e-01	0.4851	0.43370	0.1357	0.69520	0.6746
20	6.117e-02	1.6063	0.13667	1.6660	0.24533	1.5027
40	1.280e-02	2.2564	0.03404	2.0054	0.06503	1.9156
80	2.791e-03	2.1977	0.00880	1.9521	0.01737	1.9046
160	1.049e-03	1.4118	0.00370	1.2490	0.00469	1.8884

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	1.521e-02		0.54495		1.17860	
10	1.902e-02	-0.3222	0.64908	-0.2523	0.84851	0.4741
20	3.099e-03	2.6178	0.19671	1.7223	0.31638	1.4233
40	5.960e-04	2.3781	0.04254	2.2092	0.07783	2.0232
80	1.702e-04	1.8085	0.00993	2.0996	0.01977	1.9770
160	4.030e-05	2.0780	0.00252	1.9795	0.00519	1.9304

travelling vortex, $\varepsilon = 0.01, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	6.195e-04		1.08290		1.68528	
10	3.430e-04	0.8530	1.84486	-0.7686	1.53483	0.1349
20	2.029e-04	0.7577	0.84033	1.1345	0.88210	0.7991
40	4.271e-05	2.2479	0.21075	1.9954	0.22753	1.9549
80	5.768e-06	2.8882	0.03953	2.4146	0.04543	2.3244
160	7.211e-07	3.0000	0.00725	2.4460	0.00922	2.3002

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	8.241e-06		1.33334		1.98019	
10	2.533e-05	-1.6201	2.07271	-0.6365	1.87940	0.0754
20	2.904e-06	3.1249	1.87484	0.1447	1.86664	0.0098
40	2.485e-06	0.2245	0.99945	0.9076	1.00094	0.8991
80	5.507e-07	2.1742	0.25847	1.9511	0.26046	1.9422
160	7.324e-08	2.9105	0.04859	2.4113	0.04912	2.4066

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, CFL_g = 0.01, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	8.531e-10		1.37485		2.03019	
10	3.993e-09	-2.2265	1.77743	-0.3705	1.65467	0.2951
20	1.573e-10	4.6655	1.73344	0.0362	1.75130	-0.0819
40	7.507e-11	1.0674	1.74771	-0.0118	1.74695	0.0036
80	8.761e-11	-0.2228	1.83539	-0.0706	1.83462	-0.0706
160	2.917e-10	-1.7354	1.23163	0.5755	1.23193	0.5746

Table 8.10.: EOC of the second order BDFEGRUS scheme; travelling vortex test.

travelling vortex, $\varepsilon = 0.8, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.929e-01		0.33509		0.64414	
20	4.885e-02	1.9812	0.10303	1.7016	0.21995	1.5502
40	1.243e-02	1.9743	0.02671	1.9476	0.05282	2.0580
80	3.231e-03	1.9442	0.00688	1.9567	0.01307	2.0152
160	8.153e-04	1.9866	0.00177	1.9588	0.00340	1.9429

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.464e-02		0.36885		0.67845	
20	3.136e-03	2.2226	0.12028	1.6166	0.22728	1.5778
40	1.353e-03	1.2130	0.02768	2.1196	0.05440	2.0627
80	3.290e-04	2.0398	0.00747	1.8901	0.01398	1.9605
160	8.946e-05	1.8788	0.00198	1.9143	0.00370	1.9165

travelling vortex, $\varepsilon = 0.01, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	4.915e-05		0.34316		0.68431	
20	1.806e-05	1.4440	0.10971	1.6452	0.22689	1.5927
40	3.939e-06	2.1970	0.02661	2.0437	0.05297	2.0988
80	1.303e-06	1.5957	0.00685	1.9580	0.01333	1.9904
160	3.251e-07	2.0032	0.00178	1.9401	0.00350	1.9310

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	4.866e-07		0.34309		0.68414	
20	1.120e-07	2.1192	0.10957	1.6468	0.22691	1.5922
40	2.223e-08	2.3330	0.02658	2.0436	0.05295	2.0995
80	6.223e-09	1.8369	0.00685	1.9560	0.01332	1.9911
160	2.506e-09	1.3123	0.00178	1.9480	0.00349	1.9340

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	4.873e-11		0.34309		0.68414	
20	1.126e-11	2.1139	0.10957	1.6467	0.22691	1.5922
40	1.785e-12	2.6572	0.02658	2.0435	0.05295	2.0995
80	5.225e-13	1.7721	0.00685	1.9560	0.01332	1.9911
160	3.159e-13	0.7258	0.00178	1.9479	0.00349	1.9341

Table 8.11.: EOC of the second order ARSRUSELLW scheme; travelling vortex test.

travelling vortex, $\varepsilon = 0.8, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.954e-01		0.33885		0.64292	
20	4.622e-02	2.0798	0.10198	1.7324	0.21957	1.5500
40	1.216e-02	1.9268	0.02655	1.9416	0.05276	2.0571
80	3.256e-03	1.9006	0.00684	1.9559	0.01306	2.0144
160	8.339e-04	1.9651	0.00176	1.9597	0.00340	1.9419

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	2.848e-02		0.35560		0.70546	
20	4.676e-03	2.6066	0.12249	1.5376	0.23304	1.5980
40	1.564e-03	1.5803	0.03013	2.0232	0.05594	2.0587
80	4.282e-04	1.8685	0.00732	2.0418	0.01391	2.0076
160	9.413e-05	2.1857	0.00198	1.8861	0.00370	1.9119

travelling vortex, $\varepsilon = 0.01, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	7.548e-04		0.38808		0.69950	
20	5.807e-04	0.3784	0.11221	1.7902	0.23128	1.5967
40	1.339e-04	2.1162	0.02854	1.9751	0.05394	2.1001
80	4.482e-05	1.5794	0.00759	1.9101	0.01396	1.9500
160	1.234e-05	1.8604	0.00185	2.0350	0.00360	1.9560

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	7.661e-06		0.39026		0.70136	
20	9.588e-06	-0.3236	0.10253	1.9283	0.23144	1.5995
40	8.474e-06	0.1781	0.02578	1.9920	0.05373	2.1069
80	3.002e-06	1.4970	0.00721	1.8373	0.01353	1.9890
160	7.718e-07	1.9598	0.00174	2.0506	0.00355	1.9314

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	7.663e-10		0.39029		0.70138	
20	9.637e-10	-0.3308	0.10248	1.9291	0.23158	1.5987
40	1.097e-09	-0.1875	0.02526	2.0205	0.05408	2.0983
80	4.086e-10	1.4253	0.00791	1.6751	0.01390	1.9599
160	3.106e-10	0.3960	0.00204	1.9548	0.00365	1.9297

Table 8.12.: EOC of the second order RK2CNRUSELLW scheme; travelling vortex test.

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.482e-02		0.37809		0.74477	
20	2.003e-03	2.8874	0.13124	1.5265	0.25215	1.5625
40	4.508e-04	2.1515	0.03174	2.0481	0.06505	1.9547
80	2.700e-04	0.7396	0.00857	1.8884	0.01772	1.8764
160	1.678e-04	0.6859	0.00300	1.5131	0.00505	1.8101

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	7.538e-07		0.37519		0.74452	
20	9.615e-08	2.9709	0.12417	1.5953	0.25370	1.5532
40	1.841e-08	2.3851	0.03131	1.9876	0.06476	1.9699
80	3.833e-09	2.2636	0.00826	1.9223	0.01739	1.8970
160	2.918e-07	-6.2503	0.01065	-0.3661	0.01116	0.6403

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	7.537e-11		0.37519		0.74452	
20	9.616e-12	2.9705	0.12417	1.5953	0.25370	1.5532
40	1.838e-12	2.3873	0.03131	1.9876	0.06476	1.9699
80	3.838e-13	2.2596	0.00826	1.9223	0.01739	1.8970
160	3.022e-11	-6.2989	0.01070	-0.3733	0.01119	0.6353

travelling vortex, $\varepsilon = 10^{-8}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.548e-16		0.37519		0.74452	
20	1.461e-16	0.0837	0.12417	1.5953	0.25370	1.5532
40	1.849e-16	-0.3405	0.03131	1.9876	0.06476	1.9699
80	3.074e-16	-0.7330	0.00826	1.9223	0.01739	1.8970
160	9.058e-17	1.7629	0.01070	-0.3733	0.01119	0.6353

Table 8.13.: EOC of the second order BDFRUSELLW scheme; travelling vortex test.

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.4, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	9.304e-05		0.37756		0.74642	
20	9.500e-06	3.2918	0.12174	1.6329	0.24975	1.5795
40	1.668e-06	2.5098	0.03003	2.0191	0.06129	2.0266
80	3.578e-07	2.2210	0.00785	1.9357	0.01625	1.9153
160	3.260e-07	0.1341	0.00215	1.8680	0.00437	1.8938

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.4, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	9.183e-07		0.37749		0.74651	
20	9.242e-08	3.3126	0.12163	1.6340	0.24973	1.5798
40	1.666e-08	2.4715	0.03003	2.0178	0.06130	2.0265
80	3.579e-09	2.2192	0.00785	1.9357	0.01625	1.9153
160	6.942e-09	-0.9559	0.00233	1.7551	0.00443	1.8762

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.4, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	9.182e-11		0.37749		0.74651	
20	9.244e-12	3.3122	0.12162	1.6340	0.24973	1.5798
40	1.666e-12	2.4719	0.03003	2.0178	0.06130	2.0265
80	3.579e-13	2.2191	0.00785	1.9357	0.01625	1.9153
160	7.040e-13	-0.9761	0.00233	1.7518	0.00443	1.8755

Table 8.14.: EOC of the second order BDFRUSELLW scheme; travelling vortex test.

Let us now consider the elliptic approach ARSRUSELL, RK2CNRUSELL and BDFRUSELL schemes. The corresponding EOC numbers are presented in Tables 8.16-8.21. We observe that the ARSRUSELL scheme shows second order convergence for the Froude numbers $\varepsilon \in \{0.8, 0.1\}$. For lower Froude numbers it provides instabilities for $N = 160$, cf. Table 8.16. Changing the CFL_u-number also does not improve the results, cf. Table 8.17. The RK2CNRUSELL scheme shows second order convergence rates of the momentum, cf. Table 8.18. Moreover the convergence rates and errors of the momentum behave uniform with respect to the Froude number. The free surface elevation converges with second order for $\varepsilon \in \{0.8, 0.1, 0.01\}$. For smaller Froude numbers, i.e. $\varepsilon \in \{10^{-3}, 10^{-5}\}$, the convergence of the free surface elevation breaks down. However the error is then below $\Delta x^2 = \Delta y^2$. Lowering the CFL_u-number does not improve the convergence of the free surface elevation, cf. Table 8.19. For the BDFRUSELL scheme we have to lower the CFL_u-number to 0.3 to obtain stable results for low Froude numbers. Then we obtain second order convergence rates uniform with respect to the Froude number for all components, cf. Tables 8.20, 8.21. Also the errors of the momentum behave uniformly with respect to the Froude number. Thus the BDFRUSELL scheme gives better results than the RK2CNRUSELL scheme.

travelling vortex, $\varepsilon = 0.8, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.930e-01		0.33519		0.64771	
20	5.035e-02	1.9389	0.10805	1.6333	0.22831	1.5043
40	1.257e-02	2.0020	0.02783	1.9570	0.05553	2.0398
80	3.244e-03	1.9542	0.00722	1.9462	0.01411	1.9765
160	8.148e-04	1.9932	0.00186	1.9561	0.00371	1.9281

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.293e-02		0.35780		0.68710	
20	1.662e-03	2.9596	0.12116	1.5622	0.23440	1.5515
40	7.146e-04	1.2177	0.02897	2.0641	0.05645	2.0540
80	3.223e-04	1.1490	0.00743	1.9638	0.01463	1.9479
160	8.022e-05	2.0062	0.00201	1.8850	0.00395	1.8883

travelling vortex, $\varepsilon = 0.01, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	4.152e-05		0.34862		0.68990	
20	9.364e-06	2.1488	0.11410	1.6114	0.23531	1.5518
40	1.646e-06	2.5082	0.02777	2.0386	0.05563	2.0805
80	3.137e-07	2.3913	0.00719	1.9487	0.01435	1.9549
160	8.110e-08	1.9516	0.00187	1.9456	0.00379	1.9197

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	4.198e-07		0.34844		0.68987	
20	9.355e-08	2.1659	0.11412	1.6103	0.23531	1.5518
40	1.643e-08	2.5098	0.02777	2.0388	0.05563	2.0805
80	3.109e-09	2.4016	0.00720	1.9485	0.01435	1.9550
160	7.819e-10	1.9911	0.00187	1.9449	0.00379	1.9200

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	4.199e-11		0.34844		0.68987	
20	9.355e-12	2.1661	0.11412	1.6103	0.23531	1.5518
40	1.642e-12	2.5099	0.02777	2.0388	0.05563	2.0805
80	3.105e-13	2.4030	0.00720	1.9485	0.01435	1.9550
160	7.811e-14	1.9912	0.00187	1.9449	0.00379	1.9200

m_2

Table 8.15.: EOC of the second order BDFRUSELLW scheme; travelling vortex test;

travelling vortex, $\varepsilon = 0.8, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	2.159e-01		0.49039		1.11127	
10	2.009e-01	0.1038	0.34001	0.5284	0.64578	0.7831
20	4.754e-02	2.0793	0.10444	1.7029	0.22319	1.5327
40	9.492e-03	2.3242	0.02690	1.9567	0.05408	2.0450
80	2.278e-03	2.0590	0.00677	1.9903	0.01327	2.0270
160	5.932e-04	1.9412	0.00172	1.9754	0.00342	1.9567

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	2.518e-02		0.38113		1.13190	
10	5.089e-02	-1.0152	0.47132	-0.3064	0.68632	0.7218
20	1.015e-02	2.3263	0.10023	2.2333	0.23418	1.5513
40	1.593e-03	2.6716	0.02978	1.7509	0.05750	2.0261
80	3.153e-04	2.3365	0.00702	2.0846	0.01419	2.0184
160	6.094e-05	2.3713	0.00180	1.9618	0.00363	1.9660

travelling vortex, $\varepsilon = 0.01, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	2.660e-04		0.35966		1.13486	
10	1.096e-03	-2.0424	0.52784	-0.5535	0.70969	0.6773
20	1.111e-03	-0.0205	0.15908	1.7304	0.24571	1.5302
40	3.249e-04	1.7744	0.04944	1.6860	0.06729	1.8686
80	9.680e-05	1.7468	0.01670	1.5659	0.02244	1.5841
160	3.157e-04	-1.7055	0.11606	-2.7971	0.10387	-2.2105

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.148e-05		0.52906		0.70972	
20	1.625e-05	-0.5017	0.17466	1.5989	0.25590	1.4717
40	4.739e-05	-1.5441	0.05569	1.6491	0.07342	1.8013
80	1.219e-05	1.9585	0.02214	1.3310	0.02633	1.4795
160	4.331e-06	1.4935	0.12793	-2.5309	0.11251	-2.0952

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
5	2.663e-10		0.35978		1.13493	
10	1.148e-09	-2.1084	0.52908	-0.5563	0.70972	0.6773
20	1.632e-09	-0.5071	0.17489	1.5971	0.25607	1.4707
40	9.323e-09	-2.5142	0.06198	1.4965	0.07697	1.7342
80	1.857e-09	2.3279	0.02376	1.3835	0.02750	1.4848
160	5.715e-10	1.7001	0.12829	-2.4331	0.11284	-2.0368

m_2

Table 8.16.: EOC of the second order ARSRUSELL scheme; travelling vortex test.

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.7, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.019e-04		0.76571		1.05021	
20	2.081e-05	2.2920	0.18007	2.0882	0.29194	1.8469
40	6.386e-06	1.7041	0.06059	1.5714	0.08153	1.8402
80	1.262e-05	-0.9824	0.02666	1.1842	0.03435	1.2473
160	9.845e-06	0.3580	0.24672	-3.2099	0.27377	-2.9948

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	2.095e-05		0.42458		0.87083	
20	5.135e-05	-1.2933	0.18532	1.1960	0.25087	1.7955
40	1.324e-05	1.9555	0.06566	1.4969	0.07670	1.7097
80	1.176e-05	0.1706	0.03919	0.7445	0.04693	0.7087
160	2.595e-05	-1.1413	0.88972	-4.5048	1.06944	-4.5102

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.1, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.281e-04		0.94029		0.84635	
20	5.111e-05	1.3260	0.37093	1.3419	0.35547	1.2515
40	1.092e-04	-1.0953	2.33729	-2.6556	2.90459	-3.0305
80	1.457e+00	-13.7038	492.33117	-7.7186	498.21039	-7.4223
160	1.169e+00	0.3180	592.53502	-0.2673	593.99267	-0.2537

Table 8.17.: EOC of the second order ARSRUSELL scheme; travelling vortex test.

travelling vortex, $\varepsilon = 0.8, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	2.181e-01		0.34621		0.64136	
20	5.839e-02	1.9014	0.10677	1.6972	0.22316	1.5231
40	1.397e-02	2.0632	0.02790	1.9361	0.05534	2.0118
80	3.762e-03	1.8929	0.00714	1.9657	0.01383	2.0007
160	9.866e-04	1.9309	0.00184	1.9597	0.00357	1.9518

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.541e-02		0.37236		0.74939	
20	6.640e-03	1.2144	0.11131	1.7422	0.23813	1.6540
40	1.399e-03	2.2469	0.02888	1.9464	0.06036	1.9800
80	3.610e-04	1.9543	0.00761	1.9239	0.01523	1.9869
160	1.375e-04	1.3928	0.00210	1.8574	0.00400	1.9273

travelling vortex, $\varepsilon = 0.01, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	2.811e-04		0.37048		0.74110	
20	1.333e-04	1.0763	0.10799	1.7784	0.25167	1.5581
40	1.263e-04	0.0785	0.02948	1.8730	0.06041	2.0588
80	2.880e-05	2.1323	0.00838	1.8146	0.01637	1.8837
160	8.243e-06	1.8048	0.00231	1.8597	0.00434	1.9138

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	2.833e-06		0.37046		0.74103	
20	1.486e-06	0.9308	0.10847	1.7720	0.25117	1.5609
40	3.721e-06	-1.3240	0.02844	1.9311	0.06342	1.9856
80	5.990e-07	2.6348	0.00847	1.7479	0.01598	1.9883
160	3.344e-07	0.8412	0.00239	1.8261	0.00435	1.8780

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	2.833e-10		0.37046		0.74103	
20	1.488e-10	0.9293	0.10847	1.7719	0.25116	1.5609
40	3.893e-10	-1.3878	0.02847	1.9300	0.06346	1.9847
80	8.968e-11	2.1180	0.00852	1.7404	0.01613	1.9762
160	8.579e-11	0.0639	0.00239	1.8308	0.00438	1.8798

m_2

Table 8.18.: EOC of the second order RK2CNRUSELL scheme; travelling vortex test.

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.2, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	6.077e-06		0.35237		0.69582	
20	4.255e-06	0.5143	0.11850	1.5722	0.24263	1.5200
40	2.313e-06	0.8792	0.03219	1.8800	0.05993	2.0174
80	8.716e-07	1.4081	0.01010	1.6729	0.01608	1.8979
160	9.863e-07	-0.1782	0.01428	-0.4999	0.01276	0.3339

Table 8.19.: EOC of the second order RK2CNRUSELL scheme; travelling vortex test.

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.541e-02		0.37236		0.74939	
20	6.640e-03	1.2144	0.11131	1.7422	0.23813	1.6540
40	1.399e-03	2.2469	0.02888	1.9464	0.06036	1.9800
80	3.610e-04	1.9543	0.00761	1.9239	0.01523	1.9869
160	1.375e-04	1.3928	0.00210	1.8574	0.00400	1.9273

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	2.269e-06		0.47493		0.80051	
20	3.238e-07	2.8091	0.12848	1.8862	0.28160	1.5073
40	3.187e-07	0.0226	0.03703	1.7947	0.07626	1.8846
80	3.502e-08	3.1859	0.01103	1.7480	0.02103	1.8584
160	1.440e-08	1.2822	0.00696	0.6631	0.00599	1.8129

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	2.269e-10		0.47493		0.80051	
20	3.237e-11	2.8096	0.12848	1.8862	0.28160	1.5073
40	3.217e-11	0.0090	0.03703	1.7947	0.07626	1.8846
80	3.518e-12	3.1930	0.01103	1.7480	0.02103	1.8584
160	1.442e-12	1.2868	0.00696	0.6631	0.00599	1.8130

Table 8.20.: EOC of the second order BDFRUSELL scheme; travelling vortex test.

travelling vortex, $\varepsilon = 0.8, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	2.315e-01		0.33644		0.64706	
20	8.631e-02	1.4235	0.11042	1.6073	0.23367	1.4694
40	2.668e-02	1.6939	0.03087	1.8387	0.06145	1.9271
80	7.687e-03	1.7951	0.00865	1.8355	0.01643	1.9033
160	2.049e-03	1.9077	0.00232	1.8977	0.00435	1.9184

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.419e-02		0.40600		0.74278	
20	2.756e-03	2.3637	0.11675	1.7981	0.27062	1.4567
40	9.748e-04	1.4995	0.03070	1.9269	0.06784	1.9960
80	5.378e-04	0.8579	0.00992	1.6297	0.01889	1.8448
160	1.967e-04	1.4512	0.00272	1.8677	0.00494	1.9338

travelling vortex, $\varepsilon = 0.01, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.845e-04		0.43724		0.75274	
20	4.104e-05	2.1685	0.11481	1.9292	0.26617	1.4998
40	5.772e-06	2.8300	0.03175	1.8544	0.06822	1.9642
80	1.909e-06	1.5960	0.00912	1.8003	0.01813	1.9115
160	3.763e-07	2.3429	0.00251	1.8604	0.00479	1.9206

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.827e-06		0.43773		0.75284	
20	4.163e-07	2.1340	0.11454	1.9342	0.26600	1.5009
40	5.795e-08	2.8449	0.03166	1.8550	0.06808	1.9660
80	2.300e-08	1.3330	0.00908	1.8018	0.01809	1.9121
160	4.088e-09	2.4922	0.00250	1.8581	0.00478	1.9192

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.827e-10		0.43773		0.75284	
20	4.164e-11	2.1336	0.11453	1.9343	0.26600	1.5009
40	5.795e-12	2.8451	0.03166	1.8550	0.06808	1.9661
80	2.310e-12	1.3270	0.00908	1.8018	0.01809	1.9121
160	4.097e-13	2.4951	0.00250	1.8582	0.00478	1.9193

Table 8.21.: EOC of the second order BDFRUSELL scheme; travelling vortex test.

8.2.2. Asymptotic preserving property

In Chapter 7 we have discussed the asymptotic preserving property for the fully discrete IMEX R-K and IMEX multi-step schemes. Moreover we have proved it under certain assumptions, thus $\nabla_h \mathbf{Z} = \mathcal{O}(\varepsilon^2)$ and $\nabla_h \cdot \mathbf{M} = \mathcal{O}(\varepsilon^2)$ or $\mathcal{O}(\varepsilon)$ depending on if the CFD or EG numerical fluxes are used for the stiff, linear part, cf. Lemmas 7.1.7, 7.1.12. The aim of this section is to verify the results of these lemmas numerically. Further we examine if these constraints also hold for a more general setting than covered by Lemmas 7.1.7, 7.1.12. This means by applying the ARSEGRUS, RK2CNEGRUS, BDFEGRUS schemes or using a non-constant bottom topography.

Let us recall that the discrete gradient and divergence operators $\nabla_h, \nabla_h \cdot$ depend on the used spatial discretisation. More precisely we use the following differences to approximate discrete gradient and divergence. Let us point out that we did not calculate an algebraic expression for the discrete gradient and divergence used by the ARSEGRUS, RK2CNEGRUS, BDFEGRUS schemes. Instead, we have used the coefficients from our C++ implementation, cf. (8.11), (8.12).

- EGRUS scheme:

$$(z_x)_{ij} = \frac{1}{24\Delta x}\left[10(z_{i,j+1} - z_{i,j-1}) + z_{i+1,j+1} - z_{i+1,j-1} + z_{i-1,j+1} - z_{i-1,j-1}\right] \tag{8.9}$$

$$\begin{aligned}((m_1)_x)_{ij} = \frac{1}{24\Delta x}\Bigg[&\left(10 - \frac{4}{\pi}\right)\left((m_1)_{i,j+1} - (m_1)_{i,j-1}\right) \\ &+ \left(1 + \frac{2}{\pi}\right)\left((m_1)_{i+1,j+1} - (m_1)_{i+1,j-1} + (m_1)_{i-1,j+1} - (m_1)_{i-1,j-1}\right)\Bigg]\end{aligned}$$

  Note that we have described the IMEX finite volume schemes by means of the matrix operator A from (6.109), (6.111). The discrete differential operators (8.9) coincide with (6.109), (6.111).

- RUSELL, RUSELLW, ARSRUSELL, RK2CNRUSELL, BDF2RUSELL schemes:

$$(z_x)_{ij} = \frac{z_{i+1,j} - z_{i-1,j}}{2\Delta x} \tag{8.10}$$

  $(m_1)_x$ is computed analogously.

- ARSEGRUS, RK2CNEGRUS, BDFEGRUS schemes:

$$
\begin{aligned}
(z_x)_{ij} \approx \frac{1}{10\Delta x} \Big[& 6.269(z_{i,j+1} - z_{i,j-1}) \\
& - 0.893(z_{i+2,j} - z_{i-2,j}) \\
& + 0.763(z_{i+1,j+1} - z_{i-1,j+1} + z_{i+1,j-1} - z_{i-1,j-1}) \\
& - 0.168(z_{i+2,j+1} - z_{i-2,j+1} + z_{i+2,j-1} - z_{i-2,j-1}) \\
& - 0.169(z_{i+1,j+2} - z_{i-1,j+2} + z_{i+1,j-2} - z_{i-1,j-2}) \Big]
\end{aligned} \tag{8.11}
$$

$$
\begin{aligned}
((m_1)_x)_{ij} \approx \frac{1}{10\Delta x} \big[& 6.429((m_1)_{i+1,j} - (m_1)_{i-1,j}) \\
& - 1.028\,((m_1)_{i+2,j} - (m_1)_{i-2,j}) \\
& + 0.622\,((m_1)_{i+1,j+1} - (m_1)_{i-1,j+1} + (m_1)_{i+1,j-1} - (m_1)_{i-1,j-1}) \\
& - 0.103\,((m_1)_{i+2,j+1} - (m_1)_{i-2,j+1} + (m_1)_{i+2,j-1} - (m_1)_{i-2,j-1}) \\
& - 0.103\,((m_1)_{i+1,j+2} - (m_1)_{i-1,j+2} + (m_1)_{i+1,j-2} - (m_1)_{i-1,j-2}) \big]
\end{aligned} \tag{8.12}
$$

- ARSRUSELLW, RK2CNRUSELLW, BDFRUSELLW schemes:

$$
(z_x)_{ij} = \frac{6(z_{i+1,j} - z_{i-1,j}) - (z_{i+2,j} - z_{i-2,j})}{8\Delta x} \tag{8.13}
$$

  $(m_1)_x$ is computed analogously.

Tables 8.22-8.33 contain the l^1- and l^∞-norms of $\nabla_h \mathbf{Z}$ and $\nabla_h \cdot \mathbf{M}$, where the numerical solutions have been computed on $(N \times N)$-grids with $N \in \{10, 20, 40.80, 160\}$. To this end, the discrete operators ∇_h, $\nabla_h \cdot$ from (8.9)-(8.13) have been applied. They indicate that all schemes satisfy the constraint $\nabla_h \mathbf{Z} = \mathcal{O}(\varepsilon^2)$ on the free surface elevation. Moreover the schemes RUSELLW, ARSRUSELLW, BDFRUSELLW additionally satisfy the divergence constraint $\nabla_h \cdot \mathbf{M} = \mathcal{O}(\varepsilon^2)$ and the schemes EGRUS, ARSEGRUS, BDFEGRUS satisfy $\nabla_h \cdot \mathbf{M} = \mathcal{O}(\varepsilon)$. These results are in accordance with the results of the asymptotic analysis in Chapter 7 - in particular Theorems 7.1.7, 7.1.9,7.1.11 ,7.1.12, 7.2.1.

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.1851	0.3095	0.9411	3.4292
20	0.2103	0.4956	0.4716	2.8243
40	0.1771	0.6093	0.2129	1.9470
80	0.1469	0.8508	0.1082	1.5343
160	0.1285	1.0414	0.0802	1.1165

travelling vortex, $\varepsilon = 10^{-2}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.507e-03	2.620e-03	0.0752	0.2580
20	1.855e-03	4.685e-03	0.0441	0.1862
40	2.756e-03	7.241e-03	0.0339	0.1506
80	3.555e-03	9.309e-03	0.0225	0.1118
160	3.238e-03	8.540e-03	0.0102	0.0581

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	2.163e-06	4.033e-06	1.012e-03	4.226e-03
20	1.480e-06	6.264e-06	4.889e-04	3.831e-03
40	1.936e-06	1.041e-05	3.437e-04	3.383e-03
80	2.162e-06	1.206e-05	1.891e-04	1.997e-03
160	5.164e-06	1.412e-05	1.615e-04	1.156e-03

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.712e-11	4.027e-11	7.826e-06	5.213e-05
20	1.490e-10	6.043e-10	5.098e-06	3.798e-05
40	1.948e-10	1.038e-09	3.489e-06	3.367e-05
80	2.064e-10	1.192e-09	1.888e-06	1.971e-05
160	2.151e-10	1.280e-09	1.000e-06	1.066e-05

Table 8.22.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the EGRUS scheme.

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.0930	0.4053	0.2483	1.3670
20	0.0659	0.5794	0.0731	1.4973
40	0.0796	0.8765	0.0430	0.7658
80	0.0897	1.0561	0.0380	0.7671
160	0.0960	1.1696	0.0359	0.7900

travelling vortex, $\varepsilon = 10^{-2}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	5.950e-04	3.871e-03	3.811e-03	1.831e-02
20	6.122e-04	5.512e-03	8.770e-04	1.976e-02
40	7.961e-04	8.796e-03	3.490e-04	7.553e-03
80	8.945e-04	1.056e-02	7.803e-04	2.273e-02
160	9.565e-04	1.168e-02	7.946e-04	2.139e-02

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	5.922e-06	3.880e-05	3.831e-05	1.837e-04
20	6.114e-06	5.505e-05	8.883e-06	1.991e-04
40	7.961e-06	8.796e-05	3.494e-06	7.556e-05
80	8.882e-06	1.047e-04	9.944e-05	2.460e-03
160	9.546e-06	1.166e-04	1.691e-05	4.150e-04

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	5.922e-10	3.880e-09	3.831e-09	1.837e-08
20	6.114e-10	5.505e-09	8.884e-10	1.992e-08
40	7.961e-10	8.796e-09	3.494e-10	7.556e-09
80	8.848e-10	1.044e-08	1.322e-08	3.145e-07
160	9.545e-10	1.166e-08	1.740e-09	4.230e-08

travelling vortex, $\varepsilon = 10^{-6}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	5.922e-12	3.880e-11	3.831e-11	1.837e-10
20	6.114e-12	5.505e-11	8.885e-12	1.992e-10
40	7.961e-12	8.796e-11	3.496e-12	7.550e-11
80	8.848e-12	1.044e-10	1.322e-10	3.145e-09
160	9.545e-12	1.166e-10	1.741e-11	4.232e-10

Table 8.23.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the RUSELLW scheme.

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.1118	0.3684	1.6416	8.0961
20	0.0764	0.7261	0.3383	4.8550
40	0.0797	0.8716	0.0867	1.5586
80	0.0896	1.0564	0.0404	0.7769
160	0.0960	1.1697	0.0360	0.7883

travelling vortex, $\varepsilon = 10^{-2}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.0011	0.0036	1.6197	7.7476
20	0.0008	0.0075	0.2999	4.6060
40	0.0008	0.0087	0.0630	1.6190
80	0.0009	0.0112	0.0088	0.3228
160	0.0010	0.0118	0.0013	0.0451

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.066e-05	3.590e-05	1.620e+00	7.745e+00
20	7.724e-06	7.458e-05	2.996e-01	4.602e+00
40	7.980e-06	8.746e-05	6.297e-02	1.620e+00
80	1.220e-05	1.608e-04	4.096e-03	1.647e-01
160	9.626e-06	1.187e-04	5.972e-04	4.224e-02

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.066e-09	3.590e-09	1.620e+00	7.745e+00
20	7.724e-10	7.458e-09	2.996e-01	4.602e+00
40	7.980e-10	8.746e-09	6.297e-02	1.620e+00
80	1.303e-09	1.685e-08	3.664e-03	1.532e-01
160	9.628e-10	1.187e-08	5.911e-04	4.223e-02

travelling vortex, $\varepsilon = 10^{-6}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.066e-11	3.590e-11	1.620e+00	7.745e+00
20	7.724e-12	7.458e-11	2.996e-01	4.602e+00
40	7.980e-12	8.746e-11	6.297e-02	1.620e+00
80	1.303e-11	1.685e-10	3.664e-03	1.532e-01
160	9.628e-12	1.187e-10	5.911e-04	4.223e-02

Table 8.24.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the RUSELL scheme.

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.2231	0.5625	0.8875	3.2769
20	0.1944	1.0259	0.2334	1.5500
40	0.1237	1.2727	0.1017	1.0586
80	0.1088	1.2967	0.0534	0.8398
160	0.1048	1.3005	0.0370	0.7980

travelling vortex, $\varepsilon = 10^{-2}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	4.573e-03	6.618e-03	1.161e-01	4.675e-01
20	5.470e-03	1.319e-02	4.142e-02	2.038e-01
40	2.093e-03	1.149e-02	7.140e-03	6.258e-02
80	1.210e-03	1.281e-02	1.447e-03	2.188e-02
160	1.064e-03	1.298e-02	5.650e-04	7.621e-03

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.437e-04	2.502e-04	3.900e-02	1.232e-01
20	3.037e-05	5.103e-05	1.181e-03	6.274e-03
40	6.738e-05	1.687e-04	7.814e-04	3.648e-03
80	2.471e-05	1.101e-04	1.450e-04	1.494e-03
160	1.265e-05	1.277e-04	5.799e-05	9.378e-04

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	9.323e-10	1.569e-09	1.561e-05	6.034e-05
20	3.573e-10	1.674e-09	4.878e-06	2.100e-05
40	3.885e-10	1.817e-09	1.586e-06	1.406e-05
80	1.306e-09	2.293e-09	5.111e-07	8.126e-06
160	7.321e-09	1.797e-08	4.905e-07	4.045e-06

Table 8.25.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the ARSEGRUS scheme.

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.3220	0.6625	1.3264	4.4604
20	0.1740	0.9586	0.4515	2.5391
40	0.1321	1.2811	0.1115	1.0931
80	0.1091	1.3004	0.0523	0.8231
160	0.1048	1.3015	0.0371	0.7969

travelling vortex, $\varepsilon = 10^{-2}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	4.530e-02	9.170e-02	1.640e+00	8.527e+00
20	1.100e-02	4.669e-02	3.814e-01	2.739e+00
40	3.154e-03	2.080e-02	1.034e-01	1.642e+00
80	1.806e-03	1.331e-02	3.456e-02	3.662e-01
160	1.257e-03	1.322e-02	9.134e-03	1.221e-01

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	2.444e-03	4.236e-03	1.567e+00	7.536e+00
20	1.128e-03	4.422e-03	4.177e-01	4.132e+00
40	2.416e-04	1.779e-03	1.052e-01	2.305e+00
80	6.842e-05	7.696e-04	2.560e-02	7.304e-01
160	2.948e-05	5.017e-04	9.445e-03	4.730e-01

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	3.566e-07	5.925e-07	1.392e+00	1.011e+01
20	1.583e-06	4.112e-06	5.190e-01	4.175e+00
40	1.450e-06	4.349e-06	1.192e-01	2.558e+00
80	3.005e-07	3.545e-06	2.781e-02	1.307e+00
160	6.931e-08	1.952e-06	7.292e-03	6.311e-01

Table 8.26.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the RK2CNEGRUS scheme.

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.3819	0.8446	1.1739	5.0790
20	0.1456	1.0239	0.1955	1.3951
40	0.1172	1.3012	0.0676	0.9251
80	0.1080	1.3025	0.0395	0.8062
160	0.1045	1.3020	0.0347	0.7898

travelling vortex, $\varepsilon = 10^{-2}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	7.411e-03	1.360e-02	3.128e-01	1.869e+00
20	5.490e-03	1.250e-02	4.524e-02	2.427e-01
40	2.097e-03	1.155e-02	6.490e-03	6.112e-02
80	1.210e-03	1.282e-02	1.083e-03	2.110e-02
160	1.064e-03	1.299e-02	4.373e-04	8.744e-03

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	5.658e-04	1.305e-03	1.555e-01	6.759e-01
20	8.398e-05	1.553e-04	7.923e-03	5.055e-02
40	6.717e-05	1.660e-04	7.876e-04	3.625e-03
80	2.476e-05	1.105e-04	1.323e-04	1.506e-03
160	1.265e-05	1.277e-04	3.508e-05	7.484e-04

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	8.822e-08	1.896e-07	2.146e-03	8.818e-03
20	2.685e-09	6.391e-09	1.170e-05	5.153e-05
40	6.223e-10	1.895e-09	1.944e-06	1.600e-05
80	1.147e-09	2.302e-09	7.242e-07	8.087e-06
160	7.315e-09	1.795e-08	5.230e-07	4.562e-06

Table 8.27.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the BDFEGRUS scheme.

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.2928	0.9484	1.1239	6.8983
20	0.1535	1.1478	0.4316	1.5001
40	0.1379	1.2934	0.1300	1.0105
80	0.1129	1.2988	0.0788	0.8527
160	0.1063	1.3006	0.0453	0.7940

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	9.023e-06	5.725e-05	3.102e-04	1.556e-03
20	9.655e-06	1.142e-04	4.176e-05	9.358e-04
40	1.031e-05	1.271e-04	1.623e-04	2.688e-03
80	1.037e-05	1.292e-04	7.928e-05	1.458e-03
160	1.034e-05	1.301e-04	4.275e-05	8.522e-04

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	9.023e-10	5.731e-09	3.106e-08	1.557e-07
20	9.656e-10	1.142e-08	4.216e-09	9.390e-08
40	1.027e-09	1.274e-08	2.134e-08	5.056e-07
80	1.034e-09	1.295e-08	8.722e-09	2.176e-07
160	1.033e-09	1.300e-08	5.116e-09	1.285e-07

Table 8.28.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the ARSRUSELLW scheme.

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.7529	1.8733	0.9847	4.8191
20	0.2177	1.3437	0.9273	5.4738
40	0.1588	1.3100	0.2834	1.3205
80	0.1208	1.3127	0.1023	0.8351
160	0.1078	1.3130	0.0529	0.8997

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	4.429e-04	8.890e-04	2.741e+00	2.003e+01
20	4.600e-04	1.114e-03	8.374e-01	6.365e+00
40	1.944e-04	8.355e-04	2.208e-01	4.178e+00
80	1.277e-04	8.716e-04	6.301e-02	2.136e+00
160	5.513e-05	6.363e-04	2.572e-02	9.988e-01

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.798e-08	3.624e-08	2.742e+00	2.005e+01
20	1.649e-08	3.962e-08	8.347e-01	6.445e+00
40	2.840e-08	1.166e-07	2.185e-01	4.311e+00
80	1.212e-08	5.445e-08	5.414e-02	2.108e+00
160	9.576e-09	4.497e-08	1.418e-02	1.023e+00

Table 8.29.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the RK2CNRUSELLW scheme.

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.2223	0.7235	0.4700	2.1881
20	0.1186	1.1630	0.2087	0.9836
40	0.1165	1.2932	0.0919	0.8044
80	0.1117	1.2997	0.0528	0.7892
160	0.1054	1.3009	0.0416	0.7813

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	8.148e-06	4.834e-05	3.800e-05	2.143e-04
20	9.722e-06	1.134e-04	4.696e-06	1.158e-04
40	1.029e-05	1.283e-04	3.245e-06	8.182e-05
80	1.035e-05	1.296e-04	3.193e-06	8.152e-05
160	1.032e-05	1.300e-04	3.107e-06	7.826e-05

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	8.148e-10	4.834e-09	3.798e-09	2.141e-08
20	9.722e-10	1.134e-08	4.697e-10	1.158e-08
40	1.029e-09	1.283e-08	3.245e-10	8.182e-09
80	1.034e-09	1.296e-08	3.193e-10	8.160e-09
160	1.032e-09	1.300e-08	3.110e-10	7.828e-09

Table 8.30.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the BDFRUSELLW scheme.

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.8299	2.0156	4.3411	17.5331
20	0.3408	1.4017	1.0230	6.5521
40	0.1529	1.2821	0.5332	5.9985
80	0.1166	1.2905	0.1578	2.4588
160	0.1080	1.3007	0.0722	1.3166

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.997e-04	4.588e-04	5.195e+00	2.286e+01
20	3.698e-04	9.987e-04	2.438e+00	1.858e+01
40	1.437e-03	5.762e-03	1.136e+00	1.475e+01
80	3.897e-04	1.721e-03	1.921e+00	2.657e+01
160	9.079e-04	1.130e-02	3.480e+01	7.829e+02

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.998e-08	4.589e-08	5.195e+00	2.286e+01
20	3.711e-08	1.001e-07	2.444e+00	1.860e+01
40	2.417e-07	8.279e-07	1.325e+00	1.763e+01
80	5.252e-08	2.083e-07	1.926e+00	2.712e+01
160	9.251e-08	1.132e-06	3.483e+01	7.835e+02

Table 8.31.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the ARSRUSELL scheme.

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.2841	0.8061	1.2645	5.9980
20	0.2302	1.1398	1.1260	9.1596
40	0.1446	1.2841	0.3756	4.0312
80	0.1220	1.2973	0.1523	1.6795
160	0.1246	1.3070	0.1199	2.3707

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	5.100e-05	1.186e-04	1.677e+00	5.820e+00
20	4.370e-05	1.628e-04	8.289e-01	5.401e+00
40	1.334e-04	6.098e-04	2.759e-01	2.914e+00
80	3.105e-05	1.714e-04	1.119e-01	1.720e+00
160	2.223e-05	1.833e-04	1.147e-01	3.561e+00

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.45, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	5.101e-09	1.186e-08	1.677e+00	5.820e+00
20	4.373e-09	1.628e-08	8.292e-01	5.401e+00
40	1.376e-08	6.175e-08	2.805e-01	2.933e+00
80	3.581e-09	1.725e-08	1.117e-01	1.718e+00
160	3.108e-09	2.036e-08	1.149e-01	3.572e+00

Table 8.32.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the RK2CNRUSELL scheme.

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.2205	0.6129	1.6229	8.2946
20	0.1383	1.3411	0.3851	5.7592
40	0.1223	1.3032	0.1276	1.1898
80	0.1160	1.3017	0.0746	0.8153
160	0.1085	1.3014	0.0484	0.7865

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	3.004e-05	7.222e-05	1.712e+00	8.709e+00
20	1.820e-05	1.861e-04	2.957e-01	3.653e+00
40	1.122e-05	1.328e-04	4.343e-02	6.087e-01
80	1.078e-05	1.342e-04	5.799e-03	1.449e-01
160	1.040e-05	1.305e-04	1.650e-03	9.476e-02

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	3.004e-09	7.222e-09	1.712e+00	8.709e+00
20	1.820e-09	1.861e-08	2.957e-01	3.651e+00
40	1.122e-09	1.328e-08	4.342e-02	6.087e-01
80	1.078e-09	1.342e-08	5.790e-03	1.449e-01
160	1.040e-09	1.305e-08	1.649e-03	9.476e-02

Table 8.33.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the BDFRUSELL scheme.

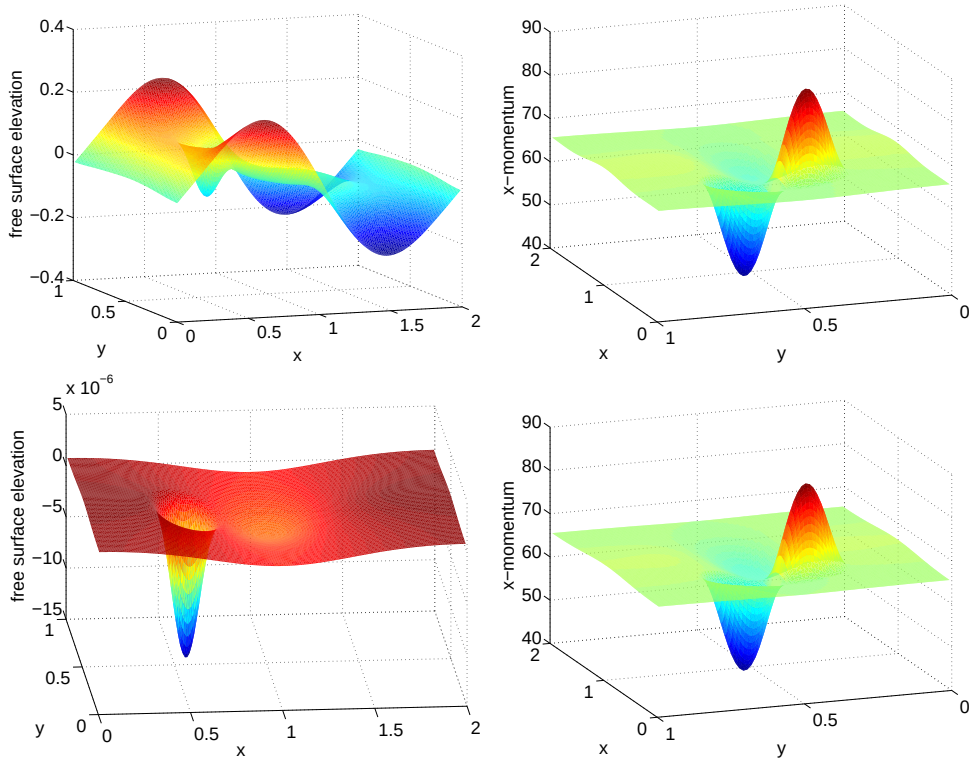

Figure 8.27.: Numerical solutions of the travelling vortex test with the non-constant bottom topography (8.14) obtained by the BDFRUSELLW scheme at time $T = 0.1$ with $CFL_u = 0.3$. Top: Froude number 0.1. Bottom: Froude number 10^{-3}.

In order to study the asymptotic preserving property in the case of a non-constant bottom topography, we consider the travelling vortex over the smooth bottom topography

$$\tilde{b}(x, y) = 10e^{-5(x-1)^2-50(y-0.5)^2} \tag{8.14}$$

on the domain $[0, 2] \times [0, 1]$ with periodic boundary conditions. We have presented numerical results for this test in [15] for the Froude number $\varepsilon = 0.05$ and slightly different boundary conditions, where the development of a periodic sine-type gravity wave due to the bottom hump and advection of the vortex have been observed. The numerical solutions obtained by the BDFRUSELLW and ARSEGRUS schemes for the Froude numbers $\varepsilon = 0.1, 0.001$ are presented in Figures 8.27, 8.28. Tables 8.34-8.36 show the l^1- and l^∞-norms of $\nabla_h \mathbf{Z}$ and $\nabla_h \cdot \mathbf{M}$ of the numerical solutions using $(2N \times N)$-grids, $N \in \{10, 20, 40, 80\}$, obtained by the BDFRUSELLW scheme using the approximations (6.63) or (6.65) and the ARSEGRUS scheme. The considered Froude numbers are $\varepsilon = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-5}$. The results are similar to the results with constant bottom topography, cf. Tables 8.25, 8.30, and indicate that the BDFRUSELLW and ARSEGRUS schemes are asymptotic preserving - also for non-constant bottom topographies. We expect similar observations for the EGRUS, BDFEGRUS, RUSELLW, ARSRUSELLW schemes.

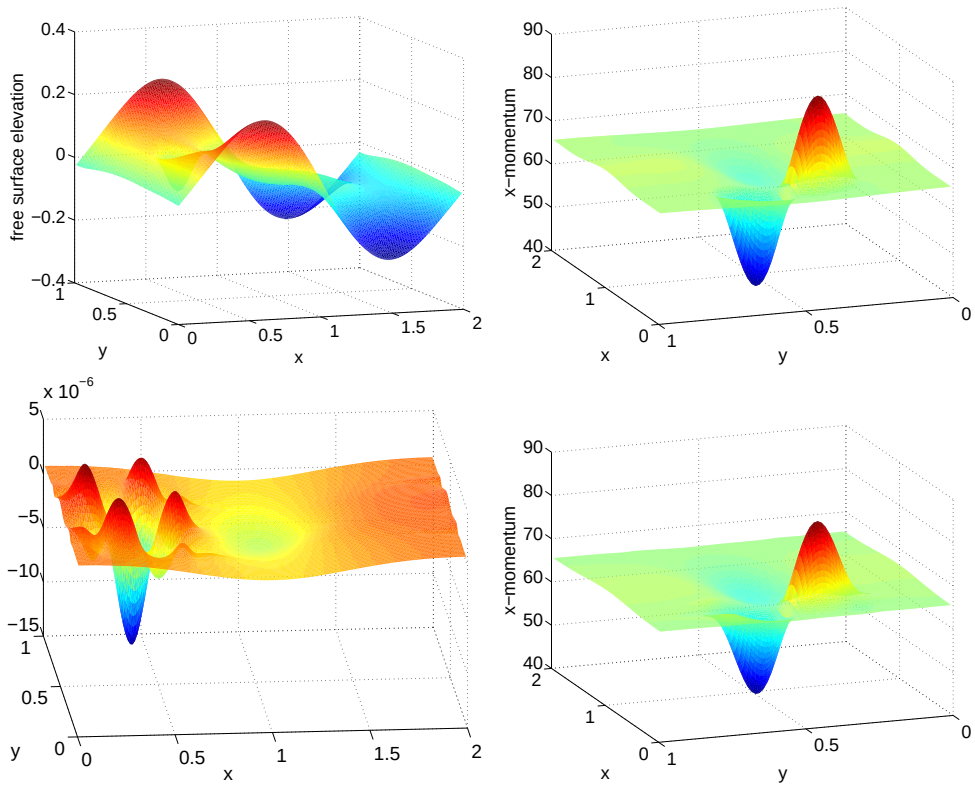

Figure 8.28.: Numerical solutions of the travelling vortex test with the non-constant bottom topography (8.14) obtained by the ARSEGRUS scheme at time $T = 0.1$ with $CFL_u = 0.45$. Top: Froude number 0.1. Bottom: Froude number 10^{-3}.

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.7129	0.7170	2.7454	5.1165
20	0.8075	1.3011	2.6893	3.5775
40	0.9501	1.4571	2.5398	3.7687
80	1.0209	1.4718	2.3068	3.8394

travelling vortex, $\varepsilon = 10^{-2}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.299e-03	4.701e-03	0.2267	0.2088
20	1.875e-03	1.109e-02	0.2307	0.1898
40	3.372e-03	1.255e-02	0.3383	0.2768
80	7.361e-03	1.449e-02	1.0355	0.8223

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.285e-05	4.985e-05	2.130e-04	3.960e-04
20	1.552e-05	1.099e-04	6.581e-06	1.180e-04
40	1.650e-05	1.300e-04	3.246e-06	8.033e-05
80	1.663e-05	1.317e-04	3.313e-06	8.397e-05

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.291e-09	5.002e-09	1.820e-08	3.704e-08
20	1.553e-09	1.099e-08	5.178e-10	1.163e-08
40	1.650e-09	1.300e-08	3.240e-10	8.030e-09
80	1.663e-09	1.317e-08	3.298e-10	8.435e-09

Table 8.34.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex with bottom topography test at time $T = 0.1$. Numerical solution is obtained by the BDFRUSELLW scheme using the approximations (6.63).

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.7124	0.7152	2.7452	5.1331
20	0.8075	1.3010	2.6893	3.5786
40	0.9501	1.4571	2.5398	3.7687
80	1.0209	1.4718	2.3068	3.8394

travelling vortex, $\varepsilon = 10^{-2}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.299e-03	4.692e-03	0.2267	0.2088
20	1.875e-03	1.109e-02	0.2307	0.1898
40	3.372e-03	1.255e-02	0.3383	0.2768
80	7.361e-03	1.449e-02	1.0355	0.8223

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.285e-05	4.976e-05	2.129e-04	3.946e-04
20	1.552e-05	1.099e-04	6.580e-06	1.179e-04
40	1.650e-05	1.300e-04	3.246e-06	8.033e-05
80	1.663e-05	1.317e-04	3.313e-06	8.397e-05

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.290e-09	4.993e-09	1.820e-08	3.689e-08
20	1.552e-09	1.099e-08	5.177e-10	1.162e-08
40	1.650e-09	1.300e-08	3.240e-10	8.030e-09
80	1.663e-09	1.317e-08	3.298e-10	8.435e-09

Table 8.35.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex with bottom topography test at time $T = 0.1$. Numerical solution is obtained by the BDFRUSELLW scheme using the approximations (6.65).

travelling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.7548	0.6485	2.7495	4.4591
20	0.9386	0.9831	2.6103	3.5203
40	1.0084	1.2500	2.2920	3.6010
80	1.0301	1.2995	2.2027	3.4190

travelling vortex, $\varepsilon = 10^{-2}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.797e-02	1.437e-02	4.256e-01	6.631e-01
20	1.877e-02	1.791e-02	3.139e-01	3.968e-01
40	1.712e-02	1.448e-02	1.456e+00	1.253e+00
80	5.172e-02	4.364e-02	7.418e-01	9.763e-01

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	1.443e-04	1.879e-04	3.818e-02	8.781e-02
20	1.128e-04	2.201e-04	6.539e-03	1.458e-02
40	8.002e-05	1.720e-04	1.466e-03	4.993e-03
80	2.951e-05	1.087e-04	2.206e-04	2.075e-03

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	7.861e-09	1.087e-08	1.212e-04	2.134e-04
20	2.933e-09	4.908e-09	1.920e-05	1.054e-04
40	3.328e-09	4.945e-09	6.234e-06	4.106e-05
80	9.889e-09	1.775e-08	2.904e-06	1.405e-05

Table 8.36.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex with bottom topography test at time $T = 0.1$. Numerical solution is obtained by the ARSEGRUS scheme.

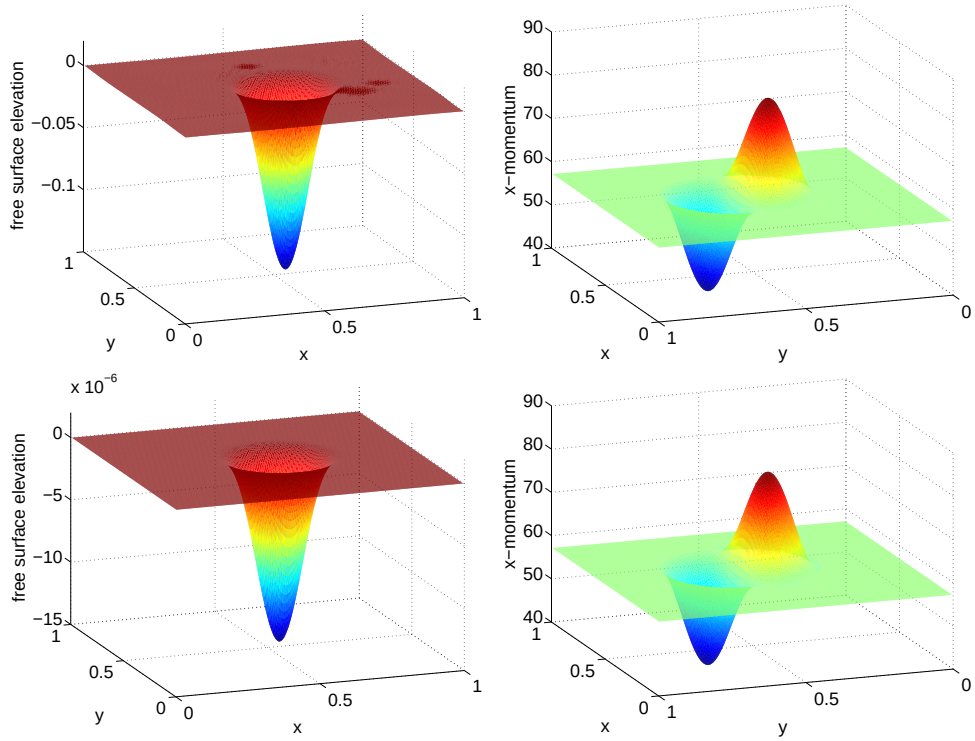

Figure 8.29.: Numerical solutions of the travelling vortex test rotated by the radian measure $\pi/6$ obtained by the BDFRUSELLW scheme at time $T = 0.1$ with $CFL_u = 0.2$. Top: Froude number 0.1. Bottom: Froude number 0.001.

8.2.3. Symmetry breaking

In order to assure that the previously obtained results are not just a consequence of grid-aligned advection of the vortex, we rotate the test setting counterclockwise by the radian measure $\pi/6$. In Table 8.37 the corresponding convergence rates at time $T = 0.1$ for Froude numbers $\varepsilon \in \{0.1, 10^{-3}, 10^{-5}\}$ and CFL_u-numbers $0.2, 0.3$ are shown - the numerical solutions have been obtained by the BDFRUSELLW scheme. For $CFL_u = 0.3$, the second order convergence breaks down, whereas for $CFL_u = 0.2$ it remains. Note that the time steps using the advective CFL number $CFL_u = 0.2$ and oblique reference velocity $\mathbf{u}_{ref} = 0.6(\cos(\pi/6), \sin(\pi/6))$ are approximately the same as for $CFL_u = 0.3$ and $\mathbf{u}_{ref} = (0.6, 0)$, cf. CFL condition (1.4). The corresponding numerical solutions for the Froude numbers $\varepsilon = 0.1, 0.001$ and $CFL_u = 0.2$ are presented in Figure 8.29.

8.2.4. Long time simulation

In this section we study the long time behaviour of numerical solutions of the travelling vortex test (8.5). To this end we compute the solutions of the travelling vortex test at the time $T = 5/3$, when the vortex has travelled one time across the whole x-axis. During this test the numerical solutions of the schemes ARSRUSELL and RK2CNRUSELL break down. The numerical solutions of the BDFRUSELL, RK2CNEGRUS, RK2CNRUSELLW schemes develop oscillations, cf. Figures 8.30-8.32. The oscillations in the solution ob-

traveling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.055e-02		0.49682		0.65269	
20	1.275e-03	3.0484	0.15369	1.6927	0.21241	1.6195
40	6.024e-04	1.0820	0.04053	1.9231	0.05510	1.9467
80	3.131e-04	0.9440	0.01089	1.8961	0.01450	1.9258
160	1.949e-04	0.6836	0.00372	1.5488	0.00475	1.6110

traveling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	4.178e-07		0.49366		0.64715	
20	1.020e-07	2.0342	0.15502	1.6710	0.21401	1.5964
40	1.655e-08	2.6239	0.03990	1.9582	0.05435	1.9773
80	3.108e-09	2.4126	0.01062	1.9090	0.01422	1.9340
160	1.011e-06	-8.3452	0.03293	-1.6322	0.03833	-1.4303

traveling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.3, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	4.177e-11		0.49366		0.64715	
20	1.020e-11	2.0338	0.15502	1.6710	0.21401	1.5964
40	1.655e-12	2.6240	0.03990	1.9582	0.05435	1.9773
80	3.119e-13	2.4072	0.01062	1.9090	0.01422	1.9340
160	1.190e-10	-8.5757	0.03320	-1.6441	0.03864	-1.4419

traveling vortex, $\varepsilon = 10^{-1}, CFL_u = 0.2, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	5.589e-03		0.45739		0.59059	
20	1.499e-03	1.8981	0.14795	1.6283	0.20144	1.5518
40	9.511e-04	0.6568	0.03769	1.9729	0.05086	1.9857
80	3.148e-04	1.5950	0.01003	1.9096	0.01310	1.9566
160	8.247e-05	1.9327	0.00271	1.8876	0.00351	1.9005

traveling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.2, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	3.676e-07		0.45453		0.58819	
20	1.019e-07	1.8507	0.14842	1.6147	0.20199	1.5420
40	1.633e-08	2.6419	0.03689	2.0083	0.05006	2.0127
80	2.827e-09	2.5302	0.00966	1.9332	0.01274	1.9739
160	6.973e-10	2.0192	0.00252	1.9371	0.00332	1.9392

traveling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.2, T = 0.1$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	3.675e-11		0.45453		0.58819	
20	1.026e-11	1.8411	0.14842	1.6147	0.20199	1.5420
40	1.633e-12	2.6512	0.03689	2.0083	0.05006	2.0127
80	2.826e-13	2.5305	0.00966	1.9332	0.01274	1.9739
160	6.960e-14	2.0218	0.00252	1.9371	0.00332	1.9392

Table 8.37.: EOC of the second order BDFRUSELLW scheme; travelling vortex test rotated by the radian measure $\pi/6$.

ε	0.8	0.1	10^{-2}	10^{-3}	10^{-4}	10^{-5}
$CFL \approx$	0.87	3.84	35	$3.4 \cdot 10^2$	$3.4 \cdot 10^3$	$3.4 \cdot 10^4$

Table 8.38.: CFL-numbers used by the second order IMEX schemes in the travelling vortex test. $CFL_u = 0.45$.

tained by the BDFRUSELL scheme can be damped by using a smaller CFL_u-number, e.g. $CFL_u = 0.1$, but they will not disappear. The other schemes remain stable. The corresponding cross-sections of the free surface elevation at $y \approx 0.5$ using the Froude numbers $10^{-1}, 10^{-3}, 10^{-5}$ are presented in the Figures 8.31, 8.32.

8.2.5. Efficiency of the IMEX finite volume schemes

The aim of this section is to discuss the efficiency of the previously presented IMEX finite volume schemes, cf. Chapter 6. In order to compare the large time steps of the IMEX schemes with the time steps of an explicit scheme, e.g. HLLC scheme, we consider the CFL-numbers (1.2) used by the IMEX schemes for the travelling vortex test, since the CFL-number is proportional to the time increment. Table 8.38 shows the used CFL-numbers for various Froude numbers ε in the range from 10^{-5} to 0.8 and $CFL_u = 0.45$. For $\varepsilon << 1$ we have $CFL \approx 0.34/\varepsilon$. Let us point out that it is not a mistake that $CFL \approx 0.34/\varepsilon$ instead of $CFL \approx CFL_u/\varepsilon = 0.45/\varepsilon$: Let u', u^*, c^* satisfy

$$u' = \max\{|u_1|, |u_2|\}, \quad u^* + c^* = \max\{|u_1| + c, |u_2| + c\}. \tag{8.15}$$

Then the CFL-number used in an IMEX scheme is

$$CFL = \frac{u^*}{u'}\left(1 + \frac{c^*}{u^*}\right) CFL_u \approx \frac{u^*}{u'\varepsilon} CFL_u. \tag{8.16}$$

For the travelling vortex test $u' \approx 0.8$. If $u^* = 0.6$ we have $CFL \approx 0.34/\varepsilon$. Consequently, we calculate much less time steps in comparison to an explicit scheme. However we pay the price by solving one or more linear systems per time step. In our previously presented numerical experiments we have used a direct solver routine - unsymmetric multifrontal sparse LU factorisation - provided by UMFPACK [31, 32, 34, 33] to solve the linear systems. This is computationally expensive due to the nonlinear complexity of the solver with respect to the cell widths $\Delta x, \Delta y$. Applying an iterative solver instead one should be able to solve linear systems with linear complexity with respect to $\Delta x, \Delta y$. Then the IMEX schemes should outperform notably standard explicit schemes.

In what follows, we compare the CPU runtimes of the explicit RK2HLLC scheme with the BDF2RUSELLW scheme, where the corresponding linear systems are either solved by the above mentioned direct solver or by the conjugate gradient (CG) method, cf. [89, 103]. Indeed we can use the CG method for the travelling vortex test, since the bottom topography is constant and therefore the matrix $\mathbb{1} + \delta^2/\varepsilon^2 E$, is symmetric and positive definite, cf. Lemma 6.8.1. The implementation of the CG method within the *"Portable, Extensible Toolkit for Scientific Computation"* (PETSc), cf. [9, 10, 11], has been used. Let us note that the matrix $\mathbb{1} + \delta^2/\varepsilon^2 E$ has the eigenvalue one as well as

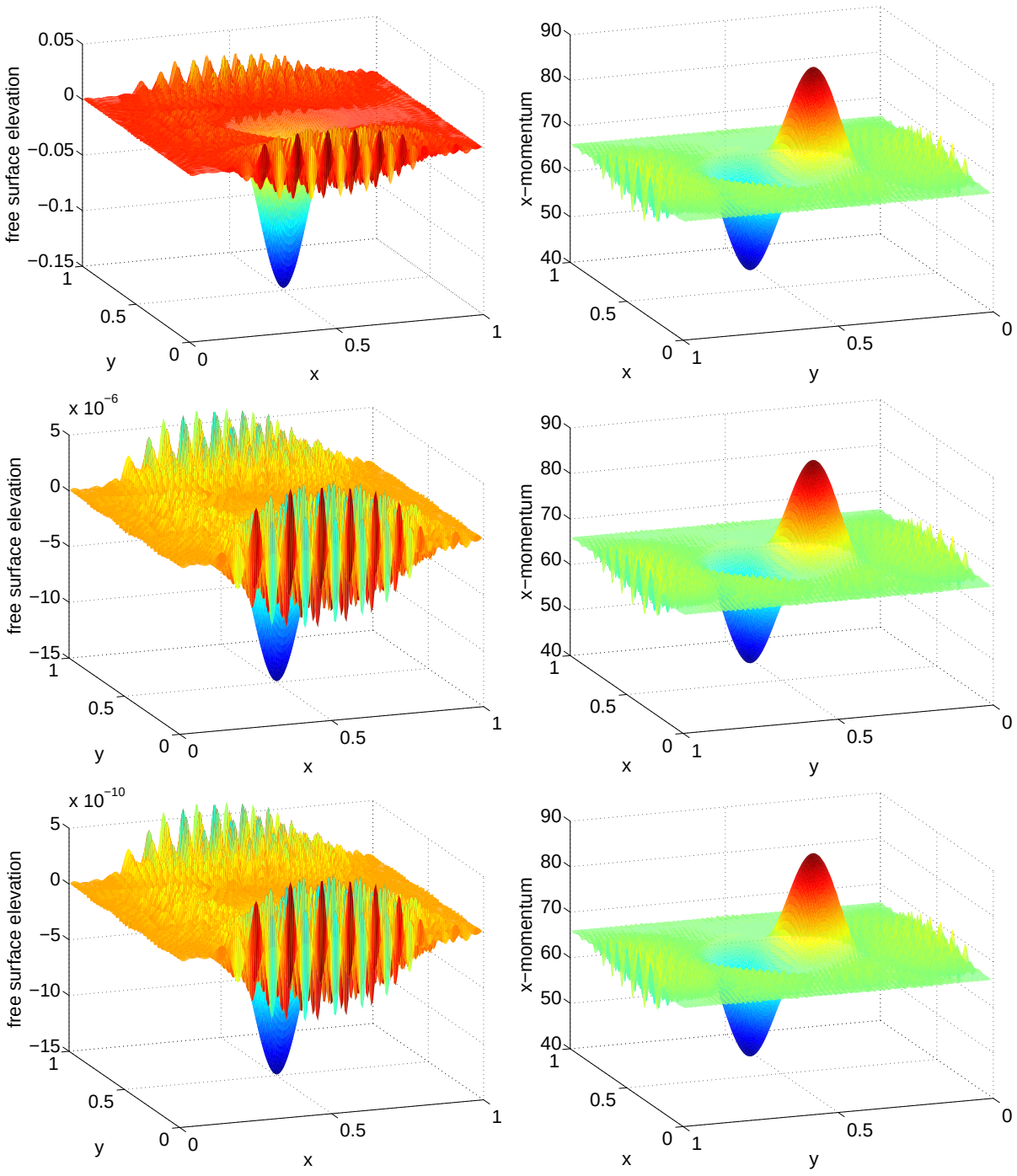

Figure 8.30.: Numerical solutions of the travelling vortex test obtained by the BD-FRUSELL scheme at time $T = 5/3$ with $CFL_u = 0.3$. From top to bottom the Froude number is $10^{-1}, 10^{-3}, 10^{-5}$.

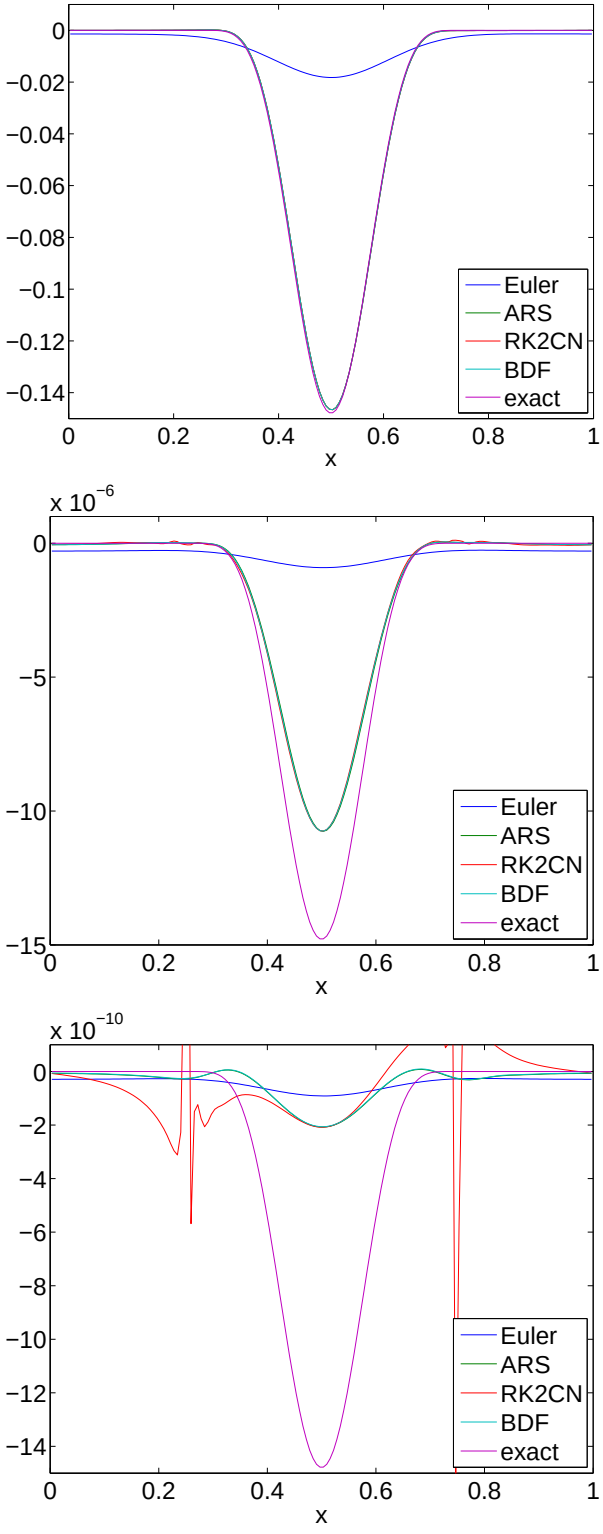

Figure 8.31.: Cross-sections of exact and numerical solutions obtained by the EGRUS, ARSEGRUS, RK2CNEGRUS and BDFEGRUS schemes of the travelling vortex test at time $T = 5/3$. The CFL_u-number 0.3 was used for the SBDF time discretisation, whereas 0.45 for the other ones. From top to bottom the Froude number is $10^{-1}, 10^{-3}, 10^{-5}$.

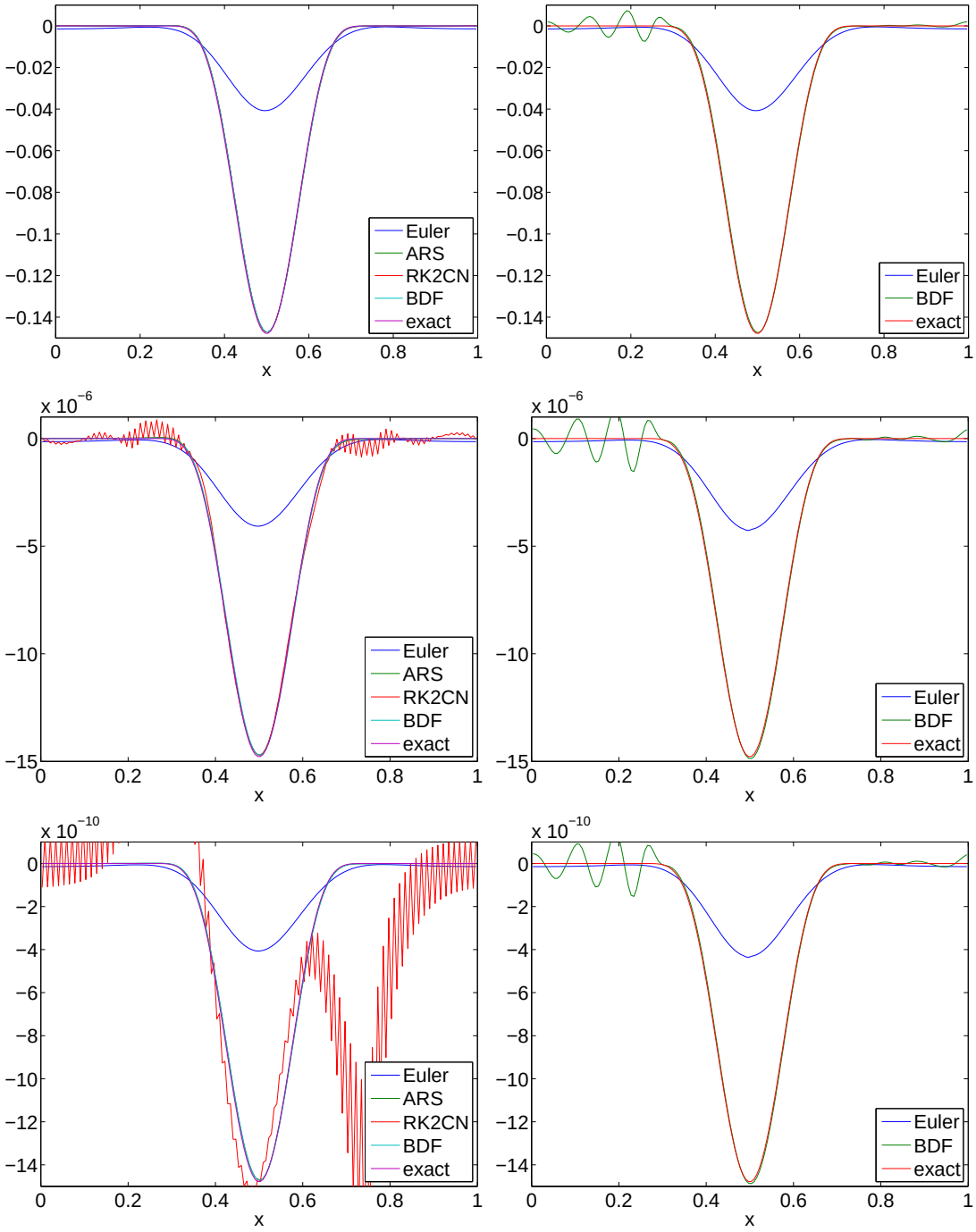

Figure 8.32.: Cross-sections of exact and numerical solutions of the travelling vortex test at time $T = 5/3$. The CFL_u-number 0.3 was used for the SBDF time discretisation, whereas 0.45 for the other ones. From top to bottom the Froude number is $10^{-1}, 10^{-3}, 10^{-5}$. Left: numerical solutions obtained by the RUSELLW, ARSRUSELLW, RK2CNRUSELLW, BDFRUSELLW schemes. Right: numerical solutions obtained by the RUSELL, BDFRUSELL schemes.

eigenvalues of the order $\mathcal{O}(1/\varepsilon^2)$. Hence, it is ill-conditioned and preconditioning plays an important role for the convergence of an iterative solver. We use the default settings of PETSc: the ILU(0) preconditioner, cf. [9, 10, 89, 103].

The following part of this section consists of two studies. In the first we present the accuracy of the BDF2RUSELLW scheme using the CG method. In the second we compare the CPU runtimes. For the sake of brevity we will refer to the BDFRUSELLW scheme where the linear systems are solved with the CG method as BDFRUSELLWCG.

By computing numerical solutions of the travelling vortex test with the BDF2RUSELLWCG scheme we have to prescribe the abort conditions for the iterative solver. Then, the CG method aborts, if the absolute or relative residuum norm is below a given threshold or the number of iterations is too high. These three parameters are crucial for the accuracy and performance of the scheme. If the prescribed norm tolerances are too high or the maximum number of iterations are too low, we will observe poor accuracy. However, if the norm tolerances are too low and the number of allowed iterations too large, the performance will be poor. After testing different values as abort criteria, we came to the following conclusion:

1. In order to obtain the second order convergence rates of the momentum and the free surface elevation the residuum norm tolerances has to decrease if $\Delta x, \Delta y, \varepsilon$ decrease. Setting the absolute residuum norm tolerance to

$$(\varepsilon \min\{\Delta x, \Delta y\})^2 \tag{8.17a}$$

   as the abort criterion we obtain second order convergence rates if the Froude number $\varepsilon \geq 10^{-3}$, cf. Table 8.39. If the Froude number $\varepsilon \leq 10^{-5}$, the second order convergence seems just to hold for sufficiently fine cell widths $\Delta x, \Delta y$.

2. If we halve the cell widths $\Delta x, \Delta y$ and use the absolute residuum tolerance (8.17a) as the abort criterion, the number of iterations often doubles, cf. Table 8.41. If we decrease the Froude number, the number of iterations seems to increase until a threshold is taken.

3. If we take the absolute residuum norm tolerance

$$\min\{\varepsilon, \Delta x, \Delta y\}^2 \tag{8.17b}$$

   as the abort criterion, we obtain second order convergence rates of the momentum. Further, we observe either second order convergence of the free surface elevation or the free surface elevation error is below the local truncation error $\min\{\Delta x, \Delta y\}^2$, cf. Table 8.40.

Tables 8.42-8.45 show the CPU runtimes of the BDFRUSELLWCG, BDFRUSELLW and RK2HLLC scheme. Let us recall that we have $\Delta x = \Delta y = 1/N$.

The CPU runtimes of the BDFRUSELLWCG scheme with the absolute residuum norm tolerances (8.17a) or (8.17b) are presented in Tables 8.42-8.43. Using the absolute residuum norm tolerance (8.17a), the CPU runtime increases if the Froude number ε or the grid width $\Delta x = \Delta y$ decreases. This is due to the larger number of iterations performed by the CG scheme, cf. Table 8.41. Since the number of iterations seems to

travelling vortex, $\varepsilon = 0.8, CFL_u = 0.3, T = 0.1, atol = (\varepsilon \min\{\Delta x, \Delta y\})^2$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.930e-001	0.0000	0.33519	0.0000	0.64771	0.0000
20	5.035e-002	1.9388	0.10805	1.6333	0.22831	1.5043
40	1.257e-002	2.0019	0.02783	1.9570	0.05553	2.0398
80	3.244e-003	1.9542	0.00722	1.9462	0.01411	1.9765
160	8.148e-004	1.9932	0.00186	1.9561	0.00371	1.9281
320	2.047e-004	1.9927	0.00048	1.9556	0.00097	1.9361

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.3, T = 0.1, atol = (\varepsilon \min\{\Delta x, \Delta y\})^2$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.293e-02	0.0000	0.35780	0.0000	0.68711	0.0000
20	1.662e-03	2.9596	0.12116	1.5622	0.23440	1.5515
40	7.146e-04	1.2179	0.02897	2.0641	0.05645	2.0540
80	3.223e-04	1.1489	0.00743	1.9638	0.01463	1.9479
160	8.022e-05	2.0062	0.00201	1.8850	0.00395	1.8883
320	2.207e-05	1.8616	0.00053	1.9145	0.00105	1.9182

travelling vortex, $\varepsilon = 10^{-2}, CFL_u = 0.3, T = 0.1, atol = (\varepsilon \min\{\Delta x, \Delta y\})^2$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	4.165e-003	0.0000	0.36985	0.0000	0.70520	0.0000
20	9.341e-006	8.8004	0.11410	1.6966	0.23531	1.5835
40	1.629e-006	2.5199	0.02777	2.0386	0.05563	2.0805
80	3.136e-007	2.3767	0.00719	1.9487	0.01435	1.9549
160	8.097e-008	1.9535	0.00187	1.9456	0.00379	1.9197
320	2.347e-008	1.7863	0.00048	1.9548	0.00099	1.9342

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1, atol = (\varepsilon \min\{\Delta x, \Delta y\})^2$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	5.956e-007	0.0000	0.34844	0.0000	0.68987	0.0000
20	1.271e-007	2.2284	0.11412	1.6103	0.23531	1.5518
40	1.813e-008	2.8094	0.02777	2.0388	0.05563	2.0805
80	3.518e-009	2.3655	0.00720	1.9485	0.01435	1.9550
160	7.928e-010	2.1499	0.00187	1.9449	0.00379	1.9200
320	2.096e-010	1.9196	0.00048	1.9542	0.00099	1.9341

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.3, T = 0.1, atol = (\varepsilon \min\{\Delta x, \Delta y\})^2$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	5.408e-011	0.0000	0.34844	0.0000	0.68987	0.0000
20	3.187e-011	0.7631	0.11412	1.6103	0.23531	1.5518
40	3.665e-011	-0.2017	0.02777	2.0388	0.05563	2.0805
80	3.984e-011	-0.1205	0.00720	1.9485	0.01435	1.9550
160	8.352e-014	8.8978	0.00187	1.9449	0.00379	1.9200
320	2.922e-014	1.5153	0.00048	1.9542	0.00099	1.9341

m_2

Table 8.39.: EOC of the second order BDFRUSELLWCG scheme; travelling vortex test. The absolute residuum norm tolerance is $atol = (\varepsilon \min\{\Delta x, \Delta y\})^2$.

travelling vortex, $\varepsilon = 0.8, CFL_u = 0.3, T = 0.1, atol = \min\{\varepsilon, \Delta x, \Delta y\}^2$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.930e-001	0.0000	0.33519	0.0000	0.64771	0.0000
20	5.035e-002	1.9388	0.10805	1.6333	0.22831	1.5043
40	1.257e-002	2.0019	0.02783	1.9570	0.05553	2.0398
80	3.244e-003	1.9542	0.00722	1.9462	0.01411	1.9765
160	8.148e-004	1.9932	0.00186	1.9561	0.00371	1.9281
320	2.047e-004	1.9927	0.00048	1.9556	0.00097	1.9361

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.3, T = 0.1, atol = \min\{\varepsilon, \Delta x, \Delta y\}^2$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	1.295e-002	0.0000	0.35758	0.0000	0.68691	0.0000
20	1.666e-003	2.9584	0.12119	1.5611	0.23441	1.5511
40	7.126e-004	1.2253	0.02898	2.0642	0.05645	2.0540
80	3.222e-004	1.1454	0.00743	1.9636	0.01463	1.9480
160	8.013e-005	2.0073	0.00201	1.8848	0.00395	1.8876
320	2.206e-005	1.8606	0.00053	1.9148	0.00105	1.9187

travelling vortex, $\varepsilon = 0.01, CFL_u = 0.3, T = 0.1, atol = \min\{\varepsilon, \Delta x, \Delta y\}^2$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	4.165e-003	0.0000	0.36985	0.0000	0.70520	0.0000
20	1.114e-005	8.5468	0.11426	1.6946	0.23530	1.5836
40	2.409e-006	2.2089	0.02773	2.0428	0.05565	2.0799
80	1.276e-006	0.9171	0.00721	1.9426	0.01434	1.9561
160	1.276e-007	3.3212	0.00187	1.9460	0.00379	1.9190
320	4.512e-008	1.5002	0.00048	1.9554	0.00099	1.9342

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1, atol = \min\{\varepsilon, \Delta x, \Delta y\}^2$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	6.447e-007	0.0000	0.34886	0.0000	0.68985	0.0000
20	3.020e-007	1.0941	0.11412	1.6121	0.23531	1.5517
40	3.333e-007	-0.1424	0.02777	2.0388	0.05564	2.0805
80	3.426e-007	-0.0397	0.00720	1.9474	0.01434	1.9558
160	2.719e-007	0.3334	0.00188	1.9359	0.00379	1.9204
320	8.790e-008	1.6292	0.00050	1.9198	0.00099	1.9323

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.3, T = 0.1, atol = \min\{\varepsilon, \Delta x, \Delta y\}^2$

N	L^1-error in z	EOC z	L^1-error in m_1	EOC m_1	L^1-error in m_2	EOC m_2
10	6.115e-011	0.0000	0.34886	0.0000	0.68985	0.0000
20	3.190e-011	0.9387	0.11411	1.6122	0.23531	1.5517
40	3.641e-011	-0.1906	0.02777	2.0388	0.05564	2.0805
80	3.989e-011	-0.1319	0.00721	1.9461	0.01434	1.9557
160	4.161e-011	-0.0607	0.00189	1.9299	0.00379	1.9199
320	4.249e-011	-0.0302	0.00050	1.9309	0.00100	1.9212

m_2

Table 8.40.: EOC of the second order BDFRUSELLWCG scheme; travelling vortex test. The absolute residuum norm tolerance is $atol = \min\{\varepsilon, \Delta x, \Delta y\}^2$.

atol	$(\varepsilon \min\{\Delta x\})^2$				$\min\{\varepsilon, \Delta x\}^2$			
ε \ N	40	80	160	320	40	80	160	320
10^{-1}	3	3	3	3	2	2	1	1
10^{-2}	7	10	13	15	1	2	3	4
10^{-3}	6	11	20	35	2	2	2	2
10^{-5}	6	11	20	36	2	2	2	2
10^{-6}	6	11	20	36	2	2	2	2

Table 8.41.: Typical number of iterations used by the CG method depending on the Froude number ε, the grid widths $\Delta x = \Delta y = 1/N$ and the absolute residuum norm tolerance *atol*.

stop increasing for sufficiently small Froude numbers ε, the CPU runtime also stops increasing with respect to the Froude number. However, the number of iterations increases for finer grid widths $\Delta x = \Delta y$. Hence, the complexity increases in a superlinear way. Using instead the absolute residuum tolerance (8.17b) the number of iterations is typically two - independent of grid widths $\Delta x = \Delta y$ and the Froude number ε. Thus, the BDF2RUSELLWCG scheme performs with linear complexity.

Table 8.44 shows the CPU runtimes of the BDFRUSELLW scheme, that seems almost independent of the Froude number ε. But since a direct solver is applied, the complexity is exponential. Thus, the computational costs rapidly increase, if the grid widths $\Delta x = \Delta y$ decrease.

The CPU runtimes of the RK2HLLC scheme are shown in Table 8.45. Since it is explicit, its complexity is linear with respect to the grid widths. But as already explained in the introduction, cf. Chapter 1, the time step is approximately proportional to $1/\varepsilon$ for low Froude numbers. Hence, the CPU runtimes notably increase, if the Froude number ε decreases.

Due to the efficient solution of the arising linear equations and the use of large time steps independent on the Froude number, the BDF2RUSELLWCG scheme clearly outperforms the BDF2RUSELLW and the RK2HLLC schemes in the weakly compressible regime, say $\varepsilon < 0.2$.

Remark 8.2.1 *Let us recall briefly that we have proved a discrete divergence constraint of a fully discrete scheme in Chapter 7 in the following way: We first showed that* $\nabla_h \mathbf{Z}^{n+1} = \mathcal{O}(\varepsilon^2)$ *due to the term* $\nabla z^{n+1}/\varepsilon^2$ *in the momentum equation. Then we followed the discrete divergence constraint from the momentum equation*

$$z_{ij}^{n+1} + \delta(\nabla \cdot \mathbf{m}^{n+1})_{ij} = z^n + \mathcal{O}(\varepsilon^2). \tag{8.18}$$

Solving the elliptic equation to obtain the free surface elevation approximately, we introduce an approximation error

$$z_{ij}^{n+1} + \delta(\nabla \cdot \mathbf{m}^{n+1})_{ij} = z^n + \mathcal{O}(\delta\varepsilon^2) + \mathcal{O}(err). \tag{8.19}$$

Consequently, we obtain the divergence constraint

$$(\nabla \cdot \mathbf{m}^{n+1})_{ij} = \max\{\mathcal{O}(\varepsilon^2), \mathcal{O}(err/\delta)\}. \tag{8.20}$$

N	10	20	40	80	160	320	ε
CPU runtime	0.027	0.108	0.589	3.808	25.594	192.152	0.8
fraction $CPU_N/CPU_{N/2}$		3.960	5.431	6.469	6.722	7.508	
CPU runtime	0.030	0.121	0.606	4.164	27.864	203.851	10^{-1}
fraction $CPU_N/CPU_{N/2}$		4.034	5.012	6.876	6.692	7.316	
CPU runtime	0.030	0.206	0.764	6.825	41.592	296.602	10^{-2}
fraction $CPU_N/CPU_{N/2}$		6.969	3.715	8.929	6.094	7.131	
CPU runtime	0.040	0.126	0.722	5.302	47.434	451.715	10^{-3}
fraction $CPU_N/CPU_{N/2}$		3.132	5.749	7.342	8.946	9.523	
CPU runtime	0.033	0.121	0.685	5.090	45.157	462.221	10^{-5}
fraction $CPU_N/CPU_{N/2}$		3.725	5.646	7.431	8.873	10.236	

Table 8.42.: CPU runtimes of the BDFRUSELLWCG scheme with the absolut residuum norm tolerance $(\varepsilon \min\{\Delta x, \Delta y\})^2$; travelling vortex test.

N	10	20	40	80	160	320	ε
CPU runtime	0.026	0.113	0.589	3.713	25.581	190.433	0.8
fraction $CPU_N/CPU_{N/2}$		4.270	5.206	6.306	6.890	7.444	
CPU runtime	0.026	0.118	0.591	3.723	25.442	190.425	10^{-1}
fraction $CPU_N/CPU_{N/2}$		4.538	5.019	6.297	6.834	7.485	
CPU runtime	0.029	0.114	0.590	3.761	26.867	209.984	10^{-2}
fraction $CPU_N/CPU_{N/2}$		3.948	5.155	6.378	7.143	7.816	
CPU runtime	0.028	0.115	0.578	3.718	26.452	198.926	10^{-3}
fraction $CPU_N/CPU_{N/2}$		4.112	5.007	6.430	7.115	7.520	
CPU runtime	0.030	0.117	0.610	3.854	26.569	199.900	10^{-5}
fraction $CPU_N/CPU_{N/2}$		3.888	5.204	6.313	6.894	7.524	

Table 8.43.: CPU runtimes of the BDFRUSELLWCG scheme with the absolut residuum norm tolerance $\min\{\varepsilon, \Delta x, \Delta y\}^2$; travelling vortex test.

N	10	20	40	80	160	320	ε
CPU runtime	0.010	0.083	1.081	16.942	287.451	5641.45	0.8
fraction $CPU_N/CPU_{N/2}$		7.973	12.976	15.669	16.966	19.626	
CPU runtime	0.012	0.105	1.645	18.974	286.556	4959.39	10^{-1}
fraction $CPU_N/CPU_{N/2}$		8.881	15.592	11.535	15.102	17.307	
CPU runtime	0.012	0.131	1.295	22.911	300.100	5110.06	10^{-2}
fraction $CPU_N/CPU_{N/2}$		11.064	9.916	17.693	13.098	17.028	
CPU runtime	0.012	0.123	1.158	17.831	326.248	4440.67	10^{-3}
fraction $CPU_N/CPU_{N/2}$		10.137	9.420	15.397	18.297	13.611	
CPU runtime	0.012	0.122	1.879	19.744	326.561	4448.95	10^{-5}
fraction $CPU_N/CPU_{N/2}$		9.891	15.392	10.510	16.540	13.624	

Table 8.44.: CPU runtimes of the BDFRUSELLW scheme; travelling vortex test.

N	10	20	40	80	160	320	ε
CPU runtime	0.004	0.029	0.212	1.620	7.155	53.860	0.8
fraction $CPU_N/CPU_{N/2}$		7.527	7.322	7.650	4.417	7.528	
CPU runtime	0.015	0.118	0.793	6.106	56.118	270.821	10^{-1}
fraction $CPU_N/CPU_{N/2}$		7.805	6.741	7.697	9.191	4.826	
CPU runtime	0.130	0.945	7.462	36.550	296.037	3274.15	10^{-2}
fraction $CPU_N/CPU_{N/2}$		7.262	7.898	4.898	8.099	11.060	
CPU runtime	1.192	7.975	49.248	380.094	2804.37	33731	10^{-3}
fraction $CPU_N/CPU_{N/2}$		6.691	6.175	7.718	7.378	12.028	
CPU runtime	82.390	734.828	5835.22	40204.3	267482		10^{-5}
fraction $CPU_N/CPU_{N/2}$		8.919	7.941	6.890	6.653		

Table 8.45.: CPU runtimes of the RK2HLLC scheme; travelling vortex test.

Let us remark that we solve a preconditioned linear system and therefore the absolute residuum error is typically not the absolute residuum error of the original linear system. Thus, the divergence constraint of the BDFRUSELLWCG scheme is unknown. Tables 8.46-8.47 present the norms of discrete free surface elevation gradients and momentum divergences of the BDF2RUSELLWCG scheme with absolute residuum norm tolerances (8.17a), (8.17b). Obviously $(\nabla z)_{ij} = \mathcal{O}(\varepsilon^2)$*. However,* $\|(\nabla \cdot \mathbf{m})_{ij}\| \neq \mathcal{O}(\varepsilon^2)$ *for all Froude numbers* ε*.*

8.3. Well-balanced property

In order to verify numerically that the fully discrete schemes, derived in Chapter 6, are well-balanced, we consider the numerical tests proposed by Canestrelli et al. in [25]. Here, we consider the lake at rest state $z = p = const., \mathbf{m} = 0$ with the following smooth and discontinuous bottom topographies

$$b_s(x,y) = z - 10 + 5\exp\left(-\frac{2}{5}\left[(x-5)^2 + (y-5)^2\right]\right), \tag{8.21a}$$

$$b_d(x,y) = z - 10 + \begin{cases} 4 & \text{if } 4 \leq x, y \leq 8 \\ 0 & \text{else} \end{cases} \tag{8.21b}$$

on the computational domain $\Omega = [0,10] \times [0,10]$ with periodic boundary conditions. In Tables 8.48-8.51 we present the L^1- and L^∞-norm of the free surface elevation difference $z - p$ and the specific discharge $q = \sqrt{m_1^2 + m_2^2}$ at time $T = 10$. Table 8.48 shows the results of the EGRUS, ARSEGRUS, RK2CNEGRUS, BDFEGRUS schemes; Table 8.49 of the RUSELL, ARSRUSELL, RK2CNRUSELL, BDFRUSELL schemes; Table 8.50, 8.51 of the RUSELLW, ARSRUSELLW, RK2CNRUSELLW, BDFRUSELLW schemes, where the averages (6.48b) and (6.50) have been used to approximate the source terms.

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.2223	0.7235	0.4699	2.1878
20	0.1186	1.1630	0.2087	0.9839
40	0.1165	1.2932	0.0919	0.8044
80	0.1117	1.2997	0.0528	0.7892
160	0.1054	1.3009	0.0416	0.7813

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	8.145e-06	4.834e-05	8.283e-05	6.141e-04
20	9.722e-06	1.134e-04	1.182e-05	1.160e-04
40	1.029e-05	1.283e-04	4.818e-06	8.153e-05
80	1.035e-05	1.296e-04	3.356e-06	8.150e-05
160	1.032e-05	1.300e-04	3.168e-06	7.822e-05

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	8.148e-10	4.834e-09	1.175e-04	6.981e-04
20	9.722e-10	1.134e-08	8.388e-06	4.052e-05
40	1.029e-09	1.283e-08	1.727e-06	1.124e-05
80	1.034e-09	1.296e-08	2.097e-07	4.196e-06
160	1.032e-09	1.300e-08	8.350e-08	2.656e-06

Table 8.46.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the BDFRUSELLWCG scheme with the absolute residuum norm tolerance $(\varepsilon \min\{\Delta x, \Delta y\})^2$.

travelling vortex, $\varepsilon = 0.1, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	0.2233	0.7266	0.4693	2.2044
20	0.1187	1.1631	0.2082	0.9800
40	0.1164	1.2934	0.0921	0.8041
80	0.1117	1.2999	0.0528	0.7888
160	0.1054	1.3009	0.0416	0.7822

travelling vortex, $\varepsilon = 10^{-3}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	8.288e-06	4.799e-05	2.450e-02	2.226e-01
20	9.765e-06	1.134e-04	1.797e-03	1.210e-02
40	1.066e-05	1.283e-04	8.586e-04	6.384e-03
80	1.070e-05	1.296e-04	1.031e-03	1.408e-02
160	1.051e-05	1.305e-04	2.455e-03	1.024e-01

travelling vortex, $\varepsilon = 10^{-5}, CFL_u = 0.3, T = 0.1$

N	$\|\nabla_h \mathbf{Z}\|_{l^1}$	$\|\nabla_h \mathbf{Z}\|_{l^\infty}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^1}$	$\|\nabla_h \cdot \mathbf{M}\|_{l^\infty}$
10	8.291e-10	4.798e-09	2.470e-02	2.245e-01
20	9.767e-10	1.134e-08	1.818e-03	1.215e-02
40	1.068e-09	1.283e-08	8.778e-04	6.471e-03
80	1.059e-09	1.318e-08	9.927e-04	2.385e-02
160	1.057e-09	1.301e-08	2.339e-03	4.988e-02

Table 8.47.: Norms of discrete momentum divergence and free surface elevation gradient of the travelling vortex test at time $T = 0.1$. Numerical solution is obtained by the BDFRUSELLWCG scheme with the absolute residuum norm tolerance $\min\{\varepsilon, \Delta x, \Delta y\}^2$.

bottom topography	scheme	p	$\|z-p\|_{L^1}$	$\|z-p\|_{L^\infty}$	$\|q\|_{L^1}$	$\|\mathbf{m}\|_{L^\infty}$
smooth	EGRUS	-1	4.68e-14	9.99e-16	2.75e-12	8.21e-14
smooth	EGRUS	0	0	0	0	0
smooth	EGRUS	1	1.73e-12	1.82e-14	2.24e-12	8.45e-14
smooth	ARSEGRUS	-1	8.60e-14	2.00e-15	3.17e-12	1.33e-13
smooth	ARSEGRUS	0	0	0	0	0
smooth	ARSEGRUS	1	9.66e-14	2.44e-15	2.90e-12	9.22e-14
smooth	RK2CNEGRUS	-1	6.63e-13	8.66e-15	5.15e-12	1.76e-13
smooth	RK2CNEGRUS	0	0	0	0	0
smooth	RK2CNEGRUS	1	2.88e-14	1.22e-15	4.21e-12	1.35e-13
smooth	BDFEGRUS	-1	2.33e-14	1.33e-15	1.11e-11	3.46e-13
smooth	BDFEGRUS	0	0	0	0	0
smooth	BDFEGRUS	1	2.58e-14	1.11e-15	1.06e-11	5.06e-13
discontinuous	EGRUS	-1	2.53e-12	2.58e-14	1.16e-12	4.86e-14
discontinuous	EGRUS	0	0	0	0	0
discontinuous	EGRUS	1	1.99e-12	2.07e-14	7.75e-13	6.36e-14
discontinuous	ARSEGRUS	-1	2.92e-13	4.44e-15	1.21e-12	9.01e-14
discontinuous	ARSEGRUS	0	0	0	0	0
discontinuous	ARSEGRUS	1	2.77e-12	2.95e-14	6.93e-13	3.58e-14
discontinuous	RK2CNEGRUS	-1	1.17e-12	1.40e-14	1.17e-12	8.41e-14
discontinuous	RK2CNEGRUS	0	0	0	0	0
discontinuous	RK2CNEGRUS	1	4.32e-12	4.47e-14	8.77e-13	7.19e-14
discontinuous	BDFEGRUS	-1	7.74e-13	8.44e-15	4.66e-12	3.23e-13
discontinuous	BDFEGRUS	0	0	0	0	0
discontinuous	BDFEGRUS	1	8.66e-12	8.72e-14	1.34e-12	6.60e-14

Table 8.48.: Comparison of the numerical solutions obtained by the EGRUS, ARSEGRUS, RK2CNEGRUS, BDFEGRUS schemes at time $T = 10$ to the initial lake at rest states.

bottom topography	scheme	p	$\|z-p\|_{L^1}$	$\|z-p\|_{L^\infty}$	$\|q\|_{L^1}$	$\|\mathbf{m}\|_{L^\infty}$
smooth	RUSELL	-1	1.02e-11	1.04e-13	4.29e-12	1.35e-13
smooth	RUSELL	0	0	0	0	0
smooth	RUSELL	1	9.84e-12	9.97e-14	4.76e-12	1.75e-13
smooth	ARSRUSELL	-1	1.42e-12	1.95e-14	1.81e-11	6.07e-13
smooth	ARSRUSELL	0	0	0	0	0
smooth	ARSRUSELL	1	2.08e-06	6.71e-07	2.54e-04	4.27e-05
smooth	RK2CNRUSELL	-1	2.55e-12	4.73e-14	1.72e-11	5.71e-13
smooth	RK2CNRUSELL	0	0	0	0	0
smooth	RK2CNRUSELL	1	2.40e-12	4.77e-14	1.94e-11	6.35e-13
smooth	BDFRUSELL	-1	3.62e-14	1.78e-15	3.12e-12	9.74e-14
smooth	BDFRUSELL	0	0	0	0	0
smooth	BDFRUSELL	1	3.99e-13	7.55e-15	5.21e-12	1.51e-13
discontinuous	RUSELL	-1	6.31e-13	8.66e-15	1.75e-12	5.38e-14
discontinuous	RUSELL	0	0	0	0	0
discontinuous	RUSELL	1	8.00e-13	1.04e-14	1.72e-12	5.74e-14
discontinuous	ARSRUSELL	-1	8.53e-14	4.55e-15	3.34e-12	1.12e-13
discontinuous	ARSRUSELL	0	0	0	0	0
discontinuous	ARSRUSELL	1	3.56e-06	5.89e-07	4.06e-04	4.74e-05
discontinuous	RK2CNRUSELL	-1	1.00e-12	3.69e-14	9.94e-12	3.14e-13
discontinuous	RK2CNRUSELL	0	0	0	0	0
discontinuous	RK2CNRUSELL	1	7.31e-13	2.45e-14	1.57e-11	5.14e-13
discontinuous	BDFRUSELL	-1	1.92e-12	2.11e-14	1.73e-12	5.54e-14
discontinuous	BDFRUSELL	0	0	0	0	0
discontinuous	BDFRUSELL	1	1.05e-12	1.27e-14	4.41e-12	1.54e-13

Table 8.49.: Comparison of the numerical solutions obtained by the RUSELL, ARSRUSELL, RK2CNRUSELL, BDFRUSELL schemes at time $T = 10$ to the initial lake at rest states.

bottom topography	scheme	p	$\|z-p\|_{L^1}$	$\|z-p\|_{L^\infty}$	$\|q\|_{L^1}$	$\|\mathbf{m}\|_{L^\infty}$
smooth	RUSELLW	-1	4.06e-12	4.22e-14	1.80e-11	6.59e-13
smooth	RUSELLW	0	0	0	0	0
smooth	RUSELLW	1	1.60e-12	1.72e-14	1.57e-11	5.14e-13
smooth	ARSRUSELLW	-1	2.12e-13	6.00e-15	1.74e-11	6.57e-13
smooth	ARSRUSELLW	0	0	0	0	0
smooth	ARSRUSELLW	1	2.19e-13	8.99e-15	1.98e-11	6.95e-13
smooth	RK2CNRUSELLW	-1	3.10e-13	1.25e-14	1.63e-11	5.94e-13
smooth	RK2CNRUSELLW	0	0	0	0	0
smooth	RK2CNRUSELLW	1	7.43e-13	2.12e-14	2.24e-11	7.72e-13
smooth	BDFRUSELLW	-1	1.74e-12	1.93e-14	2.12e-11	7.28e-13
smooth	BDFRUSELLW	0	0	0	0	0
smooth	BDFRUSELLW	1	6.34e-12	6.53e-14	3.63e-11	1.25e-12
discontinuous	RUSELLW	-1	2.28e-12	2.49e-14	1.09e-12	4.36e-14
discontinuous	RUSELLW	0	0	0	0	0
discontinuous	RUSELLW	1	4.45e-13	5.77e-15	1.55e-12	6.72e-14
discontinuous	ARSRUSELLW	-1	2.32e-12	2.73e-14	4.14e-12	1.40e-13
discontinuous	ARSRUSELLW	0	0	0	0	0
discontinuous	ARSRUSELLW	1	5.53e-12	6.04e-14	5.13e-12	1.49e-13
discontinuous	RK2CNRUSELLW	-1	4.88e-12	6.16e-14	3.98e-12	1.22e-13
discontinuous	RK2CNRUSELLW	0	0	0	0	0
discontinuous	RK2CNRUSELLW	1	5.83e-12	7.31e-14	5.82e-12	1.77e-13
discontinuous	BDFRUSELLW	-1	3.47e-11	3.50e-13	4.39e-12	1.30e-13
discontinuous	BDFRUSELLW	0	0	0	0	0
discontinuous	BDFRUSELLW	1	4.16e-11	4.18e-13	3.84e-12	1.18e-13

Table 8.50.: Comparison of the numerical solutions obtained by the RUSELLW, ARSRUSELLW, RK2CNRUSELLW, BDFRUSELLW schemes at time $T = 10$ to the initial lake at rest states. The average (6.48b) has been used.

bottom topography	scheme	p	$\|z-p\|_{L^1}$	$\|z-p\|_{L^\infty}$	$\|\mathbf{m}\|_{L^1}$	$\|\mathbf{m}\|_{L^\infty}$
smooth	RUSELLW	-1	1.83e-13	3.11e-15	1.40e-12	5.17e-14
smooth	RUSELLW	0	0	0	0	0
smooth	RUSELLW	1	6.94e-13	7.99e-15	1.68e-12	6.62e-14
smooth	ARSRUSELLW	-1	2.27e-13	6.66e-15	1.69e-11	6.16e-13
smooth	ARSRUSELLW	0	0	0	0	0
smooth	ARSRUSELLW	1	8.27e-13	1.35e-14	2.10e-11	7.52e-13
smooth	RK2CNRUSELLW	-1	3.49e-13	1.51e-14	1.51e-11	4.99e-13
smooth	RK2CNRUSELLW	0	0	0	0	0
smooth	RK2CNRUSELLW	1	2.78e-13	1.13e-14	2.39e-11	7.87e-13
smooth	BDFRUSELLW	-1	5.12e-12	5.33e-14	2.02e-11	7.07e-13
smooth	BDFRUSELLW	0	0	0	0	0
smooth	BDFRUSELLW	1	3.32e-12	3.51e-14	4.08e-11	1.33e-12
discontinuous	RUSELLW	-1	1.49e-12	1.68e-14	8.34e-13	2.62e-14
discontinuous	RUSELLW	0	0	0	0	0
discontinuous	RUSELLW	1	5.01e-13	6.22e-15	9.77e-13	2.98e-14
discontinuous	ARSRUSELLW	-1	1.43e-12	1.93e-14	4.22e-12	1.48e-13
discontinuous	ARSRUSELLW	0	0	0	0	0
discontinuous	ARSRUSELLW	1	5.56e-12	6.10e-14	5.24e-12	2.32e-13
discontinuous	RK2CNRUSELLW	-1	6.16e-12	7.43e-14	4.09e-12	1.13e-13
discontinuous	RK2CNRUSELLW	0	0	0	0	0
discontinuous	RK2CNRUSELLW	1	4.82e-12	6.44e-14	5.78e-12	1.81e-13
discontinuous	BDFRUSELLW	-1	3.68e-11	3.70e-13	3.98e-12	1.35e-13
discontinuous	BDFRUSELLW	0	0	0	0	0
discontinuous	BDFRUSELLW	1	3.70e-11	3.73e-13	4.46e-12	1.46e-13

Table 8.51.: Comparison of the numerical solutions obtained by the RUSELLW, ARSRUSELLW, RK2CNRUSELLW, BDFRUSELLW schemes at time $T = 10$ to the initial lake at rest states. The average (6.50) has been used.

8.4. Evaluation of the introduced IMEX schemes

In the previous sections of this chapter we have studied the accuracy, stability, efficiency, asymptotic preserving and well-balanced property of various IMEX finite volume schemes, as well as the behaviour in the vicinity of discontinuities. To this end we have applied the IMEX finite volume schemes for test problems with continuous and discontinuous data. The aim of this section is to evaluate the results of our investigation.

In Section 8.1 we have considered an experiment that consists of Riemann problems for the Froude numbers $\varepsilon \in \{0.8, 0.1\}$. We have observed the first order scheme to perform very diffusive, while the second order schemes suffered from oscillations near to discontinuities. In order to suppress the oscillations we have applied the minmod slope limiter to the explicit, non-linear terms. This works out nicely in the case of the Froude number $\varepsilon = 0.8$, but the numerical solution still oscillates for $\varepsilon = 0.1$. In Remark 8.1.2, we suggest a new way for using limiters in the framework of the IMEX schemes based on the defect correction approach. This approach should be further investigated in our future study.

In Sections 8.2 we have considered the almost smooth travelling vortex test with constant and non-constant bottom topography to study the accuracy, stability, efficiency and asymptotic preserving property of various IMEX finite volume schemes. In Section 8.3 we have tested the well-balanced property of the IMEX finite volume schemes. While all of our schemes have demonstrated to be well-balanced, we have observed different behaviour of the schemes with respect to the accuracy, stability, efficiency and asymptotic preserving property.

In what follows we compare the properties of the introduced IMEX finite volume schemes and suggest which one to use for continuous problems. To this end, we summarise our observations briefly in Table 8.52. Further we will not consider the CFDRUS, ARSCFDRUS, RK2CNCFDRUS, BDF2CFDRUS schemes, since they are algebraically equivalent to the RUSELLW, ARSRUSELLW, RK2CNRUSELLW, BDF2RUSELLW schemes, cf. Theorem 6.5.1, but computational more expensive due to larger linear systems to be solved, cf. Section 4.1. Also, we have observed the checkerboard instability in the numerical solutions of the travelling vortex test for the Froude number $\varepsilon = 10^{-8}$ obtained by the CFDRUS scheme, whereas the instability did not appear in the solution obtained by the RUSELLW scheme, cf. Figures 8.12-8.15.

Let us consider the first order IMEX schemes: Here, the RUSELL and RUSELLW schemes are stable and provide uniform convergence rates as well as uniform momentum errors with respect to the Froude number. For test problems presented in Section 8.2, they yield accurate results - even up to $\varepsilon = 10^{-8}$. The RUSELLW scheme is asymptotic preserving, whereas the RUSELL scheme is not. Therefore we recommend to use the RUSELLW scheme. However, it may be argued about the importance of a divergence constraint of the numerical solution. Let us also remark that the RUSELL scheme is computationally cheaper than the RUSELLW scheme due to a smaller stencil.

Comparing the second order IMEX schemes, we should not use the ARSRUSELL, RK2CNRUSELL, BDF2RUSELL schemes due to possible development of instabilities, cf. Section 8.2.4. Also, the ARSEGRUS, RK2CNEGRUS, BDF2EGRUS schemes should not be used since they do not provide uniform convergence rates on a fixed finite volume mesh. The RK2CNRUSELLW scheme develops numerical artifacts in the numerical solu-

tion of the free surface elevation for Froude numbers $\varepsilon \leq 10^{-3}$ and is not computationally cheaper than the two remaining IMEX schemes - ARSRUSELLW, BDFRUSELLW. The ARSRUSELLW, BDF2RUSELLW schemes seem to be asymptotic preserving and stable. The observed experimental convergence rates are of second order uniformly for the momentum and for $\varepsilon \geq 10^{-3}$ also for the free surface elevation. The BDF2RUSELLW scheme has even provided uniform convergence rates also for the free surface elevation. Further, let us point out that we have to solve only one linear system per time step for the BDF2RUSELLW scheme, whereas two linear systems have to be solved for the ARSRUSELLW scheme. But we have to use the CFL_u-number 0.3 for the BDF2RUSELLW scheme. If we can apply the ARSRUSELLW scheme successfully with a CFL_u-number larger than 0.6, the ARSRUSELLW scheme would be computationally cheaper than the BDF2RUSELLW scheme. Consequently, we recommend to either use the ARSRUSELLW or the BDF2RUSELLW scheme.

Though, the presented IMEX schemes has different properties that show significantly different quality of numerical results for Froude numbers smaller than 10^{-3}, all of the derived stable second order IMEX schemes, i.e. ARSEGRUS, RK2CNEGRUS, BDF2EGRUS, ARSRUSELLW, RK2CNRUSELLW, BDF2RUSELLW, can be used to describe large scale oceanographic and meteorologic phenomena due to the fact that typical Froude numbers are around 10^{-2}.

In Section 8.2, we have clearly pointed out that our three recommended schemes provide better numerical results than the HLLC, RK2HLLC schemes for the travelling vortex test. But if a direct solver is applied the computational costs increase rapidly with the number of mesh cells, cf. Table 8.44. In Section 8.2.5, we have successfully applied the CG method to solve the arising linear systems iteratively, which led to a speed-up of the factors of around $10 - 20$. To this end we have proposed two different aborting criteria for the iteration. The corresponding BDFRUSELLWCG scheme clearly outperforms the RK2HLLC scheme with regard to accuracy and computational costs, if the Froude number $\varepsilon \leq 0.1$. If the Froude number $\varepsilon = 0.8$, the CPU time of the BDFRUSELLWCG scheme seems to be around four times larger than the one of the RK2HLLC scheme. But the constraints on the free surface elevation and the divergence may weaken, if an iterative solver is applied, cf. Remark 8.2.1. This depends on how exact the arising linear systems are solved.

In a general setting with non-constant bottom topography the matrix E, cf. (7.3), is not symmetric and thus another iterative solver than the CG method has to be used. Further detailed investigations on the use of appropriate iterative solvers with some suitable preconditioners should allow us to obtain even larger speed-up than presented here. This will be a goal of future study.

Further detailed investigations on the use of appropriate iterative solvers with some suitable preconditioners should allow us to obtain even larger speed-up than presented here. This is a goal of future study.

	artifacts appear, if	unstable for	EOC $z \approx$	EOC $\mathbf{m} \approx$	uniform momentum errors with respect to the Forude number ε	$\nabla_h \mathbf{Z} = \mathcal{O}(\varepsilon^2)$, well-balanced	$\nabla_h \cdot \mathbf{M} = \mathcal{O}(\varepsilon^a)$
HLLC	$\varepsilon \leq 0.1$		1, if $\varepsilon \geq 0.1$		no	yes	$a = 1$
EGRUS	$\varepsilon \leq 0.1$		1, if $\varepsilon \geq 0.1$		no	yes	$a = 1$
RUSELL			1		yes	yes	$a = 0$
RUSELLW			1		yes	yes	$a = 2$
RK2HLLC	$\varepsilon \leq 10^{-3}$		2, if $\varepsilon \geq 10^{-2}$	2, if $\varepsilon \geq 10^{-3}$	no	yes	$a = 1$
ARSEGRUS	$\varepsilon \leq 10^{-3}$		2, if $\varepsilon \geq 10^{-3}$		no	yes	$a = 1$
RK2CNEGRUS	$\varepsilon \leq 10^{-3}$		2, if $\varepsilon \geq 10^{-3}$		no	yes	$a = 0$
BDFEGRUS	$\varepsilon \leq 10^{-3}$		2, if $\varepsilon \geq 10^{-3}$		no	yes	$a = 1$
ARSRUSELL	$\varepsilon \leq 10^{-2}$	$\varepsilon \leq 10^{-2}$, long time simulation	2, if $\varepsilon \geq 0.1$		yes	yes	$a = 0$
RK2CNRUSELL	$\varepsilon \leq 10^{-3}$	long time simulation	2, if $\varepsilon \geq 10^{-2}$	2	yes	yes	$a = 0$
BDF2RUSELL		long time simulation	2		yes	yes	$a = 0$
ARSRUSELLW			2, if $\varepsilon \geq 10^{-3}$	2	yes	yes	$a = 2$
RK2CNRUSELLW	$\varepsilon \leq 10^{-3}$		2, if $\varepsilon \geq 10^{-3}$	2	yes	yes	$a = 2$
BDF2RUSELLW			2		yes	yes	$a = 2$

Table 8.52.: Short summary of the travelling vortex tests in Section 8.2.

9. Conclusion

In the present thesis we have derived and analysed new IMEX finite volume schemes for the shallow water flows with low Froude numbers. The main idea of these schemes is to split the shallow water equations into a stiff, linear part governing the motion of fast waves and a non-stiff one that controls the remaining slow flow. To this end we have used the alternative formulation SWE (2.30) and the splitting introduced by Giraldo et al. [44], cf. Chapter 2. The splitting is crucial for the IMEX schemes, since the stiff and non-stiff subsystem are hyperbolic with "correct" signal speeds. In particular, the stiff subsystem is the first order wave equation system with the wave velocity $\sqrt{-gb} \approx \sqrt{gh}$.

The time step Δt depends on the grid widths $\Delta x, \Delta y$ and the maximum explicitly approximated signal speed via the CFL stability condition. Consequently, treating the stiff terms implicitly and the non-stiff ones explicitly, the CFL stability condition relaxes to

$$\max\left\{\frac{|u_1|}{\Delta x}, \frac{|u_2|}{\Delta y}\right\}\Delta t \leq CFL_u < 0.5, \tag{9.1}$$

since the maximum signal speed corresponding to the non-stiff terms is $2|\mathbf{u}\cdot\mathbf{n}|$, cf. Section 2.7.4. Consequently, we are able to use larger time steps that only depend on the advection velocity - more precisely the time steps enlarge approximately by the factor $1/(2\varepsilon)$, where ε denotes the Froude number. In the presented numerical calculations we have used successfully the CFL_u numbers 0.3 and 0.45. Using the CFL_u-number 0.45 the time steps was larger approximately by factor $0.34/\varepsilon$, cf. Table 8.38.

In order to obtain a semi-implicit time discretisation we have applied the IMEX Runge-Kutta and IMEX multi-step schemes, cf. Chapter 4. We have shown that the corresponding semi-discrete schemes are conditionally well-balanced and asymptotic preserving in Section 4.6 and Chapter 5. The condition on the scheme to be well-balanced is the non-negativity of the boundary integral (4.37), which is automatically satisfied for periodic, homogeneous Dirichlet or Neumann boundary conditions of the free surface elevation z. We have studied the asymptotic preserving property by applying formal asymptotic analysis. Further, we have assumed periodic boundary conditions and well-prepared initial data. We have shown that the semi-discrete IMEX Runge-Kutta methods are asymptotic preserving, if they are globally stiffly accurate, i.e. the last internal IMEX Runge-Kutta stage is already the solution at a new time step. This coincides with the results from Boscarino et al [18]. The semi-discrete consistent IMEX multi-step schemes are also asymptotic preserving. In particular, we have pointed out that the IMEX Euler, RK2CN, ARS(2,2,2) and SBDF schemes are asymptotic preserving and well-balanced for periodic boundary conditions.

The spatial discretisation is obtained by finite volumes. Thus, mass and momentum are exactly conserved on the discrete level. The fluxes of the non-stiff subsystem can be

approximated by standard consistent compressible solvers. We use the Rusanov flux for this purpose. The fluxes of the stiff, linear subsystem are approximated either by central differences or by approximate local evolution operators, that take into account multi-dimensional wave propagation. We apply the above mentioned first and second order IMEX time discretisations to obtain first and second order IMEX finite volume schemes, where the second order spatial approximation is obtained using the MUSCL approach. We have proved the corresponding IMEX finite volume schemes to be well-balanced, if the numerical solution is unique, and asymptotic preserving under certain assumptions, cf. Lemma 7.1.7, 7.1.12, 7.2.1. Thereby the spatial and time discretisations plays a crucial role to obtain an asymptotic preserving scheme. Indeed, it is not sufficient to use a asymptotic preserving time discretisation to obtain a fully discrete asymptotic preserving scheme: Though the RK2CN scheme is globally stiffly accurate, we have pointed out that the RK2CN time discretisation should not provide an asymptotic preserving scheme, unless the momentum initial data are not satisfying a discrete divergence constraint in the low Froude number limit, cf. Remark 7.2.2. However, the other considered time discretisations, IMEX Euler, ARS(2,2,2), SBDF, combined with suitable spatial finite volume discretisation yield fully discrete asymptotic preserving schemes.

In Chapter 8 we have observed first and second order numerical IMEX finite volume schemes with respect to accuracy, stability, asymptotic preserving property and efficiency. For a detailed evaluation we refer the reader to Section 8.4, which summarises the results from numerical experiments. The main results are the following: Using central finite differences to approximate the stiff terms in a suitable way leads to stable, asymptotic preserving, well-balanced IMEX finite volume schemes. Our corresponding first and second order schemes RUSELLW, ARSRUSELLW, BDFRUSELLW show uniform convergence rates with regard to the Froude number. If we use the derived approximate local evolution operator instead, the convergence rates are not uniform with respect to the Froude number. Instead, the rates depend on the used grid and Froude number. Note however, that for typical large-scale oceanographic and meteorologic applications, i.e. the Froude number is approximately 10^{-2}, the results of the second order schemes are good, though still outperformed by the central difference discretisations.

Further, all of our theoretical results with respect to the asymptotic preserving property are completely supported by the performed numerical tests. Moreover, numerical tests imply that the schemes are asymptotic preserving also for non-constant bottom topography.

A comparison of the IMEX finite volume schemes with the explicit HLLC-type schemes, both of first and second order, yields the following. The errors and convergence rates of the IMEX schemes, where the approximate local evolution operators have been applied, behave similar to those of explicit HLLC-type schemes. However, the IMEX schemes are a little more accurate and use larger time-steps. The IMEX schemes with the central differences applied are clearly more accurate. Our favored IMEX finite volume schemes, RUSELLW, BDFRUSELLW, ARSRUSELLW, outperform standard explicit schemes, if the arising linear systems are solved efficiently by an iterative solver. In Section 8.2.5 we have successfully applied the CG-method for the second order IMEX scheme BDFRUSELLW. Consequently, we have observed speed-up factors around 10 for the Froude number 0.01 in comparison to an explicit second order HLLC-type scheme. The speed-up increases for lower Froude numbers. Thus the BDFRUSELLW scheme clearly out-

performs the explicit one. We expect to observe similar results for other asymptotic preserving IMEX finite volume schemes. Moreover, our approach can be generalised to schemes of arbitrary order.

There are still several open problems for future research. If discontinuities appear in the solution the numerical fluxes of a finite volume scheme are typically limited in a non-linear way. If the Froude number is around one it seems to be sufficient to only limit the explicitly treated non-stiff parts to suppress oscillations near discontinuities. If the Froude number is small, e.g. 0.1, this does not suppress oscillations. Thus, we need to limit the stiff system, while maintaining it linear. To this end, we propose to apply artificial viscosity approach or a defect correction method.

The generalistion of the introduced approach to other systems, e.g. the Euler equations, is also of interest. First results with a different splitting has already been obtained in [92]. The construction of asymptotic preserving schemes for the non-hydrostatic Euler equations with gravity, see appendix A, is of particular interest for meteorologic applications.

A. Non-hydrostatic multidimensional Euler equations

The non-hydrostatic Euler equations used in [90, 121] read

$$\begin{aligned} &\mathbf{w}_t + \nabla \cdot \tilde{\mathcal{F}}(\mathbf{w}) = \tilde{K}(\mathbf{w}), \\ &\mathbf{w} = \begin{bmatrix} \rho \\ \rho\mathbf{u} \\ \rho\theta \end{bmatrix}, \quad \tilde{\mathcal{F}}(\mathbf{w}) = \begin{bmatrix} \rho\mathbf{u}^T \\ \rho\mathbf{u} \otimes \mathbf{u}^T + p\mathbb{1}_d \\ \rho\mathbf{u}^T\theta \end{bmatrix}, \quad \tilde{K}(\mathbf{w}) = \begin{bmatrix} 0 \\ -\rho g \mathbf{e}_d \\ 0 \end{bmatrix}, \end{aligned} \tag{A.1}$$

where ρ is the density, $\mathbf{u}$ the velocity, θ the potential temperature, p the pressure, d the spatial dimension and g the gravity constant. The state equation is

$$p = p_0 \left(\frac{R\rho\theta}{p_0} \right)^\gamma, \quad \bar{p} = p_0 \left(\frac{R\overline{\rho\theta}}{p_0} \right)^\gamma, \tag{A.2}$$

where p_0 is the pressure at height 0, R is the gas constant and γ the heat capacity ratio. We follow [90] and consider the Euler equations (A.1) in terms of perturbations of a background equilibrium state $\bar{\mathbf{w}} = (\bar{\rho}, 0, \overline{\rho\theta})$ in hydrostatic balance. More precisely we have

$$\rho = \bar{\rho} + \rho', \quad \theta = \bar{\theta} + \theta', \quad \rho\theta = \overline{\rho\theta} + (\rho\theta)', \quad \overline{\rho\theta} = \bar{\rho}\bar{\theta}, \quad p = \bar{p} + p', \tag{A.3}$$

where the the background state variables $\bar{\rho}, \bar{\theta}, \overline{\rho\theta}, \bar{p}$ are constant in time and vary in space only along the vertical x_d-axis. Moreover we have

$$\bar{p}_{x_d} = -\bar{\rho} g. \tag{A.4}$$

We introduce the background state perturbation $\mathbf{w}' := \mathbf{w} - \bar{\mathbf{w}}$ and rewrite the non-hydrostatic Euler equation (A.1) to

$$\begin{aligned} &\mathbf{w}'_t + \nabla \cdot \mathcal{F}(\mathbf{w}') = K(\mathbf{w}'), \\ &\mathbf{w}' = \begin{bmatrix} \rho' \\ \mathbf{q} := \rho\mathbf{u} \\ (\rho\theta)' \end{bmatrix}, \quad \mathcal{F}(\mathbf{w}') = \begin{bmatrix} \mathbf{q}^T \\ \dfrac{\mathbf{q} \otimes \mathbf{q}^T}{\rho' + \bar{\rho}} + p'\mathbb{1}_d \\ \mathbf{q}^T\theta \end{bmatrix}, \quad K(\mathbf{w}') = \begin{bmatrix} 0 \\ -\rho' g \mathbf{e}_d \\ 0 \end{bmatrix}, \end{aligned} \tag{A.5}$$

where

$$p' = p - \bar{p} = p_0 \left(\frac{R\rho\theta}{p_0} \right)^\gamma - p_0 \left(\frac{R\overline{\rho\theta}}{p_0} \right)^\gamma. \tag{A.6}$$

We expand the pressure perturbation (A.6) in a Taylor series at $\overline{\rho\theta}$. This gives us

$$p' = p'_L + p'_{NL}, \quad p'_L = \frac{d\bar{p}}{d\overline{\rho\theta}}(\rho\theta)' = \frac{\bar{c}^2}{\bar{\theta}}(\rho\theta)', \quad p'_{NL} = p - \bar{p} - p'_L = \mathcal{O}((\rho\theta)'^2), \tag{A.7}$$

where

$$c = \sqrt{\frac{\gamma p}{\rho}}, \quad \bar{c} = \sqrt{\frac{\gamma\bar{p}}{\bar{\rho}}} \tag{A.8}$$

are the speeds of sound of the full system and of the hydrostatic reference state.

We split the Euler equation (A.5) into a linear and nonlinear part

$$\mathcal{F}_L(\mathbf{w}') = \begin{bmatrix} \mathbf{q}^T \\ p'_L \mathbb{1}_d \\ \mathbf{q}^T\bar{\theta} \end{bmatrix}, \quad \mathcal{F}_{NL}(\mathbf{w}') = \begin{bmatrix} 0 \\ \dfrac{\mathbf{q}\otimes\mathbf{q}^T}{\rho} + p'_{NL}\mathbb{1}_d \\ \mathbf{q}^T\theta' \end{bmatrix}, \tag{A.9}$$

where

$$\theta' = \frac{(\rho\theta)' - \rho'\bar{\theta}}{\rho}. \tag{A.10}$$

Here, the equation for θ' (A.10) is obtained from

$$\rho\theta = (\rho' + \bar{\rho})(\theta' + \bar{\theta}) = \overline{\rho\theta} + \bar{\rho}\theta' + \rho'\bar{\theta} + \rho'\theta'. \tag{A.11}$$

A.1. Eigenstructure of the nonlinear subsystem

We consider the nonlinear subsystem

$$\mathbf{w}'_t + \nabla\cdot\mathcal{F}_{NL}(\mathbf{w}') = 0 \tag{A.12}$$

with $\mathbf{w}'$ and $\mathcal{F}_{NL}$ according to (A.5), (A.9). We have

$$\mathcal{F}_{NL}(\mathbf{w}')\mathbf{n} = \begin{bmatrix} 0 \\ \dfrac{\mathbf{q}}{\rho}(\mathbf{q}\cdot\mathbf{n}) + p'_{NL}\mathbf{n} \\ \theta'\mathbf{q}\cdot\mathbf{n} \end{bmatrix}, \tag{A.13}$$

$$\mathbb{P}(\mathbf{w}',\mathbf{n}) = \begin{bmatrix} 0 & 0 & 0 \\ -\dfrac{\mathbf{q}}{\rho^2}\mathbf{q}\cdot\mathbf{n} & \mathbb{1}_d\dfrac{\mathbf{q}\cdot\mathbf{n}}{\rho} + \dfrac{\mathbf{q}\cdot\mathbf{n}^T}{\rho} & \left(\dfrac{c^2}{\theta} - \dfrac{\bar{c}^2}{\bar{\theta}}\right)\mathbf{n} \\ -\dfrac{\theta}{\rho}\mathbf{q}\cdot\mathbf{n} & \theta'\mathbf{n}^T & \dfrac{\mathbf{q}\cdot\mathbf{n}}{\rho} \end{bmatrix}, \tag{A.14}$$

where $\mathbf{n}$ is a unit vector. The matrix $\mathbb{P}$ has the eigenvalues

$$\lambda_1 = 0, \quad \lambda := \lambda_2 = ... = \lambda_d = \frac{\mathbf{q}\cdot\mathbf{n}}{\rho}, \quad \lambda_{d+1} = \frac{3\lambda - \sqrt{\lambda^2 + 4\theta'\Delta c}}{2}, \tag{A.15}$$
$$\lambda_{d+2} = \frac{3\lambda + \sqrt{\lambda^2 + 4\theta'\Delta c}}{2}, \quad \Delta c = \frac{c^2}{\theta} - \frac{\bar{c}^2}{\bar{\theta}}$$

with the corresponding right eigenvectors

$$\mathbf{r}^1 = \begin{bmatrix} 2\lambda^2 - \theta'\Delta c \\ \left[\lambda^2 + \bar{\theta}\Delta c\right]\mathbf{u} - \lambda\Delta c(2\theta - \theta')\mathbf{n} \\ \lambda^2(2\theta - \theta') \end{bmatrix}, \quad \mathbf{r}^2 = \begin{bmatrix} 0 \\ \mathbf{t}^1 \\ 0 \end{bmatrix}, \quad ..., \quad \mathbf{r}_d = \begin{bmatrix} 0 \\ \mathbf{t}^{d-1} \\ 0 \end{bmatrix}, \tag{A.16}$$

$$\mathbf{r}^{d+1} = \begin{bmatrix} 0 \\ 2\mathbf{u} - (\lambda + \sqrt{\lambda^2 + 4\theta'\Delta c})\mathbf{n} \\ 2\theta' \end{bmatrix}, \quad \mathbf{r}^{d+2} = \begin{bmatrix} 0 \\ 2\mathbf{u} - (\lambda - \sqrt{\lambda^2 + 4\theta'\Delta c})\mathbf{n} \\ 2\theta' \end{bmatrix}.$$

Note that the range of λ^2 can take every value between 0 and $\|\mathbf{u}\|^2$. Consequently, the nonlinear system (A.12) is hyperbolic, if and only if $\theta'\Delta c \geq 0$. Using (A.2) and (A.11) we obtain that $\theta'\Delta c \geq 0$, if and only if $\theta'(\rho\theta' + \rho'\bar{\theta}) \geq 0$. Therefore, θ' must not be between $-\rho'\bar{\theta}/\rho$ and 0. Since θ' is the potential temperature perturbation, it may become zero. If θ' and λ are zero, the right eigenvectors are linear dependent. Thus, (A.12) is not diagonally hyperbolic.

A.2. Eigenstructure of the linear system

We consider the linear subsystem

$$\mathbf{w}'_t + \nabla \cdot \mathcal{F}_L(\mathbf{w}') = K(\mathbf{w}') \tag{A.17}$$

with $\mathbf{w}'$ and $\mathcal{F}_L, K$ according to (A.5), (A.9). We have

$$\mathcal{F}_L(\mathbf{w}')\mathbf{n} = \begin{bmatrix} \mathbf{q} \cdot \mathbf{n} \\ p'_L\mathbf{n} \\ \bar{\theta}\mathbf{q} \cdot \mathbf{n} \end{bmatrix}, \quad \mathbb{P}(\mathbf{w}', \mathbf{n}) = \begin{bmatrix} 0 & \mathbf{n}^T & 0 \\ 0 & 0 & \frac{\bar{c}^2}{\bar{\theta}}\mathbf{n} \\ 0 & \bar{\theta}\mathbf{n}^T & 0 \end{bmatrix}, \tag{A.18}$$

where $\mathbf{n}$ is a unit vector. The matrix $\mathbb{P}$ has the eigenvalues

$$\lambda_1 = -\bar{c}, \quad \lambda_2 = ... = \lambda_{d+1} = 0, \quad \lambda_{d+2} = \bar{c} \tag{A.19}$$

with the corresponding right and left eigenvectors

$$\mathbf{r}^1 = \begin{bmatrix} 1 \\ \frac{}{c} \\ -\mathbf{n} \\ \bar{\theta} \\ \frac{}{c} \end{bmatrix}, \quad \mathbf{r}^2 = \begin{bmatrix} 0 \\ \mathbf{t}^1 \\ 0 \end{bmatrix}, \quad ..., \quad \mathbf{r}_d = \begin{bmatrix} 0 \\ \mathbf{t}^{d-1} \\ 0 \end{bmatrix}, \quad \mathbf{r}^{d+1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{r}^{d+2} = \begin{bmatrix} 1 \\ \frac{}{c} \\ \mathbf{n} \\ \bar{\theta} \\ \frac{}{c} \end{bmatrix}, \tag{A.20}$$

$$\mathbf{l}^1 = \frac{1}{2}\begin{bmatrix} 0 \\ -\mathbf{n} \\ \bar{c} \\ \frac{}{\theta} \end{bmatrix}, \quad \mathbf{l}^2 = \begin{bmatrix} 0 \\ \mathbf{t}^1 \\ 0 \end{bmatrix}, \quad ..., \quad \mathbf{l}_d = \begin{bmatrix} 0 \\ \mathbf{t}^{d-1} \\ 0 \end{bmatrix}, \quad \mathbf{l}^{d+1} = \begin{bmatrix} 1 \\ 0 \\ -\frac{1}{\bar{\theta}} \end{bmatrix}, \quad \mathbf{l}^{d+2} = \frac{1}{2}\begin{bmatrix} 0 \\ \mathbf{n} \\ \bar{c} \\ \frac{}{\theta} \end{bmatrix}.$$

Here, $\left\{\mathbf{t}^1, ..., \mathbf{t}^{d-1}\right\}$ is an orthogonal basis of the tangent plane to $\mathbf{n}$. Let us point out, that the right eigenvectors $\mathbf{r}^i, i = 1, ..., d+2$, are linear independent, since $\bar{\theta}, \bar{c} > 0$. Consequently, the nonlinear system (A.17) is diagonally hyperbolic.

A.3. Exact evolution operator for the linear subsystem

The linear subsystem (A.5) reads

$$\mathbf{w}'_t + \sum_{i=1}^{d} A_i \mathbf{w}'_{x_i} = Q, \tag{A.21}$$

$$A_i = \begin{bmatrix} 0 & \mathbf{e}_i^T & 0 \\ 0 & 0 & \frac{\bar{c}^2}{\bar{\theta}}\mathbf{e}_i \\ 0 & \bar{\theta}\mathbf{e}_i^T & 0 \end{bmatrix}, \quad Q = -\begin{bmatrix} 0 \\ g\left(\rho' - (\rho\theta)'\frac{\gamma-1}{\bar{\theta}}\right)\mathbf{e}_d \\ q_d \bar{\theta}_{x_d} \end{bmatrix}, \tag{A.22}$$

in quasi-linear form. Here, additional source terms appear since $\bar{c}, \bar{\theta}$ depend on x_d. We use (A.2), (A.7) and the hydrostatic balance (A.4) to obtain the source term Q in (A.21). We follow the procedure from Section 3.2 to derive an exact evolution operator for the linearised linear subsystem of (A.5), i.e. contrary to the derivation in Section 3.3, we will use frozen Jacobians. Thus $\bar{c} = \bar{c}^a, \bar{\theta} = \bar{\theta}^a$ in the Jacobians $A_i, i = 1, ..., d$, are constant values. Consequently, the matrices consisting of right- and left-eigenvectors (A.20) read

$$R = \begin{bmatrix} \frac{1}{\bar{c}^a} & 0 & ... & 0 & 1 & \frac{1}{\bar{c}^a} \\ -\mathbf{n} & \mathbf{t}^1 & ... & \mathbf{t}^{d-1} & 0 & \mathbf{n} \\ \frac{\bar{\theta}^a}{\bar{c}^a} & 0 & ... & 0 & 0 & \frac{\bar{\theta}^a}{\bar{c}^a} \end{bmatrix}, \quad R^{-1} = \begin{bmatrix} 0 & -\frac{\mathbf{n}^T}{2} & \frac{\bar{c}^a}{2\bar{\theta}^a} \\ 0 & (\mathbf{t}^1)^T & 0 \\ \vdots & \vdots & \vdots \\ 0 & (\mathbf{t}^{d-1})^T & 0 \\ 1 & 0 & -\frac{1}{\bar{\theta}^a} \\ 0 & \frac{\mathbf{n}^T}{2} & \frac{\bar{c}^a}{2\bar{\theta}^a} \end{bmatrix}. \tag{A.23}$$

Further, we compute the characteristic variables

$$\mathbf{v} = R^{-1}\mathbf{w}' = \begin{bmatrix} \frac{1}{2}\left[\frac{\bar{c}^a}{\bar{\theta}^a}(\rho\theta)' - \mathbf{q}\cdot\mathbf{n}\right] \\ \mathbf{q}\cdot\mathbf{t}^1 \\ \mathbf{q}\cdot\mathbf{t}^2 \\ \vdots \\ \mathbf{q}\cdot\mathbf{t}^{d-1} \\ \rho' - \frac{(\rho\theta)'}{\bar{\theta}^a} \\ \frac{1}{2}\left[\frac{\bar{c}^a}{\bar{\theta}^a}(\rho\theta)' + \mathbf{q}\cdot\mathbf{n}\right] \end{bmatrix} \tag{A.24}$$

and

$$B_i = R^{-1}A_iR = \bar{c}^a \begin{bmatrix} -n_i & \frac{t_i^1}{2} & ... & \frac{t_i^{d-1}}{2} & 0 & 0 \\ t_i^1 & 0 & ... & 0 & 0 & t_i^1 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ t_i^{d-1} & 0 & ... & 0 & 0 & t_i^{d-1} \\ 0 & 0 & ... & 0 & 0 & 0 \\ 0 & \frac{t_i^1}{2} & ... & \frac{t_i^{d-1}}{2} & 0 & n_i \end{bmatrix}, \quad D_i = \ \mathrm{diag}(B_i). \quad i = 1, ..., d. \tag{A.25}$$

Due to (3.35), we get

$$S = -\frac{\bar{c}^a}{2} \begin{bmatrix} \sum_{j=1}^{d-1} \mathbf{t}^j \cdot \nabla(\mathbf{q} \cdot \mathbf{t}^j) \\ 2\frac{\bar{c}^a}{\bar{\theta}^a}\mathbf{t}^1 \cdot \nabla(\rho\theta)' \\ \vdots \\ 2\frac{\bar{c}^a}{\bar{\theta}^a}\mathbf{t}^{d-1} \cdot \nabla(\rho\theta)' \\ 0 \\ \sum_{j=1}^{d-1} \mathbf{t}^j \cdot \nabla(\mathbf{q} \cdot \mathbf{t}^j) \end{bmatrix}, \quad F = -\begin{bmatrix} -\frac{n_d}{2} \\ t_d^1 \\ \vdots \\ t_d^{d-1} \\ 0 \\ \frac{n_d}{2} \end{bmatrix} g\left(\rho' - (\gamma - 1)\frac{(\rho\theta)'}{\bar{\theta}}\right) - \begin{bmatrix} \frac{\bar{c}^a}{2} \\ 0 \\ \vdots \\ 0 \\ -1 \\ \frac{\bar{c}^a}{2} \end{bmatrix} \frac{q_d \bar{\theta}_{x_d}}{\bar{\theta}^a}. \tag{A.26}$$

Next, we calculate the bicharacteristic curves using (3.29). Thus, we have $\chi_j^i = (B_j)_{ii}$. The right-hand side of the ODEs for the normal direction is zero, since the Jacobians are frozen. Thus, the ODE systems for the $d+2$ families of bicharacteristic curves read

$$\frac{d\mathbf{x}^1}{dt} = -\bar{c}^a\mathbf{n}(\boldsymbol{\omega}), \quad \frac{d\mathbf{x}^2}{dt} = ... = \frac{d\mathbf{x}^{d+1}}{dt} = 0, \quad \frac{d\mathbf{x}^{d+2}}{dt} = \bar{c}^a\mathbf{n}(\boldsymbol{\omega}). \tag{A.27}$$

For given initial data $\mathbf{x}^i(t^{n+1}) = \mathbf{x}_P$, $\boldsymbol{\omega}$ we have

$$\begin{aligned} &\mathbf{x}^1(t;\boldsymbol{\omega}) = \mathbf{x}_P + (t^{n+1} - t)\bar{c}^a\mathbf{n}(\boldsymbol{\omega}), \quad \mathbf{x}^2(t;\boldsymbol{\omega}) = ... = \mathbf{x}^{d+1}(t;\boldsymbol{\omega}) = \mathbf{x}_P, \\ &\mathbf{x}^{d+2}(t;\boldsymbol{\omega}) = \mathbf{x}_P - (t^{n+1} - t)\bar{c}^a\mathbf{n}(\boldsymbol{\omega}). \end{aligned} \tag{A.28}$$

The next step is to compute the exact evolution operator

$$\mathbf{w}(P) = \frac{1}{|S^{d-1}|}\int_O \left\{ R\tilde{\mathbf{v}}^n + \int_{t^n}^{t^{n+1}} R(\tilde{F} + \tilde{S})(t)\ dt \right\} |dS^{d-1}|(\boldsymbol{\omega})\ d\boldsymbol{\omega}. \tag{A.29}$$

Let us recall that tilde denotes that the i-th component is considered along the i-th bicharacteristic. Due to the symmetry of the first and last bicharacteristic in (A.28), we can rewrite the integrals along the last bicharacteristic in analogous integrals along the

first one. To this end, we shift $\boldsymbol{\omega} \mapsto \boldsymbol{\omega} + \pi \mathbf{e}_1$, cf. Section 3.3. Thus, we obtain

$$\int_O R\tilde{v}^n \ |dS^{d-1}|d\boldsymbol{\omega} = \int_O \left\{ \begin{bmatrix} 1 \\ -\bar{c}^a \mathbf{n}(\boldsymbol{\omega}) \\ \bar{\theta}^a \end{bmatrix} \left[\frac{(\rho\theta)'}{\bar{\theta}^a} - \frac{\mathbf{q} \cdot \mathbf{n}(\boldsymbol{\omega})}{\bar{c}^a} \right] \right\} (\mathbf{x}^n, t^n) \ |dS^{d-1}(\boldsymbol{\omega})|d\boldsymbol{\omega} \tag{A.30}$$

$$+ \begin{bmatrix} |S^{d-1}| \left(\rho' - \dfrac{(\rho\theta)'}{\bar{\theta}} \right) \\ (|S^{d-1}| - N_d^2)\mathbf{q} \\ 0 \end{bmatrix} (\mathbf{x}_P, t^n),$$

where $\mathbf{x}^n = \mathbf{x}^1(t^n; \boldsymbol{\omega})$, and

$$\int_{t^n}^{t^{n+1}} \int_O R(\tilde{F} + \tilde{S})(t; \boldsymbol{\omega}) |dS^{d-1}|(\boldsymbol{\omega}) \ d\boldsymbol{\omega} \ dt \tag{A.31}$$

$$= \int_{t^n}^{t^{n+1}} \int_O f(\mathbf{x}, t, \boldsymbol{\omega}) \begin{bmatrix} 1 \\ -\bar{c}^a \mathbf{n}(\omega) \\ \bar{\theta}^a \end{bmatrix} |dS^{d-1}|(\boldsymbol{\omega}) \ d\boldsymbol{\omega} \ dt$$

$$+ \int_{t^n}^{t^{n+1}} \begin{bmatrix} |S^{d-1}| \dfrac{q_d \bar{\theta}_{x_d}}{\bar{\theta}^a} \\ -(|S^{d-1}| - N_d^2) \left(\dfrac{(\bar{c}^a)^2}{\bar{\theta}^a} \nabla(\rho\theta)' + \mathbf{e}_d g \left(\rho' - (\gamma - 1) \dfrac{(\rho\theta)'}{\bar{\theta}} \right) \right) \\ 0 \end{bmatrix} (\mathbf{x}_P, t) \ dt$$

with $\mathbf{x} = \mathbf{x}^1(t; \boldsymbol{\omega})$ and

$$f(\mathbf{x}, t, \boldsymbol{\omega}) = -\sum_{j=1}^{d-1} \mathbf{t}^j(\boldsymbol{\omega}) \cdot \nabla(\mathbf{q} \cdot \mathbf{t}^j(\boldsymbol{\omega})) + \frac{n_d(\boldsymbol{\omega}) g}{\bar{c}^a} \left(\rho' - (\gamma - 1) \frac{(\rho\theta)'}{\bar{\theta}} \right) - \frac{q_d \bar{\theta}_{x_d}}{\bar{\theta}^a}. \tag{A.32}$$

Here, N_d^2 is according to (3.8). Let us recall that $N_2^2 = \pi$, $N_3^2 = 4\pi/3$.

Analogously to the derivation of the evolution operator in Section 3.3, we can integrate the frozen momentum equation along the second bicharacteristic to simplify representation of the evolution operator (A.29)

$$\mathbf{q}(\mathbf{x}_P, t^{n+1}) = \mathbf{q}(\mathbf{x}_P, t^n) - \int_{t^n}^{t^{n+1}} \left[\ \frac{(\bar{c}^a)^2}{\bar{\theta}^a} \nabla(\rho\theta)' + g\mathbf{e}_d \left(\rho' - (\gamma - 1) \frac{(\rho\theta)'}{\bar{\theta}} \right) \ \right] (\mathbf{x}_P, t) \ dt. \tag{A.33}$$

Combining (A.29), (A.30), (A.31) and (A.33) we have

$$\rho'(\mathbf{x}_P, t^{n+1}) = \rho'(\mathbf{x}_P, t^n) - \frac{(\rho\theta)'}{\bar{\theta}^a}(\mathbf{x}_P, t^n) + \frac{(\rho\theta)'}{\bar{\theta}^a}(\mathbf{x}_P, t^{n+1}) + \int_{t^n}^{t^{n+1}} \frac{q_d \bar{\theta}_{x_d}}{\bar{\theta}^a}\, dt, \tag{A.34a}$$

$$q_i(\mathbf{x}_P, t^{n+1}) = -\frac{1}{N_d^2} \int_O \left[\frac{(\rho\theta)'}{\bar{\theta}^a} - \frac{\mathbf{q}\cdot\mathbf{n}(\boldsymbol{\omega})}{\bar{c}^a} \right] (\mathbf{x}^n, t^n) n_i(\boldsymbol{\omega}) \bar{c}^a |dS^{d-1}|(\boldsymbol{\omega})\, d\boldsymbol{\omega} \tag{A.34b}$$

$$- \frac{1}{N_d^2} \int_{t^n}^{t^{n+1}} \int_O f(\mathbf{x}, t, \boldsymbol{\omega}) n_i(\boldsymbol{\omega}) \bar{c}^a |dS^{d-1}|(\boldsymbol{\omega})\, d\boldsymbol{\omega}\, dt,$$

$$(\rho\theta)'(\mathbf{x}_P, t^{n+1}) = \frac{1}{|S^{d-1}|} \int_O \left[\frac{(\rho\theta)'}{\bar{\theta}^a} - \frac{\mathbf{q}\cdot\mathbf{n}(\boldsymbol{\omega})}{\bar{c}^a} \right] (\mathbf{x}^n, t^n) \bar{\theta}^a |dS^{d-1}|(\boldsymbol{\omega})\, d\boldsymbol{\omega} \tag{A.34c}$$

$$+ \frac{1}{|S^{d-1}|} \int_{t^n}^{t^{n+1}} \int_O f(\mathbf{x}, t, \boldsymbol{\omega}) |dS^{d-1}|(\boldsymbol{\omega})\, d\boldsymbol{\omega} \bar{\theta}^a\, dt$$

with $\mathbf{x}^n = \mathbf{x}^1(t^n; \boldsymbol{\omega}), \mathbf{x} = \mathbf{x}^1(t; \boldsymbol{\omega})$ and f from (A.32). For $d = 3$ and the normal vector

$$\mathbf{n}(\omega_1, \omega_2) = \begin{pmatrix} \sin(\omega_1)\sin(\omega_2) \\ \sin(\omega_1)\cos(\omega_2) \\ \cos(\omega_1) \end{pmatrix} \tag{A.35}$$

the exact evolution operator (A.34) reads

$$\rho'(\mathbf{x}_P, t^{n+1}) = \rho'(\mathbf{x}_P, t^n) - \frac{(\rho\theta)'}{\bar{\theta}^a}(\mathbf{x}_P, t^n) + \frac{(\rho\theta)'}{\bar{\theta}^a}(\mathbf{x}_P, t^{n+1}) + \int_{t^n}^{t^{n+1}} \frac{q_3 \bar{\theta}_{x_3}}{\bar{\theta}^a}\, dt, \tag{A.36a}$$

$$q_1(\mathbf{x}_P, t^{n+1}) = -\frac{3}{4\pi} \int_0^{2\pi} \int_0^{\pi} \left[\frac{(\rho\theta)'}{\bar{\theta}^a} - \frac{\mathbf{q}\cdot\mathbf{n}(\omega_1, \omega_2)}{\bar{c}^a} \right] (\mathbf{x}^n, t^n) \sin^2(\omega_1)\sin(\omega_2) \bar{c}^a\, d\omega_1\, d\omega_2 \tag{A.36b}$$

$$- \frac{3}{4\pi} \int_{t^n}^{t^{n+1}} \int_0^{2\pi} \int_0^{\pi} f(\mathbf{x}, t, \omega_1, \omega_2) \sin^2(\omega_1)\sin(\omega_2) \bar{c}^a\, d\omega_1\, d\omega_2\, dt,$$

$$q_2(\mathbf{x}_P, t^{n+1}) = -\frac{3}{4\pi} \int_0^{2\pi} \int_0^{\pi} \left[\frac{(\rho\theta)'}{\bar{\theta}^a} - \frac{\mathbf{q}\cdot\mathbf{n}(\omega_1, \omega_2)}{\bar{c}^a} \right] (\mathbf{x}^n, t^n) \sin^2(\omega_1)\cos(\omega_2) \bar{c}^a\, d\omega_1\, d\omega_2 \tag{A.36c}$$

$$- \frac{3}{4\pi} \int_{t^n}^{t^{n+1}} \int_0^{2\pi} \int_0^{\pi} f(\mathbf{x}, t, \omega_1, \omega_2) \sin^2(\omega_1)\cos(\omega_2) \bar{c}^a\, d\omega_1\, d\omega_2\, dt,$$

$$q_3(\mathbf{x}_P, t^{n+1}) = -\frac{3}{4\pi} \int_0^{2\pi} \int_0^{\pi} \left[\frac{(\rho\theta)'}{\bar{\theta}^a} - \frac{\mathbf{q}\cdot\mathbf{n}(\omega_1, \omega_2)}{\bar{c}^a} \right] (\mathbf{x}^n, t^n) \cos(\omega_1)\sin(\omega_1) \bar{c}^a\, d\omega_1\, d\omega_2 \tag{A.36d}$$

$$- \frac{3}{4\pi} \int_{t^n}^{t^{n+1}} \int_0^{2\pi} \int_0^{\pi} f(\mathbf{x}, t, \omega_1, \omega_2) \cos(\omega_1)\sin(\omega_1) \bar{c}^a\, d\omega_1\, d\omega_2\, dt,$$

$$(\rho\theta)'(\mathbf{x}_P, t^{n+1}) = \frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}\left[\frac{(\rho\theta)'}{\bar{\theta}^a} - \frac{\mathbf{q}\cdot\mathbf{n}(\omega_1,\omega_2)}{\bar{c}^a}\right](\mathbf{x}^n, t^n)\bar{\theta}^a \sin(\omega_1)\; d\omega_1\; d\omega_2 \tag{A.36e}$$
$$+ \frac{1}{4\pi}\int_{t^n}^{t^{n+1}}\int_0^{2\pi}\int_0^{\pi} f(\mathbf{x}, t, \omega_1, \omega_2)\sin(\omega_1)\; d\omega_1\; d\omega_2\bar{\theta}^a\; dt$$

with $\mathbf{x}^n = \mathbf{x}^1(t^n;\boldsymbol{\omega}), \mathbf{x} = \mathbf{x}^1(t;\boldsymbol{\omega})$ and f from (A.32). Further, we have

$$\mathbf{q}\cdot\mathbf{n}(\omega_1,\omega_2) = \sin(\omega_1)\sin(\omega_2)q_1 + \sin(\omega_1)\cos(\omega_2)q_2 + \cos(\omega_1)q_3, \tag{A.36f}$$

$$\begin{aligned} f(\mathbf{x}, t, \omega_1, \omega_2) = &-\cos(\omega_1)\sin(\omega_2)\frac{\partial}{\partial x_1}\left[\cos(\omega_1)\sin(\omega_2)q_1 + \cos(\omega_1)\cos(\omega_2)q_2 - \sin(\omega_1)q_3\right] \\ &-\cos(\omega_1)\cos(\omega_2)\frac{\partial}{\partial x_2}\left[\cos(\omega_1)\sin(\omega_2)q_1 + \cos(\omega_1)\cos(\omega_2)q_2 - \sin(\omega_1)q_3\right] \\ &+\sin(\omega_1)\sin(\omega_2)\frac{\partial}{\partial x_3}\left[\cos(\omega_1)\sin(\omega_2)q_1 + \cos(\omega_1)\cos(\omega_2)q_2 - \sin(\omega_1)q_3\right] \\ &-\cos(\omega_2)\frac{\partial}{\partial x_1}\left[\cos(\omega_2)q_1 - \sin(\omega_2)q_2\right] \\ &+\sin(\omega_2)\frac{\partial}{\partial x_2}\left[\cos(\omega_2)q_1 - \sin(\omega_2)q_2\right] \\ &+\frac{\cos(\omega_1)g}{\bar{c}^a}\left(\rho' - (\gamma-1)\frac{(\rho\theta)'}{\bar{\theta}}\right) - \frac{q_3\bar{\theta}_{x_3}}{\bar{\theta}^a}. \end{aligned} \tag{A.36g}$$

A.4. Approximate evolution operator for the linear subsystem

As already explained in Section 3.4, we are not able to evaluate the exact evolution operator (A.34) due to its implicit nature. The aim of this section is therefore to derive an approximate evolution operator To this end we follow the procedure from Section 3.4.

We apply the rectangle rule to approximate the time integrals. Moreover we use (3.70), (3.71), which leads us to the approximate evolution operator

$$\rho'(\mathbf{x}_P, t^{n+1}) = \rho'(\mathbf{x}_P, t^n) - \frac{(\rho\theta)'}{\bar{\theta}^a}(\mathbf{x}_P, t^n) + \frac{(\rho\theta)'}{\bar{\theta}^a}(\mathbf{x}_P, t^{n+1}) + \Delta t\frac{q_d\bar{\theta}_{x_d}}{\bar{\theta}^a}(x_P, t^n), \tag{A.37a}$$

$$q_i(\mathbf{x}_P, t^{n+1}) = -\frac{1}{N_d^2}\int\limits_O \left[\frac{(\rho\theta)'}{\bar\theta^a} + \frac{q_i}{\bar c^a n_i(\boldsymbol{\omega})}\right](\mathbf{x}^n, t^n) n_i(\boldsymbol{\omega})\bar c^a |dS^{d-1}|(\boldsymbol{\omega})\; d\boldsymbol{\omega}$$
$$+ \frac{1}{N_d^2}\int\limits_O (d+1)\frac{\mathbf{q}(\mathbf{x}^n, t^n)\cdot\mathbf{n}(\boldsymbol{\omega})}{\bar c^a} n_i(\boldsymbol{\omega})\bar c^a |dS^{d-1}|(\boldsymbol{\omega})\; d\boldsymbol{\omega} \tag{A.37b}$$
$$- \frac{\Delta t}{N_d^2}\int\limits_O h(\mathbf{x}^n, t^n, \boldsymbol{\omega}) n_i(\boldsymbol{\omega})\bar c^a |dS^{d-1}|(\boldsymbol{\omega})\; d\boldsymbol{\omega}$$
$$+ \sum_{l=1}^{N}\sum_{j=1}^{d-1}\int\limits_{\boldsymbol{\alpha}^l(j)}^{\beta^l(j)} \left[\frac{\mathbf{t}^j(\boldsymbol{\omega})\cdot\mathbf{q}(\mathbf{x}^n, t^n) n_i(\boldsymbol{\omega})|dS^{d-1}|(\boldsymbol{\omega})}{k_j(\boldsymbol{\omega})}\right]_{\omega_j=\alpha_j^l}^{\omega_j=\beta_j^l} d\boldsymbol{\omega}(j),$$
$$(\rho\theta)'(\mathbf{x}_P, t^{n+1}) = \frac{1}{|S^{d-1}|}\int\limits_O \left[\frac{(\rho\theta)'}{\bar\theta^a} - d\frac{\mathbf{q}\cdot\mathbf{n}(\boldsymbol{\omega})}{\bar c^a}\right](\mathbf{x}^n, t^n)\bar\theta^a |dS^{d-1}|(\boldsymbol{\omega})\; d\boldsymbol{\omega} \tag{A.37c}$$
$$+ \frac{\Delta t}{|S^{d-1}|}\int\limits_O h(\mathbf{x}^n, t, \boldsymbol{\omega})\bar\theta^a |dS^{d-1}|(\boldsymbol{\omega})\; d\boldsymbol{\omega}$$
$$- \sum_{l=1}^{N}\sum_{j=1}^{d-1}\int\limits_{\boldsymbol{\alpha}^l(j)}^{\beta^l(j)} \left[\frac{\mathbf{t}^j(\boldsymbol{\omega})\cdot\mathbf{q}(\mathbf{x}^n, t^n)\bar\theta^a |dS^{d-1}|(\boldsymbol{\omega})}{\bar c^a k_j(\boldsymbol{\omega})}\right]_{\omega_j=\alpha_j^l}^{\omega_j=\beta_j^l} d\boldsymbol{\omega}(j),$$

with $\mathbf{x}^n = \mathbf{x}^1(t^n)$ and

$$h(\mathbf{x}, t, \boldsymbol{\omega}) = \frac{n_d(\boldsymbol{\omega}) g}{\bar c^a}\left(\rho' - (\gamma-1)\frac{(\rho\theta)'}{\bar\theta}\right) - \frac{q_d\bar\theta_{x_d}}{\bar\theta^a}. \tag{A.37d}$$

The coefficients $\boldsymbol{\alpha}^l, \boldsymbol{\beta}^l$, $l = 1, ..., N$, correspond to a decomposition of the domain O into subdomains O_l, $l = 1, ..., N$, cf. Section 3.4. We recall briefly that

$$O_l = [\alpha_1^l, \beta_1^l] \times ... \times [\alpha_{d-1}^l, \beta_{d-1}^l] =: [\boldsymbol{\alpha}^l, \boldsymbol{\beta}^l] \tag{3.65}$$

with

$$\boldsymbol{\alpha}^l = (\alpha_1^l, ..., \alpha_{d-1}^l)^T, \quad \boldsymbol{\beta}^l = (\beta_1^l, ..., \beta_{d-1}^l)^T. \tag{3.66}$$

Further, we use the notation

$$\boldsymbol{\alpha}(i) := (\alpha_1, ..., \alpha_{i-1}, \alpha_{i+1}, ..., \alpha_{d-1}), \quad d\boldsymbol{\omega}(i) = d\omega_1 ... d\omega_{i-1} d\omega_{i+1} ... d\omega_{d-1}, \tag{3.67}$$

for $i = 1, ..., d-1$, and

$$\int\limits_{O_l} f(\boldsymbol{\omega})\; d\boldsymbol{\omega} = \int\limits_{\alpha_1^l}^{\beta_1^l} ... \int\limits_{\alpha_{d-1}^l}^{\beta_{d-1}^l} f(\boldsymbol{\omega})\; d\omega_1 ... d\omega_{d-1} =: \int\limits_{\boldsymbol{\alpha}^l}^{\boldsymbol{\beta}^l} f(\boldsymbol{\omega})\; d\boldsymbol{\omega}. \tag{3.68}$$

In particular, for $d = 3$ the approximated evolution operator (A.37) reads

$$\rho'(\mathbf{x}_P, t^{n+1}) = \rho'(\mathbf{x}_P, t^n) - \frac{(\rho\theta)'}{\bar\theta^a}(\mathbf{x}_P, t^n) + \frac{(\rho\theta)'}{\bar\theta^a}(\mathbf{x}_P, t^{n+1}) + \Delta t\frac{q_3\bar\theta_{x_3}}{\bar\theta^a}(x_P, t^n), \tag{A.38a}$$

$$q_1(\mathbf{x}_P, t^{n+1}) = -\frac{3}{4\pi}\int_0^{2\pi}\int_0^{\pi}\left[\frac{(\rho\theta)'}{\bar{\theta}^a}\sin(\omega_1)\sin(\omega_2) + \frac{q_1}{\bar{c}^a}\right](\mathbf{x}^n, t^n)\bar{c}^a\sin(\omega_1)\ d\omega_1\ d\omega_2 \tag{A.38b}$$

$$+ \frac{3}{4\pi}\int_0^{2\pi}\int_0^{\pi}(d+1)\frac{\mathbf{q}(\mathbf{x}^n, t^n)\cdot\mathbf{n}(\omega_1,\omega_2)}{\bar{c}^a}\sin^2(\omega_1)\sin(\omega_2)\bar{c}^a\ d\omega_1\ d\omega_2$$

$$- \frac{3\Delta t}{4\pi}\int_0^{2\pi}\int_0^{\pi}h(\mathbf{x}, t, \omega_1, \omega_2)\sin^2(\omega_1)\sin(\omega_2)\bar{c}^a\ d\omega_1\ d\omega_2$$

$$+ \sum_{l=1}^{N}\int_{\alpha_2^l}^{\beta_2^l}\left[\mathbf{t}^1(\omega_1,\omega_2)\cdot\mathbf{q}(\mathbf{x}^n, t^n)\sin^2(\omega_1)\sin(\omega_2)\right]_{\omega_1=\alpha_1^l}^{\omega_1=\beta_1^l}\ d\omega_2$$

$$+ \sum_{l=1}^{N}\int_{\alpha_1^l}^{\beta_1^l}\left[\mathbf{t}^2(\omega_2)\cdot\mathbf{q}(\mathbf{x}^n, t^n)\sin(\omega_1)\sin(\omega_2)\right]_{\omega_2=\alpha_2^l}^{\omega_2=\beta_2^l}\ d\omega_1,$$

$$q_2(\mathbf{x}_P, t^{n+1}) = -\frac{3}{4\pi}\int_0^{2\pi}\int_0^{\pi}\left[\frac{(\rho\theta)'}{\bar{\theta}^a}\sin(\omega_1)\cos(\omega_2) + \frac{q_2}{\bar{c}^a}\right](\mathbf{x}^n, t^n)\bar{c}^a\sin(\omega_1)\ d\omega_1\ d\omega_2 \tag{A.38c}$$

$$+ \frac{3}{4\pi}\int_0^{2\pi}\int_0^{\pi}(d+1)\frac{\mathbf{q}(\mathbf{x}^n, t^n)\cdot\mathbf{n}(\omega_1,\omega_2)}{\bar{c}^a}\sin^2(\omega_1)\cos(\omega_2)\bar{c}^a\ d\omega_1\ d\omega_2$$

$$- \frac{3\Delta t}{4\pi}\int_0^{2\pi}\int_0^{\pi}h(\mathbf{x}, t, \omega_1, \omega_2)\sin^2(\omega_1)\cos(\omega_2)\bar{c}^a\ d\omega_1\ d\omega_2$$

$$+ \sum_{l=1}^{N}\int_{\alpha_2^l}^{\beta_2^l}\left[\mathbf{t}^1(\omega_1,\omega_2)\cdot\mathbf{q}(\mathbf{x}^n, t^n)\sin^2(\omega_1)\cos(\omega_2)\right]_{\omega_1=\alpha_1^l}^{\omega_1=\beta_1^l}\ d\omega_2$$

$$+ \sum_{l=1}^{N}\int_{\alpha_1^l}^{\beta_1^l}\left[\mathbf{t}^2(\omega_2)\cdot\mathbf{q}(\mathbf{x}^n, t^n)\sin(\omega_1)\cos(\omega_2)\right]_{\omega_2=\alpha_2^l}^{\omega_2=\beta_2^l}\ d\omega_1,$$

$$q_3(\mathbf{x}_P, t^{n+1}) = -\frac{3}{4\pi}\int_0^{2\pi}\int_0^{\pi}\left[\frac{(\rho\theta)'}{\bar{\theta}^a}\cos(\omega_1) + \frac{q_3}{\bar{c}^a}\right](\mathbf{x}^n, t^n)\bar{c}^a\sin(\omega_1)\ d\omega_1\ d\omega_2 \tag{A.38d}$$

$$+ \frac{3}{4\pi}\int_0^{2\pi}\int_0^{\pi}(d+1)\frac{\mathbf{q}(\mathbf{x}^n, t^n)\cdot\mathbf{n}(\omega_1,\omega_2)}{\bar{c}^a}\cos(\omega_1)\sin(\omega_1)\bar{c}^a\ d\omega_1\ d\omega_2$$

$$- \frac{3\Delta t}{4\pi}\int_0^{2\pi}\int_0^{\pi}h(\mathbf{x}, t, \omega_1, \omega_2)\cos(\omega_1)\sin(\omega_1)\bar{c}^a\ d\omega_1\ d\omega_2$$

$$+ \sum_{l=1}^{N}\int_{\alpha_2^l}^{\beta_2^l}\left[\mathbf{t}^1(\omega_1,\omega_2)\cdot\mathbf{q}(\mathbf{x}^n, t^n)\cos(\omega_1)\sin(\omega_1)\right]_{\omega_1=\alpha_1^l}^{\omega_1=\beta_1^l}\ d\omega_2$$

$$+ \sum_{l=1}^{N}\int_{\alpha_1^l}^{\beta_1^l}\left[\mathbf{t}^2(\omega_2)\cdot\mathbf{q}(\mathbf{x}^n, t^n)\cos(\omega_1)\right]_{\omega_2=\alpha_2^l}^{\omega_2=\beta_2^l}\ d\omega_1,$$

$$(\rho\theta)'(\mathbf{x}_P, t^{n+1}) = \frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}\left[\frac{(\rho\theta)'}{\bar{\theta}^a} - \frac{\mathbf{q}\cdot\mathbf{n}(\omega_1,\omega_2)}{\bar{c}^a}\right](\mathbf{x}^n, t^n)\bar{\theta}^a \sin(\omega_1)\; d\omega_1\; d\omega_2 \tag{A.38e}$$

$$- \frac{\bar{\theta}^a}{\bar{c}^a}\sum_{l=1}^{N}\int_{\alpha_2^l}^{\beta_2^l}\left[\mathbf{t}^1(\omega_1,\omega_2)\cdot\mathbf{q}(\mathbf{x}^n,t^n)\cos(\omega_1)\sin(\omega_1)\right]_{\omega_1=\alpha_1^l}^{\omega_1=\beta_1^l}\; d\omega_2$$

$$- \frac{\bar{\theta}^a}{\bar{c}^a}\sum_{l=1}^{N}\int_{\alpha_1^l}^{\beta_1^l}\left[\mathbf{t}^2(\omega_2)\cdot\mathbf{q}(\mathbf{x}^n,t^n)\cos(\omega_1)\right]_{\omega_2=\alpha_2^l}^{\omega_2=\beta_2^l}\; d\omega_1,$$

$$+ \frac{\Delta t}{4\pi}\int_0^{2\pi}\int_0^{\pi} h(\mathbf{x},t,\omega_1,\omega_2)\sin(\omega_1)\; d\omega_1\; d\omega_2\bar{\theta}^a$$

with $\mathbf{x} = \mathbf{x}^1(t^n)$ and

$$h(\mathbf{x},t,\boldsymbol{\omega}) = \frac{\cos(\omega_1)g}{\bar{c}^a}\left(\rho' - (\gamma-1)\frac{(\rho\theta)'}{\bar{\theta}}\right) - \frac{q_3\bar{\theta}_{x_3}}{\bar{\theta}^a}. \tag{A.38f}$$

List of Figures

List of Tables

Bibliography

[1] D. Alevras. Simulations of the indian ocean tsunami with realistic bathymetry using a high-order triangular discontinuous galerkin shallow water model. Master's thesis, Naval Postgraduate School, 2008.

[2] D. Arcas and V. Titov. Sumatra tsunami: lessons from modeling. *Surveys in Geophysics*, 27(6):679–705, 2006.

[3] K.R. Arun, M. Kraft, M. Lukáčová-Medvid'ová, and Ph. Prasad. Finite volume evolution galerkin method for hyperbolic conservation laws with spatially varying flux functions. *J. Comput. Phys.*, 228:565–590, 2009.

[4] K.R. Arun and S. Noelle. An asymptotic preserving scheme for low froude number shallow flows. In *Proccedings of the 14th international conference on hyperbolic problems. Theory, numerics and applications*. American institute of mathematical sciences, 2013.

[5] U.M. Ascher, S.J. Ruuth, and R.J. Spiteri. Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. *Appl. Numer. Math.*, 25(2-3):151–167, 1997.

[6] U.M. Ascher, S.J. Ruuth, and B.T.R. Wetton. Implicit-explicit methods for time-dependent partial differential equations. *J. Numer. Anal.*, 32(3):797–823, 1995.

[7] E. Audusse, F. Bouchut, M. Bristeau, R. Klein, and B. Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *J. Sci. Comput.*, 25(6):2050–2065, 2004.

[8] E. Audusse, R. Klein, D.D. Nguyen, and S. Vater. Preservation of the discrete geostrophic equilibrium in shallow water flows. In *Finite volumes for complex applications VI: Problems and perspectives. FVCA 6, international symposium, Prague, Czech Republich, June 6–10, 2011. Vol. 1 and 2.*, pages 59–67. Berlin: Springer, 2011.

[9] S. Balay, S. Abhyankar, M.F. Adams, J. Brown, P. Brune, K. Buschelman, V. Eijkhout, W.D. Gropp, D. Kaushik, M.G. Knepley, L.C. McInnes, K. Rupp, B.F. Smith, and H. Zhang. PETSc users manual. Technical Report ANL-95/11 - Revision 3.5, Argonne National Laboratory, 2014.

[10] S. Balay, S. Abhyankar, M.F. Adams, J. Brown, P. Brune, K. Buschelman, V. Eijkhout, W.D. Gropp, D. Kaushik, M.G. Knepley, L.C. McInnes, K. Rupp, B.F. Smith, and H. Zhang. PETSc Web page. `http://www.mcs.anl.gov/petsc`, 2014.

[11] S. Balay, W.D. Gropp, L.C. McInnes, and B.F. Smith. Efficient management of parallelism in object oriented numerical software libraries. In E. Arge, A. M. Bruaset, and H. P. Langtangen, editors, *Modern Software Tools in Scientific Computing*, pages 163–202. Birkhäuser Press, 1997.

[12] J.T. Beale, T. Kato, and A. Majda. Remarks on the breakdown of smooth solutions for the 3-D Euler equations. *Commun. Math. Phys.*, 94:61–66, 1984.

[13] A. Bermúdez and M.E. Vázquez. Upwind methods for hyperbolic conservation laws with source terms. *Comput. Fluids*, 23(8):1049–1071, 1994.

[14] H. Bijl and P. Wesseling. A unified method for computing incompressible and compressible flows in boundary-fitted coordinates. *J. Comput. Phys.*, 141(2):153–173, art. no. cp985914, 1998.

[15] G. Bispen, K.R. Arun, M. Lukáčová-Medvid'ová, and S. Noelle. Imex large time step finite volume methods for low froude number shallow water flows. *Commun. Comp. Phys.*, 16:307–347, 2014.

[16] A. Bollermann, M. Lukáčová-Medvid'ová, and S. Noelle. Well-balanced finite volume evolution Galerkin methods for the 2D shallow water equations on adaptive grids. Handlovičová, Angela (ed.) et al., Algoritmy 2009. 18th conference on scientific computing, Vysoké Tatry – Podbsanské, Slovakia, March 15–20, 2009. Proceedings of contributed papers and posters. Bratislava: Slovak University of Technology, Faculty of Civil Engineering, Department of Mathematics and Descriptive Geometry. 81-90 (2009)., 2009.

[17] S. Boscarino. Error analysis of IMEX Runge-Kutta methods derived from differential-algebraic systems. *J. Numer. Anal.*, 45(4):1600–1621, 2007.

[18] S. Boscarino, L. Pareschi, and G. Russo. Implicit-explicit Runge–Kutta schemes for hyperbolic systems and kinetic equations in the diffusion limit. *J. Sci. Comput.*, 35(1):a22–a51, 2013.

[19] S. Boscarino and G. Russo. On a class of uniformly accurate IMEX Runge-Kutta schemes and applications to hyperbolic systems with relaxation. *J. Sci. Comput.*, 31(3):1926–1945, 2009.

[20] N. Botta, R. Klein, S. Langenberg, and S. Lützenkirchen. Well balanced finite volume methods for nearly hydrostatic flows. *J. Comput. Phys.*, 196(2):539–565, 2004.

[21] D. Bresch, R. Klein, and C. Lucas. Multiscale analyses for the shallow water equations. In Egon Krause, Yurii Shokin, Michael Resch, Dietmar Kröner, and Nina Shokina, editors, *Computational Science and High Performance Computing IV*, volume 115 of *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, pages 149–164. Springer Berlin Heidelberg, 2011.

[22] A. Bressan and R. M. Colombo. The semigroup generated by 2×2 conservation laws. *Arch. Ration. Mech. Anal.*, 133(1):1–75, 1995.

[23] A. Bressan, G. Crasta, and B. Piccoli. Well-posedness of the Cauchy problem for $n \times n$ systems of conservation laws. *Mem. Am. Math. Soc.*, 694:134, 2000.

[24] A. Bressan, T.-P. Liu, and T. Yang. L^1 stability estimates for $n \times n$ conservation laws. *Arch. Ration. Mech. Anal.*, 149(1):1–22, 1999.

[25] A. Canestrelli, M. Dumbser, A. Siviglia, and E.F. Toro. Well-balanced high-order centered schemes on unstructured meshes for shallow water equations with fixed and mobile bed. *Advanced in Water Resources*, 33:291–303, 2010.

[26] B. Cockburn, G. Karniadakis, and C.-W. Shu. *Discontinuous Galerkin Methods.* Springer Berlin Heidelberg, 2000.

[27] F. Cordier, P. Degond, and A. Kumbaro. An asymptotic-preserving all-speed scheme for the Euler and Navier-Stokes equations. *J. Comput. Phys.*, 231(17):5685–5704, 2012.

[28] R. Courant, K. Friedrichs, and H. Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.*, 100:32–74, 1928.

[29] R. Courant and D. Hilbert. *Methods of Mathematical Physics. Volume II: Partial Differential Equations. Transl. and rev. from the German Original. Reprint of the 1st Engl. ed. 1962.* New York etc.: John Wiley &— Sons/Interscience Publishers, reprint of the 1st engl. ed. 1962 edition, 1989.

[30] P.J. Davis. *Circulant matrices. 2nd ed.* New York, NY: AMS Chelsea Publishing, 2nd ed. edition, 1994.

[31] T.A. Davis. Algorithm 832: Umfpack v4.3—an unsymmetric-pattern multifrontal method. *ACM Trans. Math. Softw.*, 30(2):196–199, June 2004.

[32] T.A. Davis. A column pre-ordering strategy for the unsymmetric-pattern multifrontal method. *ACM Trans. Math. Softw.*, 30(2):165–195, June 2004.

[33] T.A. Davis and I.S. Duff. An unsymmetric-pattern multifrontal method for sparse LU factorization. *J. Matrix Anal. Appl.*, 18(1):140–158, 1997.

[34] T.A. Davis and I.S. Duff. A combined unifrontal/multifrontal method for unsymmetric sparse matrices. *ACM Trans. Math. Softw.*, 25(1):1–20, March 1999.

[35] C. De Lellis and L. jun. Székelyhidi. On admissibility criteria for weak solutions of the Euler equations. *Arch. Ration. Mech. Anal.*, 195(1):225–260, 2010.

[36] C. De Lellis and L. jun. Székelyhidi. The h-principle and the equations of fluid dynamics. *Bull. Am. Math. Soc., New Ser.*, 49(3):347–375, 2012.

[37] P. Degond, S. Jin, and J.-G. Liu. Mach-number uniform asymptotic-preserving gauge schemes for compressible flows. *Bull. Inst. Math., Acad. Sin. (N.S.)*, 2(4):851–892, 2007.

[38] P. Degond and M. Tang. All speed scheme for the low mach number limit of the isentropic euler equations. *Commun. Comput. Phys.*, 10:1–31, 2011.

[39] J. J. Dronkers. *Tidal Computations in Rivers and Coastal Waters.* North-Holland, Amsterdam, 1964.

[40] D.G. Ebin. Motion of slightly compressible fluids in a bounded domain. I. *Commun. Pure Appl. Math.*, 35:451–485, 1982.

[41] M. Feistauer, J. Felcman, and I. Straškraba. *Mathematical and Computational Methods for Compressible Flow.* Numerical Mathematics and Scientific Computation. Oxford: Oxford University Press. xiii, 535 p., 2003.

[42] J.H. Ferziger and M. Perić. *Computational Methods for Fluid Dynamics. 2nd rev. ed.* Berlin: Springer. xiv, 389 p., 1999.

[43] S.A. Gershgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Bull. Acad. Sci. URSS*, 1931(6):749–754, 1931.

[44] F.X. Giraldo and M. Restelli. High-order semi-implicit time-integrators for a triangular discontinuous galerkin oceanic shallow water model. *Int. J. Numer. Methods Fluids*, 63(9):1077–1102, 2010.

[45] F.X. Giraldo, M. Restelli, and M. Läuter. Semi-implicit formulations of the Navier-Stokes equations: application to nonhydrostatic atmospheric modeling. *J. Sci. Comput.*, 32(6):3394–3425, 2010.

[46] J. Glimm. Solutions in the large for nonlinear hyperbolic systems of equations. *Commun. Pure Appl. Math.*, 18:697–715, 1965.

[47] E. Godlewski and P.-A. Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws.* New York, NY: Springer, 1996.

[48] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Rev.*, 43(1):89–112, 2001.

[49] J.M. Greenberg and A.-Y. Le Roux. A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. *J. Numer. Anal.*, 33(1):1–16, 1996.

[50] H.P. Greenspan. *The Theory of Rotating Fluids.* Cambridge U.P London, 1968.

[51] J. Haack, S. Jin, and J. Liu. An all-speed asymptotic-preserving method for the isentropic euler and navier-stokes equations. *Commun. Comput. Phys.*, 12:955–980, 2012.

[52] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations. I: Nonstiff Problems. 2. rev. ed.* Springer Series in Computational Mathematics. 8. Berlin: Springer-Verlag. xv, 528 p. DM 138.00 , 1993.

[53] M. Hanke-Bourgeois. *Grundlagen der numerischen Mathematik und des wissenschaftlichen Rechnens.* Wiesbaden: Teubner, 2nd revised and expanded ed. edition, 2006.

[54] L.B. Hoffmann. *Ein Zeitlich Selbstadaptives Numerisches Verfahren zur Berechnung von Strömungen Aller Mach-Zahlen Basierend auf Mehrskalenasymptotik und Diskreter Datenanalys.* PhD thesis, Universität Hamburg, 2000.

[55] A. Hundertmark-Zaušková, M. Lukáčová-Medvid'ová, and F. Prill. Large time step finite volume evolution Galerkin methods. *J. Sci. Comput.*, 48(1-3):227–240, 2011.

[56] J. Jang and N. Masmoudi. Well-posedness for compressible Euler equations with physical vacuum singularity. *Commun. Pure Appl. Math.*, 62(10):1327–1385, 2009.

[57] S. Jin. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *J. Sci. Comput.*, 21(2):441–454, 1999.

[58] S. Jin. Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review. *Riv. Mat. Univ. Parma (N.S.)*, 3(2):177–216, 2012.

[59] G.E. Karniadakis, M. Israeli, and S.A. Orszag. High-order splitting methods for the incompressible Navier-Stokes equations. *J. Comput. Phys.*, 97(2):414–443, 1991.

[60] C.A. Kennedy and M.H. Carpenter. Additive Runge-Kutta schemes for convection-diffusion-reaction equations. *Appl. Numer. Math.*, 44(1-2):139–181, 2003.

[61] J.K. Kevorkian and J.D. Cole. *Multiple Scale and Singular Perturbation Methods.* Berlin: Springer, 1996.

[62] S. Klainerman and A. Majda. Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids. *Commun. Pure Appl. Math.*, 34:481–524, 1981.

[63] S. Klainerman and A. Majda. Compressible and incompressible fluids. *Commun. Pure Appl. Math.*, 35:629–651, 1982.

[64] R. Klein. Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics. I: One-dimensional flow. *J. Comput. Phys.*, 121(2):213–237, 1995.

[65] R. Klein, N. Botta, T. Schneider, C.-D. Munz, S. Roller, A. Meister, L. Hoffmann, and T. Sonar. Asymptotic adaptive methods for multi-scale problems in fluid mechanics. *J. Eng. Math.*, 39(1-4):261–343, 2001.

[66] J.B. Klemp, W.C. Skamarock, and J. Dudhia. Conservative split-explicit time integration methods for the compressible nonhydrostatic equations. *Monthly Weather Rev.*, 135:2897–2913, 2007.

[67] I. Kra and S.R: Simanca. On circulant matrices. *Notices Am. Math. Soc.*, 59(3):368–377, 2012.

[68] D. Kröner. *Numerical schemes for conservation laws.* Chichester: Wiley; Stuttgart: Teubner, 1997.

[69] D. Kröner, S. Noelle, and M. Rokyta. Convergence of higher order upwind finite volume schemes on unstructured grids for scalar conservation laws in several space dimensions. *Numer. Math.*, 71(4):527–560, 1995.

[70] D. Kröner and M. Rokyta. Convergence of upwind finite volume schemes for scalar conservation laws in two dimensions. *SIAM J. Numer. Anal.*, 31(2):324–343, 1994.

[71] S.N. Kruzhkov. First order quasilinear equations in several independent variables. *Math. USSR, Sb.*, 10:217–243, 1970.

[72] A. Kurganov and D. Levy. Central-upwind schemes for the Saint-Venant system. *M2AN, Math. Model. Numer. Anal.*, 36(3):397–425, 2002.

[73] O. Le Maître, J. Levin, M. Iskandarani, and O.M. Knio. A multiscale pressure splitting of the shallow-water equations. I: Formulation and 1D tests. *J. Comput. Phys.*, 166(1):116–151, 2001.

[74] B. LéMéhauté. *An Introduction to Hydrodynamics and Water Waves.* Springer-Verlag, New York, Heidelber, Berlin, 1976.

[75] R.J. LeVeque. *Numerical methods for conservation laws. 2nd ed.* Basel: Birkhäuser, 2nd ed. edition, 1992.

[76] R.J. LeVeque. Balancing source terms and flux gradients in high-resolution Godunov methods: The quasi-steady wave-propagation algorithm. *J. Comput. Phys.*, 146(1):346–365, art. no. cp986058, 1998.

[77] R.J. Leveque. *Finite Volume Methods for Hyperbolic Problems.* Cambridge: Cambridge University Press, 2002.

[78] C.D. Levermore, M. Oliver, and E.S. Titi. Global well-posedness for the lake equations. *Physica D*, 98(2-4):492–509, 1996.

[79] Q. Liang and A.G.L. Borthwick. Adaptive quadtree simulation of shallow flows with wet-dry fronts over complex topography. *Comput. Fluids*, 38(2):221–234, 2009.

[80] Q. Liang and F. Marche. Numerical resolution of well-balanced shallow water equations with complex source terms. *Advances in Water Resources*, 32(6):873 – 884, 2009.

[81] M. Lukácová-Medvid'ová, K.W. Morton, and G. Warnecke. Evolution Galerkin methods for hyperbolic systems in two space dimensions. *Math. Comput.*, 69(232):1355–1384, 2000.

[82] M. Lukácová-Medvidová, K.W. Morton, and G. Warnecke. Finite volume evolution Galerkin methods for hyperbolic systems. *J. Sci. Comput.*, 26(1):1–30, 2004.

[83] M. Lukáčová-Medvid'ová and K.W. Morton. Finite volume evolution Galerkin methods – A survey. *Indian J. Pure Appl. Math.*, 41(2):329–361, 2010.

[84] M. Lukáčová-Medvid'ová, K.W. Morton, and G. Warnecke. Finite volume evolutio galerkin methods for euler equations of gas dynamics. *Int. J. Numer. Meth. Fluids*, 40:425–234, 2002.

[85] M. Lukáčová-Medvid'ová, S. Noelle, and M. Kraft. Well-balanced finite volume evolution Galerkin methods for the shallow water equations. *J. Comput. Phys.*, 221(1):122–147, 2007.

[86] M. Lukáčová-Medvid'ová, J. Saibertova, and G. Warnecke. Finite volume evolution galerkin methods for nonlinear hyperbolic systems. *J. Comput. Phys.*, 183:533–562, 2002.

[87] M. Lukáčová-Medvid'ová and J. Saibertová-Zatočilová. Finite volume schemes for multi-dimensional hyperbolic systems based on the use of bicharacteristics. *Appl. Math., Praha*, 51(3):205–228, 2006.

[88] A. Majda. *Introduction to PDEs and Waves for the Atmosphere and Ocean.* Providence, RI: American Mathematical Society (AMS); New York, NY: Courant Institute of Mathematical Sciences, 2003.

[89] A. Meister. *Numerik linearer Gleichungssysteme. Eine Einführung in moderne Verfahren. Mit MATLAB-Implementierungen von C. Vömel.* Wiesbaden: Vieweg, 2nd revised ed. edition, 2005.

[90] A. Müller, J. Behrens, F.X. Giraldo, and V. Wirth. Comparision between adaptive and uniform discontinuous galerkin simulations in dry 2d bubble experiments. *J. Comput. Phys.*, 235:371–393, 2013.

[91] C.-D. Munz, S. Roller, R. Klein, and K.J. Geratz. The extension of incompressible flow solvers to the weakly compressible regime. *Comput. Fluids*, 32(2):173–196, 2003.

[92] S. Noelle, G. Bispen, K.R. Arun, M. Lukáčová-Medvid'ová, and C.-D. Munz. A weakly asymptotic preserving low mach number scheme for the euler equations of gas dynamics. *J. Sci. Comput.*, 36:B989–B1024, 2014.

[93] A. Novotný and I. Straškraba. *Introduction to the mathematical theory of compressible flow.* Oxford: Oxford University Press, 2004.

[94] L. Pareschi and G. Russo. Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. *J. Sci. Comput.*, 25(1-2):129–155, 2005.

[95] J.H. Park and C.-D. Munz. Multiple pressure variables methods for fluid flow at all Mach numbers. *Int. J. Numer. Methods Fluids*, 49(8):905–931, 2005.

[96] Ph. Prasad. *Nonlinear Hyperbolic Waves in Multidimensions.* Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics. 121. Boca Raton, FL: Chapman & Hall/CRC. xvi, 338 p., 2001.

[97] Ph. Prasad. Ray theories for hyperbolic waves, kinematical conservation laws (KCL) and applications. *Indian J. Pure Appl. Math.*, 38(5):467–490, 2007.

[98] M. Restelli. *Semi-Lagrangian and Semi-Implicit Discontinuous Galerkin Methods for Atmospheric Modeling Applications*. PhD thesis, Politecnico di Milano, 2007.

[99] M. Restelli and F.X. Giraldo. A conservative discontinuous Galerkin semi-implicit formulation for the Navier-Stokes equations in nonhydrostatic mesoscale modeling. *J. Sci. Comput.*, 31(3):2231–2257, 2009.

[100] M. Ricchiuto and A. Bollermann. Stabilized residual distribution for shallow water simulations. *J. Comput. Phys.*, 228(4):1071–1115, 2009.

[101] B.D. Rogers, A.G.L. Borthwick, and P.H. Taylor. Mathematical balancing of flux gradient and source terms prior to using Roe's approximate Riemann solver. *J. Comput. Phys.*, 192(2):422–451, 2003.

[102] B.D. Rogers, M. Fujihara, and A.G.L. Borthwick. Adaptive q-tree godunov-type scheme for shallow water equations. *Int. J. Numer. Meth. Fluids*, 35(3):247–280, 2001.

[103] Y. Saad. *Iterative methods for sparse linear systems. 2nd ed.* Philadelphia, PA: SIAM Society for Industrial and Applied Mathematics. xviii, 528 p. $ 89.00 , 2003.

[104] J. Sesterhenn, B. Müller, and H. Thomann. On the cancellation problem in calculating compressible low Mach number flows. *J. Comput. Phys.*, 151(2):597–615, art. no. jcph.1999.6211, 1999.

[105] G. Stecca, A. Siviglia, and E.F. Toro. A finite volume upwind-biased centred scheme for hyperbolic conservation laws. application to shallow water equations. *Commun. Comput. Phys.*, 12(4):1183–1214, 2012.

[106] H. J. Stetter. The defect correction principle and discretization methods. *Numer. Math.*, 29:425–443, 1978.

[107] Y. Sun and Y. Ren. The finite volume local evolution Galerkin method for solving the hyperbolic conservation laws. *J. Comput. Phys.*, 228(13):4945–4960, 2009.

[108] M. Tang. Second order all speed method for the isentropic euler equations. *Kinet. Relat. Models*, 5(1):155–184, 2012.

[109] E.F. Toro. *Riemann solvers and numerical methods for fluid dynamics. A practical introduction. 3rd ed.* Berlin: Springer, 3rd ed. edition, 2009.

[110] E.F. Toro, M. Spruce, and W. Speares. Restoration of the contact surface in the HLL-Riemann solver. *Shock Waves*, 4(1):25–34, 1994.

[111] Eleuterio F. Toro. *Shock-capturing Methods for Free-Surface Shallow Flows*. Chichester: Wiley, 2001.

[112] G.K. Vallis. *Atmospheric and Oceanic Fluid Dynamics*. Cambridge University Press, Cambridge, U.K., 2006.

[113] B. van Leer. Towards the ultimate conservative difference scheme. i: The quest of monotonicity . *Springer Lecture Notes Phys.*, 18:163–168, 1973.

[114] B. van Leer. Towards the ultimate conservative difference scheme. ii: Monotonicity and conservation combined in a second-order scheme. *J. Comput. Phys.*, 14:361–370, 1974.

[115] B. van Leer. Towards the ultimate conservative difference scheme. iii: Upstream-centered finite-difference schemes for ideal compressible flow. *J. Comput. Phys.*, 23:263–275, 1977.

[116] B. van Leer. Towards the ultimate conservative difference scheme. IV: A new approach to numerical convection. *J. Comput. Phys.*, 23:276–299, 1977.

[117] B. van Leer. Towards the ultimative conservative difference scheme. V: A second-order sequel to Godunov's method. *J. Comput. Phys.*, 135(2):229–248, art. no. cp975704, 1979.

[118] S. Vater. *A Multigrid-based Multiscale Numerical Scheme for Shallow Water Flows at Low Froude Number*. PhD thesis, Freie Universität Berlin, 2012.

[119] C.B. Vreugdenhil. *Numerical Methods for Shallow Water Flow*. Kluwer Academic Publishers, The Netherlands, 1994.

[120] K. Wu and H. Tang. Finite volume local evolution galerkin method for two-dimensional relativistic hydrodynamics. *J. Comput. Phys.*, 256(0):277 – 307, 2014.

[121] L. Yelash, A. Müller, M. Lukáčová-Medvid'ová, F.X. Giraldo, and V. Wirth. Adaptive discontinuous evolution galerkin method for dry atmospheric flow. *J. Comput. Phys.*, 268:106–133, 2014.

Danksagung

An dieser Stelle danke ich zunächst meiner Doktormutter vom Institut für Mathematik der JGU Mainz, die mich während der Erstellung dieser Arbeit betreute, mir hilfreich mit Rat und Tat zur Seite stand und die es mir ermöglichte an diesem interessanten Thema zu arbeiten.

Weiterhin danke ich den Kollegen der Arbeitsgruppen Numerik und Funktionalanalysis der JGU Mainz für eine freundliche und sehr angenehme Arbeitsatmosphäre, sowie die vielen netten Gespräche und Kickerpartien.

Ein besonderer Dank geht auch an meinen Brokollegen, der immer sehr hilfsbereit und diskussionsfreudig war. Seine Programmierkenntnisse waren eine enorme Hilfe bei der Umsetzung der Algorithmen.

Ebenso danke ich dem Institut für Mathematik der JGU Mainz für die ausgezeichnete Mathematikausbildung während meines Studiums, sowie meinen Kommilitonen die mit mir diesen Weg beschritten.

Ein Dank geht auch an meinen Zweitgutachter und die Koauthoren vom IGPM an der RWTH Aachen und vom IISER Thiruvananthapuram für die erfolgreiche und sehr lehrreiche Zusammenarbeit.

Ferner danke ich meiner Freundin, die mir stets Mut zugesprochen und mich in meiner Arbeit bestärkt hat.

Ganz besonders möchte ich an dieser Stelle meinen Eltern danken, die mich mein ganzes Leben in jeder möglichen Hinsicht unterstützt haben und ohne die ich heute nicht an diesem Punkt angekommen wäre.