Stefan Vater

A New Projection Method for the Zero Froude Number Shallow Water Equations

Diplomarbeit

eingereicht am
Fachbereich Mathematik und Informatik
der Freien Universität Berlin

15. Dezember 2004

Betreuer: Prof. Dr. Rupert Klein

Abstract

For non-zero Froude numbers the shallow water equations are a hyperbolic system of partial differential equations. In the zero Froude number limit, they are of mixed hyperbolic-elliptic type, and the velocity field is subject to a *divergence constraint*.

A new semi-implicit projection method for the zero Froude number shallow water equations is presented. This method enforces the divergence constraint on the velocity field, in two steps. First, the numerical fluxes of an auxiliary hyperbolic system are computed with a standard second order method. Then, these fluxes are corrected by solving two Poisson-type equations. These corrections guarantee that the new velocity field satisfies a discrete form of the above-mentioned divergence constraint. The main feature of the new method is a unified discretization of the two Poisson-type equations, which rests on a *Petrov-Galerkin* finite element formulation with piecewise bilinear ansatz functions for the unknown variable. This discretization naturally leads to piecewise linear ansatz functions for the momentum components. The projection method is derived from a semi-implicit finite volume method for the zero Mach number Euler equations, which uses standard discretizations for the solution of the Poisson-type equations.

The new scheme can be formulated as an *approximate* as well as an *exact* projection method. In the former case, the divergence constraint is not exactly satisfied. The "approximateness" of the method can be estimated with an asymptotic upper bound of the velocity divergence at the new time level, which is consistent with the method's second-order accuracy. In the exact projection method, the piecewise linear components of the momentum are employed for the computation of the numerical fluxes of the auxiliary system at the new time level.

In order to show the stability of the new projection step, a *primal-dual mixed* finite element formulation is derived, which is equivalent to the Poisson-type equations of the new scheme. Using the abstract theory of Nicolaïdes for generalized saddle

point problems, existence and uniqueness of the continuous problem are proven. Furthermore, preliminary results regarding the stability of the discrete method are presented.

The numerical results obtained with the new exact method show significant accuracy improvements over the version that uses standard discretizations for the solution of the Poisson-type equations. In the L^2 as well as the L^∞ norm, the global error is about four times smaller for smooth solutions. Simulating the advection of a vortex with discontinuous vorticity field, the new method yields a more accurate position of the center of the vortex.

Zusammenfassung

Die Flachwassergleichungen bilden für positive Froude-Zahlen ein hyperbolisches System von Differentialgleichungen. Im Limes Froude-Zahl gegen Null wechseln sie ihren Typ zu einem elliptisch-hyperbolischen System. Darüber hinaus unterliegt das Geschwindigkeitsfeld einer *Divergenzbedingung*.

Im Rahmen dieser Arbeit wird eine neue semi-implizite Projektionsmethode zur Lösung der Flachwassergleichungen im Limes einer verschwindenden Froude-Zahl präsentiert. In diesem Verfahren wird die Divergenzbedingung an das Geschwindigkeitsfeld in zwei Schritten erzwungen: Zuerst werden die numerischen Flüsse eines hyperbolischen Hilfssystems mit einer Standardmethode zweiter Ordnung berechnet. Im zweiten Schritt werden diese durch die Lösung zweier Poisson-Typ-Gleichungen korrigiert. Die Korrekturen garantieren, dass das Geschwindigkeitsfeld eine diskrete Form der oben genannten Divergenzbedingung erfüllt. Das Hauptmerkmal der neuen Methode ist ein vereinheitlichter Ansatz bei der Diskretisierung der Poisson-Typ-Gleichungen, die auf einer *Petrov-Galerkin* Finite-Elemente-Formulierung mit stückweise bilinearen Ansatzfunktionen für die Unbekannte basiert. Diese Diskretisierung führt in natürlicher Weise zu stückweise linearen Ansatzfunktionen für die Impuls-Variable. Die vorgestellte Projektionsmethode beruht auf einem semi-impliziten Finite-Volumen-Verfahren zur Lösung der Euler-Gleichungen im Limes einer verschwindenden Mach-Zahl, welches klassische Diskretisierungen zur Lösung der Poisson-Typ-Gleichungen verwendet.

Das neue Verfahren kann sowohl als *approximative* als auch als *exakte* Methode formuliert werden. Im ersten Fall wird die Divergenzbedingung nicht exakt erfüllt. Die „Approximiertheit“ der Methode ist durch eine asymptotische obere Schranke der Geschwindigkeitsdivergenz zu Beginn des neuen Zeitschrittes abschätzbar. Damit wird gewährleistet, dass die zweite Ordnung des Verfahrens erhalten bleibt. In der exakten Projektionsmethode werden die stückweise linearen Verteilungen des Impulses zur

Berechnung der numerischen Flüsse des Hilfssystems im darauf folgenden Zeitschritt verwendet.

Für den Beweis der Stabilität des neuen Projektionsschrittes wird eine *primal-duale gemischte* Finite-Elemente-Formulierung hergeleitet, die äquivalent zu der zweiten Poisson-Typ-Gleichung des neuen Verfahrens ist. Unter Benutzung der abstrakten Theorie von NICOLAÏDES für generalisierte Sattelpunkt-Probleme wird die Existenz und Eindeutigkeit des kontinuierlichen Problems gezeigt. Außerdem werden erste Ergebnisse in Bezug auf die Stabilität der diskreten Methode vorgestellt.

Die numerischen Resultate der neuen exakten Methode weisen signifikante Verbesserungen in der Genauigkeit gegenüber der Version auf, die klassische Diskretisierungen zur Lösung der Poisson-Typ-Gleichungen benutzt. Für glatte Lösungen ist der globale Fehler in der L^2- sowie der L^∞-Norm um das Vierfache geringer. Bei der Simulation eines Wirbels mit unstetigem Wirbelstärke-Feld ergibt die neue Methode eine wesentlich genauere Position des Wirbelzentrums.

Acknowledgements

First of all, I would like to express my gratitude to Prof. Dr. R. Klein for providing the highly interesting research topic and constant encouragement. He also gave me the opportunity to prepare this thesis at the Potsdam Institute for Climate Impact Research (PIK), a very stimulating and comfortable place to work. Moreover, I want to thank all members of Prof. Klein's group at PIK for many fruitful discussions and good ideas. In particular, Dr. N. Botta patiently answered my countless questions and introduced me to his scientific software for the numerical implementation of my algorithms. The many espresso breaks were often accompanied by discussions, which gave me a lot of insight into related research fields. Dr. J. Gerlach, with whom I shared an office, was always open to discussing with me several detailed questions, which arose during my writing.

Special thanks to my partner Theresia Petrow, who provided me with many valuable comments and suggestions. I would also like to thank my family and friends. They have always supported me in my work and believed in me and the success of this thesis. Last, but not least, I am indebted to Dr. N. Botta, Jennifer Elrick and Martin Weiser for revising large parts of the manuscript.

This work benefitted greatly from free software products. Without these tools – such as LaTeX, the GNU C/C++ compiler and the linux operating system – a lot of tasks would not have been so easy to realize. It is my pleasure to thank all developers for their excellent products.

Contents

List of Figures

List of Tables

List of Symbols

The following list summarizes the symbols, which have been used throughout this work. The list is not complete, however. Symbols that are only used in delimited parts, are omitted. Bold type has been used for vectors and matrices. Calligraphic type has been mainly used for non-common function and finite element spaces. Furthermore, subequations are denoted by indices, i.e. $(2.5)_2$.

Vector and function spaces

$\mathbb{N}$	natural numbers $(0, 1, 2, 3, \ldots)$
$\mathbb{R}$, $\mathbb{R}_0^+$	real numbers, positive real numbers with zero
$L^2(\Omega)$, $\|\cdot\|_{0,\Omega}$	space of square integrable functions on Ω and its norm
$H^1(\Omega)$, $\|\cdot\|_{1,\Omega}$	first order Sobolev space on Ω and its norm
$\|\cdot\|_{1,\Omega}$	semi norm on $H^1(\Omega)$
$H(\mathrm{div};\Omega)$, $\|\cdot\|_{\mathrm{div},\Omega}$	space of square integrable vector functions with square integrable divergence on Ω and its norm
$\mathcal{U}$, $\mathcal{H}$, $\mathcal{X}$, $\mathcal{M}$, ...	other function spaces
$\mathcal{U}^h$, $\mathcal{H}^h$, $\mathcal{Q}^h$, ...	finite element spaces

Variables

t, $\boldsymbol{x} = (x, y)$	time and space coordinates
h', $\boldsymbol{v}'$	dimensional height and velocity
h, $\boldsymbol{v}$	nondimensional height and velocity
h^*, $\boldsymbol{v}^*$	variables of the auxiliary system
h_0, $h^{(2)}$	uniform background height and second order height perturbation
ω	nondimensional vorticity
g'	gravitational constant
$\mathbf{u}(\boldsymbol{x}, t)$	vector of conserved variables
$\mathbf{f}(\mathbf{u}, \boldsymbol{n})$	flux function

Asymptotic and dimensional analysis

ε	(small) parameter in asymptotic analysis
$u^{(i)}$	asymptotic function
$\{\phi_n(\varepsilon)\}_{n\in\mathbb{N}}$	asymptotic sequence
Fr	Froude number
Sr	Strouhal number
t'_{ref}, ℓ'_{ref}, h'_{ref}, v'_{ref}	dimensional reference units

Discretizations and partitions

V, I	control volume, interface of primary discretization
$\bar{V}$, $\bar{I}$	control volume, interface of dual discretization
$\mathcal{V}$, $\mathcal{I}$	set of control volumes, interfaces of primary discretization
$\bar{\mathcal{V}}$, $\bar{\mathcal{I}}$	set of control volumes, interfaces of dual discretization
δx, δy	grid spacings

Discrete operators

$G_{\mathcal{I}}^{\mathcal{V}}$, $D_{\bar{\mathcal{V}}}^{\mathcal{I}}$	gradient and divergence for the first projection, original projection method
$G_{\bar{\mathcal{V}}}^{\bar{\mathcal{V}}}$, $D_{\bar{\mathcal{V}}}^{\mathcal{V}}$	gradient and divergence for the second projection, original projection method
$\mathsf{L}_{\bar{\mathcal{V}}}^{\bar{\mathcal{V}}}$, $\mathsf{G}_{\bar{\mathcal{V}}}^{\bar{\mathcal{V}}}$, $\mathsf{D}_{\bar{\mathcal{V}}}^{\mathcal{V}}$	Laplacian, gradient and divergence for the second projection, new projection method

Miscellaneous

$\mathcal{O}(\cdot)$, $o(\cdot)$	Landau symbols
$\mathscr{J}(q)$	energy functional
$\mathscr{I}(\boldsymbol{v})$	complementary energy functional
$\mathscr{L}(\boldsymbol{v}, q)$	Lagrangian
$\boldsymbol{n}$	normal vector
$\boldsymbol{I}$	2×2 identity matrix
χ_U	characteristic function on $U \subset \mathbb{R}^n$
δ_{ij}	Kronecker symbol

1 Introduction

Many phenomena of interest in geophysical fluid mechanics can be modeled with the shallow water equations. This system of equations is an appropriate approximation of processes acting on large horizontal length scales in relation to the considered vertical length scale. It describes flows of an incompressible fluid with a free surface. The shallow water equations are interesting not only from the geophysical, but also from the numerical point of view. On the one hand, they can describe the important aspects of atmospheric and oceanic phenomena. While ignoring the presence of stratification, the shallow water equations incorporate the effects of gravity and can account for the earth's rotation and for bottom topography. They are, for instance, a suitable approximation for large scale midlatitude motions [Majda, 2003]. On the other hand, the shallow water system is characterized as a hyperbolic system of only two equations. Its nonlinear structure is fairly simple, but it is similar to those of more complex examples, such as the Euler equations of gas dynamics [LeVeque, 2002].

The physical processes in the ocean act on very different spatial and temporal scales. Gravity waves on the surface of the ocean, which carry energy and momentum over large distances, are among the fastest of these kind of processes [Le Maître et al., 2001]. These waves can travel at speeds exceeding 200 meters per second in deep waters, whereas the advection velocity of the water is normally less than 5 meters per second. Obviously, there are two different scales within one system, and the great disparity between the scales is expressed by a small *Froude number*, the ratio between the velocity of flow and the speed of gravity waves.

In the *zero* Froude number shallow water equations, we consider the case in which the ratio between the gravity wave speed and the characteristic advection velocity of the fluid becomes infinitely large. This limit process brings with it considerable changes to the mathematical properties of the governing equations: While the shal-

low water equations are a hyperbolic system of partial differential equations, they are of mixed hyperbolic-elliptic type in the limit of a vanishing Froude number. Furthermore, the velocity field of the fluid has to satisfy, in the limit, a *divergence constraint* (e.g. $\nabla \cdot \boldsymbol{v} = 0$ for cases with no flux across the boundary). Clearly, these circumstances require different numerical methods for the computation of approximate solutions in both regimes.

1.1 The shallow water equations

The following assumptions form the basis for the derivation of the shallow water equations. We consider an incompressible, inviscid fluid, which is shallow and homogeneous. Given a characteristic depth d'_{ref} and a characteristic length scale for the horizontal motion ℓ'_{ref}, the "shallowness" of the fluid can be expressed by the ratio $d'_{\text{ref}}/\ell'_{\text{ref}} \ll 1$. Its homogeneity is manifested in a constant and uniform density ϱ'.[1] Moreover, the hydrostatic approximation

$$\frac{\partial p'}{\partial z'} = -\varrho'\, g'$$

is assumed to be valid. Here, p' is the pressure, z' the vertical coordinate and g' the gravitational constant. The axis of rotation of the fluid is considered to coincide with the vertical axis, and the frequency of rotation is given by the (Coriolis) parameter f'. With these assumptions, the *quasilinear form* of the two-dimensional rotating shallow water equations is given by [Majda, 2003, Chapter 4]

$$\begin{aligned} \frac{Dh'}{Dt'} &+ h'\,\nabla' \cdot \boldsymbol{v}' &&= 0 \\ \frac{D\boldsymbol{v}'}{Dt'} &+ f'\,\boldsymbol{v}'^{\perp} &&= -g'\,\nabla' h' \quad . \end{aligned} \tag{1.1}$$

In these equations, $\boldsymbol{v}' = (u'(\boldsymbol{x}', t'), v'(\boldsymbol{x}', t'))$ is the horizontal component of the fluid velocity, and $\boldsymbol{v}'^{\perp} = (-v', u')$ is the "orthogonal velocity". The total depth $h' = h'_{\text{T}} - h'_{\text{B}}$ is given as the difference between the top of the fluid $h'_{\text{T}}(\boldsymbol{x}', t')$ and the bottom

[1] Variables with primes are always dimensional, while those without primes are nondimensional.

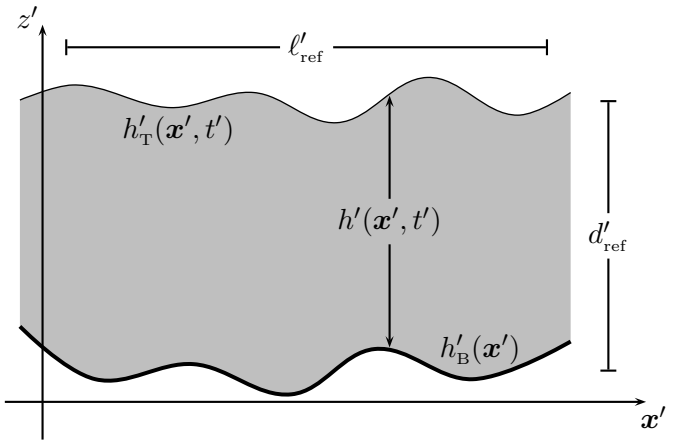

Figure 1.1: The shallow water model.

topography $h'_{\text{B}}(\boldsymbol{x}')$ (see Figure 1.1). Furthermore, $\frac{D}{Dt'} \coloneqq \frac{\partial}{\partial t'} + \boldsymbol{v}' \cdot \nabla'$ is the material derivative. For a complete derivation of the shallow water equations from the three-dimensional incompressible Euler equations, the reader is referred to PEDLOSKY [1987, pp. 59-63].

The shallow water equations are a system of first order partial differential equations. The system (1.1) can be also written in *conservation form*, which is given by

$$\begin{array}{rcl} \dfrac{\partial h'}{\partial t'} + \nabla' \cdot (h' \boldsymbol{v}') & = & 0 \\ \dfrac{\partial (h' \boldsymbol{v}')}{\partial t'} + \nabla' \cdot \left(h' \boldsymbol{v}' \circ \boldsymbol{v}' + \dfrac{g'}{2} h'^2 \, \boldsymbol{I} \right) & = & -(f' \, \boldsymbol{v}'^{\perp} + g' \, \nabla' h'_{\text{B}}) \, h' \quad , \end{array} \tag{1.2}$$

where $\boldsymbol{I}$ is the 2×2 identity matrix. In the conservation of "momentum" $(1.2)_2$, we have written the contributions made by rotation and bottom topography as source terms on the right hand side of the equation. In this thesis, bottom topography and rotational effects are not considered. Therefore, we omit the right hand side of $(1.2)_2$ in the following.

Remark 1.1 *The shallow water equations have the same mathematical structure as the Euler equations of compressible isentropic gas dynamics* [MAJDA, 2003, p. 50].

In fact, the latter are given by

$$\begin{aligned} \frac{\partial \varrho'}{\partial t'} &+ \nabla' \cdot (\varrho' \boldsymbol{v}') &&= 0 \\ \frac{\partial (\varrho' \boldsymbol{v}')}{\partial t'} &+ \nabla' \cdot (\varrho' \boldsymbol{v}' \circ \boldsymbol{v}' + \hat{\kappa} \, (\varrho')^{\gamma} \, \boldsymbol{I}) &&= 0 \end{aligned}$$

with constants $\hat{\kappa}$ and $\gamma > 1$. By replacing ϱ' with h', $\hat{\kappa}$ with $g'/2$, and setting $\gamma = 2$, the shallow water equations without source terms are recovered. ◁

Therefore, similar numerical methods can be used for the approximate solution of both systems of equations.

1.1.1 Dimensional analysis

The aim of any concrete (experimental or theoretical) physical study is to understand the relationship between the characterizing quantities of the problem under consideration. To measure the relative importance of the different terms in the shallow water equations, we undertake a *dimensional analysis*. In this analysis, reference quantities of the dependent and independent variables in the problem have to be identified. By taking the ratio between these parameters, a well defined number of nondimensional *characteristic numbers* can be deduced, which specify the problem's nature. This connection is described by the so-called Π-theorem [Barenblatt, 1996, Section 1.2][2].

Let us introduce, besides the length scale ℓ'_{ref}, a typical time scale t'_{ref} of the problem under consideration, and denote by h'_{ref} and v'_{ref} reference units for the height and velocity, respectively (e.g. given by the initial conditions). Then, we can define the nondimensional variables

$$\boldsymbol{x} \coloneqq \frac{\boldsymbol{x}'}{\ell'_{\text{ref}}} \, , \quad t \coloneqq \frac{t'}{t'_{\text{ref}}} \, , \quad h \coloneqq \frac{h'}{h'_{\text{ref}}} \quad \text{and} \quad \boldsymbol{v} \coloneqq \frac{\boldsymbol{v}'}{v'_{\text{ref}}} \, .$$

The substitution of the dimensional variables by their nondimensional counterparts

[2] For an introduction to this topic, the reader is also referred to Klein and Vater [2003, Section 2.1.2 and Chapter 3].

in (1.2) leads to

$$\begin{aligned} \frac{h'_{\text{ref}}}{t'_{\text{ref}}}\,\frac{\partial h}{\partial t} &+ \frac{h'_{\text{ref}}\,v'_{\text{ref}}}{\ell'_{\text{ref}}}\,\nabla\cdot(h\boldsymbol{v}) &&= 0 \\ \frac{h'_{\text{ref}}\,v'_{\text{ref}}}{t'_{\text{ref}}}\,\frac{\partial(h\boldsymbol{v})}{\partial t} &+ \frac{h'_{\text{ref}}\,v'^2_{\text{ref}}}{\ell'_{\text{ref}}}\,\nabla\cdot\left(h\boldsymbol{v}\circ\boldsymbol{v}+\frac{g'\,h'_{\text{ref}}}{2\,v'^2_{\text{ref}}}\,h^2\,\boldsymbol{I}\right) &&= 0 \quad . \end{aligned}$$

If the first equation is multiplied by $\ell'_{\text{ref}}/(h'_{\text{ref}}v'_{\text{ref}})$ and the second one by $\ell'_{\text{ref}}/(h'_{\text{ref}}v'^2_{\text{ref}})$, the nondimensional shallow water equations are obtained:

$$\begin{aligned} \mathsf{Sr}\,\frac{\partial h}{\partial t} &+ \nabla\cdot(h\boldsymbol{v}) &&= 0 \\ \mathsf{Sr}\,\frac{\partial(h\boldsymbol{v})}{\partial t} &+ \nabla\cdot\left(h\boldsymbol{v}\circ\boldsymbol{v}+\frac{1}{2\,\mathsf{Fr}^2}\,h^2\,\boldsymbol{I}\right) &&= 0 \quad . \end{aligned} \tag{1.3}$$

Here, we have introduced the dimensionless characteristic numbers

$$\mathsf{Sr} \coloneqq \frac{\ell'_{\text{ref}}}{t'_{\text{ref}}\,v'_{\text{ref}}} \quad \text{and} \quad \mathsf{Fr} \coloneqq \frac{v'_{\text{ref}}}{\sqrt{g'\,h'_{\text{ref}}}} \quad ,$$

which are known as the *Strouhal* and the *Froude number*, respectively. In this thesis, we do not consider external forces which could assign an additional time scale to the problem. Thus, we are interested in a reference time scale equal to the advection time scale of the fluid, so that $t'_{\text{ref}} = \ell'_{\text{ref}}/v'_{\text{ref}}$ and the Strouhal number becomes one ($\mathsf{Sr} = 1$).

In Remark 1.1, we saw that the shallow water system is equivalent to a special case of the Euler equations of gas dynamics. The importance of compressibility in the Euler equations is given by the *Mach number*, which is defined by the ratio between the representative fluid velocity v'_{ref} and the speed of sound $c'_{\text{ref}} \coloneqq \sqrt{p'_{\text{ref}}/\varrho'_{\text{ref}}}$. As mentioned earlier, in the shallow water model we consider an incompressible fluid, but the analogue of the Mach number is given by the Froude number. Thus, the associated "compressibility" effects are given by the ratio of the typical fluid velocity v'_{ref} and the *gravity wave speed* $\sqrt{g'\,h'_{\text{ref}}}$, that is the speed at which long wave perturbations of the depth travel [Pedlosky, 1987].

For two-dimensional velocity fields, the vorticity ω is given by

$$\omega \coloneqq \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \quad .$$

An evolution equation for this quantity can be derived by taking the curl of the quasilinear form of the momentum equation in (1.3), which is given by

$$\mathsf{Sr}\ \frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{v} \cdot \nabla \boldsymbol{v} + \frac{1}{\mathsf{Fr}^2} \nabla h = 0 \quad .$$

By these means, we obtain

$$\mathsf{Sr}\ \frac{\partial \omega}{\partial t} + \nabla \cdot (\omega\, \boldsymbol{v}) = 0 \quad . \tag{1.4}$$

1.1.2 Characteristic structure

The introduction of the shallow water equations is completed with a characteristic analysis. By integrating the governing equations (1.3) over an arbitrary bounded volume $\Omega \subset \mathbb{R}^2$ and using the divergence theorem, we obtain a *conservation law* of the form

$$\frac{\partial}{\partial t} \int_\Omega \mathbf{u}\ d\boldsymbol{x} + \int_{\partial\Omega} \mathbf{f}(\mathbf{u}, \boldsymbol{n}; \mathsf{Fr})\ d\sigma = \mathbf{0} \quad \forall\, t > 0 \quad .$$

It describes the interplay between the *density function* of conserved variables

$$\mathbf{u} : \Omega \times [0, \infty) \to \mathbb{R}^3 \quad \text{with} \quad \mathbf{u}(\boldsymbol{x}, t) \coloneqq \begin{pmatrix} h \\ h\boldsymbol{v} \end{pmatrix}$$

and the *flux function* $\mathbf{f} : U \times \Omega \times [0, \infty) \to \mathbb{R}^3$, where $U \subset \mathbb{R}^3$ is an open set (see Figure 1.2). The flux function is given by

$$\mathbf{f}(\mathbf{u}(\boldsymbol{x}, t), \boldsymbol{n}(\boldsymbol{x}); \mathsf{Fr}) \coloneqq \begin{pmatrix} h(\boldsymbol{v} \cdot \boldsymbol{n}) \\ h\boldsymbol{v}(\boldsymbol{v} \cdot \boldsymbol{n}) \quad + \quad \frac{1}{2\mathsf{Fr}^2}\, h^2\, \boldsymbol{n} \end{pmatrix} \quad .$$

The *Jacobian* matrix $\frac{d}{d\mathbf{u}} \mathbf{f}(\mathbf{u}, \boldsymbol{n})$ has real eigenvalues $\boldsymbol{v} \cdot \boldsymbol{n}$ and $\boldsymbol{v} \cdot \boldsymbol{n} \pm \sqrt{h}/\mathsf{Fr}$ and a complete set of eigenvectors. Therefore, the matrix is diagonalizable, and the

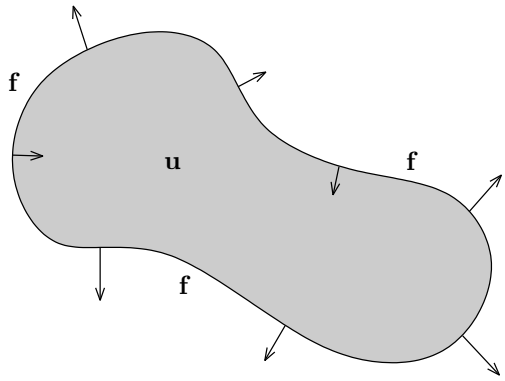

Figure 1.2: Density $\mathbf{u}$ of conserved variables and flux $\mathbf{f}(\mathbf{u}, \boldsymbol{n})$ (denoted by arrows) on the boundary of a control volume.

shallow water equations are a hyperbolic system of partial differential equations. These eigenvalues become singular in the limit $\mathsf{Fr} \to 0$.

1.2 Purpose and objectives

The main objective of this thesis is to derive a new semi-implicit projection method for the zero Froude number shallow water equations. This method is based on a finite volume method for the zero Mach number Euler equations, originally proposed in Schneider et al. [1999]. The new scheme features two elliptic projections, which are based on a *Petrov-Galerkin* finite element formulation. In the course of this work, the following questions will be addressed:

- Which modifications of the original scheme have to be done to implement the new projection method?
- Is it possible to utilize the finite element formulation of the new projection to show analytically the stability of this part of the method?
- What is the behavior of the new projection method compared to the original scheme?
- Numerical methods for the solution of hyperbolic problems typically consist of a reconstruction step followed by the computation of numerical fluxes. Can

the numerical solution be improved, when the reconstruction is constrained by auxiliary equations?

For the investigation of the zero Froude number limit of the shallow water equations, we undertake an asymptotic analysis in Chapter 2. This analysis results in a divergence constraint for the velocity field, which is a major ingredient of the numerical method presented in Chapter 3. The finite element formulation of the new projection rests on bilinear ansatz functions for the nonuniform component of the height. Two different versions of the new method are presented.

The divergence constraint, in conjunction with the momentum update, leads to the formulation of a saddle point problem, which is equivalent to the new projection. This formulation is derived in Chapter 4 and provides the basis for the subsequent stability analysis of the new projection. Numerical results, which are obtained with the original method as well as the new projection method are presented in Chapter 5. In the final part, open questions are discussed and we outline possible approaches for their solution.

2 Asymptotic Analysis

In asymptotic analysis we exploit the fact that the problem under consideration incorporates at least one very small or very large dimensionless characteristic quantity. Often, such circumstances can be used to simplify the equations describing the physical system considerably. In this chapter, the shallow water equations are investigated in their "incompressible" limit. Thus, the Froude number has the role of the small parameter, in which the asymptotic analysis is undertaken. The analysis of the shallow water equations in the low Froude number limit is done in analogy to the study of the low Mach number Euler equations by Klein [1995], using a two space scale, single time scale ansatz.

First, we shortly introduce the most important principles of asymptotic analysis, and the required properties for the multiple scales ansatz are proven in analogy to the work of Meister [1997].[1] After the identification of the small parameter and the formulation of the asymptotic ansatz, an analysis of the asymptotic limit equations is performed. Two different regimes of flow are investigated. In the first one, only one space scale is considered, while in the second regime two space scales are taken into account. In the final part of this chapter the results for the single scale regime, which coincides with the zero Froude number shallow water equations, are summarized.

2.1 Basic principles

For the discussion of the basic principles in asymptotic analysis let us define an interval $I \coloneqq (0, \varepsilon']$, in which $\varepsilon' > 0$ is a positive real number. Also, we consider an

[1] See also Schneider [1978] and Kevorkian and Cole [1996] for an introduction to asymptotic analysis.

n-dimensional set $D \subset \mathbb{R}^n$ and a scalar real valued function

$$u \;:\; D \times I \to \mathbb{R} \quad \text{with} \quad (x_1, \ldots, x_n; \varepsilon) \mapsto u(x_1, \ldots, x_n; \varepsilon) \quad .$$

The basis for any asymptotic analysis is an asymptotic sequence, which is defined as follows.

Definition 2.1 *A sequence of functions* $\{\phi_n(\varepsilon)\}_{n\in\mathbb{N}}$ *with* $\phi_n : \mathbb{R}^+ \to \mathbb{R}$ *for all* n *is called an* asymptotic sequence*, if*

$$\phi_{n+1}(\varepsilon) = o(\phi_n(\varepsilon)) \quad \text{as } \varepsilon \to 0$$

is valid for all $n \in \mathbb{N}$*.*

A simple example of such a sequence is $\{\varepsilon^n\}_{n\in\mathbb{N}}$. In general, asymptotic sequences can also consist of fractional powers, logarithmic functions, etc. With a given asymptotic sequence the notion of an asymptotic expansion can be defined.

Definition 2.2 *Let* $u : D \times I \to \mathbb{R}$ *with* $(\boldsymbol{x}; \varepsilon) \mapsto u(\boldsymbol{x}; \varepsilon)$ *and* $\{\phi_n(\varepsilon)\}_{n\in\mathbb{N}}$ *be an asymptotic sequence. We define for* $N \in \mathbb{N}$ *a series of the form*

$$\sum_{i=0}^{N} \phi_i(\varepsilon)\, u^{(i)}(\boldsymbol{x}) \tag{2.1}$$

to be an asymptotic N-term expansion *of* u*, if for each* $\boldsymbol{x} \in D$

$$u(\boldsymbol{x}; \varepsilon) - \sum_{i=0}^{N} \phi_i(\varepsilon)\, u^{(i)}(\boldsymbol{x}) = o(\phi_N(\varepsilon)) \quad \text{as } \varepsilon \to 0 \quad .$$

The functions $u^{(i)} : \tilde{D} \to \mathbb{R}$ $(\tilde{D} \in \mathbb{R}^n)$ *are called* asymptotic functions*.*

The idea is then, to replace the unknown of the problem under consideration by an asymptotic expansion, and to find subsequently solutions for the asymptotic functions $u^{(i)}$. This (hopefully) leads to an approximate solution of the problem. An asymptotic expansion we get for a given problem strongly depends on the prescribed asymptotic sequence. Additionally, such an expansion might not even exist. For example, let us

assume that $u(x;\varepsilon) = 1+\sqrt{\varepsilon}$ is the exact solution of a given differential equation and that we have used $\{\varepsilon^n\}_{n\in\mathbb{N}}$ as asymptotic sequence. In this case, we could not expect to obtain an adequate asymptotic expansion, because u cannot be represented by a series of the form (2.1). On the other hand, the sequence $\{\varepsilon^{n/2}\}_{n\in\mathbb{N}}$ would clearly reproduce the exact solution for $N = 1$.

If N approaches infinity, the asymptotic expansion might converge for ε being in a certain range. However, in many circumstances we lack information about the convergence properties of the calculated asymptotic expansion, and it is only reasonable to compute the first one or two terms of the expansion. Thus, in most cases it is irrelevant how the series behaves for $N \to \infty$ and ε finite; the more important question is how the expansion behaves for $\varepsilon \to 0$ given a fixed N [Schneider, 1978, p. 67]. The usefulness of an asymptotic expansion is given by the property that ε has only to be chosen small enough to approximate the unknown solution sufficiently well.

Any N-term expansion of a given function u incorporates the k-term expansions of u with $k \in \mathbb{N}, k < N$ [cf. Meister, 1997]. Using the following fundamental property of asymptotic analysis, we will outline how to use the tool of asymptotic analysis in solving differential equations, at least up to a certain order of accuracy.

Proposition 2.1 *Let $\{\phi_n(\varepsilon)\}_{n\in\mathbb{N}}$ be an asymptotic sequence and $L^{(i)}$, $i = 0,\dots,N$ arbitrary terms independent of ε (e.g. real valued functions on D). Then*

$$\sum_{i=0}^{N} \phi_i(\varepsilon)\, L^{(i)} = o(\phi_N(\varepsilon)) \quad \text{as } \varepsilon \to 0 \tag{2.2}$$

is equivalent to

$$L^{(i)} = 0 \quad \text{for } i = 0,\dots,N \quad .$$

Proof. *Assuming $L^{(i)} = 0$ for $i = 0,\dots,N$ we immediately get the first statement. For the opposite direction let us assume that there is at least one $L^{(n)} \neq 0$, $0 \le n \le N$. W.l.o.g. we can assume that $L^{(m)} = 0$ for $0 \le m < n$ to obtain with (2.2)*

$$\sum_{i=0}^{n} \phi_i(\varepsilon)\, L^{(i)} = o(\phi_n(\varepsilon)) \quad \text{as } \varepsilon \to 0 \ .$$

This leads to

$$0 = \lim_{\varepsilon \to 0} \frac{\sum_{i=0}^{n} \phi_i(\varepsilon)\, L^{(i)}}{\phi_n(\varepsilon)} = \lim_{\varepsilon \to 0} \frac{\sum_{i=0}^{n-1} \phi_i(\varepsilon)\, L^{(i)}}{\phi_n(\varepsilon)} + L^{(n)} = L^{(n)}$$

and thus, contradicts our assumption $L^{(n)} \neq 0$. □

Using this idea, the following steps have to be performed in the analysis of a given homogeneous differential equation. First, an asymptotic sequence $\{\phi_n(\varepsilon)\}_{n \in \mathbb{N}}$ is chosen and an ansatz of the form

$$u(\boldsymbol{x}; \varepsilon) = \sum_{i=0}^{N} \phi_i(\varepsilon)\, u^{(i)}(\boldsymbol{x}) + \mathcal{O}(\phi_N(\varepsilon)) \quad \text{as } \varepsilon \to 0 \tag{2.3}$$

is specified for the unknown u. By inserting this ansatz into the differential equation, the problem is reformulated to obtain

$$\sum_{j=0}^{M} \psi_j(\varepsilon)\, L^{(j)}(u^{(0)}, \ldots, u^{(N)}) = \mathcal{O}(\psi_M(\varepsilon)) \quad \text{as } \varepsilon \to 0 \tag{2.4}$$

with $\psi_{j+1}(\varepsilon) = \mathcal{O}(\psi_j(\varepsilon))$ for $j = 0, \ldots, M-1$. In each $L^{(j)}(u^{(0)}, \ldots, u^{(N)})$ we have merged terms, which are multiplied by equal powers in ε. The $L^{(j)}$ are independent of ε. Thus, Proposition 2.1 can be applied, and by solving the system of differential equations

$$L^{(m)}(u^{(0)}, \ldots, u^{(N)}) = 0 \quad \text{for } m = 0, \ldots, M \quad , \tag{2.5}$$

we finally obtain an (approximate) solution in the form (2.3) for u.

A lot of applications include phenomena, which act on different scales in time or space. In this case, even a well chosen asymptotic sequence might not result in a satisfying expansion. Often, this happens if the associated differential equation under consideration loses an order or changes its type in the limit $\varepsilon \to 0$ [Meister, 1997]. An example for such a behavior is a linear oscillator with small mass which is driven by a sinusoidal background force[2]. This system can be described by the initial value

[2]For a description of the weakly damped case, the reader is referred to [Klein and Vater, 2003, Chapter 2].

problem

$$\varepsilon y'' + y = \cos\tau,\ y(0) = y_0,\ y'(0) = y_0' \tag{2.6}$$

and has for $y_0 = 1 + \varepsilon$ and $y_0' = 0$ the exact solution

$$y(\tau;\varepsilon) = \frac{1}{1-\varepsilon}\left(\cos\tau - \varepsilon^2 \cos\frac{\tau}{\sqrt{\varepsilon}}\right) \quad . \tag{2.7}$$

For a fixed ε, two well defined frequencies are present in this solution. Although the limit equation of (2.6) looses two orders, the boundedness of the cosine function implies uniform convergence of (2.7) towards the solution of the unperturbed problem. Using $\{\varepsilon^n\}_{n\in\mathbb{N}}$ as asymptotic sequence, an asymptotic analysis as described above would result in the (single scale) solution

$$y_{\mathrm{ss}}(\tau;\varepsilon) = (1 + \varepsilon + \varepsilon^2)\cos\tau + \mathcal{O}\big(\varepsilon^2\big) \quad . \tag{2.8}$$

Despite the fact that y_{ss} also converges to the solution of the unperturbed problem as $\varepsilon \to 0$, this is not an asymptotic expansion in the sense of Definition 2.2. The asymptotic solution (2.8) only reproduces the behavior of the external force. This is due to the fact that the asymptotic functions $u^{(i)}$ just depend on τ. Therefore, they can only represent long wave components of the solution and the influence of ε on the frequency is lost [Meister, 1997].

The concept of an asymptotic expansion in Definition 2.2 is given as an ansatz with *separation of variables*. Obviously, this approach is not comprehensive enough for all purposes. On the other hand, an ansatz of the kind $u^{(i)} = u^{(i)}(\boldsymbol{x},\varepsilon)$ without other side constraints might be too general. Therefore, we introduce the notion of multiple scales expansions.

Definition 2.3 *Let $u : D\times I \to \mathbb{R}$ with $(\boldsymbol{x};\varepsilon) \mapsto u(\boldsymbol{x};\varepsilon)$, $\{\phi_n(\varepsilon)\}_{n\in\mathbb{N}}$ be an asymptotic sequence and $\boldsymbol{g} : D \times I \to \tilde{D} \subset \mathbb{R}^m$. The series*

$$\sum_{i=0}^{N} \phi_i(\varepsilon)\, u^{(i)}(\boldsymbol{g}(\boldsymbol{x},\varepsilon))$$

is called an asymptotic N-term multiple scales expansion *of* u*, if*

$$u(\boldsymbol{x};\varepsilon) - \sum_{i=0}^{N} \phi_i(\varepsilon)\, u^{(i)}(\boldsymbol{g}(\boldsymbol{x},\varepsilon)) = \mathcal{O}(\phi_N(\varepsilon)) \quad \text{as } \varepsilon \to 0 \quad .$$

With this definition, the domain of the asymptotic functions $u^{(i)}$ has changed. Furthermore, the function $\boldsymbol{g}$ couples the considered scales of the problem. To find a multiple scales asymptotic solution of a given differential equation, it is still sufficient to proceed in the same way as outlined above. However, in (2.4) the coefficients $L^{(i)}$ are now dependent on ε. Consequently, it is not clear that all of them have to vanish, but if we found a solution for (2.5) independent of ε, then the asymptotic expansion would be valid in either case.

For our example (2.6) of the linear oscillator, a multiple scales analysis with $\boldsymbol{g}(\tau;\varepsilon) = (\tau, \tau/\sqrt{\varepsilon})$ would result in the approximate solution

$$y_{\mathrm{ms}}(\tau;\varepsilon) = (1+\varepsilon+\varepsilon^2)\cos\tau - \varepsilon^2 \cos\frac{\tau}{\sqrt{\varepsilon}} + \mathcal{O}\big(\varepsilon^2\big) \quad , \tag{2.9}$$

where we have used the same asymptotic sequence as before. In contrast to the single scale expansion, this solution is an asymptotic expansion and thus also tends to the unperturbed solution as $\varepsilon \to 0$. The second term in (2.9) represents the fast time scale of the problem, which was missing in the single scale expansion. Through the mapping $\boldsymbol{g}$ the asymptotic functions $u^{(i)}$ are now dependent on the two different physical scales.

2.2 Ansatz for the low Froude number limit

Looking at the low Froude number limit of the shallow water equations, we identify a small parameter ε with the Froude number. We seek solutions to the nondimensional shallow water equations (1.3) (including suitable initial and boundary conditions) by using a multiple scales expansion of the unknowns. Thus, let $\mathsf{Fr} = \varepsilon \in I \coloneqq (0, \varepsilon']$

with $\varepsilon' \ll 1$ and

$$\boldsymbol{g} : \mathbb{R}^{d+1} \times I \to \mathbb{R}^{2d+1} \quad \text{with} \quad \boldsymbol{g}(\boldsymbol{x}, t; \varepsilon) = (\boldsymbol{x}, \varepsilon \boldsymbol{x}, t) =: (\boldsymbol{\eta}, \boldsymbol{\xi}, t) \quad .$$

A quantity $w(\boldsymbol{x}, t; \varepsilon)$ with ε fixed but arbitrary small shall then be representable as

$$\begin{aligned} w(\boldsymbol{x}, t; \varepsilon) &= \sum_{i=0}^{N} \phi_i(\varepsilon)\ w^{(i)}(\boldsymbol{g}(\boldsymbol{x}, t; \varepsilon)) + o(\phi_N(\varepsilon)) \\ &= \sum_{i=0}^{N} \varepsilon^i\ w^{(i)}(\boldsymbol{\eta}, \boldsymbol{\xi}, t) + o\big(\varepsilon^N\big) \quad \text{as } \varepsilon \to 0 \end{aligned} \tag{2.10}$$

uniformly for all $(\boldsymbol{x}, t) \in \mathbb{R}^d \times \mathbb{R}_0^+$. Also, the expansion should include all the k-term expansions with $0 \leq k < N$. Because two space coordinates are considered in this expansion, the differentiation in space yields for an asymptotic function $w^{(j)}$, $j = 0, \ldots, N$

$$\left.\frac{\partial w^{(j)}}{\partial x_i}\right|_{\varepsilon} (\boldsymbol{g}(\boldsymbol{x}, t; \varepsilon)) = \frac{\partial w^{(j)}}{\partial \eta_i}(\boldsymbol{\eta}, \boldsymbol{\xi}, t) + \varepsilon \frac{\partial w^{(j)}}{\partial \xi_i}(\boldsymbol{\eta}, \boldsymbol{\xi}, t) \quad .$$

With the notation $\frac{\partial w^{(j)}}{\partial x_i}|_{\varepsilon}$ it should be stressed that the parameter ε is a fixed quantity. Otherwise, also the considered scales of the problem and thus the Froude number would change.

By inserting the ansatz (2.10) into (1.3) we obtain for the dimensionless shallow water equations

$$\begin{aligned} \Big[h_t^{(0)} + \nabla_{\boldsymbol{\eta}} \cdot (h\boldsymbol{v})^{(0)}\Big](\boldsymbol{\eta}, \boldsymbol{\xi}, t) \quad + \\ \varepsilon \Big[h_t^{(1)} + \nabla_{\boldsymbol{\eta}} \cdot (h\boldsymbol{v})^{(1)} + \nabla_{\boldsymbol{\xi}} \cdot (h\boldsymbol{v})^{(0)}\Big](\boldsymbol{\eta}, \boldsymbol{\xi}, t) \quad + \quad o(\varepsilon) \quad = \quad 0 \end{aligned} \tag{2.11}$$

and

$$\begin{aligned} &\frac{1}{\varepsilon^2}\left[(h\nabla_{\boldsymbol{\eta}} h)^{(0)}\right](\boldsymbol{\eta}, \boldsymbol{\xi}, t) + \frac{1}{\varepsilon}\left[(h\nabla_{\boldsymbol{\eta}} h)^{(1)} + (h\nabla_{\boldsymbol{\xi}} h)^{(0)}\right](\boldsymbol{\eta}, \boldsymbol{\xi}, t) \quad + \\ &\Big[(h\boldsymbol{v})_t^{(0)} + \nabla_{\boldsymbol{\eta}} \cdot (h\boldsymbol{v} \circ \boldsymbol{v})^{(0)} + (h\nabla_{\boldsymbol{\eta}} h)^{(2)} + (h\nabla_{\boldsymbol{\xi}} h)^{(1)}\Big](\boldsymbol{\eta}, \boldsymbol{\xi}, t) \quad + \quad o(1) \quad = \quad \mathbf{0} \end{aligned} \tag{2.12}$$

as $\varepsilon \to 0$. For further conclusions we try to find solutions for which the terms in brackets are independent of ε. As stated earlier, under these circumstances we can use Proposition 2.1. By indicating with $\mathcal{O}(\varepsilon^i)$ the corresponding order in ε of the equation, the continuity equation (2.11) is equivalent to

$$\begin{aligned} \mathcal{O}(1): \quad & \frac{\partial h^{(0)}}{\partial t} + \nabla_{\boldsymbol{\eta}} \cdot (h\boldsymbol{v})^{(0)} &&= 0 \\ \mathcal{O}(\varepsilon): \quad & \frac{\partial h^{(1)}}{\partial t} + \nabla_{\boldsymbol{\eta}} \cdot (h\boldsymbol{v})^{(1)} + \nabla_{\boldsymbol{\xi}} \cdot (h\boldsymbol{v})^{(0)} &&= 0 \end{aligned} \tag{2.13}$$

and the momentum equation (2.12) is equivalent to

$$\begin{aligned} \mathcal{O}(\varepsilon^{-2}): \quad & h^{(0)} \nabla_{\boldsymbol{\eta}} h^{(0)} &&= \mathbf{0} \\ \mathcal{O}(\varepsilon^{-1}): \quad & h^{(0)} \nabla_{\boldsymbol{\eta}} h^{(1)} + h^{(1)} \nabla_{\boldsymbol{\eta}} h^{(0)} + h^{(0)} \nabla_{\boldsymbol{\xi}} h^{(0)} &&= \mathbf{0} \\ \mathcal{O}(1): \quad & (h\boldsymbol{v})_t^{(0)} + \nabla_{\boldsymbol{\eta}} \cdot (h\boldsymbol{v} \circ \boldsymbol{v})^{(0)} + (h \nabla_{\boldsymbol{\eta}} h)^{(2)} + (h \nabla_{\boldsymbol{\xi}} h)^{(1)} &&= \mathbf{0} \quad . \end{aligned} \tag{2.14}$$

For an asymptotic function $w^{(i)}$, a *sub-linear growth* condition is imposed: We assume that

$$w^{(i)}(\boldsymbol{\eta}, \boldsymbol{\xi}, t) = o(r) \text{ for } \boldsymbol{\eta} \in \partial B(\mathbf{0}, r) \text{ as } r \to \infty$$

for all $(\boldsymbol{\xi}, t) \in \mathbb{R}^d \times \mathbb{R}_0^+$. In this formula, $B(\mathbf{0}, r) \coloneqq \{\boldsymbol{\eta} \in \mathbb{R}^d \big| |\boldsymbol{\eta}| \leq r\}$ is the ball with radius r about the origin.

2.3 Analysis of the asymptotic system

The equations (2.13) and (2.14) are now analyzed to obtain further information about solutions in the low Froude number limit. The nondimensional equations (1.3) change their type as $\mathsf{Fr} \to 0$ from a hyperbolic to a mixed elliptic-hyperbolic system. This is already visible in the momentum equation, in which the gradient of the height is divided by the square of the Froude number. In the asymptotic analysis, this relationship becomes evident in the equations $(2.14)_1$ and $(2.14)_2$ for the two leading order terms of the height. These equations are also the starting point for this analysis. In particular, it will be shown that $h^{(0)}$ is only dependent on time and that $h^{(1)}$ is independent of the short space scale $\boldsymbol{\eta}$.

The assumption of a positive height h implies that the leading order term $h^{(0)}$ is also greater than zero. Thus, $(2.14)_1$ can be divided by $h^{(0)}$ to obtain $\nabla_{\boldsymbol{\eta}} h^{(0)} = \mathbf{0}$, meaning that

$$h^{(0)}(\boldsymbol{\eta}, \boldsymbol{\xi}, t) = \tilde{h}^{(0)}(\boldsymbol{\xi}, t) \quad .$$

Using the independence of $h^{(0)}$ from the short space scale, $(2.14)_2$ simplifies to

$$\nabla_{\boldsymbol{\eta}} h^{(1)} + \nabla_{\boldsymbol{\xi}} h^{(0)} = \mathbf{0} \quad . \tag{2.15}$$

To derive that $h^{(0)}$ is only dependent on time, equation (2.15) is integrated over $B(\mathbf{0}, r) \coloneqq \{\boldsymbol{\eta} \in \mathbb{R}^d \big| |\boldsymbol{\eta}| \leq r\}$. Applying the divergence theorem we get for all $(\boldsymbol{\xi}, t) \in \mathbb{R}^d \times \mathbb{R}_0^+$

$$\begin{aligned} \int\limits_{\partial B(\mathbf{0},r)} h^{(1)}(\boldsymbol{\eta}, \boldsymbol{\xi}, t)\, \boldsymbol{n}(\boldsymbol{\eta})\, d\sigma &= -\int\limits_{B(\mathbf{0},r)} \nabla_{\boldsymbol{\xi}} h^{(0)}(\boldsymbol{\eta}, \boldsymbol{\xi}, t)\, d\boldsymbol{\eta} \\ &= -|B(\mathbf{0}, r)|\ \nabla_{\boldsymbol{\xi}} \tilde{h}^{(0)}(\boldsymbol{\xi}, t) \quad , \end{aligned}$$

where $\boldsymbol{n}$ is the outward pointing unit normal vector on $\partial B(\mathbf{0}, r)$. From the sub-linear growth condition for $h^{(1)}$ in $\boldsymbol{\eta}$ it follows that

$$\begin{aligned} \nabla_{\boldsymbol{\xi}} \tilde{h}^{(0)}(\boldsymbol{\xi}, t) &= -\frac{1}{|B(\mathbf{0}, r)|} \int\limits_{\partial B(\mathbf{0},r)} h^{(1)}(\boldsymbol{\eta}, \boldsymbol{\xi}, t)\, \boldsymbol{n}(\boldsymbol{\eta})\, d\sigma \\ &= \mathcal{O}\big(r^{-d}\big) \cdot \mathcal{O}\big(r^{d-1}\big) \cdot o(r) \\ &= o(1) \quad \text{as } r \to \infty \quad . \end{aligned}$$

Consequently, $h^{(0)}$ is just dependent on time, and using this result in (2.15) we obtain

$$\begin{aligned} h^{(0)}(\boldsymbol{\eta}, \boldsymbol{\xi}, t) &= h_0(t) \\ h^{(1)}(\boldsymbol{\eta}, \boldsymbol{\xi}, t) &= h_1(\boldsymbol{\xi}, t) \quad . \end{aligned}$$

This means that short wave length components of the height only have an influence of order $\mathcal{O}(\varepsilon^2)$ on the solution and that long wave length fluctuations are of $\mathcal{O}(\varepsilon)$. The conclusions for the height variable also imply a requirement for the velocity field.

From $(2.13)_1$ the divergence constraint

$$\nabla_{\boldsymbol{\eta}} \cdot \boldsymbol{v}^{(0)}(\boldsymbol{\eta}, \boldsymbol{\xi}, t) = -\frac{1}{h_0(t)} \frac{dh_0}{dt}(t) \tag{2.16}$$

is obtained. Also, the momentum equation of order $\mathcal{O}(1)$ simplifies to

$$(h_0 \boldsymbol{v}^{(0)})_t + h_0 \nabla_{\boldsymbol{\eta}} \cdot (\boldsymbol{v} \circ \boldsymbol{v})^{(0)} + h_0(\nabla_{\boldsymbol{\eta}} h^{(2)} + \nabla_{\boldsymbol{\xi}} h^{(1)}) = \mathbf{0} \quad . \tag{2.17}$$

Two different regimes of flow will be considered in the remaining discussion. In the first case, a system with only a single length scale is considered. Thus, for any asymptotic function $w^{(i)}$, $i = 0, 1, \ldots, N$, we set

$$\nabla_{\boldsymbol{\xi}} w^{(i)} = \mathbf{0} \quad ,$$

so that information on the $\boldsymbol{\xi}$-scale becomes void. This regime can be interpreted as a system with dimensions comparable to our reference length, in which long wave components would have an infinitely large wavelength compared to the system dimensions. In the second regime, both space scales are considered. The dimensions of such a system are large compared to the reference length scale, and long wave length components of the solution cannot be neglected in this regime any more.

For further analysis of the first regime (2.16) is integrated in $\boldsymbol{\eta}$ over the whole domain of the system. Using the divergence theorem we obtain

$$\frac{d}{dt}(\ln h_0)(t) = -\frac{1}{|\Omega|} \int_{\partial\Omega} \boldsymbol{v}^{(0)}(\boldsymbol{\eta}, \boldsymbol{\xi}, t) \cdot \boldsymbol{n}(\boldsymbol{\eta}) \, d\sigma \tag{2.18}$$

with the same notations as above. This equation states that $\mathcal{O}(1)$ changes in height can only be induced by flux across the boundary. Another interpretation of (2.18) follows from the spatial homogeneity of h_0. If this quantity is given, an integral constraint for the normal velocity on $\partial\Omega$ is obtained.

Combining (2.18) with (2.16) yields

$$\nabla_{\boldsymbol{\eta}} \cdot \boldsymbol{v}^{(0)}(\boldsymbol{\eta}, \boldsymbol{\xi}, t) = -\frac{1}{|\Omega|} \int_{\partial\Omega} \boldsymbol{v}^{(0)}(\boldsymbol{\eta}, \boldsymbol{\xi}, t) \cdot \boldsymbol{n}(\boldsymbol{\eta}) \, d\sigma \quad .$$

Thus, the divergence of $\boldsymbol{v}^{(0)}$ is uniform in space and varies in time upon a volume flux across the boundary of the system. Moreover, the momentum equation (2.17) becomes

$$(h_0\boldsymbol{v}^{(0)})_t + h_0\nabla_{\boldsymbol{\eta}}\cdot(\boldsymbol{v}\circ\boldsymbol{v})^{(0)} + h_0\nabla_{\boldsymbol{\eta}}h^{(2)} = 0 \quad .$$

As mentioned above, in the second regime the solution also has long wave components. We will see that the analysis of this regime reveals an evolution equation for $h^{(1)}$. The first result is obtained by integrating (2.16) over the ball $B(\mathbf{0},\varepsilon^{-1})$. The sub-linear growth condition for $\boldsymbol{v}^{(0)}$ in $\boldsymbol{\eta}$ in conjunction with the divergence theorem leads to

$$\begin{aligned} -\frac{1}{h_0(t)}\,\frac{dh_0}{dt}(t) &= \frac{1}{|B(\mathbf{0},\varepsilon^{-1})|}\int\limits_{\partial B(\mathbf{0},\varepsilon^{-1})} \boldsymbol{v}^{(0)}(\boldsymbol{\eta},\boldsymbol{\xi},t)\cdot\boldsymbol{n}(\boldsymbol{\eta})\,d\sigma \\ &= \mathcal{O}\big(\varepsilon^{d}\big)\cdot\mathcal{O}\big(\varepsilon^{1-d}\big)\cdot o\big(\varepsilon^{-1}\big) \\ &= o(1) \quad \text{as } \varepsilon\to 0 \quad . \end{aligned}$$

Consequently, h_0 is constant with respect to the time scale considered and

$$\frac{1}{h_0(t)}\,\frac{dh_0}{dt}(t) = 0 \quad . \tag{2.19}$$

When (2.19) is inserted into (2.16) we get the local divergence constraint

$$\nabla_{\boldsymbol{\eta}}\cdot\boldsymbol{v}^{(0)}(\boldsymbol{\eta},\boldsymbol{\xi},t) = 0 \quad .$$

To obtain information for the long wave components of the solution the asymptotic equations $(2.13)_2$ and (2.17) are averaged over the short space scale $\boldsymbol{\eta}$ in the limit $\varepsilon\to 0$. For this reason let us define

$$\overline{w^{(i)}}^{\boldsymbol{\eta}}(\boldsymbol{\xi},t) \coloneqq \lim_{\varepsilon\to 0}\frac{1}{|B(\mathbf{0},\varepsilon^{-1})|}\int\limits_{B(\mathbf{0},\varepsilon^{-1})} w^{(i)}(\boldsymbol{\eta},\boldsymbol{\xi},t)\,d\boldsymbol{\eta}$$

for any asymptotic function $w^{(i)}$. Taking the average of the momentum equation (2.17), the terms $\nabla_{\boldsymbol{\eta}}\cdot(\boldsymbol{v}\circ\boldsymbol{v})^{(0)}$ and $\nabla_{\boldsymbol{\eta}}h^{(2)}$ vanish because of the sub-linear growth of

$\boldsymbol{v}^{(0)}$ and $h^{(2)}$ in $\boldsymbol{\eta}$. The equation becomes

$$\overline{(h_0\boldsymbol{v}^{(0)})_t}^{\boldsymbol{\eta}} + h_0\nabla_{\boldsymbol{\xi}}h^{(1)} = \mathbf{0} \quad . \tag{2.20}$$

The application of the averaging procedure to the continuity equation $(2.13)_2$ yields, in conjunction with the independence of $h^{(1)}$ from $\boldsymbol{\eta}$,

$$\frac{\partial h^{(1)}}{\partial t} + h_0\overline{\nabla_{\boldsymbol{\xi}} \cdot \boldsymbol{v}^{(0)}}^{\boldsymbol{\eta}} = 0 \quad . \tag{2.21}$$

Assuming now that the averaging can be interchanged both with differentiation in time and with differentiation in the large space scale, (2.20) and (2.21) become

$$\begin{aligned} \frac{\partial h^{(1)}}{\partial t} + h_0\nabla_{\boldsymbol{\xi}} \cdot \overline{\boldsymbol{v}^{(0)}}^{\boldsymbol{\eta}} &= 0 \\ \frac{\partial}{\partial t}\overline{\boldsymbol{v}^{(0)}}^{\boldsymbol{\eta}} + \nabla_{\boldsymbol{\xi}}h^{(1)} &= 0 \quad . \end{aligned} \tag{2.22}$$

This is a linear system of differential equations with constant coefficients, from which the evolution of $h^{(1)}$ can be computed. If the exchangeability of differentiation in time and the large space scale holds for all t and $\boldsymbol{\xi}$ as well, we can finally combine the two equations from above to get the wave equation

$$\frac{\partial^2 h^{(1)}}{\partial t^2} - h_0\Delta_{\boldsymbol{\xi}}h^{(1)} = 0 \quad .$$

This completes the asymptotic analysis of the shallow water equations in the zero Froude number limit. Before the construction of the new scheme for the numerical solution of the zero Froude number shallow water equations is presented, the results for the first regime are summarized in the following section.

2.4 The zero Froude number limit

The equations, which are derived from (1.3) in the limit of a vanishing Froude number, are identical to those obtained in the first regime. Hence, the zero Froude number

shallow water equations are given by

$$\begin{aligned} h_t &+ \nabla\cdot(h\boldsymbol{v}) &&= 0 \\ (h\boldsymbol{v})_t &+ \nabla\cdot(h\boldsymbol{v}\circ\boldsymbol{v}) + h\nabla h^{(2)} &&= \mathbf{0} \\ & h &&= h_0(t) \quad . \end{aligned} \tag{2.23}$$

This system of equations is no longer hyperbolic, but of mixed elliptic-hyperbolic type. An additional variable $h^{(2)}$ is introduced and the height is split into a time dependent zero-gradient part h_0 and a second order perturbation $\varepsilon^2 h^{(2)}$. Having prescribed the normal velocity field on the boundary of the domain of integration, i.e.

$$\boldsymbol{v}(\boldsymbol{x},t)\cdot\boldsymbol{n}(\boldsymbol{x}) = b(\boldsymbol{x},t) \quad \text{on } \partial\Omega \quad ,$$

the change of height is given by

$$|\Omega|\frac{dh_0}{dt} = -h_0\int_{\partial\Omega} b\,d\sigma \quad . \tag{2.24}$$

If, on the other hand, $(h_0)_t$ is prescribed, the above equation implies a condition for the normal velocity field on the boundary of Ω. Integrating $(2.23)_1$ over an arbitrary volume $V\subset\Omega$ yields

$$\int_{\partial V}(h\boldsymbol{v})\cdot\boldsymbol{n}\,d\sigma = -|V|\frac{dh_0}{dt} \quad . \tag{2.25}$$

Thus, (2.25) in conjunction with the uniformity of $h = h_0$ in space implies an integral constraint for the velocity divergence in V.

We will see that the numerical scheme is constructed by solving a slightly different system compared to (2.23) as a predictor for the flow field at the new time step. This auxiliary system is given by

$$\begin{aligned} h_t^* &+ \nabla\cdot(h\boldsymbol{v})^* &&= 0 \\ (h\boldsymbol{v})_t^* &+ \nabla\cdot\left((h\boldsymbol{v}\circ\boldsymbol{v})^* + \frac{(h^*)^2}{2}\boldsymbol{I}\right) &&= \mathbf{0} \end{aligned} \tag{2.26}$$

with the flux function

$$\mathbf{f}^*(\mathbf{u}^*(\boldsymbol{x},t),\boldsymbol{n}(\boldsymbol{x})) = \begin{pmatrix} h(\boldsymbol{v}\cdot\boldsymbol{n}) \\ h\boldsymbol{v}(\boldsymbol{v}\cdot\boldsymbol{n}) \quad + \quad \frac{1}{2}\,h^2\,\boldsymbol{n} \end{pmatrix}^* \quad . \tag{2.27}$$

The auxiliary system is hyperbolic and has the same convective fluxes as (2.23). The eigenvalues of the Jacobian of the flux function $\mathbf{f}^*$ are $\boldsymbol{v}^*\cdot\boldsymbol{n}$ and $\boldsymbol{v}^*\cdot\boldsymbol{n}\pm\sqrt{h^*}$. Having constant height h^* and a velocity field $\boldsymbol{v}^*$ with zero divergence at time t_0, solutions of (2.26) satisfy at time $t_0+\delta t$ (cf. Appendix A.1)

$$\begin{aligned} \nabla\cdot\boldsymbol{v}^* &= \mathcal{O}(\delta t) \\ (h^*\nabla h^*) &= \mathcal{O}(\delta t^2) \end{aligned}$$

for $\delta t \to 0$.

Remark 2.1 *System* (2.26) *can be interpreted as another system of shallow water equations with* $\mathsf{Fr}=1$. ◁

3 The Numerical Scheme

In this chapter the new numerical scheme for the solution of the zero Froude number shallow water equations is described. The basis is a semi-implicit method originally proposed in SCHNEIDER ET AL. [1999] for the zero Mach number Euler equations. Several modifications are proposed in order to improve the latter ones accuracy and stability.

The scheme consists of two steps. First, the auxiliary system (2.26) is integrated over one time step using a standard second order method for hyperbolic conservation laws. In this step, predictions for the nonlinear convective flux components are calculated. The next step consist of two projections of this flux, each of them involving the solution of one Poisson-type equation for the height $h^{(2)}$. The solution of the first equation is used to correct the predictions of the convective fluxes in order to satisfy a discrete version of the divergence condition (2.25). In the second projection, the additional non-convective components of the fluxes are computed. This correction guarantees that the discrete velocity field at the new time step satisfies another discretization of (2.25).

In the original finite volume method, the unknowns are averages over control volumes, and standard discretizations are used to solve the Poisson-type equations. The description of this scheme concerning its application to the zero Froude number shallow water equations is given in the first section of this chapter. A new discretization for the two elliptic corrections is introduced in Section 3.2. It is based on a finite element formulation, in which $h^{(2)}$ is approximated by means of bilinear ansatz functions. This approach involves the introduction of piecewise linear velocity distributions. The resulting scheme can be formulated as an *approximate* projection method [cf. ALMGREN ET AL., 1996] as well as an *exact* method. Furthermore, additional constraints on the gradient of the momentum components in each cell, which are based on consistency considerations, are proposed in Section 3.3.

3.1 Original projection method

The original numerical method rests on a divergence constraint on the velocity field, which is derived in an asymptotic analysis of the low Mach number Euler equations. This constraint is equivalent to the one that is obtained in the asymptotic analysis in Chapter 2, and the scheme can be derived in a similar way for the zero Froude number shallow water equations.

The asymptotic analysis demonstrates the singular behavior of the governing equations as $\mathsf{Fr} \to 0$. In the nondimensional equations this singularity is manifested in an infinitely large gravity wave speed. Furthermore, the gradient of the height vanishes in the limit, but the term $\nabla h/\mathsf{Fr}^2$ in (1.3) becomes $\nabla h^{(2)}$, where $h^{(2)}$ is the second order height perturbation from the asymptotic analysis. In addition to the aforementioned divergence constraint, these characteristics have to be considered for the construction of a numerical method for the solution of the zero Froude number shallow water equations. The terms involving the propagation of gravity waves have to be treated implicitly to allow a *Courant-Friedrichs-Lewy (CFL)* time step restriction [Courant et al., 1928], which is dictated by the flow velocity.[1] Besides the spatially uniform background height, a second height variable has to be introduced to account for the contributions of $h^{(2)}$.

3.1.1 Construction of the scheme

Throughout this work we assume a regular space discretization of the computational domain Ω. In this discretization, the volume of a cell V is expressed as $|V|$, and two neighboring cells are separated by an interface I with area $|I|$ (cf. Figure 3.1). $\mathcal{V}$ and $\mathcal{I}$ are defined as the collection of all cells and interfaces, respectively. We denote the set of all interfaces, which are part of the boundary of a cell V, by $\mathcal{I}_{\partial V} \subset \mathcal{I}$.

For the construction of the method, a finite volume scheme in conservation form is considered, i.e.

$$\mathbf{U}_V^{n+1} = \mathbf{U}_V^n - \frac{\delta t}{|V|} \sum_{I \in \mathcal{I}_{\partial V}} |I| \, \mathbf{F}_I \quad . \tag{3.1}$$

[1] The CFL condition is a necessary condition for stability. It states that the numerical domain of dependence has to contain the domain of dependence of the continuous partial differential equation. See also LeVeque [2002, p. 68].

In (3.1) $\mathbf{U}_V^n$ is a numerical approximation to the average of the exact solution $\mathbf{u}(\boldsymbol{x}, t)$ of the problem over cell V at time t^n:

$$\mathbf{U}_V^n \approx \frac{1}{|V|} \int_V \mathbf{u}(\boldsymbol{x}, t^n)\ dV \quad , \quad \mathbf{u}(\boldsymbol{x}, t) \coloneqq \begin{pmatrix} h \\ h\boldsymbol{v} \end{pmatrix} \quad .$$

$\mathbf{F}_I$ approximates the average of the flux function

$$\mathbf{f}(\mathbf{u}(\boldsymbol{x}, t), \boldsymbol{n}(\boldsymbol{x})) \coloneqq \begin{pmatrix} h(\boldsymbol{v} \cdot \boldsymbol{n}) \\ h\boldsymbol{v}(\boldsymbol{v} \cdot \boldsymbol{n}) \quad + \quad h_0\, h^{(2)}\, \boldsymbol{n} \end{pmatrix}$$

of the zero Froude number shallow water equations. In this case, the average is taken over one time step $[t^n, t^{n+1}]$, with $t^{n+1} \coloneqq t^n + \delta t$, and over the interface I between two cells, i.e.

$$\mathbf{F}_I(\mathbf{u}_I, \boldsymbol{n}_I) \coloneqq \begin{pmatrix} h(\boldsymbol{v} \cdot \boldsymbol{n}) \\ h\boldsymbol{v}(\boldsymbol{v} \cdot \boldsymbol{n}) \quad + \quad h_0\, h^{(2)}\, \boldsymbol{n} \end{pmatrix}_I \approx \frac{1}{\delta t\ |I|} \int_{t^n}^{t^{n+1}} \int_I \mathbf{f}(\mathbf{u}, \boldsymbol{n})\ d\sigma\, dt \quad . \tag{3.2}$$

We will refer to such an $\mathbf{F}_I$ by using the term *numerical flux*. Addressing the difficulties mentioned above, for the construction of a numerical flux we define the following rules:

- $\mathbf{F}_I$ is constructed using the fluxes of a standard finite volume scheme for hyperbolic systems;
- the interface velocities used in the numerical flux satisfy a discrete version of the divergence constraint (2.25);
- for smooth solutions, the average of the exact flux is approximated by $\mathbf{F}_I$ up to errors of order $\mathcal{O}(\delta t^2)$; and
- after each time step the divergence constraint is also satisfied by the new cell velocities.

To achieve second order accuracy in time for the numerical fluxes, the integral over $[t^n, t^{n+1}]$ in (3.2) can be replaced by a suitable quadrature rule. Using the midpoint rule, the integral is approximated by δt times the exact flux evaluated at time $t^{n+1/2} \coloneqq t^n + \delta t/2$. Hence, the numerical scheme is motivated by integrating

the zero Froude number shallow water system (2.23) in time, and by approximating the time integral over the flux function with the mid point rule. This leads to the semi-discrete equations

$$h(\boldsymbol{x}, t^{n+1}) = h(\boldsymbol{x}, t^n) - \delta t \left[\nabla \cdot (h\boldsymbol{v})(\boldsymbol{x}, t^{n+1/2})\right] + \mathcal{O}\big(\delta t^3\big) \tag{3.3}$$

and

$$\begin{aligned} (h\boldsymbol{v})(\boldsymbol{x}, t^{n+1}) \;&=\; (h\boldsymbol{v})(\boldsymbol{x}, t^n) - \delta t \left[\nabla \cdot (h\boldsymbol{v} \circ \boldsymbol{v})(\boldsymbol{x}, t^{n+1/2}) + \right. \\ &\qquad \left. (h_0 \nabla h^{(2)})(\boldsymbol{x}, t^{n+1/2})\right] + \mathcal{O}(\delta t^3) \quad . \end{aligned} \tag{3.4}$$

for $\delta t \to 0$.[2] The accuracy requirements for the numerical fluxes are satisfied, if we compute second order accurate approximations of the values in the brackets after half a time step.

Let us assume that appropriate approximations of the fluxes for the auxiliary system (2.26) have been computed with initial height and velocity fields at time t^n, which are constant and divergence free, respectively. Using Taylor series expansion of momentum and velocity about $t^{n+1/2}$, leads to

$$\begin{aligned} (h\boldsymbol{v})(\boldsymbol{x}, t^{n+1/2}) \;&=\; (h\boldsymbol{v})^*(\boldsymbol{x}, t^{n+1/2}) - \frac{\delta t}{2} (h_0 \nabla h^{(2)})(\boldsymbol{x}, t^{n+1/4}) + \mathcal{O}\big(\delta t^3\big) \\ \boldsymbol{v}(\boldsymbol{x}, t^{n+1/2}) \;&=\; \boldsymbol{v}^*(\boldsymbol{x}, t^{n+1/2}) - \frac{\delta t}{2} \nabla h^{(2)}(\boldsymbol{x}, t^{n+1/4}) + \mathcal{O}\big(\delta t^3\big) \quad , \end{aligned} \tag{3.5}$$

(cf. Appendix A.2). The variables with stars denote those of the auxiliary system. Note that the second term on the right hand side of both equations could have also been approximated at another time in the interval $[t^n, t^{n+1/2}]$ to achieve second order accuracy. This provides some flexibility in the interpretation of the associated numerical variables, which are introduced in the next part.

In order to ensure that the interface velocities in the resulting numerical scheme fulfill a discrete analogue of the divergence constraint (2.25), we impose this condition at time $t^{n+1/2}$ and insert our approximation of the momentum $(3.5)_1$:

$$\frac{\delta t}{2} \nabla \cdot (h_0 \nabla h^{(2)})(\boldsymbol{x}, t^{n+1/4}) = \nabla \cdot (h\boldsymbol{v})^*(\boldsymbol{x}, t^{n+1/2}) + \frac{dh_0}{dt}(t^{n+1/2}) + \mathcal{O}\big(\delta t^3\big) \quad . \tag{3.6}$$

[2] In the remaining discussion of this part, the investigated limit behavior is always $\delta t \to 0$.

This is a Poisson-type equation for $h^{(2)}$ and, by applying (3.5), the solution of this problem can be used to compute both, the right hand side of (3.3) and the first term in the brackets of (3.4).

The second term in (3.4) is computed by satisfying a discrete version of the divergence constraint at the new time level as well. Let

$$(h\boldsymbol{v})^{**}(\boldsymbol{x}) \coloneqq (h\boldsymbol{v})(\boldsymbol{x}, t^n) - \delta t \left[\nabla \cdot (h\boldsymbol{v} \circ \boldsymbol{v})(\boldsymbol{x}, t^{n+1/2})\right] \tag{3.7}$$

be an intermediate momentum update. Then, the momentum at time t^{n+1} can be expressed as

$$(h\boldsymbol{v})(\boldsymbol{x}, t^{n+1}) = (h\boldsymbol{v})^{**}(\boldsymbol{x}) - \delta t \, (h_0 \nabla h^{(2)})(\boldsymbol{x}, t^{n+1/2}) + \mathcal{O}\big(\delta t^3\big) \quad . \tag{3.8}$$

The divergence constraint is imposed once more at a half time step, but this time by interpolating the divergence of the momentum with the values at the full time level. This leads to

$$\frac{1}{2}\left[\nabla \cdot (h\boldsymbol{v})(\boldsymbol{x}, t^{n+1}) + \nabla \cdot (h\boldsymbol{v})(\boldsymbol{x}, t^n)\right] = -\frac{dh_0}{dt}(t^{n+1/2}) + \mathcal{O}\big(\delta t^2\big) \tag{3.9}$$

and, with the combination of (3.8) and (3.9), a second Poisson-type problem for $h^{(2)}$ is obtained:

$$\begin{aligned} \delta t \, \nabla \cdot (h_0 \nabla h^{(2)})(\boldsymbol{x}, t^{n+1/2}) \;\; = \;\; & \nabla \cdot (h\boldsymbol{v})^{**}(\boldsymbol{x}) + \nabla \cdot (h\boldsymbol{v})(\boldsymbol{x}, t^n) \; + \\ & 2 \, \frac{dh_0}{dt}(t^{n+1/2}) + \mathcal{O}\big(\delta t^2\big) \quad . \end{aligned} \tag{3.10}$$

Although $h^{(2)}$ could be interpreted as being calculated at a half time step in the first elliptic problem (3.6), it has to be computed twice to obtain a zero divergence velocity field at the new time level.

Hence, three problems have to be solved in the numerical scheme to obtain a solution with the desired properties mentioned on page 25. First, the auxiliary system is solved with a standard second order method for hyperbolic conservation laws. In a following step, the fluxes of this system are corrected by solving the two elliptic equations (3.6) and (3.10).

3.1.2 Calculation of the numerical fluxes

To obtain a finite volume scheme in conservation form, the equations (3.3) and (3.4) are integrated over a volume V of the given discretization. Then, by omitting the higher order terms, the numerical fluxes $\mathbf{F}_{\mathcal{I}}$ are given by

$$\mathbf{F}_I = \mathbf{F}_I^* - \frac{\delta t}{2} \begin{pmatrix} h_0^{n+1/4} \nabla h^{(2)} \cdot \boldsymbol{n} \\ (h\boldsymbol{v})^* \nabla h^{(2)} \cdot \boldsymbol{n} + h_0^{n+1/4} \nabla h^{(2)} \boldsymbol{v}^* \cdot \boldsymbol{n} \end{pmatrix}_I + h_0^{n+1/2} \begin{pmatrix} 0 \\ h^{(2)} \boldsymbol{n} \end{pmatrix}_I . \quad (3.11)$$

In this formulation, $\mathbf{F}_I^*$ is the numerical flux of the auxiliary system

$$\mathbf{F}_I^*(\mathbf{u}_I^*, \boldsymbol{n}_I) \coloneqq \begin{pmatrix} h(\boldsymbol{v} \cdot \boldsymbol{n}) & & \\ h\boldsymbol{v}(\boldsymbol{v} \cdot \boldsymbol{n}) & + & \frac{1}{2} h^2 \boldsymbol{n} \end{pmatrix}_I^*$$

across the interface I. The interface values of momentum and velocity in (3.11) have been replaced by

$$\begin{aligned} (h\boldsymbol{v})_I &= (h\boldsymbol{v})_I^* - \frac{\delta t}{2} h_0^{n+1/4} (\nabla h^{(2)})_I \\ \boldsymbol{v}_I &= \boldsymbol{v}_I^* - \frac{\delta t}{2} (\nabla h^{(2)})_I \quad , \end{aligned} \quad (3.12)$$

which represent approximations to the integral over $I \times [t^n, t^{n+1}]$ of momentum and velocity. They are the discretizations of the semi-discrete equations (3.5). Note that $h^{(2)}$ has actually two different meanings in (3.11): In the first brace it is the solution of the first Poisson-type problem, while in the second brace it is the solution of the second Poisson-type problem. Because we know that $(h^* \nabla h^*) = \mathcal{O}(\delta t^2)$ (cf. Appendix A.1), the approximation of the flux function (3.11) is accurate up to terms of order $\mathcal{O}(\delta t^2)$.

The computation of the numerical fluxes for the auxiliary system (2.26) is done using an explicit high resolution upwind method for hyperbolic conservation laws [van Leer, 1979]. In contrast to Schneider et al. [1999], our implementation is based on a semi-discrete method with Runge-Kutta time stepping [Osher, 1985]. This approach is often referred to as the *method of lines*. The stability of the numerical solution of the auxiliary system strongly depends on a CFL time step restriction [Courant et al., 1928]. As mentioned in Section 2.4, the eigenvalues (characteristic speeds) of this system do not depend on the Froude number. Thus, they are of

order $\mathcal{O}(1)$ as $\mathsf{Fr} \to 0$, leading to $\delta t = \mathcal{O}(\delta x)$ on a regular discretization with grid spacing δx.

After the computation of the $\mathbf{F}^*_{\mathcal{I}}$, the values of $(\nabla h^{(2)})_I$ can be derived with a discrete version of the first Poisson-type equation (3.6), which has been obtained in the previous section. The gradient is discretized at the interfaces with a linear rule based on the yet unknown cell averages $h^{(2)}_{\mathcal{V}}$:

$$\nabla h^{(2)}|_I \coloneqq G^{\mathcal{V}}_I(h^{(2)}_{\mathcal{V}}) = G^{\mathcal{V}}_{\mathcal{I}}(h^{(2)}_{\mathcal{V}})|_I \quad .$$

The operator $G^{\mathcal{V}}_{\mathcal{I}}(h^{(2)}_{\mathcal{V}})$ maps cell-centered values of the height to interface values of its gradient vector field. The discrete Poisson-type problem is then obtained by the integration of (3.6) over a volume $V \in \mathcal{V}$. Using the divergence theorem, it can be written as

$$\begin{aligned} \frac{\delta t}{2} \sum_{I \in \mathcal{I}_{\partial V}} |I|\, h_0^{n+1/4}\, G^{\mathcal{V}}_I(h^{(2)}_{\mathcal{V}}) \cdot \boldsymbol{n}_I = \\ \sum_{I \in \mathcal{I}_{\partial V}} |I|\, (h\boldsymbol{v})^*_I \cdot \boldsymbol{n}_I + |V|\, \frac{dh_0}{dt}(t^{n+1/2}) \quad \forall\, V \in \mathcal{V} \; . \end{aligned} \tag{3.13}$$

Furthermore, a discrete divergence can be defined by

$$D^{\mathcal{I}}_{\mathcal{V}}(\cdot) : \; D^{\mathcal{I}}_{\mathcal{V}}(\boldsymbol{a}_{\mathcal{I}})|_V = D^{\mathcal{I}}_V(\boldsymbol{a}_{\mathcal{I}}) \coloneqq \frac{1}{|V|} \sum_{I \in \mathcal{I}_{\partial V}} |I|\, \boldsymbol{a}_I \cdot \boldsymbol{n}_I \quad \forall\, V \in \mathcal{V} \quad , \tag{3.14}$$

which is a linear mapping from vector fields of interface averages to scalar cell averages. With this definition, the linear system of equations (3.13) can be written as

$$\frac{\delta t}{2}\, D^{\mathcal{I}}_{\mathcal{V}}\Big(h_0^{n+1/4}\, G^{\mathcal{V}}_{\mathcal{I}}(h^{(2)}_{\mathcal{V}})\Big) = D^{\mathcal{I}}_{\mathcal{V}}((h\boldsymbol{v})^*_{\mathcal{I}}) + \frac{dh_0}{dt}(t^{n+1/2}) \quad . \tag{3.15}$$

In the special case of the shallow water equations h_0 can be taken out of the divergence operator, because it only depends on time. The discrete gradient is defined in such a way that the Laplacian $D^{\mathcal{I}}_{\mathcal{V}} G^{\mathcal{V}}_{\mathcal{I}}$ has compact stencil and that standard iterative methods can be applied to solve (3.15). In the method by Schneider et al. [1999], $G^{\mathcal{V}}_{\mathcal{I}}$ and $D^{\mathcal{I}}_{\mathcal{V}}$ are defined to yield the standard five point finite differences

Laplacian on a two dimensional Cartesian grid with constant grid spacing in both coordinate directions (cf. Appendix B.1 and Figure 3.2). To obtain a well posed problem, suitable boundary conditions have to be specified for (3.15). These are discussed in the next part of this section.

With the solution of (3.15), the convective parts $(h\boldsymbol{v} \cdot \boldsymbol{n})_I$ and $(h\boldsymbol{v}\boldsymbol{v} \cdot \boldsymbol{n})_I$ of the numerical fluxes can be computed. This first correction is closely related to a MAC projection [Harlow and Welch, 1965]. In contrast to the Euler equations, the height $h = h_0$ does not have to be updated in each cell, because it is constant in space and uniquely defined by the boundary conditions through (2.24). To obtain the final flux of the momentum equation, we still have to consider the contribution of $h^{(2)}$, i.e. the last term of (3.11). In analogy to (3.7), intermediate cell averages of the momentum are computed by

$$(h\boldsymbol{v})_V^{**} \coloneqq (h\boldsymbol{v})_V^n - \frac{\delta t}{|V|} \sum_{I \in \mathcal{I}_{\partial V}} |I| F_{h\boldsymbol{v},I}^{**} \tag{3.16}$$

with the numerical flux

$$F_{h\boldsymbol{v},I}^{**} \coloneqq F_{h\boldsymbol{v},I}^{*} - \frac{\delta t}{2} \left((h\boldsymbol{v})_I^* \, G_I^{\mathcal{V}}(h_{\mathcal{V}}^{(2)}) \cdot \boldsymbol{n}_I + h_0^{n+1/4} \, G_I^{\mathcal{V}}(h_{\mathcal{V}}^{(2)}) \, \boldsymbol{v}_I^* \cdot \boldsymbol{n}_I \right) \ . \tag{3.17}$$

Note that, in general, the full velocity vector cannot be obtained from the numerical fluxes of the auxiliary system. Thus, the interface values of the velocity in (3.17) are interpolated on the basis of the cell averages

$$\boldsymbol{v}_{\mathcal{I}}^* \coloneqq L_{\mathcal{I}}^{\mathcal{V}}(\boldsymbol{v}_{\mathcal{V}}^*) \quad ,$$

in which $L_{\mathcal{I}}^{\mathcal{V}}$ is a linear operator mapping cell centered vector fields to interface values. Using (3.16), the momentum at the new time step is given by

$$(h\boldsymbol{v})_V^{n+1} = (h\boldsymbol{v})_V^{**} - \frac{\delta t}{|V|} \sum_{I \in \mathcal{I}_{\partial V}} |I| \, h_0^{n+1/2} \, h_I^{(2)} \, \boldsymbol{n}_I \quad . \tag{3.18}$$

An efficient way to compute the interface values $h_{\mathcal{I}}^{(2)}$ is to calculate $h^{(2)}$ in the grid nodes and then to use a suitable quadrature rule. Thus, for the numerical solution

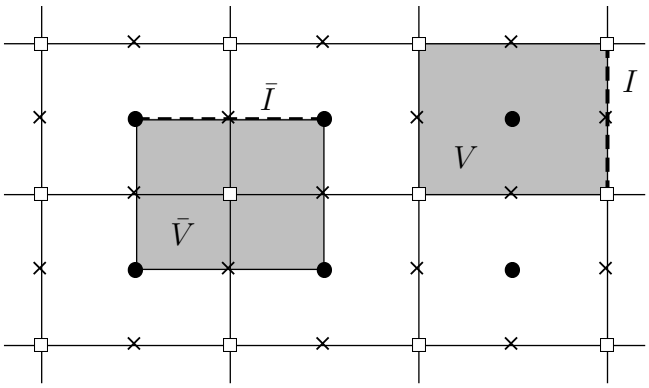

Figure 3.1: Control volume V and interface I of the primary discretization and those ($\bar{V}$ and $\bar{I}$) of the dual discretization. Cell centers are denoted by circles, nodes by squares and midpoints of the interfaces by crosses.

of (3.10) we introduce a *dual discretization* of the computational domain Ω. We define $\bar{\mathcal{V}}$ to be the set of control volumes $\bar{V}$ centered about nodes of the original grid. Let $\bar{I}$ refer to interfaces between cells of $\bar{\mathcal{V}}$, and $\bar{\mathcal{I}}$ be the set of all such $\bar{I}$ (see Figure 3.1). Using these notations, the quadrature rule for the calculation of an $h_I^{(2)}$ can be expressed by the linear operator $L_{\mathcal{I}}^{\bar{\mathcal{V}}}$ with

$$h_{\mathcal{I}}^{(2)} \coloneqq L_{\mathcal{I}}^{\bar{\mathcal{V}}}(h_{\bar{\mathcal{V}}}^{(2)}) \quad . \tag{3.19}$$

The integration of (3.10) over $\bar{V} \in \bar{\mathcal{V}}$, in conjuction with the divergence theorem, yields

$$\begin{aligned} \frac{\delta t}{|\bar{V}|} \int_{\partial \bar{V}} (h_0 \, \nabla h^{(2)}) \cdot \boldsymbol{n} \; d\sigma \;\; &= \;\; \frac{1}{|\bar{V}|} \int_{\partial \bar{V}} (h\boldsymbol{v})^{**} \cdot \boldsymbol{n} \; d\sigma + \\ &\quad \frac{1}{|\bar{V}|} \int_{\partial \bar{V}} (h\boldsymbol{v})^{n} \cdot \boldsymbol{n} \; d\sigma + 2 \, \frac{dh_0}{dt}(t^{n+1/2}) \end{aligned} \tag{3.20}$$

up to second order accuracy. This time, the integrals of $(h\boldsymbol{v})^n$ and $(h\boldsymbol{v})^{**}$ over the interfaces of the dual discretization have to be approximated. These integrals have to be computed from the cell averages of the primary discretization. Once again, a linear quadrature rule

$$(h\boldsymbol{v})_{\bar{\mathcal{I}}}^{n} \coloneqq L_{\bar{\mathcal{I}}}^{\mathcal{V}}((h\boldsymbol{v})_{\mathcal{V}}^{n}) \tag{3.21}$$

is used, which maps cell averages to interface values of the dual discretization. The resulting discrete divergence is then defined by

$$D^{\mathcal{V}}_{\bar{\mathcal{V}}}(\cdot):\ D^{\mathcal{V}}_{\bar{\mathcal{V}}}(\boldsymbol{a}_{\mathcal{V}})|_{\bar{V}} = D^{\mathcal{V}}_{\bar{V}}(\boldsymbol{a}_{\mathcal{V}}) \coloneqq \frac{1}{|\bar{V}|}\sum_{\bar{I}\in\bar{\mathcal{I}}_{\partial\bar{V}}} |\bar{I}| L^{\mathcal{V}}_{\bar{I}}(\boldsymbol{a}_{\mathcal{V}})\cdot\boldsymbol{n}_{\bar{I}} \quad .$$

Consequently, an approximation of the gradient at the cell centers of the grid is needed. This approximation has to be given in terms of the unknown node values of $h^{(2)}$. Therefor, let us define the discrete gradient

$$G^{\bar{\mathcal{V}}}_{\mathcal{V}}(\cdot):\ G^{\bar{\mathcal{V}}}_{\mathcal{V}}(a_{\bar{\mathcal{V}}})|_{V} = G^{\bar{\mathcal{V}}}_{V}(a_{\bar{\mathcal{V}}}) \coloneqq \sum_{I\in\mathcal{I}_{\partial V}} \frac{|I|}{|V|} L^{\bar{\mathcal{V}}}_{I}(a_{\bar{\mathcal{V}}})\,\boldsymbol{n}_{I} \quad , \tag{3.22}$$

which maps node centered values to vector field averages of the primary discretization. Using these definitions, the discrete version of the second Poisson-type problem (3.10) for the unknowns $h^{(2)}_{\bar{\mathcal{V}}}$ can be written as

$$\delta t\, D^{\mathcal{V}}_{\bar{\mathcal{V}}}\Big(h_0^{n+1/2}\, G^{\bar{\mathcal{V}}}_{\mathcal{V}}(h^{(2)}_{\bar{\mathcal{V}}})\Big) = D^{\mathcal{V}}_{\bar{\mathcal{V}}}((h\boldsymbol{v})^{**}_{\mathcal{V}}) + D^{\mathcal{V}}_{\bar{\mathcal{V}}}((h\boldsymbol{v})^{n}_{\mathcal{V}}) + 2\frac{dh_0}{dt}(t^{n+1/2}) \quad . \tag{3.23}$$

Also in this case, the linear operators $L^{\bar{\mathcal{V}}}_{\mathcal{I}}$ and $L^{\mathcal{V}}_{\bar{\mathcal{I}}}$ are defined in order to obtain a discrete Laplacian $D^{\mathcal{V}}_{\bar{\mathcal{V}}}G^{\bar{\mathcal{V}}}_{\mathcal{V}}$ with compact stencil such that the linear system (3.23) can be solved by standard iterative methods (cf. Appendix B.2 and Figure 3.2). A discussion of the boundary conditions for this problem will be given in the next section.

Using (3.18), the second flux correction is finally given by

$$(h\boldsymbol{v})^{n+1}_{V} = (h\boldsymbol{v})^{**}_{V} - \delta t\, h_0^{n+1/2}\, G^{\bar{\mathcal{V}}}_{V}(h^{(2)}_{\bar{\mathcal{V}}}) \quad . \tag{3.24}$$

For flows without change in the background height h_0, the last term of equation (3.23) vanishes. In this case, an initially divergence free velocity field has also zero divergence at the new time step. This is verified by the following lemma.

Lemma 3.1 *Let us consider a velocity field at time t^n, which has zero divergence in the sense that $D^{\mathcal{V}}_{\bar{\mathcal{V}}}(\boldsymbol{v}^n_{\mathcal{V}}) = 0$. Assuming a constant background height (i.e. $\partial_t h_0 \equiv 0$), the velocity field at the new time step satisfies $D^{\mathcal{V}}_{\bar{\mathcal{V}}}(\boldsymbol{v}^{n+1}_{\mathcal{V}}) = 0$ as well.*

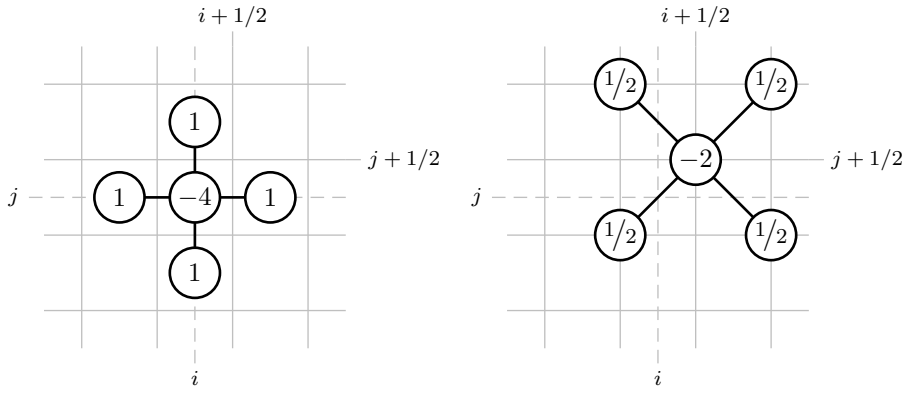

Figure 3.2: Stencils of the original discrete Laplacians for the case $\delta x = \delta y$. First projection (left) and second projection (right).

Proof. *With the assumption that* $\partial_t h_0 \equiv 0$*, equation* (3.23) *becomes*

$$D_{\bar{\mathcal{V}}}^{\mathcal{V}}((h\boldsymbol{v})_{\mathcal{V}}^{**}) + D_{\bar{\mathcal{V}}}^{\mathcal{V}}((h\boldsymbol{v})_{\mathcal{V}}^{n}) - \delta t\, D_{\bar{\mathcal{V}}}^{\mathcal{V}}\Big(h_0^{n+1/2}\, G_{\mathcal{V}}^{\bar{\mathcal{V}}}(h_{\bar{\mathcal{V}}}^{(2)})\Big) = 0 \quad . \tag{3.25}$$

Using the uniformity of $h^n = h_0(t^n)$ *in space, the height can be taken out of the divergence in the second term of* (3.25)*. Because the divergence of the remaining velocity field is zero at time* t^n*, the whole term vanishes and we obtain with* (3.24)

$$0 = D_{\bar{\mathcal{V}}}^{\mathcal{V}}\Big((h\boldsymbol{v})_{\mathcal{V}}^{**} - \delta t\, h_0^{n+1/2}\, G_{\mathcal{V}}^{\bar{\mathcal{V}}}(h_{\bar{\mathcal{V}}}^{(2)})\Big) = D_{\bar{\mathcal{V}}}^{\mathcal{V}}\big((h\boldsymbol{v})_{\mathcal{V}}^{n+1}\big) = h_0^{n+1} D_{\bar{\mathcal{V}}}^{\mathcal{V}}\big(\boldsymbol{v}_{\mathcal{V}}^{n+1}\big) \quad . \qquad \square$$

3.1.3 Initial and boundary conditions

The review of the original scheme is completed by a discussion of the initial and boundary conditions. The former are essential for the solution of the auxiliary system, while the latter are also needed for the solution of the two elliptic problems (3.15) and (3.23).

Initial conditions

The asymptotic analysis of the zero Froude number equations reveals that the background height h_0 is uniform in space. This condition also has to hold for the initial

conditions. The divergence constraint (2.25) implies an analogous discrete condition for the initial velocity field

$$\boldsymbol{v}_V^0 \coloneqq \frac{(h\boldsymbol{v})_V^0}{h_0(0)} \quad ,$$

which shall be given by

$$D_{\bar{V}}^{\mathcal{V}}(\boldsymbol{v}_{\mathcal{V}}^0) = -\frac{1}{h_0}\frac{dh_0}{dt}\bigg|_{t=0} \quad \forall\, \bar{V} \in \bar{\mathcal{V}} \quad . \tag{3.26}$$

Therefore, for a problem with no change in h_0 we have to ensure that the initial velocity field is divergence free in the sense of the discrete operator defined in (3.22).

Boundary conditions

Boundary conditions for finite volume schemes are constraints for the (numerical) fluxes at the boundary of the domain. The numerical fluxes $\mathbf{F}_{\mathcal{I}}$ of the method presented above are computed by the fluxes of the auxiliary system and by two implicit corrections, which are given by the Poisson-type equations (3.15) and (3.23). Thus, we have to formulate suitable boundary conditions for each of these three problems to satisfy the boundary conditions for the whole problem consistently. We restrict our discussion to periodic boundary conditions and rigid non-permeable walls.

Periodic boundary conditions for the whole system can be satisfied by imposing them on all three flux components. For rigid walls on the boundary of the computational domain, the convective part of $\mathbf{F}_{\mathcal{I}}$ has to vanish. Thus, the numerical fluxes of the auxiliary system have to satisfy

$$\mathbf{F}_I^* \coloneqq \begin{pmatrix} 0 \\ \frac{1}{2}\, h^2\, \boldsymbol{n} \end{pmatrix}_I^* \quad \forall\, I \in \mathcal{I}_w \quad ,$$

where $\mathcal{I}_w$ is the collection of all interfaces at walls in the boundary. Let us remark that

$$\nabla h^{(2)}|_I \cdot \boldsymbol{n}_I \coloneqq G_I^{\mathcal{V}}(h_{\mathcal{V}}^{(2)}) \cdot \boldsymbol{n}_I$$

has to be computed on all interfaces belonging to the boundary of the domain. By

imposing the same boundary conditions for the velocity fields, i.e.

$$\boldsymbol{v}_I \cdot \boldsymbol{n}_I = \boldsymbol{v}_I^* \cdot \boldsymbol{n}_I \quad \forall\, I \subset \partial\Omega \quad ,$$

we obtain with $(3.12)_2$

$$\frac{\delta t}{2} (\nabla h^{(2)})_I \cdot \boldsymbol{n}_I = (\boldsymbol{v}_I^* - \boldsymbol{v}_I) \cdot \boldsymbol{n}_I = 0 \quad \forall\, I \subset \partial\Omega \quad .$$

This condition implies an integral constraint for the right hand side of equation (3.15) for the solution $h^{(2)}$ to exist. The constraint is given by

$$\sum_{V \in \mathcal{V}} |V|\, D_V^{\mathcal{I}}(F_{h,I}^*) + |\Omega| \frac{dh_0}{dt} = 0 \quad .$$

Note that this is the discrete counterpart of (2.24) and specifies in which discrete sense this constraint has to be satisfied.

In the second Poisson-type equation the unknown $h^{(2)}$ is defined on control volumes centered around nodes of the given grid. Along the boundary, a part of these control volumes is outside the domain of integration. We can solve this problem for periodic boundary conditions, because on a regular Cartesian grid, each such volume corresponds to a volume on the other side of the computational domain (cf. Figure 3.3). Therefore, all control volumes of the dual discretization are in fact inside the domain Ω.

In the case of rigid wall boundary conditions, the control volumes have to be split by $\partial\Omega$, and only the part inside Ω is used for the computation of the discrete divergence field $D_{\mathcal{V}}^{\mathcal{V}}$. The normal derivatives

$$(h \nabla h^{(2)})_{\bar{I}} \cdot \boldsymbol{n}_{\bar{I}}$$

and the scalar products

$$(h\boldsymbol{v})_{\bar{I}}^{**} \cdot \boldsymbol{n}_{\bar{I}} \quad \text{and} \quad (h\boldsymbol{v})_{\bar{I}}^{n} \cdot \boldsymbol{n}_{\bar{I}}$$

at the new boundaries $\bar{I} \subset \partial\Omega$ are set to zero in (3.20), and the linear operator $L_{\bar{\mathcal{I}}}^{\mathcal{V}}$ is modified to incorporate only cell averages inside the domain.

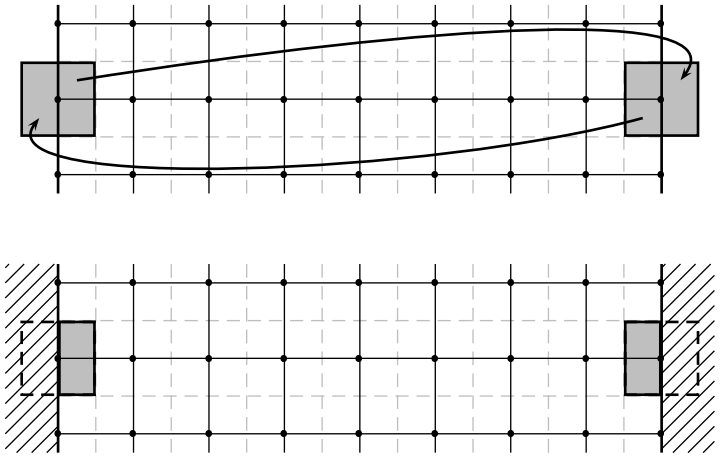

Figure 3.3: Boundary conditions for the dual discretization: In the periodic case (top), control volumes at opposite boundaries coincide. For rigid non-permeable walls (bottom), the dual control volumes are split.

3.2 A new projection method

In the original scheme described above, the discrete gradient of the second projection maps the values of the height perturbation at the four vertices of a cell into one average value corresponding to this cell (cf. equation (B.1)). This discretization produces a local decoupling, and the kernel of this gradient has a dimension greater than one.[3] In the following, a new projection method is introduced that is based on a finite element formulation of the Poisson-type problem (3.10). This discretization was originally introduced by Süli [1991], who proved stability and convergence of the scheme in a mesh-dependent H^1 norm.

Interpreting the cell averages of the original projection method as piecewise constant functions on the primary grid, the main difference is that the discrete velocity space is enriched with piecewise linear functions. The nodal values of the scalar variable $h^{(2)}$ uniquely define a finite element space consisting of piecewise bilinear functions on the primary grid cells, which are continuous over the whole domain.

[3] For example, on a Cartesian grid, in which the dual control volumes are denoted by $\bar{V}_{i+1/2,j+1/2}$, a distribution of a scalar variable q with $q|_{\bar{V}_{i+1/2,j+1/2}} = C_1$ for $i+j$ even and $q|_{\bar{V}_{i+1/2,j+1/2}} = C_2$ otherwise results in a zero gradient field.

This type of discretization for the unknowns has been used before in similar applications [e.g. in ALMGREN ET AL., 1996]. It has the advantage that the continuous gradient of the scalar variable is within the velocity space. However, in contrast to the work mentioned above, we do not use a classical finite element formulation with identical trial and test spaces. The problem is discretized as a *Petrov-Galerkin* finite element method, using a test space with piecewise constant functions on the dual discretization of the computational domain.

For the derivation of the new projection method, a Poisson problem with natural (Neumann) boundary conditions is considered. These boundary conditions are motivated by the analysis of rigid wall boundary conditions in the previous section. Thus, we are interested in the solution of

$$\left\{ \begin{array}{rcll} -\nabla \cdot \nabla p & = & f & \text{in } \Omega \\ \dfrac{\partial p}{\partial \boldsymbol{n}} & = & 0 & \text{on } \partial\Omega \end{array} \right. \quad . \tag{3.27}$$

Given $f \in L^2(\Omega)$ with $\int_\Omega f \, d\boldsymbol{x} = 0$, this problem has a unique solution $p \in H^1(\Omega)/\mathbb{R}$. In the presented numerical scheme the right hand side f is of the type $-\nabla \cdot \boldsymbol{v}$, with a given velocity field $\boldsymbol{v}$. Therefore, f is substituted with this term in the following discussion.

The corresponding weak formulation is given by multiplying (3.27) with a test function ψ and by integrating the new equation over the whole domain Ω. Thus, we have to find p with

$$\int_\Omega \psi \; \nabla \cdot \nabla p \; d\boldsymbol{x} = \int_\Omega \psi \; \nabla \cdot \boldsymbol{v} \; d\boldsymbol{x} \quad \forall \, \psi \quad . \tag{3.28}$$

With the choice of piecewise constant functions for the test space, *Green's formula* cannot be used any longer. Instead, the divergence theorem is applied.

For further analysis, let us define the test space as

$$\mathcal{Q}^h \coloneqq \left\{ q \in L^2(\Omega) \mid \forall \, \bar{V} \in \bar{\mathcal{V}} : q|_{\bar{V}} \in \mathcal{P}_0(\bar{V}) \right\} \quad ,$$

in which

$$\mathcal{P}_k(U) \coloneqq \left\{ p \in C^\infty(U) \;\middle|\; p(x,y) = \sum_{\substack{i+j\le k \\ i,j\ge 0}} c_{ij}\, x^i\, y^j \right\} \tag{3.29}$$

is the space of polynomial functions of degree less than or equal to k on $U \subset \mathbb{R}^2$. A basis of $\mathcal{Q}^h$ is given by $\bigcup_{\bar{V}\in\bar{\mathcal{V}}}\{\chi_{\bar{V}}\}$, whereby $\chi_{\bar{V}}$ is the characteristic function on the cell $\bar{V}$.[4] Thus, if $\psi \in \mathcal{Q}^h$, ψ has a volumewise representation

$$\psi(x,y) = \sum_{\bar{V}\in\bar{\mathcal{V}}} \psi_{\bar{V}}\, \chi_{\bar{V}}(x,y) \quad , \tag{3.30}$$

and the discrete problem, corresponding to (3.28), is to find $p \in \mathcal{H}^h \subset H^1(\Omega)$, such that

$$\sum_{\bar{V}\in\bar{\mathcal{V}}} \psi_{\bar{V}} \left(\int_{\bar{V}} \nabla\cdot\nabla p \; d\boldsymbol{x} - \int_{\bar{V}} \nabla\cdot\boldsymbol{v} \; d\boldsymbol{x} \right) = 0 \quad \forall\, \psi \in \mathcal{Q}^h$$

with ψ as in (3.30). Furthermore, by applying the divergence theorem, the problem can be rewritten as

$$\sum_{\bar{V}\in\bar{\mathcal{V}}} \psi_{\bar{V}} \left(\int_{\partial\bar{V}} \nabla p\cdot\boldsymbol{n} \; d\boldsymbol{x} - \int_{\partial\bar{V}} \boldsymbol{v}\cdot\boldsymbol{n} \; d\boldsymbol{x} \right) = 0 \quad \forall\, \psi \in \mathcal{Q}^h \quad . \tag{3.31}$$

This problem is a linear combination of the local problems to find $p \in \mathcal{H}^h$, such that

$$\int_{\partial\bar{V}} \nabla p\cdot\boldsymbol{n} \; d\boldsymbol{x} - \int_{\partial\bar{V}} \boldsymbol{v}\cdot\boldsymbol{n} \; d\boldsymbol{x} = 0 \quad \forall\, \bar{V} \in \bar{\mathcal{V}} \quad , \tag{3.32}$$

and the solution p satisfies (3.31), if and only if it satisfies (3.32).

The finite element spaces for the unknown p and the velocity $\boldsymbol{v}$ still have to be defined. As mentioned above, $\boldsymbol{v}$ is approximated by linear functions on the control volumes $V \in \mathcal{V}$, i.e. it is in the space

$$\mathcal{U}^h \coloneqq \left\{ \boldsymbol{v} = (u,v) \in (L^2(\Omega))^2 \mid \forall\, V \in \mathcal{V} : u|_V, v|_V \in \mathcal{P}_1(V) \right\} \quad .$$

On a Cartesian grid, a function $\boldsymbol{v} \in \mathcal{U}^h$ can be represented on a cell $V_{i,j}$ by

$$\boldsymbol{v}(x,y)|_{V_{i,j}} = \boldsymbol{v}_{i,j} + (x - x_i)\boldsymbol{v}_{x,i,j} + (y - y_j)\boldsymbol{v}_{y,i,j} \quad ,$$

[4] I.e. $\chi_{\bar{V}}(x,y) = 1$ if $(x,y) \in \bar{V}$, and 0 otherwise.

in which $\boldsymbol{v}_{i,j}$ is the cell average of $\boldsymbol{v}(x,y)$ and $\boldsymbol{v}_{x,i,j}$ and $\boldsymbol{v}_{y,i,j}$ are the partial derivatives of $\boldsymbol{v}$ on $V_{i,j}$. These vector coefficients uniquely define an element of $\mathcal{U}^h$.

Remark 3.1 *An orthogonal decomposition of the space $\mathcal{U}^h$ can be given as follows. For each $\boldsymbol{v} \in \mathcal{U}^h$, let us define its piecewise constant component as*

$$\bar{\boldsymbol{v}} \coloneqq \sum_{V \in \mathcal{V}} \chi_V \bar{\boldsymbol{v}}_V = \sum_{V \in \mathcal{V}} \frac{1}{|V|} \int_V \boldsymbol{v} \, d\boldsymbol{x}$$

and the variation as

$$\tilde{\boldsymbol{v}} \coloneqq \boldsymbol{v} - \bar{\boldsymbol{v}} \quad .$$

For each cell, this implies that $\int_V \tilde{\boldsymbol{v}} \, d\boldsymbol{x} = 0$ and that the two components are orthogonal in $L^2(\Omega)$. ◁

The space of piecewise bilinear functions on $V \in \mathcal{V}$, which are continuous at the interfaces between control volumes, is given by

$$\mathcal{H}^h \coloneqq \left\{ p \in H^1(\Omega) \mid \forall\, V \in \mathcal{V} : p|_V \in \mathcal{P}_2(V), \forall\, I \in \mathcal{I} : p|_I \in \mathcal{P}_1(I) \right\} \quad .$$

Using this definition, an element of $\mathcal{H}^h$ can be written as

$$p(x,y) = \sum_{\bar{V} \in \bar{\mathcal{V}}} p_{\bar{V}} \, \varphi_{\bar{V}}(x,y) \quad ,$$

in which $\varphi_{\bar{\mathcal{V}}}$ are the standard basis functions for $\mathcal{H}^h$. In our framework of a Cartesian grid, these functions are piecewise bilinear on each $V_{i,j}$ and have node values $\varphi_{\bar{V}_{i+1/2,j+1/2}}(x_{k+1/2}, y_{l+1/2}) = \delta_{ik}\delta_{jl}$ (cf. Appendix C.1 and Figure 3.4). By definition of the finite element spaces, $\nabla p \cdot \boldsymbol{n}$ and $\boldsymbol{v} \cdot \boldsymbol{n}$ are piecewise linear along the boundary of $\bar{V}$. Therefore, the line integrals in (3.32) can be calculated analytically to obtain a linear system for the unknown "vector" $(p_{i+1/2,j+1/2})$.

Using a suitable normalization, the integrals in (3.32) define a discrete Laplacian and a divergence of p and $\boldsymbol{v}$ on the dual discretization, respectively. Specifically, let us define on a Cartesian grid for $p_{\bar{\mathcal{V}}} \in \mathcal{H}^h$

$$\mathsf{L}^{\bar{\mathcal{V}}}_{\bar{\mathcal{V}}}(\cdot) : \; \mathsf{L}^{\bar{\mathcal{V}}}_{\bar{V}_{i+1/2,j+1/2}}(p_{\bar{\mathcal{V}}}) \coloneqq \frac{1}{|\bar{V}_{i+1/2,j+1/2}|} \int\limits_{\partial \bar{V}_{i+1/2,j+1/2}} \nabla p_{\bar{\mathcal{V}}} \cdot \boldsymbol{n} \; d\boldsymbol{x} \tag{3.33}$$

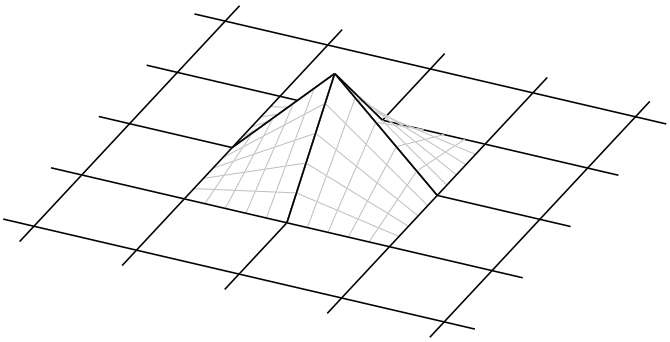

Figure 3.4: Bilinear basis function of the space $\mathcal{H}^h$.

and for $\boldsymbol{v}_\mathcal{V} \in \mathcal{U}^h$

$$\mathsf{D}^{\mathcal{V}}_{\bar{\mathcal{V}}}(\cdot):\ \mathsf{D}^{\mathcal{V}}_{\bar{V}_{i+1/2,j+1/2}}(\boldsymbol{v}_\mathcal{V}) \coloneqq \frac{1}{|\bar{V}_{i+1/2,j+1/2}|} \int\limits_{\partial \bar{V}_{i+1/2,j+1/2}} \boldsymbol{v}_V \cdot \boldsymbol{n}\ d\boldsymbol{x} \quad . \tag{3.34}$$

In these definitions, grid functions are identified with functions defined on the whole domain Ω. The resulting stencil of the Laplacian is given in Figure 3.5. To distinguish them from the original ones, a different font is used for the new operators. Note again that the gradient of $p \in \mathcal{H}^h$ is in the space $\mathcal{U}^h$. In particular, on a control volume $V_{i,j}$ of the primary discretization p can also be represented by

$$p(x,y)|_{V_{i,j}} = p_{i,j} + (x-x_i)p_{x,i,j} + (y-y_j)p_{y,i,j} + (x-x_i)(y-y_j)p_{xy,i,j} \quad , \tag{3.35}$$

in which $p_{i,j}$ is the mean value of p on $V_{i,j}$, and $p_{x,i,j}$, $p_{y,i,j}$ and $p_{xy,i,j}$ are the partial and mixed derivatives of p in (x_i, y_j), respectively. These values can be given in terms of the nodal values of p (cf. Appendix C.2). Using this notation, the gradient of p is given by

$$\nabla p(x,y)|_{V_{i,j}} = \begin{pmatrix} p_{x,i,j} + (y-y_j)p_{xy,i,j} \\ p_{y,i,j} + (x-x_i)p_{xy,i,j} \end{pmatrix} \quad ,$$

and a discrete gradient operator is defined by

$$\mathsf{G}^{\bar{\mathcal{V}}}_{\mathcal{V}}(\cdot):\ \mathsf{G}^{\bar{\mathcal{V}}}_{V_{i,j}}(p_{\bar{\mathcal{V}}}) \coloneqq \nabla p(x,y)|_{V_{i,j}} \quad . \tag{3.36}$$

These discrete operators satisfy $\mathsf{L}^{\bar{\mathcal{V}}}_{\bar{\mathcal{V}}} = \mathsf{D}^{\mathcal{V}}_{\bar{\mathcal{V}}}(\mathsf{G}^{\bar{\mathcal{V}}}_{\mathcal{V}})$ as well, which becomes evident through

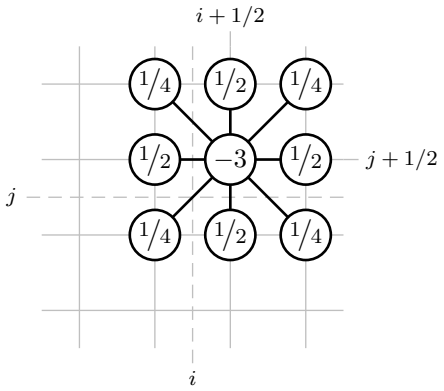

Figure 3.5: Stencil of the new discrete Laplacian for the case $\delta x = \delta y$.

a comparison of (3.33) and (3.34). For a uniform Cartesian grid in two space dimensions, the particular formulas for $\mathsf{L}_{\mathcal{V}}^{\bar{\mathcal{V}}}$, $\mathsf{D}_{\bar{\mathcal{V}}}^{\mathcal{V}}$ and $\mathsf{G}_{\mathcal{V}}^{\bar{\mathcal{V}}}$ are given in Appendix C.2.

Remark 3.2 *Using piecewise constant functions for the test space, the new projection method could also be interpreted as a finite volume method. This fact is obvious from the problem associated with* (3.32). ◁

3.2.1 Approximate second projection

The original scheme is constructed using variables defined as cell averages of either the primary or the dual discretization. This is in contrast to the new projection outlined above, in which vector functions with piecewise linear variations on the primary discretization are used. An easy way to embed the new projection into the original scheme is given as follows. First, the new projection is applied to piecewise constant functions, which represent the cell averages of the velocity field. The result of this procedure is in $\mathcal{U}^h$ and not necessarily piecewise constant any more. Thus, the vector field has to be projected back onto the space of piecewise constant functions. From Remark 3.1 it follows that this procedure can be characterized as an exact discrete projection onto the enriched velocity space $\mathcal{U}^h$, followed by an orthogonal (L^2) projection onto the subspace of piecewise constant functions. Note that in this case the divergence constraint is no longer exactly satisfied. However, we have an analytic characterization of the "appoximateness" as well as the stability of the final approximation [Almgren et al., 2000].

The combination of the exact and the L^2 projection can also be interpreted as one projection, in which an "incorrect" gradient has been used at the end of the procedure. This type of projection with a Laplacian being only an approximation of $D(G)$ is called an *approximate projection* and was first introduced in Almgren et al. [1996].

In the case of a Cartesian grid, the new exact projection combined with the L^2 projection can be implemented by using the new discretizations for the divergence and the Laplacian, but the gradient of the original method. Thus, (3.23) and (3.24) are modified, and with the solution of the discrete Poisson-type problem for $h^{(2)}_{\bar{\mathcal{V}}}$

$$\delta t\, \mathsf{D}^{\mathcal{V}}_{\bar{\mathcal{V}}}(h_0^{n+1/2}\, \mathsf{G}^{\bar{\mathcal{V}}}_{\mathcal{V}}(h^{(2)}_{\bar{\mathcal{V}}})) = \mathsf{D}^{\mathcal{V}}_{\bar{\mathcal{V}}}((h\boldsymbol{v})^{**}_{\mathcal{V}}) + \mathsf{D}^{\mathcal{V}}_{\bar{\mathcal{V}}}((h\boldsymbol{v})^{n}_{\mathcal{V}}) + 2\frac{dh_0}{dt}(t^{n+1/2}) \quad , \tag{3.37}$$

the momentum is corrected by

$$\begin{aligned} (h\boldsymbol{v})^{n+1}_{V} &= (h\boldsymbol{v})^{**}_{V} - \delta t\, h_0^{n+1/2}\, \overline{\mathsf{G}^{\bar{\mathcal{V}}}_{V}(h^{(2)}_{\bar{\mathcal{V}}})} \\ &= (h\boldsymbol{v})^{**}_{V} - \delta t\, h_0^{n+1/2}\, G^{\bar{\mathcal{V}}}_{V}(h^{(2)}_{\bar{\mathcal{V}}}) \quad . \end{aligned} \tag{3.38}$$

Here, the bar denotes the L^2 projection onto the space of piecewise constant functions. The equality $\overline{\mathsf{G}^{\bar{\mathcal{V}}}_{\mathcal{V}}} = G^{\bar{\mathcal{V}}}_{\mathcal{V}}$ is true, because the difference between the gradients is given by the terms involving the mixed derivatives of the scalar variable. These are eliminated by the L^2 projection.

By the following lemma an upper bound for the "appoximateness" of the divergence at the new time level can be given.

Lemma 3.2 *Let h_0 be uniform in space with no change in time. Furthermore, the velocity field at time t^n shall satisfy $\mathsf{D}^{\mathcal{V}}_{\bar{\mathcal{V}}}(\boldsymbol{v}^n_{\mathcal{V}}) = 0$. Then, the divergence of the momentum at the new time level is controlled up to terms of order $\mathcal{O}(\delta t\,(\delta x^2 + \delta y^2))$.*

Proof. *The momentum update* (3.38) *can be written as*

$$(h\boldsymbol{v})^{n+1}_{V} = (h\boldsymbol{v})^{**}_{V} - \delta t\, h_0^{n+1/2} \left[\mathsf{G}^{\bar{\mathcal{V}}}_{V}(h^{(2)}_{\bar{\mathcal{V}}}) - \left(\mathsf{G}^{\bar{\mathcal{V}}}_{V} - G^{\bar{\mathcal{V}}}_{V}\right)(h^{(2)}_{\bar{\mathcal{V}}})\right] \quad .$$

On a Cartesian grid, the second term in the brackets is given by

$$\left(\mathsf{G}_{i,j}^{\bar{\mathcal{V}}} - G_{i,j}^{\bar{\mathcal{V}}}\right)(h_{\bar{\mathcal{V}}}^{(2)}) = \begin{pmatrix} y - y_j \\ x - x_i \end{pmatrix} h_{xy,i,j}^{(2)} \quad ,$$

and (using Lemma 3.1) the divergence of the momentum at the new time level is

$$\begin{aligned}
\mathsf{D}_{\bar{V}_{i+1/2,j+1/2}}^{\mathcal{V}}((h\boldsymbol{v})_{\mathcal{V}}^{n+1}) &= \underbrace{\mathsf{D}_{\bar{V}_{i+1/2,j+1/2}}^{\mathcal{V}}\Big((h\boldsymbol{v})_{\mathcal{V}}^{**} - \delta t\, h_0^{n+1/2}\, \mathsf{G}_{\mathcal{V}}^{\bar{\mathcal{V}}}(h_{\bar{\mathcal{V}}}^{(2)})\Big)}_{=0} + \\
&\quad \delta t\, h_0^{n+1/2}\, \mathsf{D}_{\bar{V}_{i+1/2,j+1/2}}^{\mathcal{V}}\Big(\Big(\mathsf{G}_{\mathcal{V}}^{\bar{\mathcal{V}}} - G_{\mathcal{V}}^{\bar{\mathcal{V}}}\Big)(h_{\bar{\mathcal{V}}}^{(2)})\Big) \\
&= \delta t\, h_0^{n+1/2} \left(\frac{\delta x^2 + \delta y^2}{8\,\delta x\,\delta y}\right) \cdot \\
&\quad \Big(h_{xy,i+1,j}^{(2)} - h_{xy,i,j}^{(2)} - h_{xy,i+1,j+1}^{(2)} + h_{xy,i,j+1}^{(2)}\Big) \\
&= -\frac{\delta t\,\delta \boldsymbol{x}^2}{8} h_0^{n+1/2}\, h_{xxyy,i+1/2,j+1/2}^{(2)} + \mathcal{O}\big(\delta t\,\delta \boldsymbol{x}^2\big) \quad ,
\end{aligned}$$

where $\delta \boldsymbol{x}^2$ is an abbreviation for $(\delta x^2 + \delta y^2)$. □

Remark 3.3 *Since $(h\boldsymbol{v})_V^{**}$ and $(h\boldsymbol{v})_V^n$ have no variation for $V \in \mathcal{V}$, in the case of the approximate projection described above, the new divergence reduces to the original version described in Section 3.1.* ◁

3.2.2 Exact second projection

To derive an exact projection method the piecewise linear functions for the momentum have to be used throughout the whole scheme. In the semi-discrete implementation for the solution of the auxiliary system Heun's method is applied for the integration in time, i.e.

$$\begin{aligned}
\mathbf{U}^{*,\mathrm{int}} &= \mathbf{U}^n + \delta t\, f(\mathbf{U}^n) \\
\mathbf{U}^{*} &= \mathbf{U}^n + \frac{\delta t}{2}\left(f(\mathbf{U}^n) + f(\mathbf{U}^{*,\mathrm{int}})\right) \quad ,
\end{aligned}$$

where $\delta t \coloneqq t^{n+1} - t^n$. This approach leads to second-order accuracy in time. To obtain second-order accuracy in space as well, the cell average values in $\mathbf{U}^n$ and $\mathbf{U}^{*,\mathrm{int}}$ have to be reconstructed as piecewise linear functions on each cell. The numerical fluxes are then evaluated with the reconstructed values on the two sides of any particular interface.

Therefore, the following modifications are applied to the original scheme to obtain the new exact projection method. A new reconstruction step is introduced after the first projection, which reconstructs piecewise linear functions from cell averages of the intermediate momentum components $(hu)_V^{**}$ and $(hv)_V^{**}$. The new projection method is then applied to this vector field to obtain a final momentum distribution. In the new time step, the gradients of the momentum components $(h\boldsymbol{v})_{\boldsymbol{x},V}^{n+1}$ are used for the calculation of the numerical fluxes of the auxiliary system. The variation is not only used for $\mathbf{U}^n$, but for $\mathbf{U}^{*,\mathrm{int}}$ as well. This can be done, because a Taylor series expansion yields

$$\mathbf{U}_{\boldsymbol{x},V}^{*,\mathrm{int}} = \mathbf{U}_{\boldsymbol{x},V}^{n} + \mathcal{O}(\delta t) \quad .$$

In this scheme $\mathbf{U}_{\boldsymbol{x},V}$ is always multiplied by δx to yield the numerical fluxes of the auxiliary system. Therefore, the second order accuracy in space and time is retained. Also for this projection method Lemma 3.1 is valid. Note that the reconstruction procedure is no longer *total variation diminishing (TVD)*, regardless of the limiter function being used in the reconstruction of the previous time step.[5]

Besides the introduction of the new Poisson-type problem (3.37), the numerical scheme is also modified in the final momentum update (3.24), in which the new discrete gradient $\mathsf{G}_{\mathcal{V}}^{\bar{\mathcal{V}}}$ has been used. It has to be stressed that this update not only involves the cell mean values, but also the gradient within a cell. Thus, because of the identity $\overline{\mathsf{G}_{\mathcal{V}}^{\bar{\mathcal{V}}}} = G_{\mathcal{V}}^{\bar{\mathcal{V}}}$, this update can be defined by the two equations

$$\begin{aligned} \overline{(h\boldsymbol{v})_V^{n+1}} &= \overline{(h\boldsymbol{v})_V^{**}} - \delta t\, h_0^{n+1/2}\, G_V^{\bar{\mathcal{V}}}(h_{\bar{\mathcal{V}}}^{(2)}) \\ \widetilde{(h\boldsymbol{v})_V^{n+1}} &= \widetilde{(h\boldsymbol{v})_V^{**}} - \delta t\, h_0^{n+1/2}\, \left(\mathsf{G}_V^{\bar{\mathcal{V}}} - G_V^{\bar{\mathcal{V}}}\right)(h_{\bar{\mathcal{V}}}^{(2)}) \quad . \end{aligned} \tag{3.39}$$

In (3.39), the bar once again denotes the L^2 projection onto the space of piecewise

[5] For a definition of TVD methods see LeVeque [2002, p. 109]

constant functions, and the tilde denotes the projection onto the orthogonal complement (the linear portion). The final momentum is then given by

$$(h\boldsymbol{v})_V^{n+1} = \overline{(h\boldsymbol{v})_V^{n+1}} + \widetilde{(h\boldsymbol{v})_V^{n+1}} \quad .$$

3.2.3 Application for the first projection

The finite element formulation presented above for the second projection can be adapted for the first projection. However, the situation is slightly different in this case. The trial spaces for the unknown and the velocity are spanned by piecewise bilinear scalar functions and piecewise linear vector functions on the dual discretization. The test functions are piecewise constant on primary control volumes. In order to correct the convective part of the numerical fluxes, the gradient of the height $h^{(2)}$ has to be integrated over the boundary of a control volume of the primary discretization. This can be done analytically again, because the gradient of $h^{(2)}$ is piecewise linear on control volumes of the dual discretization in this case.

3.3 Additional consistency considerations

The result of the piecewise linear reconstruction strongly depends on the particular limiter function that has been selected. To avoid this arbitrariness, new rules for the reconstruction based on additional consistency considerations are introduced in this section. In particular, the divergence constraint (2.25) and the transport property of the vorticity are employed to further control the gradients of the reconstructed quantities.

Let us analyze the discretization of the new divergence (3.34) on a Cartesian grid (cf. equation (C.1)). The application of this operator on a vector field includes additional degrees of freedom, which enable us to modify the result of the second projection. For example, the divergence is not changed, as long as the partial derivatives $u_{y,i,j}$ and $v_{x,i,j}$ within a cell $V_{i,j}$ fulfill the condition

$$\frac{\delta y}{\delta x} u_{y,i,j} + \frac{\delta x}{\delta y} v_{x,i,j} = C_{i,j} \quad , \tag{3.40}$$

where $C_{i,j}$ is a constant for each cell $V_{i,j}$. The values of $u_{x,i,j}$ and $v_{y,i,j}$ do not influence the discrete divergence at all.

To control these values, the divergence constraint (2.25) could also be imposed within each cell $V_{i,j}$, i.e.

$$\nabla \cdot \boldsymbol{v}(\boldsymbol{x}, t)|_{V_{i,j}} = u_{x,i,j} + v_{y,i,j} = -\frac{1}{h_0(t)} \frac{dh_0}{dt}(t) \quad .$$

Moreover, the evolution equation (1.4) for the vorticity $\omega \coloneqq v_x - u_y$ can be reformulated as

$$\frac{\partial \omega}{\partial t} + \boldsymbol{v} \cdot \nabla \omega = -\omega \, \nabla \cdot \boldsymbol{v} \quad .$$

Thus, in the zero Froude number case, the vorticity satisfies an advection equation with a source term known from the boundary conditions. For the case of no flux across the boundary, the vorticity is just an advected quantity.

To compute the advection of vorticity, we have to extend the numerical scheme for the zero Froude number shallow water equations by an additional equation for ω. Note that this equation is not independent of the continuity and momentum equations, and the vorticity acts as a tracer. The auxiliary system has to be solved with an additional tracer equation and the numerical flux $F^*_{\omega,I}$ has to be corrected by the first projection. Thus, the vorticity in the new time step is computed by

$$\omega_V^{n+1} \coloneqq \omega_V^n - \frac{\delta t}{|V|} \sum_{I \in \mathcal{I}_{\partial V}} |I| F_{\omega,I}$$

with the numerical flux

$$F_{\omega,I} \coloneqq F^*_{\omega,I} - \frac{\delta t}{2} \left(\omega^*_I \, G^{\mathcal{V}}_I(h^{(2)}_{\mathcal{V}}) \cdot \boldsymbol{n}_I \right) \quad .$$

Here, we associate the vorticity with the one we obtain from the auxiliary system, i.e. $\omega_I = \omega^*_I$. Because this variable cannot be obtained from the numerical fluxes in general, we interpolate it on the basis of the cell averages:

$$\omega^*_{\mathcal{I}} \coloneqq L^{\mathcal{V}}_{\mathcal{I}}(\omega^*_{\mathcal{V}}) \quad .$$

Equation (3.40) together with the advected vorticity yield for each cell a linear system of two equations for the partial derivatives $u_{y,i,j}$ and $v_{x,i,j}$. Therefore, these values are uniquely defined. By using the additional restriction implied by the application of the divergence constraint within a control volume as well, the only undefined quantity remains to be

$$u_{x,i,j} - v_{y,i,j} \quad .$$

Clearly, additional (evolution) equations could have also be derived for $u_{x,i,j}$ or $v_{y,i,j}$. However, the vorticity equation is characterized by its simplicity and the evolution of vorticity is straight forward to compute. The divergence constraint naturally arises in the zero Froude number limit of the shallow water equations and is easy to apply as well.

4 Stability of the New Projection

In order to prove stability of our semi-implicit method, the stability of the second projection step is an important prerequisite. This issue will be addressed in the following discussion.

In the second projection, we compute the height perturbation $h^{(2)}$ to correct the intermediate momentum update $(h\boldsymbol{v})^{**}$ in a post-processing step. Thus, $h^{(2)}$ is only an auxiliary variable, and we are interested rather in the momentum at the new time step. The associated Poisson-type problem is derived by imposing the additional requirement that the momentum at the new time step shall satisfy a discrete version of the divergence constraint (2.25). In the context of finite element methods, this leads to the theory of *saddle point problems*, which arise from minimization problems with additional side conditions. Starting with the fundamental work of Babuška [1971] and Brezzi [1974], this theory provides conditions for existence and uniqueness of solutions and for stable discretizations of such problems.

After having introduced the fundamental functional framework, we briefly present and review basic formulations for the discretization of saddle point problems. This discussion serves as a basis for providing an overview of the different approaches as well as establishing the fundamental theoretical results on these methods. Furthermore, the discrete Poisson-type problem (3.37) is reformulated for the new projection method as a generalized saddle point problem, which is the starting point for the subsequent *stability analysis*. Existence and uniqueness are shown for the continuous problem, and preliminary results concerning the stability of the new method are given in the last part of this chapter.

4.1 Approximation of saddle point problems

For the survey on saddle point problems and their approximation by finite element methods, first the function spaces are introduced, which are needed for the analysis. The notion of *mixed* and *hybrid* finite element formulations are motivated by deriving them from the Poisson problem (3.27), and the necessary and sufficient conditions for unique solvability of the continuous problem as well as for the stability of the corresponding discrete approximation are stated. Finally, we introduce a generalization of such problems.

A thorough analysis of particular approximations of boundary value problems by mixed and hybrid finite element methods can be found in Brezzi and Fortin [1991] and Roberts and Thomas [1991]. For an introduction to mixed problems, the reader is also referred to Brenner and Scott [1994, pp. 237–260] and Braess [2003, pp. 123–162].

For simplicity it is always assumed that Ω is a bounded open subset of $\mathbb{R}^n$, which is connected and has a Lipschitz-continuous boundary $\partial\Omega$. The theory of finite element methods heavily benefits from the utilization of *Sobolev spaces*.[1] These are based on $L^2(\Omega)$, the space of square integrable vector functions on Ω. The latter is defined by

$$L^2(\Omega) \coloneqq \left\{ q \;\middle|\; \int_\Omega |q(\boldsymbol{x})|^2 \, d\boldsymbol{x} < +\infty \right\} \quad ,$$

and a norm on this space is given by

$$\|q\|_{0,\Omega} \coloneqq \left(\int_\Omega |q(\boldsymbol{x})|^2 d\boldsymbol{x} \right)^{1/2} \quad .$$

Then, the first order Sobolev space is

$$H^1(\Omega) \coloneqq \left\{ q \in L^2(\Omega) \mid \nabla q \in (L^2(\Omega))^n \right\} \quad .$$

We put

$$|q|_{1,\Omega} \coloneqq \left(\int_\Omega |\nabla q(\boldsymbol{x})|^2 d\boldsymbol{x} \right)^{1/2} \quad \text{and} \quad \|q\|_{1,\Omega} \coloneqq \left(\|q\|_{0,\Omega}^2 + |q|_{1,\Omega}^2 \right)^{1/2} \quad ,$$

[1] For the general definition of Sobolev spaces see Werner [2000, pp. 180, 193].

which define a semi-norm and a norm on $H^1(\Omega)$, respectively. Note that $|\cdot|_{1,\Omega}$ defines a norm on the quotient space

$$\mathcal{H} \coloneqq H^1(\Omega)/\mathbb{R} \quad ,$$

in which functions are only uniquely defined up to an additive constant. We also refer to spaces of vector valued functions. For this reason, let us introduce

$$H(\mathrm{div};\Omega) \coloneqq \{\boldsymbol{v} \in (L^2(\Omega))^n \mid \nabla \cdot \boldsymbol{v} \in L^2(\Omega)\} \quad .$$

For a vector function $\boldsymbol{v} \in H(\mathrm{div};\Omega)$ it is possible to define its normal component on the boundary $\partial\Omega$ [Girault and Raviart, 1986, Chapter I, Theorem 2.5 and Corollary 2.8], and the subspace with vanishing normal component on $\partial\Omega$ is denoted by

$$\mathcal{U} \coloneqq H_0(\mathrm{div};\Omega) = \{\boldsymbol{v} \in H(\mathrm{div};\Omega) \mid \boldsymbol{v} \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega\} \quad .$$

These spaces are equipped with the Hilbertian graph norm

$$\|\boldsymbol{v}\|_{\mathrm{div},\Omega} \coloneqq \left(\|\boldsymbol{v}\|_{0,\Omega}^2 + \|\nabla \cdot \boldsymbol{v}\|_{0,\Omega}^2\right)^{1/2} \quad .$$

For $\boldsymbol{v} \in H(\mathrm{div};\Omega)$ and $q \in H^1(\Omega)$ the following *Green's formula* is valid [Girault and Raviart, 1986, p. 28]:

$$\int_\Omega \boldsymbol{v} \cdot \nabla q \, d\boldsymbol{x} + \int_\Omega q \, \nabla \cdot \boldsymbol{v} \, d\boldsymbol{x} = \int_{\partial\Omega} q \, \boldsymbol{v} \cdot \boldsymbol{n} \, d\sigma \quad .$$

4.1.1 Mixed and hybrid formulations

The theory of saddle point problems deals with minimization problems, which are constrained by additional side conditions. We illustrate some basic examples by the application of the Poisson problem (3.27) that has been introduced for the derivation of the new projection. Before dealing with multi-field formulations, two single-field formulations of the problem are introduced.

Primal and dual single field formulations

Let us consider again the Poisson problem (3.27) with $f \in L^2(\Omega)$ and $\int_\Omega f \, d\boldsymbol{x} = 0$. Then, the solution $p \in \mathcal{H}$ of (3.27) is the unique minimizer of the *energy* functional [Braess, 2003]:

$$p = \inf_{q \in \mathcal{H}} \mathscr{J}(q) \quad \text{with} \quad \mathscr{J}(q) \coloneqq \frac{1}{2} \int_\Omega |\nabla q|^2 d\boldsymbol{x} - \int_\Omega f \, q \, d\boldsymbol{x} \quad . \tag{4.1}$$

Equivalently, p is characterized by the weak formulation

$$\int_\Omega \nabla p \cdot \nabla q \, d\boldsymbol{x} = \int_\Omega f \, q \, d\boldsymbol{x} \quad \forall \, q \in \mathcal{H} \quad . \tag{4.2}$$

Problem (4.2) is often referred to as the *primal weak formulation* and the unknown p as the *primal unknown* of problem (3.27).

It has been already pointed out that we are particularly interested in the variable $\boldsymbol{u} \coloneqq \nabla p$, rather than in p itself. Despite the possibility of calculating $\boldsymbol{u}$ from the solution of the Poisson problem (3.27), $\boldsymbol{u}$ is also given by the minimization of the so-called *complementary energy* functional. The minimization problem is given by [Quarteroni and Valli, 1997]

$$\boldsymbol{u} = \inf_{\boldsymbol{v} \in \mathcal{W}^f} \mathscr{I}(\boldsymbol{v}) \quad \text{with} \quad \mathscr{I}(\boldsymbol{v}) \coloneqq \frac{1}{2} \int_\Omega |\boldsymbol{v}|^2 d\boldsymbol{x} \quad , \tag{4.3}$$

where

$$\mathcal{W}^f \coloneqq \{\boldsymbol{v} \in \mathcal{U} \mid \nabla \cdot \boldsymbol{v} + f = 0\} \quad .$$

The relationship between $\boldsymbol{u}$ and p is given by $\boldsymbol{u} = \nabla p$. Furthermore, $\boldsymbol{u}$ is referred to as the *dual unknown* of problem (3.27) and is characterized by the *dual weak formulation*

$$\int_\Omega \boldsymbol{u} \cdot \boldsymbol{v} \, d\boldsymbol{x} = 0 \quad \forall \, \boldsymbol{v} \in \mathcal{W}^0 \quad . \tag{4.4}$$

This formulation has the advantage that $\boldsymbol{u}$ is calculated as an independent variable and that it exactly satisfies the divergence relation $\nabla \cdot \boldsymbol{u} + f = 0$. In general, this relation cannot be fulfilled for a $\boldsymbol{u}$, which has been numerically computed from the solution of the primal formulation as a post-processed quantity. Furthermore, the

latter procedure usually leads to numerical inaccuracies [Causin, 2002]. On the other hand, it is often difficult to construct a basis for a discrete subspace of $\mathcal{W}^0$, which would be needed for the discretization of problem (4.4). This difficulty will be circumvented by the technique of *Lagrangian multipliers*.

Dual-mixed formulation

By relaxing the divergence constraint $\nabla \cdot \boldsymbol{u} + f = 0$, the minimization problem (4.3) can be rewritten as a saddle point problem. For this reason, let us introduce the Lagrange multiplier $q \in L^2(\Omega)$. Then, the problem is given by

$$\inf_{\boldsymbol{v} \in \mathcal{U}} \sup_{q \in L^2(\Omega)} \mathscr{L}_{\mathrm{DM}}(\boldsymbol{v}, q)$$

with the *Lagrangian*

$$\mathscr{L}_{\mathrm{DM}}(\boldsymbol{v}, q) \coloneqq \mathscr{I}(\boldsymbol{v}) + \int_\Omega (\nabla \cdot \boldsymbol{v} + f) q \, d\boldsymbol{x} \quad .$$

The unique saddle point $(\boldsymbol{u}, p)$ is characterized by the variational system

$$\left\{ \begin{array}{lcll} \displaystyle\int_\Omega \boldsymbol{u} \cdot \boldsymbol{v} \, d\boldsymbol{x} + \int_\Omega p \, (\nabla \cdot \boldsymbol{v}) \, d\boldsymbol{x} & = & 0 & \forall \, \boldsymbol{v} \in \mathcal{U} \\ \displaystyle\int_\Omega (\nabla \cdot \boldsymbol{u} + f) q \, d\boldsymbol{x} & = & 0 & \forall \, q \in L^2(\Omega) \end{array} \right. . \tag{4.5}$$

Furthermore, $(\boldsymbol{u}, p)$ is the solution of (4.5) if, and only if, p is the solution of the Poisson problem (3.27) [Causin, 2002]. The relation between the two unknowns is again given by $\boldsymbol{u} = \nabla p$. We refer to (4.5) as the *dual mixed formulation* of the Poisson problem.

Primal-hybrid formulation

Hybrid formulations are based on a partition $\mathcal{T}_h$ of $\bar{\Omega}$ into disjoint subsets T. This partition can be chosen independently of any discretization. In all cases, we deal with variables which are defined either on the interior of each subdomain T or on their boundary (hybrid variables).

Given the partition $\mathcal{T}_h$, the energy functional of the primal formulation can be written as

$$\mathscr{J}(q) = \sum_{T \in \mathcal{T}_h} \left(\frac{1}{2} \int_T |\nabla q|^2 d\boldsymbol{x} - \int_T f\, q\, d\boldsymbol{x} \right) \quad .$$

Moreover, with the partition $\mathcal{T}_h$, the space $\mathcal{H}$ can also be characterized as being a subset of the "broken" space

$$\mathcal{Y} \coloneqq \{q \in L^2(\Omega) \mid \forall\, T \in \mathcal{T}_h : q|_T \in H^1(T)\} = \prod_{T \in \mathcal{T}_h} H^1(T) \ ,$$

in which a function $q \in \mathcal{H}$ is characterized by

$$\sum_{T \in \mathcal{T}_h} \int_{\partial T} q\, (\boldsymbol{v} \cdot \boldsymbol{n})\, d\sigma = 0 \quad \forall\, \boldsymbol{v} \in \mathcal{U} \quad .$$

This constraint expresses the continuity of q at the interfaces between subdomains of $\mathcal{T}_h$. The variable $\boldsymbol{v} \cdot \boldsymbol{n}$ may be interpreted as the Lagrange multiplier and the minimization problem (4.1) can be replaced by

$$\inf_{q \in \mathcal{Y}} \sup_{\boldsymbol{v} \in \mathcal{U}} \mathscr{L}_{\mathrm{PH}}(q, \boldsymbol{v})$$

with the Lagrangian

$$\mathscr{L}_{\mathrm{PH}}(q, \boldsymbol{v}) \coloneqq \mathscr{J}(q) - \sum_{T \in \mathcal{T}_h} \int_{\partial T} q\, (\boldsymbol{v} \cdot \boldsymbol{n})\, d\sigma \quad .$$

Thus, we seek $(p, \boldsymbol{u}) \in \mathcal{Y} \times \mathcal{U}$, such that

$$\begin{cases} \displaystyle\sum_{T \in \mathcal{T}_h} \left(\int_T \nabla p \cdot \nabla q\, d\boldsymbol{x} - \int_{\partial T} q\, (\boldsymbol{u} \cdot \boldsymbol{n})\, d\sigma \right) &= \displaystyle\int_\Omega f\, q\, d\boldsymbol{x} \quad \forall\, q \in \mathcal{Y} \\ \displaystyle\sum_{T \in \mathcal{T}_h} \int_{\partial T} p\, (\boldsymbol{v} \cdot \boldsymbol{n})\, d\sigma &= 0 \qquad\qquad \forall\, \boldsymbol{v} \in \mathcal{U} \end{cases} \quad . \tag{4.6}$$

Note that in this *primal-hybrid weak formulation* only the normal trace of the variable $\boldsymbol{u}$ is uniquely defined. Therefore, (4.6) can be reformulated with the hybrid variable $\mu \coloneqq \boldsymbol{u} \cdot \boldsymbol{n}$, which represents the Lagrangian multiplier and is only defined on the interfaces between the subdomains of $\mathcal{T}_h$.

Of course, it is also possible to formulate primal-mixed and dual-hybrid methods for the Poisson problem (3.27). Here, the motivation was to present the basic principles of the approximation of saddle point problems by finite element methods. We conclude this part with a remark about the terminology regarding the methods mentioned above.

Remark 4.1 *Formulations derived from the minimization of the energy functional $\mathscr{J}(q)$ are called* primal*, whereas methods which are based on the minimization of $\mathscr{I}(\boldsymbol{v})$ are called* dual*. A* mixed *method is given in the case of saddle point problems, when constraints are relaxed by the application of Lagrangian multipliers. If the relaxation of constraints arises from a partitioning of the domain, the formulation is called* hybrid*.* ◁

4.1.2 Existence and uniqueness of solutions

The weak formulations given above can be written in a common general form. We are always interested in finding the saddle point $(u, p) \in \mathcal{X} \times \mathcal{M}$, such that

$$\begin{cases} a(u,v) + b(v,p) &= \langle f, v\rangle \quad \forall\, v \in \mathcal{X} \\ b(u,q) &= \langle g, q\rangle \quad \forall\, q \in \mathcal{M} \end{cases} . \tag{4.7}$$

In this formulation, $\mathcal{X}$ and $\mathcal{M}$ are two (real) Hilbert spaces with norms $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{M}}$. Furthermore,

$$a : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad \text{and} \quad b : \mathcal{X} \times \mathcal{M} \to \mathbb{R}$$

are suitably defined bilinear forms. The linear functionals f and g are in the dual spaces $\mathcal{X}'$ and $\mathcal{M}'$, respectively, and $\langle\cdot,\cdot\rangle$ denotes the dual pairing between a Hilbert space and its dual. The stability and convergence properties of such problems are given by the theory of saddle point problems, which originates from the work of Babuška [1971] and Brezzi [1974].

To guarantee existence, uniqueness and stability for (4.7), both of the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ have to satisfy in a certain *coercivity condition*. In particular,

$a(\cdot,\cdot)$ is *coercive* on the subspace

$$\mathcal{K} \coloneqq \{v \in \mathcal{X} \mid b(v,q) = 0 \; \forall \, q \in \mathcal{M}\} \quad , \tag{4.8}$$

if there exists a constant $\alpha > 0$ with

$$a(v,v) \geq \alpha \, \|v\|_{\mathcal{X}}^2 \quad \forall \, v \in \mathcal{K} \quad .$$

Moreover, $b(\cdot,\cdot)$ satisfies the *inf-sup condition*, if there exists a constant $\beta > 0$, such that

$$\inf_{q \in \mathcal{M}} \sup_{v \in \mathcal{X}} \frac{b(v,q)}{\|v\|_{\mathcal{X}} \, \|q\|_{\mathcal{M}}} \geq \beta > 0 \quad . \tag{4.9}$$

The following theorem can be stated [cf. Braess, 2003, Chapter III, Proposition 4.3].

Theorem 4.1 *Let us assume that $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are bounded on $\mathcal{X} \times \mathcal{X}$ and $\mathcal{X} \times \mathcal{M}$, respectively. Furthermore, let $a(\cdot,\cdot)$ be coercive on $\mathcal{K}$ and $b(\cdot,\cdot)$ satisfy the inf-sup condition* (4.9)*. Then, problem* (4.7) *has a unique solution (u,p) for all $f \in \mathcal{X}'$ and $g \in \mathcal{M}'$. The solution satisfies the stability bound*

$$\|u\|_{\mathcal{X}} + \|p\|_{\mathcal{M}} \leq c \, (\|f\|_{\mathcal{X}'} + \|g\|_{\mathcal{M}'}) \quad .$$

For the approximation by finite elements, appropriate finite dimensional subspaces $\mathcal{X}^h \subset \mathcal{X}$ and $\mathcal{M}^h \subset \mathcal{M}$ have to be chosen and (4.7) is reformulated with $\mathcal{X}^h$ and $\mathcal{M}^h$ instead of $\mathcal{X}$ and $\mathcal{M}$. Therefore, we seek the solution $(u_h, p_h) \in \mathcal{X}^h \times \mathcal{M}^h$, such that

$$\left\{ \begin{array}{lcll} a(u_h, v_h) + b(v_h, p_h) & = & \langle f, v_h \rangle & \forall \, v_h \in \mathcal{X}^h \\ b(u_h, q_h) & = & \langle g, q_h \rangle & \forall \, q_h \in \mathcal{M}^h \end{array} \right. \quad . \tag{4.10}$$

For the stability of problem (4.10), the discrete spaces $\mathcal{X}^h$ and $\mathcal{M}^h$ cannot be chosen independently from each other. They have to be compatible in some sense. This statement is supported by the following theorem [Braess, 2003, Chapter III, Proposition 4.5].

Theorem 4.2 *Let us assume that $a(\cdot,\cdot)$ is coercive with coercivity constant α_h on the subspace $\mathcal{K}^h \subset \mathcal{X}^h$, analogously defined to* (4.8)*. Furthermore, let $b(\cdot,\cdot)$ satisfy the*

discrete inf-sup condition

$$\inf_{q_h \in \mathcal{M}^h} \sup_{v_h \in \mathcal{X}^h} \frac{b(v_h, q_h)}{\|v_h\|_{\mathcal{X}} \, \|q_h\|_{\mathcal{M}}} \geq \beta_h > 0 \quad . \tag{4.11}$$

Then, (4.10) *has a unique solution* (u_h, p_h)*, which satisfies*

$$\|u_h\|_{\mathcal{X}} + \|p_h\|_{\mathcal{M}} \leq c_h \, (\|f\|_{\mathcal{X}'} + \|g\|_{\mathcal{M}'}) \quad ,$$

and stability is obtained, if the constant $c_h = c_h(\alpha_h, \beta_h)$ *is independent of* h*. Additionally, the solution satisfies the error estimate*

$$\|u - u_h\|_{\mathcal{X}} + \|p - p_h\|_{\mathcal{M}} \leq C \left(\inf_{v_h \in \mathcal{X}^h} \|u - v_h\|_{\mathcal{X}} + \inf_{q_h \in \mathcal{M}^h} \|p - q_h\|_{\mathcal{M}} \right) \quad .$$

It is important to observe that the coercivity of $a(\cdot, \cdot)$ on $\mathcal{K}$ does not imply its coercivity on $\mathcal{K}^h$, since, in general, $\mathcal{K}^h \nsubseteq \mathcal{K}$. Likewise, the discrete inf-sup condition for $b(\cdot, \cdot)$ is not necessarily implied by its continuous counterpart. This is due to the fact that in the majority of cases $\mathcal{X}^h$ is a proper subspace of $\mathcal{X}$.

Remark 4.2 *Condition* (4.9) *is often referred to as the* Babuška-Brezzi compatibility *condition or as the* Ladyzhenskaya-Babuška-Brezzi (LBB) *condition. The discrete condition* (4.11) *is also called the* discrete LBB *condition. However, these names are not always employed in the same way. We will refer to* (4.9) *as the* inf-sup *condition and to* (4.11) *as the* discrete inf-sup *condition, respectively.* ◁

4.1.3 Generalized problems

The results of the preceding section can be easily extended to more general problems. In particular, in the analysis of the new projection we will be interested in formulations with three distinct bilinear forms instead of two. That is, find $(u, p) \in (\mathcal{X}_2 \times \mathcal{M}_1)$, such that

$$\begin{cases} a(u, v) + b_1(v, p) &= \langle f, v \rangle \quad \forall \, v \in \mathcal{X}_1 \\ b_2(u, q) &= \langle g, q \rangle \quad \forall \, q \in \mathcal{M}_2 \quad . \end{cases} \tag{4.12}$$

In this formulation, $\mathcal{X}_i$ and $\mathcal{M}_i$ $(i = 1, 2)$ are four Hilbert spaces with norms $\|\cdot\|_{\mathcal{X}_i}$ and $\|\cdot\|_{\mathcal{M}_i}$. The bilinear form $a(\cdot, \cdot)$ is defined on $\mathcal{X}_2 \times \mathcal{X}_1$ and the bilinear forms $b_i(\cdot, \cdot)$ are defined on $\mathcal{X}_i \times \mathcal{M}_i$ $(i = 1, 2)$. Furthermore, f and g are elements of $\mathcal{X}_1'$ and $\mathcal{M}_2'$, the dual spaces of $\mathcal{X}_1$ and $\mathcal{M}_2$. The abstract theory of such problems is given in NICOLAÏDES [1982] and furtherly developed in BERNARDI ET AL. [1988].

To obtain conditions for existence, uniqueness and stability of problem (4.12), let us introduce for any $g \in \mathcal{M}_i'$ $(i = 1, 2)$ the closed affine spaces

$$\mathcal{K}_i(g) \coloneqq \{v \in \mathcal{X}_i \mid \forall\, q \in \mathcal{M}_i : b_i(v, q) = \langle g, q \rangle\} \quad .$$

We denote by $\mathcal{K}_i \coloneqq \mathcal{K}_i(0)$ the kernel of the operator induced by $b_i(\cdot, \cdot)$.

Theorem 4.3 *Let $a(\cdot, \cdot)$ and $b_i(\cdot, \cdot)$ $(i = 1, 2)$ be bounded. Assume that there exists a constant $\alpha > 0$, such that*

$$\inf_{u \in \mathcal{K}_2} \sup_{v \in \mathcal{K}_1} \frac{a(u, v)}{\|u\|_{\mathcal{X}_2} \, \|v\|_{\mathcal{X}_1}} \geq \alpha \tag{4.13}$$

and

$$\sup_{u \in \mathcal{K}_2} a(u, v) > 0 \quad \forall\, v \in \mathcal{K}_1 \setminus \{0\} \quad . \tag{4.14}$$

Furthermore, assume that $b_i(\cdot, \cdot)$ $(i = 1, 2)$ satisfies the inf-sup condition

$$\inf_{q \in \mathcal{M}_i} \sup_{v \in \mathcal{X}_i} \frac{b_i(v, q)}{\|v\|_{\mathcal{X}_i} \, \|q\|_{\mathcal{M}_i}} \geq \beta_i > 0 \quad . \tag{4.15}$$

Then, problem (4.12) has a unique solution (u, p) for all $f \in \mathcal{X}_1'$ and $g \in \mathcal{M}_2'$ and the following estimate holds:

$$\|u\|_{\mathcal{X}_2} + \|p\|_{\mathcal{M}_1} \leq c \left(\|f\|_{\mathcal{X}_1'} + \|g\|_{\mathcal{M}_2'} \right) \quad . \tag{4.16}$$

Remark 4.3 *In case of finite-dimensional spaces $\mathcal{K}_1$ and $\mathcal{K}_2$, the conditions (4.13) and (4.14) are equivalent to the requirement [BERNARDI ET AL., 1988]*

$$\dim \mathcal{K}_1 = \dim \mathcal{K}_2 \quad . \qquad \triangleleft$$

For the discretization of problem (4.12), it is assumed that there are finite-dimensional subspaces $\mathcal{X}_i^h \subset \mathcal{X}_i$ and $\mathcal{M}_i^h \subset \mathcal{M}_i$ $(i = 1, 2)$. We are looking for the solution $(u_h, p_h) \in (\mathcal{X}_2^h \times \mathcal{M}_1^h)$ of the approximation

$$\begin{cases} a(u_h, v_h) + b_1(p_h, v_h) &= \langle f, v_h \rangle \quad \forall\, v_h \in \mathcal{X}_1^h \\ b_2(u_h, q_h) &= \langle g, q_h \rangle \quad \forall\, q_h \in \mathcal{M}_2^h \end{cases} \quad . \tag{4.17}$$

With the definition of the discrete affine spaces

$$\mathcal{K}_i^h(g) \coloneqq \{v_h \in \mathcal{X}_i^h \mid \forall\, q_h \in \mathcal{M}_i^h : b_i(v_h, q_h) = \langle g, q_h \rangle\} \quad ,$$

in which $g \in \mathcal{M}_i^{h\prime}$ $(i = 1, 2)$, Theorem 4.3 can be applied to problem (4.17), and existence, uniqueness and stability are obtained given the constant c in (4.16) is independent of h. Moreover, the following error estimate concerning the approximate solution can be stated.

Theorem 4.4 *Assuming that Theorem 4.3 holds for the continuous problem* (4.12) *as well as for its approximation* (4.17)*, the error is bounded by*

$$\begin{aligned} &\|u - u_h\|_{\mathcal{X}_2} + \|p - p_h\|_{\mathcal{M}_1} \leq \\ &\qquad C \left(\inf_{w_h \in \mathcal{K}_2^h(g)} \|u - w_h\|_{\mathcal{X}_2} + \inf_{v_h \in \mathcal{X}_2^h} \|u - v_h\|_{\mathcal{X}_2} + \inf_{q_h \in \mathcal{M}_1^h} \|p - q_h\|_{\mathcal{M}_1} \right) \quad . \end{aligned} \tag{4.18}$$

This completes the review of finite element methods for the approximation of saddle point problems. As we have seen, the different approaches all lead to a similar abstract formulation, for which the theory can be applied. In the following, such a formulation is derived for the new projection in order to analyze its stability concerning the corrected momentum field.

4.2 Reformulation of the problem

The derivation of a mixed formulation equivalent to the Poisson-type problem (3.37) is easily established. The continuous counterpart of this equation is obtained by a

combination of the momentum update (3.8) and the divergence constraint (3.9), i.e.

$$\begin{aligned} (h\boldsymbol{v})^{n+1} &= (h\boldsymbol{v})^{**} - \delta t\,(h_0 \nabla h^{(2)}) \\ \frac{1}{2}\left[\nabla\cdot(h\boldsymbol{v})^{n+1} + \nabla\cdot(h\boldsymbol{v})^{n}\right] &= -\frac{dh_0}{dt} \quad . \end{aligned} \tag{4.19}$$

A variational formulation of these two equations is derived by the usual procedure: $(4.19)_1$ and $(4.19)_2$ are multiplied with test functions $\boldsymbol{\varphi}$ and ψ and the resulting equations are integrated over the whole domain Ω. This leads to

$$\begin{aligned} \int_\Omega \left((h\boldsymbol{v})^{n+1}\cdot\boldsymbol{\varphi} + \delta t\, h_0\, \nabla h^{(2)}\cdot\boldsymbol{\varphi}\right) d\boldsymbol{x} &= \int_\Omega (h\boldsymbol{v})^{**}\cdot\boldsymbol{\varphi}\, d\boldsymbol{x} \\ \int_\Omega \psi\, \nabla\cdot(h\boldsymbol{v})^{n+1}\, d\boldsymbol{x} &= -\int_\Omega \psi\left(\nabla\cdot(h\boldsymbol{v})^{n} + 2\,\frac{dh_0}{dt}\right) d\boldsymbol{x}\ . \end{aligned} \tag{4.20}$$

Note that this formulation can be already interpreted as a generalized problem as formulated in (4.12). The discrete method – equivalent to the Poisson-type problem (3.37) – is derived by introducing appropriate finite dimensional trial and test spaces. For the choice of the trial spaces, we are confined to our selection for the momentum $(h\boldsymbol{v})$ and the height $h^{(2)}$ in Section 3.2. In the new projection method, the momentum distribution is approximated by piecewise linear functions belonging to the space

$$\mathcal{U}^h \coloneqq \left\{\boldsymbol{v} = (u,v) \in (L^2(\Omega))^2 \mid \forall\, V \in \mathcal{V} : u|_V, v|_V \in \mathcal{P}_1(V)\right\}$$

in which $\mathcal{P}_k(U)$ is the space of k-th order polynomials defined in (3.29). The height perturbation $h^{(2)}$ is given by piecewise bilinear functions. This space was defined by

$$\mathcal{H}^h \coloneqq \left\{p \in H^1(\Omega) \mid \forall\, V \in \mathcal{V} : p|_V \in \mathcal{P}_2(V), \forall\, I \in \mathcal{I} : p|_I \in \mathcal{P}_1(I)\right\} \quad .$$

To obtain the same divergence as in Section 3.2, also the test functions ψ for the

second equation of (4.20) are fixed. As noted above, these functions span the space

$$\mathcal{Q}^h \coloneqq \left\{ q \in L^2(\Omega) \mid \forall\, \bar{V} \in \bar{\mathcal{V}} : q|_{\bar{V}} \in \mathcal{P}_0(\bar{V}) \right\} \quad ,$$

and a basis of $\mathcal{Q}^h$ is given by $\bigcup_{\bar{V}\in\bar{\mathcal{V}}}\{\chi_{\bar{V}}\}$. The selection of the test space for the first equation is yet undetermined. Let us choose $\mathcal{U}^h$, the space which is also used for the momentum variable. A basis of $\mathcal{U}^h$ is given by

$$\begin{aligned} \bigcup_{V\in\mathcal{V}} \Bigg\{ \begin{pmatrix} \chi_V \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \chi_V \end{pmatrix}, \begin{pmatrix} (x-x_V)\chi_V \\ 0 \end{pmatrix}, \begin{pmatrix} (y-y_V)\chi_V \\ 0 \end{pmatrix}, \\ \begin{pmatrix} 0 \\ (x-x_V)\chi_V \end{pmatrix}, \begin{pmatrix} 0 \\ (y-y_V)\chi_V \end{pmatrix} \Bigg\} \quad , \end{aligned} \tag{4.21}$$

where (x_V, y_V) is the center of the cell V.

The following discussion is focused on Cartesian grids with cells $V_{i,j}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, and cell centers (x_i, y_j). Because of the linearity of the equations (4.20) in $\boldsymbol{\varphi}$ and ψ, it is sufficient to "test" them with only a basis of $\mathcal{U}^h$ and $\mathcal{Q}^h$, respectively. Let us consider the first equation in conjunction with the test function $\boldsymbol{\varphi} = (\chi_{V_{i,j}}, 0)^T$. Because the second component of $\boldsymbol{\varphi}$ is zero and its support is $V_{i,j}$, this yields

$$\int_{V_{i,j}} (hu)^{n+1} \, d\boldsymbol{x} + \delta t \, h_0 \int_{V_{i,j}} \frac{\partial h^{(2)}}{\partial x} \, d\boldsymbol{x} = \int_{V_{i,j}} (hu)^{**} \, d\boldsymbol{x} \quad . \tag{4.22}$$

Furthermore, by expanding the height $h^{(2)}$ in a volumewise representation, as in (3.35), the calculation of the second integral in (4.22) leads to

$$\int_{V_{i,j}} \frac{\partial h^{(2)}}{\partial x} \, d\boldsymbol{x} = \int_{V_{i,j}} \left(h^{(2)}_{x,i,j} + (y - y_j) h^{(2)}_{xy,i,j} \right) d\boldsymbol{x} = \delta x \, \delta y \, h^{(2)}_{x,i,j} \quad .$$

The integral of the second term vanishes, because it is an odd function in y with respect to y_j. With similar results for the other terms in (4.22), we finally obtain

$$(hu)^{n+1}_{i,j} + \delta t \, h_0 \, h^{(2)}_{x,i,j} = (hu)^{**}_{i,j} \quad . \tag{4.23}$$

By using the other five test functions in (4.21), this procedure yields the equations

$$\begin{aligned}
(hv)^{n+1}_{i,j} &+ \delta t\, h_0\, h^{(2)}_{y,i,j} &&= (hv)^{**}_{i,j} \\
(hu)^{n+1}_{x,i,j} & && = (hv)^{**}_{x,i,j} \\
(hu)^{n+1}_{y,i,j} &+ \delta t\, h_0\, h^{(2)}_{xy,i,j} &&= (hu)^{**}_{y,i,j} \\
(hv)^{n+1}_{x,i,j} &+ \delta t\, h_0\, h^{(2)}_{xy,i,j} &&= (hv)^{**}_{x,i,j} \\
(hv)^{n+1}_{y,i,j} & && = (hv)^{**}_{y,i,j} \quad .
\end{aligned} \tag{4.24}$$

Therefore, six equations are obtained for each cell $V_{i,j}$. They represent the discretization of $(4.20)_1$.

The discretization of the second equation in (4.20) is similar to the discretization of the right hand side of the Poisson problem (3.28). The application of the test function $\psi = \chi_{\bar{V}_{i+1/2,j+1/2}}$ yields for the terms involving the momentum the discrete divergence from Section 3.2 multiplied by $|\bar{V}_{i+1/2,j+1/2}|$. Thus, dividing this equation by $|\bar{V}_{i+1/2,j+1/2}|$ leads to

$$\mathsf{D}^{\mathcal{V}}_{i+1/2,j+1/2}\big((h\boldsymbol{v})^{n+1}_{\mathcal{V}}\big) = -\mathsf{D}^{\mathcal{V}}_{i+1/2,j+1/2}((h\boldsymbol{v})^{n}_{\mathcal{V}}) - 2\,\frac{dh_0}{dt} \quad . \tag{4.25}$$

Let us recall that $h^{(2)}$ is uniquely defined by its node values and that each velocity component has three degrees of freedom per cell. Then there are $7 \cdot m \cdot n$ unknowns in case of periodic boundary conditions. The analysis above yielded the same number of linear equations, leading to a well-defined problem. Finally, by inserting the equations from (4.23) and (4.24) into (4.25), the second discrete Poisson-type problem from our new projection method is obtained. We have derived a *Petrov-Galerkin* mixed formulation, which utilizes different trial and test spaces for the scalar variables.

Remark 4.4 *The original second discrete Poisson-type problem* (3.23) *described by* Schneider et al. [1999] *is obtained with the same procedure, but with the trial and*

test spaces

$$\tilde{\mathcal{U}}^h \coloneqq \{\boldsymbol{v} = (u, v) \in (L^2(\Omega))^2 \mid \forall V \in \mathcal{V} : u|_V \in \mathcal{P}_0(V), v|_V \in \mathcal{P}_0(V)\} \tag{4.26}$$

for the first equation in (4.20)*. Thus, essentially the space for the momentum variable has been enriched by linear functions on the cells* $V \in \mathcal{V}$ *to obtain the new projection method for the second discrete Poisson-type problem.* ◁

In the primal-hybrid formulation derived in Section 4.1.1 the constraints on the function space were relaxed by the introduction of the "broken" space $\mathcal{Y}$. This leads to a discretization, in which the discrete spaces are in general not contained in the original continuous spaces.[2] It will be outlined in the stability analysis of our mixed formulation that also in our case some of the above defined spaces are not in their continuous counterparts, leading to a nonconforming finite element method. Thus, the possibility of formulating our discrete Poisson-type problem as a hybrid method was investigated. However, under the given time constraints it was not possible to find a suitable formulation. It remains to be analyzed, whether or not this is possible.

4.3 Stability analysis of the mixed formulation

In order to apply the theory from Section 4.1.3 to the mixed formulation (4.20), the corresponding continuous problem is defined which can be shown to have a unique solution. The section concludes with an investigation of the discrete mixed formulation.

For the derivation of the continuous problem the function spaces for the trial and test functions have to be chosen. In the Poisson-type problem (3.10) – the continuous counterpart of (3.37) – the height perturbation $h^{(2)}$ is only determined up to an additive constant. This constant can be fixed by the additional condition of a zero mean value, i.e. $\int_\Omega h^{(2)} d\boldsymbol{x} = 0$. Thus, a suitable space is given by the previously defined $\mathcal{H} \coloneqq H^1(\Omega)/\mathbb{R}$. An appropriate space for the momentum should also bound the divergence of the unknown variable. Furthermore, the boundary conditions are given by the integral constraint (2.25). For simplicity, let us assume, that there is

[2] See also [Causin, 2002] for further examples.

no flux across the boundary, i.e. there are non-permeable rigid walls and $dh_0/dt \equiv 0$. Therefore, the momentum is sought in the space $\mathcal{U} = H_0(\mathrm{div};\Omega)$. The test functions of the discrete problem are discontinuous at the interfaces either of the primal or of the dual discretization. Therefore, no regularity is assumed for the test functions in the continuous problem as well.

With the definition of the bilinear forms

$$\begin{aligned} a &: \mathcal{U} \times (L^2(\Omega))^2 \to \mathbb{R} \quad \text{with} \quad a(\boldsymbol{u},\boldsymbol{v}) \coloneqq \int_\Omega \boldsymbol{u}\cdot\boldsymbol{v}\; d\boldsymbol{x} \\ b_1 &: (L^2(\Omega))^2 \times \mathcal{H} \to \mathbb{R} \quad \text{with} \quad b_1(\boldsymbol{v},q) \coloneqq \delta t\, h_0 \int_\Omega \boldsymbol{v}\cdot\nabla q\; d\boldsymbol{x} \\ b_2 &: \mathcal{U} \times L^2(\Omega) \to \mathbb{R} \quad \text{with} \quad b_2(\boldsymbol{v},q) \coloneqq \int_\Omega q\,(\nabla\cdot\boldsymbol{v})\; d\boldsymbol{x} \quad , \end{aligned} \tag{4.27}$$

problem (4.20) can be reformulated to obtain the following continuous saddle point problem. Find $((h\boldsymbol{v})^{n+1}, h^{(2)}) \in (\mathcal{U}\times\mathcal{H})$, such that

$$\begin{aligned} a\big((h\boldsymbol{v})^{n+1},\boldsymbol{\varphi}\big) + b_1\big(\boldsymbol{\varphi},h^{(2)}\big) &= \langle (h\boldsymbol{v})^{**},\boldsymbol{\varphi}\rangle \qquad \forall\,\boldsymbol{\varphi}\in(L^2(\Omega))^2 \\ b_2\big((h\boldsymbol{v})^{n+1},\psi\big) &= \langle -\nabla\cdot(h\boldsymbol{v})^n,\psi\rangle \quad \forall\,\psi\in L^2(\Omega) \quad . \end{aligned} \tag{4.28}$$

By setting $\mathcal{X}_1 \coloneqq (L^2(\Omega))^2$, $\mathcal{X}_2 \coloneqq \mathcal{U}$, $\mathcal{M}_1 \coloneqq \mathcal{H}$ and $\mathcal{M}_2 \coloneqq L^2(\Omega)$ a problem of the form (4.12) is defined. This formulation is also referred to as a *primal-dual* formulation [Thomas and Trujillo, 1999]. In order to show existence and uniqueness of the solution, the bilinear forms have to be bounded.

Lemma 4.5 *Let $a(\cdot,\cdot)$, $b_1(\cdot,\cdot)$ and $b_2(\cdot,\cdot)$ be defined as above. Then, all three bilinear forms are bounded.*

Proof. *Using the Cauchy-Schwarz inequality we obtain for arbitrary $\boldsymbol{u}\in\mathcal{U}$, $\boldsymbol{v}\in(L^2(\Omega))^2$*

$$a(\boldsymbol{u},\boldsymbol{v}) = \int_\Omega \boldsymbol{u}\cdot\boldsymbol{v}\; d\boldsymbol{x} \le \left(\int_\Omega |\boldsymbol{u}|^2\; d\boldsymbol{x}\right)^{1/2}\left(\int_\Omega |\boldsymbol{v}|^2\; d\boldsymbol{x}\right)^{1/2} \le \|\boldsymbol{u}\|_{\mathrm{div},\Omega}\;\|\boldsymbol{v}\|_{0,\Omega} \quad .$$

Similarly, it follows from $\boldsymbol{v} \in (L^2(\Omega))^2$, $q \in \mathcal{H}$ *that*

$$b_1(\boldsymbol{v}, q) \leq \delta t\, h_0 \left(\int_\Omega |\boldsymbol{v}|^2 \, d\boldsymbol{x} \right)^{1/2} \left(\int_\Omega |\nabla q|^2 \, d\boldsymbol{x} \right)^{1/2} = \delta t\, h_0 \, \|\boldsymbol{v}\|_{0,\Omega} \, |q|_{1,\Omega}$$

and from $\boldsymbol{v} \in \mathcal{U}$, $q \in L^2(\Omega)$ *that*

$$b_2(\boldsymbol{v}, q) \leq \left(\int_\Omega (\nabla \cdot \boldsymbol{v})^2 \, d\boldsymbol{x} \right)^{1/2} \left(\int_\Omega q^2 \, d\boldsymbol{x} \right)^{1/2} \leq \|\boldsymbol{v}\|_{\mathrm{div},\Omega} \, \|q\|_{0,\Omega} \quad . \qquad \square$$

In the following, it is shown that the bilinear forms satisfy the assumptions of Theorem 4.3, and therefore, problem (4.28) has a unique solution. For this purpose, let us define the subspaces

$$\begin{aligned} \mathcal{K}_1 &\coloneqq \{\boldsymbol{v} \in (L^2(\Omega))^2 \mid \forall\, q \in \mathcal{H} : b_1(\boldsymbol{v}, q) = 0\} \\ \mathcal{K}_2 &\coloneqq \{\boldsymbol{v} \in \mathcal{U} \mid \forall\, q \in L^2(\Omega) : b_2(\boldsymbol{v}, q) = 0\} \quad . \end{aligned}$$

These spaces can be characterized more precisely. An orthogonal decomposition of $(L^2(\Omega))^2$ is given by

$$(L^2(\Omega))^2 = \{\boldsymbol{v} \in \mathcal{U} \mid \nabla \cdot \boldsymbol{v} = 0\} \oplus \{\nabla q \mid q \in H^1(\Omega)\}$$

[Girault and Raviart, 1986, Chapter I, Theorem 2.7]. This is a generalization of the Helmholtz-Decomposition-Principle, which states that every smooth vector field can be uniquely decomposed into an irrotational (no vorticity) and a solenoidal (no divergence) part. Thus, by the definition of $b_1(\cdot, \cdot)$, $\mathcal{K}_1$ can also be written as

$$\mathcal{K}_1 = \{\boldsymbol{v} \in \mathcal{U} \mid \nabla \cdot \boldsymbol{v} = 0\} \quad .$$

Furthermore, since $\nabla \cdot \boldsymbol{v} \in L^2(\Omega)$ for all $\boldsymbol{v} \in H(\mathrm{div}; \Omega)$, the definition of the bilinear form $b_2(\cdot, \cdot)$ implies that the divergence also has to vanish for all functions in $\mathcal{K}_2$, resulting in $\mathcal{K}_1 = \mathcal{K}_2$.

With this characterization of $\mathcal{K}_i$ $(i = 1, 2)$ the identity $\|\boldsymbol{v}\|_{\mathrm{div},\Omega} = \|\boldsymbol{v}\|_{0,\Omega}$ is obtained for $\boldsymbol{v} \in \mathcal{K}_i$. Additionally, the following estimates can be derived. For each $\boldsymbol{u} \in \mathcal{K}_2$,

$\|\boldsymbol{u}\|_{0,\Omega} \neq 0$, $a(\cdot,\cdot)$ satisfies

$$\sup_{\boldsymbol{v}\in\mathcal{K}_1} \frac{a(\boldsymbol{u},\boldsymbol{v})}{\|\boldsymbol{v}\|_{0,\Omega}} \geq \frac{a(\boldsymbol{u},\boldsymbol{u})}{\|\boldsymbol{u}\|_{0,\Omega}} = \frac{\|\boldsymbol{u}\|^2_{0,\Omega}}{\|\boldsymbol{u}\|_{0,\Omega}} = \|\boldsymbol{u}\|_{\mathrm{div},\Omega} \quad ,$$

and for $\boldsymbol{v} \in \mathcal{K}_1 \setminus \{0\}$ we obtain

$$\sup_{\boldsymbol{u}\in\mathcal{K}_2} a(\boldsymbol{u},\boldsymbol{v}) \geq a(\boldsymbol{v},\boldsymbol{v}) > 0 \quad .$$

Therefore, the conditions (4.13) and (4.14) are satisfied.

Using the fact that $p \in \mathcal{H}$ implies $\nabla p \in (L^2(\Omega))^2$, leads to

$$\sup_{\boldsymbol{v}\in(L^2(\Omega))^2} \frac{b_1(\boldsymbol{v},p)}{\|\boldsymbol{v}\|_{0,\Omega}} \geq \frac{b_1(\nabla p,p)}{\|\nabla p\|_{0,\Omega}} = \delta t\, h_0\ |p|_{1,\Omega} \quad ,$$

which satisfies condition (4.15) for $b_1(\cdot,\cdot)$.

The condition (4.15) for $b_2(\cdot,\cdot)$ is established using the auxiliary problem to find $\varphi_q \in H^1/\mathbb{R}$ with $\Delta\varphi_q = q$ in Ω and $\partial\varphi_q/\partial\nu = 0$ on $\partial\Omega$ [cf. Roberts and Thomas, 1991]. For $q \in L^2(\Omega)$, $\int_\Omega q\, d\boldsymbol{x} = 0$, this problem has a unique solution with $|\varphi_q|_{1,\Omega} \leq c\ \|q\|_{0,\Omega}$. The function $\boldsymbol{v}_q \coloneqq \nabla\varphi_q$ belongs to $H(\mathrm{div};\Omega)$, which leads to $\nabla\cdot\boldsymbol{v}_q = q$. Additionally, $\boldsymbol{v}_q$ fulfills

$$\begin{aligned} \|\boldsymbol{v}_q\|^2_{\mathrm{div},\Omega} &= \int_\Omega |\boldsymbol{v}_q|^2 + (\nabla\cdot\boldsymbol{v}_q)^2\ d\boldsymbol{x} \\ &= |\varphi_q|^2_{1,\Omega} + \|q\|^2_{0,\Omega} \\ &\leq C\ \|q\|^2_{0,\Omega} \quad . \end{aligned}$$

Thus, $b_2(\cdot,\cdot)$ satisfies the estimate

$$\sup_{\boldsymbol{u}\in\mathcal{U}} \frac{b_2(\boldsymbol{u},q)}{\|\boldsymbol{u}\|_{\mathrm{div},\Omega}} \geq \frac{b_2(\boldsymbol{v}_q,q)}{\|\boldsymbol{v}_q\|_{\mathrm{div},\Omega}} \geq \frac{\|q\|^2_{0,\Omega}}{C\ \|q\|_{0,\Omega}} = \frac{1}{C}\ \|q\|_{0,\Omega} \quad ,$$

and the following theorem can be concluded from the results above.

Theorem 4.6 *The generalized saddle point problem* (4.28) *has a unique solution* $((h\boldsymbol{v})^{n+1}, h^{(2)})$ *in* $(\mathcal{U}\times\mathcal{H})$.

With the definition of the bilinear forms in (4.27), the mixed formulation derived in Section 4.2 is to find $((h\boldsymbol{v})^{n+1}, h^{(2)}) \in (\mathcal{U}^h \times \mathcal{H}^h)$, such that

$$\begin{aligned} a\big((h\boldsymbol{v})^{n+1}, \boldsymbol{\varphi}\big) + b_1\big(\boldsymbol{\varphi}, h^{(2)}\big) &= \langle (h\boldsymbol{v})^{**}, \boldsymbol{\varphi}\rangle \qquad \forall\, \boldsymbol{\varphi} \in \mathcal{U}^h \\ b_2\big((h\boldsymbol{v})^{n+1}, \psi\big) &= \langle -\nabla\cdot(h\boldsymbol{v})^{n}, \psi\rangle \quad \forall\, \psi \in \mathcal{Q}^h \quad . \end{aligned} \tag{4.29}$$

Note that the trial space $\mathcal{U}^h$ is not contained in its continuous counterpart $\mathcal{U}$. The use of common discrete subspaces of $H(\mathrm{div};\Omega)$ like Raviart-Thomas [Raviart and Thomas, 1977] or BDFM elements [Brezzi et al., 1987], which restrict the degrees of freedom by imposing additional constraints on the boundary of each element, is unsuitable for our purposes. The piecewise linear versions of these spaces demand continuity of the normal velocity components at the boundary of an element. This is in contrast to the idea of solving Riemann problems in the predictor step of our projection method. Therefore, the discrete problem (4.29) is an approximation using *nonconforming elements*, and an error estimate like (4.18) would have to be modified using a similar statement as the *second Strang Lemma* [Braess, 2003, Chapter III, Proposition 1.2].

For the stability analysis of the mixed formulation, let us define the spaces

$$\begin{aligned} \mathcal{K}_1^h &\coloneqq \{\boldsymbol{v}_h \in \mathcal{U}^h \mid \forall\, q_h \in \mathcal{H}^h : b_1(\boldsymbol{v}_h, q_h) = 0\} \\ \mathcal{K}_2^h &\coloneqq \{\boldsymbol{v}_h \in \mathcal{U}^h \mid \forall\, q_h \in \mathcal{Q}^h : b_2(\boldsymbol{v}_h, q_h) = 0\} \quad . \end{aligned}$$

The characterization of these spaces is slightly more complicated than it was for their continuous counterparts. A preliminary analysis reveals that $\mathcal{K}_1^h$ contains those elements $\boldsymbol{v} \in \mathcal{U}^h$, which satisfy on $V_{i,j}$

$$u_{i,j} = v_{i,j} = 0 \;, \quad \delta y^2\, u_{y,i,j} = \delta x^2\, v_{x,i,j} \quad .$$

In $\mathcal{K}_2^h$, there are at least the elements $\boldsymbol{v} \in \mathcal{U}^h$ with

$$u_{i,j} = v_{i,j} = 0 \;, \quad u_{y,i,j} = v_{x,i,j} \quad .$$

This suggests an application of Remark 4.3, but it has to be further analyzed, if these are the only functions contained in $\mathcal{K}_1^h$ and $\mathcal{K}_2^h$, respectively.

Since this is a nonconforming finite element method, the $H(\mathrm{div};\Omega)$ norm is no longer appropriate for the space $\mathcal{U}^h$, and a suitable mesh dependent norm has to be introduced [cf. Braess, 2003, p. 101]. Thus, the characterization of $\mathcal{K}_1^h$ and $\mathcal{K}_2^h$ and the choice of the norm for $\mathcal{U}^h$ are the necessary requirements to show the conditions (4.13) and (4.14) for the discrete case.

For the discrete inf-sup condition concerning $b_1(\cdot,\cdot)$ the following can be stated. It has been already pointed out that $p \in \mathcal{H}^h$ implies $\nabla p \in \mathcal{U}^h$. Thus, as in the continuous case, we have for arbitrary $p \in \mathcal{H}^h$

$$\sup_{\boldsymbol{v}\in\mathcal{U}^h} \frac{b_1(\boldsymbol{v},p)}{\|\boldsymbol{v}\|_{0,\Omega}} \geq \frac{b_1(\nabla p,p)}{\|\nabla p\|_{0,\Omega}} = \frac{\delta t\, h_0\ |p|_{1,\Omega}^2}{|p|_{1,\Omega}} = \delta t\, h_0\ |p|_{1,\Omega} \quad .$$

Note that, if our projection is considered as part of a time step method, δt goes to zero as δx does, and the above inf-sup estimate is not independent of the grid size. Here, $\delta t\, h_0$ is assumed to be fixed.

Remark 4.5 *A simple counter-example can be presented, which shows that the original projection method does not satisfy condition* (4.15) *for* $b_1(\cdot,\cdot)$*. Let us consider a scalar function* $q \in \mathcal{H}^h$ *with*

$$q(x_{i+1/2}, y_{j+1/2}) = \begin{cases} 1 & \text{if } i+j \text{ is even,} \\ -1 & \text{if } i+j \text{ is odd.} \end{cases}$$

On $V_{i,j}$*, this function is given by*

$$q(x,y)|_{V_{i,j}} = \pm(x-x_i)(y-y_j)\frac{4}{\delta x\,\delta y} \quad .$$

For any $\boldsymbol{v}$ *being in the space of piecewise constant functions* (4.26)*,* $\boldsymbol{v}$ *is given on* $V_{i,j}$

by $\boldsymbol{v}|_{V_{i,j}} =: \boldsymbol{v}_{i,j} = (u_{i,j}, v_{i,j})$*, which leads to*

$$\begin{aligned} b_1(\boldsymbol{v}, q) &= \delta t\, h_0 \sum_{i,j} \int_{V_{i,j}} \nabla q \cdot \boldsymbol{v}\ d\boldsymbol{x} \\ &= \delta t\, h_0 \sum_{i,j} \left(\pm \frac{4}{\delta x\, \delta y} \int_{V_{i,j}} (y - y_j) u_{i,j} + (x - x_i) v_{i,j}\ d\boldsymbol{x} \right) \\ &= 0 \quad , \end{aligned}$$

because the integral of both terms vanishes on $V_{i,j}$*, regardless of the value of* $\boldsymbol{v}_{i,j}$*.* ◁

The bilinear form $b_2(\cdot, \cdot)$ acts on functions defined on the primary discretization as well as on functions defined on the dual discretization, and therefore complicating its analysis. A possible strategy for a prove of the discrete inf-sup condition (4.15) is outlined as follows: Since each element of $\mathcal{Q}^h$ has a volumewise representation as given in (3.30), for $\psi \in \mathcal{Q}^h$ and $\boldsymbol{v} \in \mathcal{U}^h$ we can rewrite

$$b_2(\boldsymbol{v}, \psi) = b_2\left(\boldsymbol{v}, \sum_{\bar{V} \in \bar{\mathcal{V}}} \psi_{\bar{V}}\ \chi_{\bar{V}}\right) = \sum_{\bar{V} \in \bar{\mathcal{V}}} \psi_{\bar{V}}\ b_2(\boldsymbol{v}, \chi_{\bar{V}}) \quad .$$

Then, (4.15) is clearly satisfied, if for each $\psi \in \mathcal{Q}^h$ there exists a $\boldsymbol{v} \in \mathcal{U}^h$ with

$$b_2(\boldsymbol{v}, \chi_{\bar{V}}) = \psi_{\bar{V}} \quad . \tag{4.30}$$

This leads to a linear system for $u_{i,j}$, $u_{y,i,j}$, $v_{i,j}$ and $v_{x,i,j}$ $(i = 1, \ldots, n;\ j = 1, \ldots, m)$ that has to be analyzed in order to complete the prove.

Under the given time constraints a more profound analysis of the discrete problem was not possible. However, we have successfully established a mixed formulation equivalent to the second projection of the new scheme presented in Section 3.2. Using this formulation for the stability analysis of the projection, existence and uniqueness have been shown for the associated continuous saddle point problem. In the discrete case, one out of four discrete conditions on the three bilinear forms (4.27) in the formulation has been shown to hold.

5 Numerical Tests and Simulations

To illustrate the performance of the new projection method, the results of two test cases are presented. The main goal is to assess its accuracy and to compare it with the original scheme which rests on standard discretizations. In the first case, the second-order convergence of the method is demonstrated for smooth solutions. The second test deals with the translation of a vortex. Furthermore, we verify that our implementation is consistent with the theoretical estimate about the velocity divergence of the approximate projection method, which has been derived in Section 3.2.1.

For both test cases the exact solution of the particular problem is known, and the error of the numerical approximation can be computed. The computations are performed on a uniform Cartesian grid with equal grid spacing $\delta x = \delta y$. The boundary conditions are those discussed in Section 3.1.3. So far, we have only investigated the case of constant background height $h_0 \equiv 1$. Thus, in all calculations, the term dh_0/dt is set to zero. To start with initial data, which have zero divergence in the sense of (3.26), the given values for the momentum are corrected by the solution of a Poisson problem

$$D_{\bar{\mathcal{V}}}^{\mathcal{V}}\Big(h_0(0)\, G_{i,j}^{\bar{\mathcal{V}}}(h_{\bar{\mathcal{V}}}^{(2),0})\Big) = D_{\bar{\mathcal{V}}}^{\mathcal{V}}\Big(\overline{(h\boldsymbol{v})(x,y,0)}^{V_{i,j}}\Big)$$

for the initial height $h_{\bar{\mathcal{V}}}^{(2),0}$. Here, $\overline{(h\boldsymbol{v})}^{V_{i,j}}$ is the average of the exact solution $(h\boldsymbol{v})$ on $V_{i,j}$. The momentum distribution is then given by

$$(h\boldsymbol{v})_{i,j}^{0} = \overline{(h\boldsymbol{v})(x,y,0)}^{V_{i,j}} - h_0(0)\, G_{i,j}^{\bar{\mathcal{V}}}(h_{\bar{\mathcal{V}}}^{(2),0}) \quad .$$

A similar procedure is used for the new projection method with the operators $\mathsf{D}_{\bar{\mathcal{V}}}^{\mathcal{V}}$ and $\mathsf{G}_{\mathcal{V}}^{\bar{\mathcal{V}}}$ instead of $D_{\bar{\mathcal{V}}}^{\mathcal{V}}$ and $G_{\mathcal{V}}^{\bar{\mathcal{V}}}$.

As mentioned earlier, the auxiliary system is solved using an explicit standard second-order method for hyperbolic conservation laws. The stability of this method

strongly relies on a CFL time step restriction. In all the computations presented in this chapter, a time step has been chosen, which is $C = 0.8$ times smaller than the maximum allowed by the CFL condition.

The discrete divergence and gradient operators, which are used in the two elliptic correction steps, are those given in the Appendices B.1 and B.2 for the original projection method and in Appendix C.2 for the new projection method. The linear systems for computing the height $h^{(2)}$ on the primary and on the dual discretizations are solved using the Bi-CGSTAB algorithm [VAN DER VORST, 1992]. In each iteration, the Euclidean norm

$$\|a_{\mathcal{V}}\|_2 \coloneqq \left(\sum_{V \in \mathcal{V}} a_V^2 \right)^{1/2}$$

(similarly for $\|a_{\bar{\mathcal{V}}}\|_2$) of the residual vector

$$\begin{aligned} r_{\mathrm{P1}}\left(h_{\mathcal{V}}^{(2)}\right) &\coloneqq D_{\mathcal{V}}^{\mathcal{I}}((h\boldsymbol{v})_{\mathcal{I}}^{*}) - \frac{\delta t}{2} D_{\mathcal{V}}^{\mathcal{I}}\left(h_0^{n+1/4} G_{\mathcal{I}}^{\mathcal{V}}(h_{\mathcal{V}}^{(2)})\right) \\ r_{\mathrm{P2}}\left(h_{\bar{\mathcal{V}}}^{(2)}\right) &\coloneqq D_{\bar{\mathcal{V}}}^{\mathcal{V}}((h\boldsymbol{v})_{\mathcal{V}}^{**}) + D_{\bar{\mathcal{V}}}^{\mathcal{V}}((h\boldsymbol{v})_{\mathcal{V}}^{n}) - \delta t\, D_{\bar{\mathcal{V}}}^{\mathcal{V}}\left(h_0^{n+1/2} G_{\mathcal{V}}^{\bar{\mathcal{V}}}(h_{\bar{\mathcal{V}}}^{(2)})\right) \end{aligned}$$

is calculated. The algorithm is terminated when either this absolute value or the ratio between the norm of the current residual and that of the initial residual is less than 10^{-11}.

5.1 Convergence studies

The first test case demonstrates the second-order convergence of numerical solutions to the exact solution for smooth data. This test was originally proposed in MINION [1996] and ALMGREN ET AL. [1998] for the incompressible flow equations. Here it has been adapted for the zero Froude number shallow water equations.

For constant height h_0 and an initial velocity distribution

$$\begin{aligned} u_0(x,y) &= 1 - 2\cos(2\pi x)\sin(2\pi y) \\ v_0(x,y) &= 1 + 2\sin(2\pi x)\cos(2\pi y) \quad , \end{aligned}$$

the exact solution of the zero Froude number shallow water equations is given by

$$\begin{aligned} u(x,y,t) &= 1 - 2\cos(2\pi(x-t))\sin(2\pi(y-t)) \\ v(x,y,t) &= 1 + 2\sin(2\pi(x-t))\cos(2\pi(y-t)) \\ h^{(2)}(x,y,t) &= -\cos(4\pi(x-t)) - \cos(4\pi(y-t)) \quad . \end{aligned}$$

The problem is solved on the unit square with $(x, y) \in [0, 1] \times [0, 1]$ and periodic boundary conditions. The piecewise linear reconstruction of the momentum field components is done using central differences with no slope limiter.

The numerical solution is computed on three different grids with 32×32, 64×64 and 128×128 cells. We start the calculation at $t = 0$, and the error vector in the velocity $\boldsymbol{e}^N$ with elements

$$e^N_{i,j} \coloneqq \left| \overline{u(x,y,t^N)}^{V_{i,j}} - u^N_{i,j} \right| + \left| \overline{v(x,y,t^N)}^{V_{i,j}} - v^N_{i,j} \right|$$

is evaluated at time $t^N = 3$. This corresponds to approximately 735, 1460 and 2900 time steps, respectively. Note that we could have also incorporated the linear variation of the velocity on each cell in the error analysis of the new projection. We do not choose this alternative in favor of a better comparison with the original method. The global error is measured using the discrete L^2 norm and the L^∞ norm. These are defined by

$$\left\|\boldsymbol{e}^N\right\|_0 \coloneqq \left(\sum_{i,j} \left(|V_{i,j}|\, e^N_{i,j} \right)^2 \right)^{1/2} \quad \text{and} \quad \left\|\boldsymbol{e}^N\right\|_\infty \coloneqq \max_{i,j} \{e^N_{i,j}\} \quad .$$

We have summarized these error measures for the original projection method as well as for the new approximate and the new exact projection methods in Table 5.1. Additionally, the corresponding convergence rate γ is given, which is calculated by

$$\gamma \coloneqq \frac{\log(\left\|\boldsymbol{e}^N_c\right\| / \left\|\boldsymbol{e}^N_f\right\|)}{\log(\delta x_c / \delta x_f)} \quad . \tag{5.1}$$

In this definition, $\boldsymbol{e}^N_c$ and $\boldsymbol{e}^N_f$ are the computed error vectors of the solution on the coarse and the fine grid and δx_c and δx_f are the corresponding grid spacings. Clearly,

Method	Norm	32x32	Rate γ	64x64	Rate γ	128x128
original projection	L^2	0.292947	2.16	0.065641	2.16	0.014645
	L^∞	0.420732	2.15	0.094521	2.18	0.020871
new approximate projection	L^2	0.292943	2.16	0.065641	2.16	0.014645
	L^∞	0.420726	2.15	0.094521	2.18	0.020871
new exact projection	L^2	0.081603	2.64	0.013051	2.17	0.002898
	L^∞	0.127741	2.45	0.023417	2.32	0.004687

Table 5.1: Errors and convergence rates for the original and the new projection method.

second order accuracy is obtained in the L^2 as well as in the L^∞ norm. Also note that the absolute error obtained with the new exact projection is about four times smaller than the one obtained with the original method.

Table 5.2 shows the same calculations for the computations, in which the piecewise linear components of the velocity field are modified based on additional consistency considerations (cf. Section 3.3). Second-order accuracy is also retained for these cases. By solving an auxiliary equation for the vorticity field, similar results are obtained as in the case, in which the new exact projection is applied without correction. This is in contrast to the case, where the divergence constraint (2.25) is applied within a cell. The latter procedure produces an error of the same order as for the original projection.

Method	Norm	32x32	Rate γ	64x64	Rate γ	128x128
vorticity correction	L^2	0.070770	2.50	0.012498	1.94	0.003257
	L^∞	0.107118	2.47	0.019346	2.08	0.004585
divergence correction	L^2	0.331333	2.11	0.076824	2.29	0.015671
	L^∞	0.473625	2.11	0.109463	2.29	0.022382
both corrections	L^2	0.366079	1.82	0.103773	2.09	0.024291
	L^∞	0.518214	1.82	0.146918	2.09	0.034472

Table 5.2: Errors and convergence rates for the new exact projection method with correction based on additional consistency considerations.

5.2 Advection of a vortex

Let us consider the advection of a vortex by a constant background flow. For the implementation of this test case, originally proposed by Gresho and Chan [1990], a rectangular domain with size $[0, 4] \times [0, 1]$ is examined. The domain has periodic boundary conditions at the short sides and walls at the long sides. The initial conditions are defined to be

$$u(x, y, 0) = 1 - v_\theta(r) \sin\theta \quad \text{and} \quad v(x, y, 0) = v_\theta(r) \cos\theta \quad ,$$

in which

$$v_\theta(r) = \begin{cases} 5r\, v_{\max} & \text{for } 0 \leq r < \frac{1}{5} \\ (2 - 5r)\, v_{\max} & \text{for } \frac{1}{5} \leq r < \frac{2}{5} \\ 0 & \text{for } \frac{2}{5} \leq r \end{cases} \tag{5.2}$$

and

$$r = \sqrt{\left(x - \tfrac{1}{2}\right)^2 + \left(y - \tfrac{1}{2}\right)^2} \quad .$$

In equation (5.2) $v_{\max}$ is the maximum tangential velocity of the vortex. The height $h^{(2)}$ must then satisfy the constraint $\partial_r h^{(2)} = v_\theta^2 / r$. This relationship is visualized in Figure 5.1.

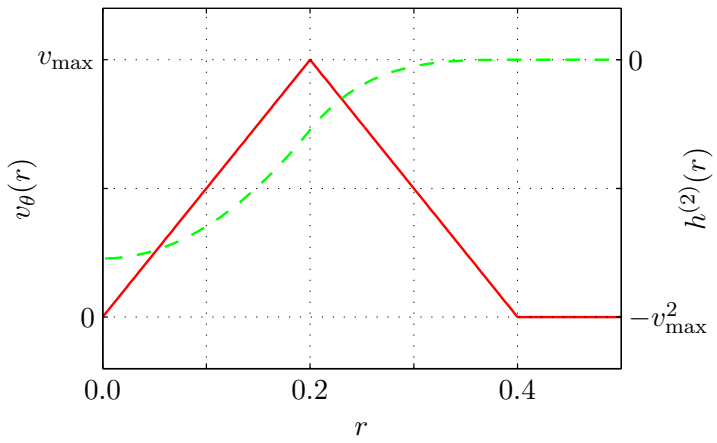

Figure 5.1: Advection of a vortex: tangential velocity (solid red) and height profile (dashed green) with respect to the distance r from the center of the vortex.

The test is set up with $v_{\max} = 1$ and background height $h_0 \equiv 1$. The computational domain consists of 80×20 grid cells. Three different strategies for the linear reconstruction of the components of the momentum variable are investigated. In particular, we consider central differences (no limiter), the *monotonized central difference (MC)* limiter and *Sweby's* limiter [Schulz-Rinne, 1993] with $k = 1.8$, the latter being a convex combination of the *minmod* ($k = 1$) and the *superbee* limiter ($k = 2$).

The results for the original scheme are given in Figure 5.2, in which the streamfunction of the velocity distribution is displayed at four different times of the simulation.

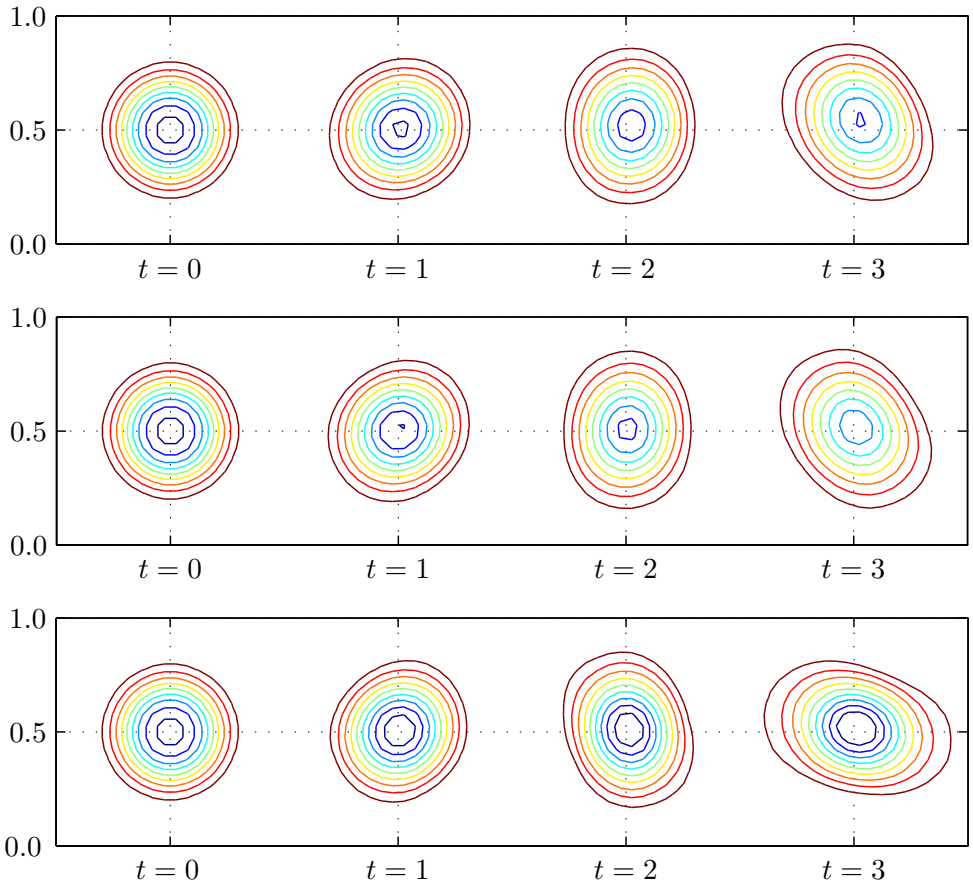

Figure 5.2: Advection of a vortex at times $t = 0, 1, 2$ and 3 for the original method. Contour lines of the streamfunction are shown at [-0.02, -0.04, ..., -0.18] starting from outside of the vortex. Top: unlimited slopes, middle: monotonized central difference (MC) limiter, bottom: Sweby's limiter ($k = 1.8$).

Similar to the results in SCHNEIDER ET AL. [1999] for the incompressible Euler equations, the core is advected almost along the center line of the channel. Also, the vortex experiences a considerable deformation due to the coarse discretization we have chosen for this test.

As in the convergence studies, the new exact projection method shows a significant improvement in the numerical results for this test (cf. Figure 5.3). All reconstruction strategies show less deviation from the center line of the channel than in the original method. Furthermore, the loss in vorticity is slightly reduced. These features become more evident in Figure 5.4, which shows the advected vortex at time $t = 10$ for the case of unlimited slopes. The results of the new approximate projection method (not shown) are comparable to the ones of the original projection.

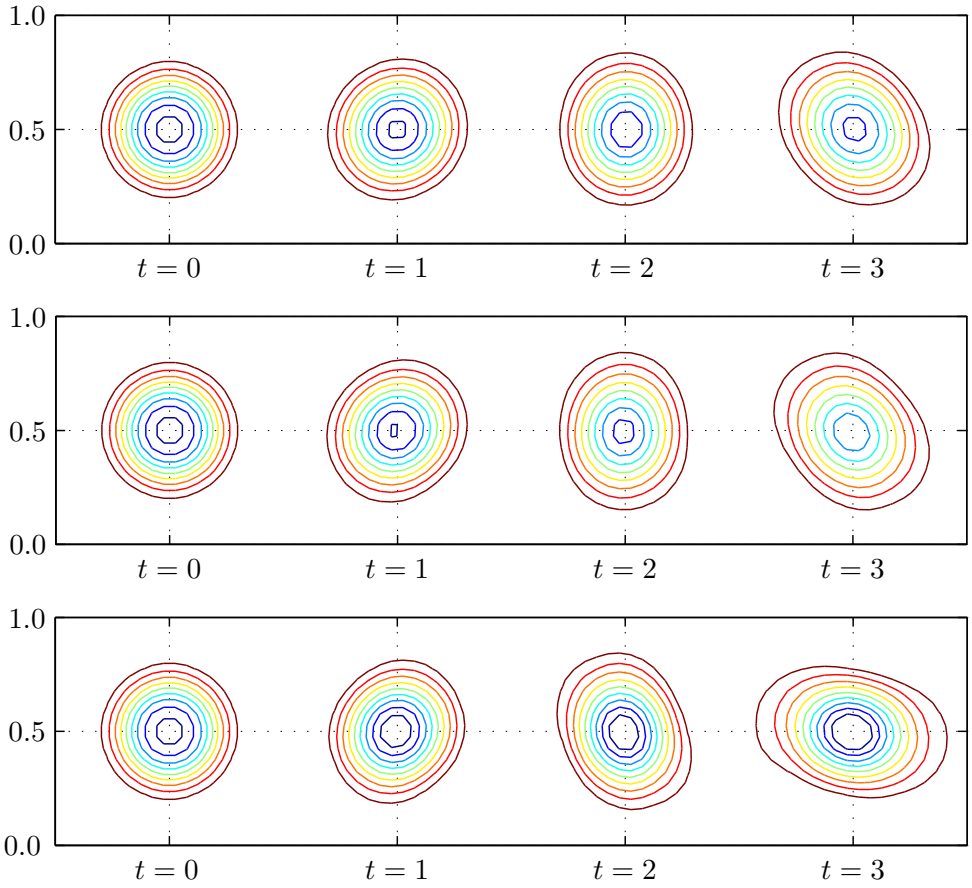

Figure 5.3: Same as Figure 5.2 for the new exact projection method.

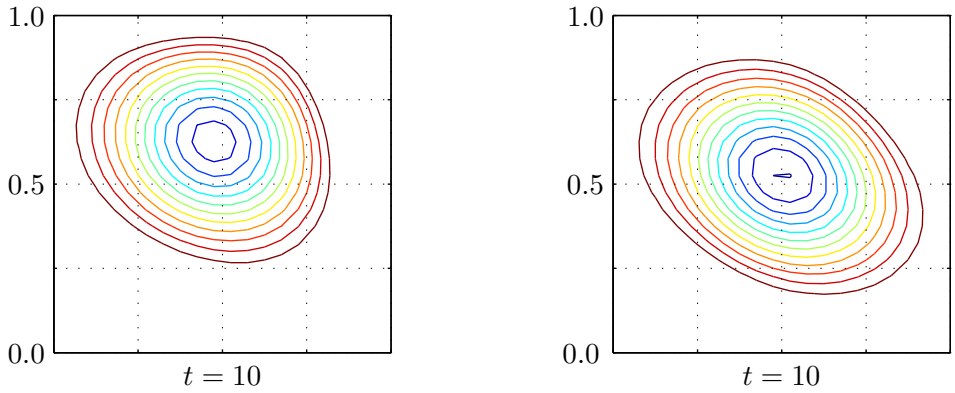

Figure 5.4: Advection of a vortex at time $t = 10$ for the original (left) and the new exact projection method (right), unlimited slopes. Contour lines of the streamfunction are shown at [-0.02, -0.03, . . . , -0.13].

Figure 5.5 presents the results, in which the outcome of the new second projection is corrected based on the additional consistency considerations. For these cases, only the computations with unlimited slopes are displayed. Using the advected vorticity as a constraint on the piecewise linear velocity components, the vortex becomes highly distorted and the maximum of the vorticity is increased (from initially 0.188 to 0.218 at time $t = 3$). Unlike the vorticity correction, the correction due to the application of the divergence constraint within a cell does not affect the quality of the solution in this case. The combination of both corrections mostly reproduces the behavior of the computation in which only the advected vorticity was used for the correction.

5.3 Divergence of the new approximate projection

To verify the consistency of our implementation with Lemma 3.2 concerning the behavior of the velocity divergence in the approximate projection method, we reuse the test case described in Section 5.1 for the convergence studies. This time, numerical solutions for four different grids with 32×32, 64×64, 128×128 and 256×256 cells are computed on the unit square after one time step at t^1. The initial velocity field has been corrected using the original projection method. Note that this yields a zero divergence field with piecewise constant vector functions in terms of the new projection as well. The divergence of the resulting velocity field is measured with

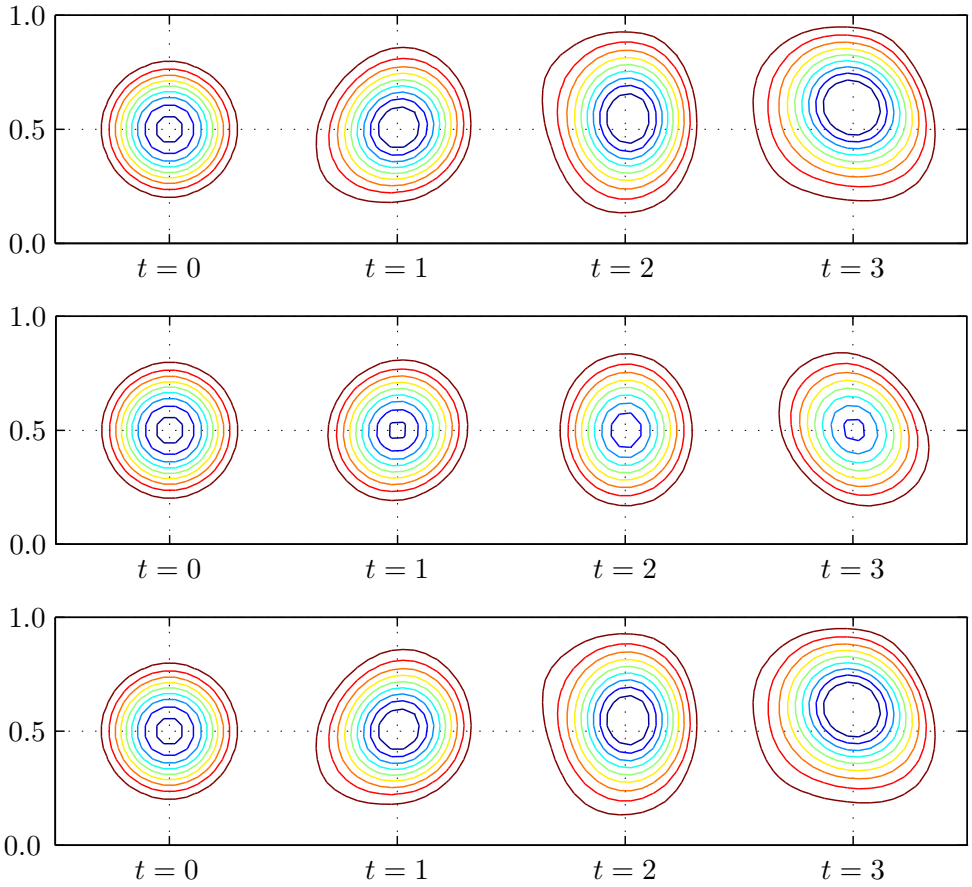

Figure 5.5: Same as Figure 5.2 for the new exact projection method with correction based on additional consistency considerations, unlimited slopes. Top: vorticity correction, middle: divergence correction, bottom: both corrections.

Norm	32x32	Rate μ	64x64	Rate μ	128x128	Rate μ	256x256
L^2	4.504e-05	3.84	3.152e-06	3.54	2.713e-07	3.45	2.474e-08
L^∞	1.839e-04	2.54	3.163e-05	2.84	4.418e-06	2.94	5.742e-07

Table 5.3: L^2 and L^∞ norm of the divergence in the new approximate projection method. Additionally, the convergence rates μ for $\delta x \sim \delta t \to 0$ are given.

the L^2 and the L^∞ norm, respectively. These values are given in Table 5.3. For the estimation of how the divergence behaves for $\delta x \sim \delta t \to 0$, the rates

$$\mu \coloneqq \frac{\log(\|\nabla \cdot \boldsymbol{v}_c^1\| \,/\, \|\nabla \cdot \boldsymbol{v}_f^1\|)}{\log(\delta x_c / \delta x_f)} \quad ,$$

are calculated similarly to (5.1).

In both norms, the asymptotic behavior of the velocity divergence approaches third order rates as the grid spacing goes to zero. This is consistent with the theoretical results derived in Section 3.2.1.

6 Discussion

In this thesis, we have introduced a new projection method for the zero Froude number shallow water equations. The method is based on the solution of two Poisson-type equations using a Petrov-Galerkin finite element formulation with piecewise bilinear ansatz functions for the unknown height variable. This discretization naturally leads to a piecewise linear approximation for the velocity variable.

The following section provides a discussion about the specific properties of the two different versions of the new projection method. In particular, the numerical results from Chapter 5 are analyzed in more detail. The stability of the new projection is discussed in Section 6.2, and directions for its further analysis are proposed. We conclude this chapter with an outlook for possible future research paths on this topic.

6.1 Comparison of the different methods

The convergence test in Section 5.1 and the test about the advection of a vortex in Section 5.2 reveal two major results. First, the approximate method, described in Section 3.2.1, and the "original" scheme, which rests on standard discretizations for the solution of the elliptic equations, yield almost indistinguishable results. By the choice of an approximate method, no deteriorations could be observed in the computational results. Second, the new projection method (cf. Section 3.2.2) shows results with considerable improvements in the accuracy.

The numerical evidence is supported by the theoretical analysis of the new projection method. The discrete gradient of the new method does not result in a local decoupling as described on page 36 for the original second projection. This was the main motivation for the development of the new method. Further results have been derived in the course of this work:

It has been outlined, that the approximate projection can be also seen as an exact projection followed by a L^2 projection onto the space of piecewise constant vector functions. This interpretation together with the upper bound for the divergence at the new time level, given in Lemma 3.2, characterize the "approximateness" of the method and show stability of the approximation compared to the outcome of the exact projection. However, approximate projection methods have been also found to produce poor results when applied to "difficult" problems [Almgren et al., 2000]. It has to be analyzed, what are the consequences of the unprecise control of the discrete divergence in the case of the presented method. In the exact projection method, the gradients of the momentum components, which are computed in the second projection of the method, are used for the calculation of the numerical fluxes of the auxiliary system at the new time level. This is done in order to obtain an exact projection method, but with this ansatz, the TVD property of the reconstruction is lost. This is a delicate issue, because it concerns the stability of the method for the solution of the auxiliary system.

We have outlined in Section 3.2.3 that the discretization for the new projection can be also used for the first projection of the method, yielding a unified discretization for both Poisson-type problems. Furthermore, the linear systems associated with the Poisson-type equations, can be solved with the same algorithms that are used for the original discretizations. These facts support the application of the new discretization instead of the old approach.

The numerical simulations, in which the additional consistency constraints were imposed on the partial derivatives of the velocity components within a cell, show no further improvement of the method. On the contrary, in the convergence test the application of the divergence constraint within a cell deteriorates the quality of the numerical solution, yielding a performance comparable to the original method (cf. Table 5.2). In the second test, in which a vortex with discontinuous vorticity distribution is advected, the results are considerably worse, when the advected vorticity constraints the reconstruction (cf. Figure 5.5). These results do not allow to draw clear-cut conclusions and the approach has to be further analyzed. The outcome of the second test case with correction based on the vorticity equation might be due to a poor representation of the vorticity advection. This is supported by the result

of another simulation, in which the gradients within a cell were corrected using the analytically derived vorticity (not shown), and that resulted in considerably better results.

6.2 The question of stability

The numerical results presented in Chapter 5 suggest that the new numerical method is stable and converges towards smooth solutions of the zero Froude number limit of the shallow water equations with second order accuracy. Of course, we would like to prove that this is indeed the case. For this, some form of stability is needed. In this thesis, we propose a formulation of the stability problem by deriving a *mixed Petrov-Galerkin* finite element formulation (cf. equation (4.28)), which is equivalent to the Poisson-type problem (3.23) of our semi-implicit method. Using the theory of Nicolaïdes [1982], we have proven that the associated continuous saddle point problem has a unique solution. To do this, three inf-sup conditions (and one additional condition for $a(\cdot,\cdot)$) have been derived for the bilinear forms $a(\cdot,\cdot)$ and $b_i(\cdot,\cdot)$ $(i = 1, 2)$, given in (4.27).

The finite element spaces that have to be used for the *discrete* mixed formulation, lead to a nonconforming method, and thus complicating the convergence analysis. In the stability analysis, only the inf-sup condition for the bilinear form $b_1(\cdot,\cdot)$ could be proven. The conditions on $a(\cdot,\cdot)$ and the inf-sup condition on $b_2(\cdot,\cdot)$ have to be further investigated. For the condition on $a(\cdot,\cdot)$, the subspaces $\mathcal{K}_1^h$ and $\mathcal{K}_2^h$ have to be specified. A possible solution for this purpose is to follow Micheletti and Sacco [2001], who have formulated a discrete Helmholtz decomposition principle for a similar generalized saddle point problem. This approach leads to the same argumentation that has been used in the proof for existence and uniqueness of our continuous saddle point problem. Furthermore, since the discrete finite element space for the momentum is not contained in the function space, which has been used in the continuous problem (i.e. $\mathcal{U}^h \not\subseteq \mathcal{U}$), a suitable mesh dependent norm has to be defined on $\mathcal{U}^h$. Because the second projection of the original method, interpreted as a mixed formulation, turned out to be unstable in this analysis (cf. Remark 4.5), we consider the new projection to be more stable.

Other approaches have been investigated in order to find an appropriate formulation equivalent to the Poisson-type problem (3.37). Following Causin [2002], it was attempted to use the technique of hybrid finite element methods. These were used to relax the constraints implied by the continuous function spaces and to obtain a formulation with piecewise linear functions for the discrete momentum components. In contrast to Causin, who investigated only one discretization, our problem is based on a primary discretization, which consists of the cells of the given grid, as well as a dual discretization with node centered control volumes, making it difficult to adapt the approach for our purposes.

6.3 Conclusion and future research prospects

The new semi-implicit projection method yields significant accuracy enhancements compared to the original scheme. Both, the numerical results as well as the theoretical analyses suggest that the new method has better stability properties. The mixed finite element formulation that we have derived, provides the necessary analytical framework for a stability proof.

The applied test cases were straightforward, and in the future the behavior of the method in more complex simulations has to be tested. For the stability analysis of the method, also other approaches should be considered. The finite element space for the approximation of the velocity variable is based on local ansatz functions, whose support is only one grid cell. This approach is also utilized in *discontinuous Galerkin* and *interior penalty* methods [see e.g. Arnold et al., 1998]. It would be interesting to apply analyses derived for these methods to our approach.

Besides a stability and convergence proof of the new method, there are several other open research paths. An extension of the scheme to the low Froude number regime is certainly desirable. Such a method would be capable of computing free-surface waves on the ocean. Similar extensions have been proposed by Le Maître et al. [2001] and for the weakly compressible (low Mach number) Euler equations by Geratz [1997], Munz et al. [2003] and recently by Klein and Geratz [2004]. To explicitly account for the results of the asymptotic analysis from Chapter 2, we suggest to consider the application of *multigrid methods* in such an approach.

Another possible research direction concerns the piecewise linear ansatz functions of the velocity space in the second projection of our scheme. The proposed exact projection method consists of a Godunov-type finite volume solver for the solution of the auxiliary system and the initial gradients are determined by a reconstruction based on local cell averages. We proposed two approaches to account for the evolution of the gradients in Section 3.3. Similar ideas are also pursued in *discontinuous Galerkin* methods for convection dominated problems [see e.g. COCKBURN, 1999], and it would be worthwhile to study these approaches in order to improve the new scheme.

Currently, our method is only formulated for Cartesian grids. For the application of the scheme to realistic geometries, the scheme should be extended to more general grids (e.g. triangulations). The finite element formulation of the second projection promises to be a first step into this direction, because its formulation is more flexible. An equivalent version for a discretization consisting of triangular elements might consist of piecewise linear instead of piecewise bilinear elements for the unknown $h^{(2)}$.

Appendix A

The Role of the Auxiliary System

A.1 Error of the predicted variables

Let us assume that $(h, \boldsymbol{v})^*(\boldsymbol{x}, t_0 + \delta t)$ is a smooth solution of the auxiliary system (2.26) with initial values $(h, \boldsymbol{v})^*(\boldsymbol{x}, t_0)$. With the additional constraints

$$\nabla \cdot \boldsymbol{v}^*(\boldsymbol{x}, t_0) = 0 \quad \text{and} \quad \nabla h^*(\boldsymbol{x}, t_0) = \mathbf{0}$$

it follows that

$$\begin{aligned} h_t^*(\boldsymbol{x}, t_0) &= -\nabla \cdot ((h\boldsymbol{v})^*(\boldsymbol{x}, t_0)) \\ &= -h^*(\boldsymbol{x}, t_0) \underbrace{\nabla \cdot \boldsymbol{v}^*(\boldsymbol{x}, t_0)}_{=0} - \boldsymbol{v}^*(\boldsymbol{x}, t_0) \cdot \underbrace{\nabla h^*(\boldsymbol{x}, t_0)}_{=\mathbf{0}} = 0 \quad . \end{aligned}$$

Expanding $\boldsymbol{v}^*(\boldsymbol{x}, t)$ and $h^*(\boldsymbol{x}, t)$ about t_0 yields

$$\nabla \cdot \boldsymbol{v}^*(\boldsymbol{x}, t_0 + \delta t) = \nabla \cdot \left(\boldsymbol{v}^*(\boldsymbol{x}, t_0) + \delta t\, \boldsymbol{v}_t^*(\boldsymbol{x}, t_0) + \mathcal{O}\left(\delta t^2\right)\right) = \mathcal{O}(\delta t)$$

and

$$\nabla h^*(\boldsymbol{x}, t_0 + \delta t) = \nabla \left(h^*(\boldsymbol{x}, t_0) + \delta t\, h_t^*(\boldsymbol{x}, t_0) + \mathcal{O}\left(\delta t^2\right)\right) = \mathcal{O}\left(\delta t^2\right) \quad .$$

Finally, from $h^*(\boldsymbol{x}, t) = \mathcal{O}(1)$ we obtain

$$(h^* \nabla h^*)(\boldsymbol{x}, t_0 + \delta t) = \mathcal{O}\left(\delta t^2\right) \quad . \tag{A.1}$$

A.2 Relationship to the unknown variables

If the zero Froude number shallow water equations (2.23) are applied to the same flow field as the auxiliary system, the difference between the changes in momentum is

$$((h\boldsymbol{v})_t - (h\boldsymbol{v})_t^*)(x,t_0) = -(h\nabla h^{(2)})(x,t_0) \quad . \tag{A.2}$$

Similarly we get for the difference between the changes of the velocity fields

$$(\boldsymbol{v}_t - \boldsymbol{v}_t^*)(x,t_0) = -\nabla h^{(2)}(x,t_0) \quad . \tag{A.3}$$

To obtain a representation of the momentum at a half time step $t_0 + \delta t/2$, a Taylor series expansion of $(h\boldsymbol{v})$ and $(h\boldsymbol{v})^*$ is performed about $(x, t_0 + \delta t/2)$:

$$(h\boldsymbol{v})(x,t_0+\tfrac{\delta t}{2}) = (h\boldsymbol{v})(x,t_0) + \frac{\delta t}{2}\frac{\partial}{\partial t}(h\boldsymbol{v})(x,t_0) + \frac{\delta t^2}{8}\frac{\partial^2}{\partial t^2}(h\boldsymbol{v})(x,t_0) + \mathcal{O}\big(\delta t^3\big)$$

$$(h\boldsymbol{v})^*(x,t_0+\tfrac{\delta t}{2}) = (h\boldsymbol{v})^*(x,t_0) + \frac{\delta t}{2}\frac{\partial}{\partial t}(h\boldsymbol{v})^*(x,t_0) + \frac{\delta t^2}{8}\frac{\partial^2}{\partial t^2}(h\boldsymbol{v})^*(x,t_0) + \mathcal{O}\big(\delta t^3\big)$$

Using (A.2), the combination of these two expansions yields

$$\begin{aligned}(h\boldsymbol{v})(x,t_0+\delta t/2) \;\; = \;\; & (h\boldsymbol{v})^*(x,t_0+\delta t/2) - \frac{\delta t}{2}(h\nabla h^{(2)})(x,t_0) \\ & -\frac{\delta t^2}{8}\left(\frac{\partial}{\partial t}(h\nabla h^{(2)})(x,t_0)\right) + \mathcal{O}\big(\delta t^3\big) \quad .\end{aligned}$$

Finally, another Taylor series expansion of $(h\nabla h^{(2)})$ about (x,t_0) yields

$$(h\boldsymbol{v})(x,t_0+\delta t/2) = (h\boldsymbol{v})^*(x,t_0+\delta t/2) - \frac{\delta t}{2}(h\nabla h^{(2)})(x,t_0+\delta t/4) + \mathcal{O}\big(\delta t^3\big) \quad .$$

The same procedure applied to the velocities of the two systems together with (A.3) leads to

$$\boldsymbol{v}(x,t_0+\delta t/2) = \boldsymbol{v}^*(x,t_0+\delta t/2) - \frac{\delta t}{2}\nabla h^{(2)}(x,t_0+\delta t/4) + \mathcal{O}\big(\delta t^3\big) \quad .$$

Appendix B

Discretization of the Original Projections

The discrete gradients and divergences of the two projections from Section 3.1 are given for a two-dimensional Cartesian grid with constant grid spacings δx and δy.

B.1 First projection

The double index (i, j) is used to refer to a cell value, while the indices $(i + 1/2, j)$ and $(i, j + 1/2)$ are used for interface values between the cells (i, j)-$(i + 1, j)$ and (i, j)-$(i, j + 1)$, respectively. In the original first projection, the gradient $G_{\mathcal{I}}^{\mathcal{V}}$ is given by

$$G_{I_{i+1/2,j}}^{\mathcal{V}}(p_{\mathcal{V}}) \coloneqq \begin{pmatrix} \dfrac{p_{i+1,j} - p_{i,j}}{\delta x} \\ \dfrac{p_{i,j+1} - p_{i,j-1} + p_{i+1,j+1} - p_{i+1,j-1}}{4\delta y} \end{pmatrix}$$

and

$$G_{I_{i,j+1/2}}^{\mathcal{V}}(p_{\mathcal{V}}) \coloneqq \begin{pmatrix} \dfrac{p_{i+1,j} - p_{i-1,j} + p_{i+1,j+1} - p_{i-1,j+1}}{4\delta x} \\ \dfrac{p_{i,j+1} - p_{i,j}}{\delta y} \end{pmatrix} .$$

According to (3.14), the discrete divergence $D_{\mathcal{V}}^{\mathcal{I}}$ is defined as

$$D_{V_{i,j}}^{\mathcal{I}}(\boldsymbol{v}_{\mathcal{I}}) \coloneqq \frac{u_{i+1/2,j} - u_{i-1/2,j}}{\delta x} + \frac{v_{i,j+1/2} - v_{i,j-1/2}}{\delta y}$$

with $\boldsymbol{v}_{\mathcal{I}} \coloneqq (u_{\mathcal{I}}, v_{\mathcal{I}})$. With these definitions $D^{\mathcal{I}}_{\mathcal{V}} G^{\mathcal{V}}_{\mathcal{I}}$ is the standard 5-points Laplacian (cf. Figure 3.2)

$$D^{\mathcal{I}}_{V_{i,j}}\left(G^{\mathcal{V}}_{\mathcal{I}}\left(p_{\mathcal{V}}\right)\right) \coloneqq \frac{p_{i+1,j} - 2p_{i,j} + p_{i-1,j}}{\delta x^2} + \frac{p_{i,j+1} - 2p_{i,j} + p_{i,j-1}}{\delta y^2} \quad .$$

B.2 Second projection

Additionally to the notation from the previous section the double index $(i+1/2, j+1/2)$ is used for node values. The indices $(i+1, j+1/2)$ and $(i+1/2, j+1)$ are used for interface values of the dual discretization, which are between the control volumes $(i+1/2, j+1/2)$-$(i+3/2, j+1/2)$ and $(i+1/2, j+1/2)$-$(i+1/2, j+3/2)$, respectively. The linear operators $L^{\bar{\mathcal{V}}}_{\mathcal{I}}(p_{\bar{\mathcal{V}}})$ from (3.19) and $L^{\mathcal{V}}_{\bar{\mathcal{I}}}(\boldsymbol{v}_{\mathcal{V}})$ from (3.21) are defined as follows

$$\begin{aligned}
L^{\bar{\mathcal{V}}}_{I_{i+1/2,j}}\left(p_{\bar{\mathcal{V}}}\right) &\coloneqq \frac{1}{2}\left(p_{i+1/2,j+1/2} + p_{i+1/2,j-1/2}\right) \\
L^{\bar{\mathcal{V}}}_{I_{i,j+1/2}}\left(p_{\bar{\mathcal{V}}}\right) &\coloneqq \frac{1}{2}\left(p_{i-1/2,j+1/2} + p_{i+1/2,j+1/2}\right) \\
L^{\mathcal{V}}_{\bar{I}_{i+1,j+1/2}}\left(\boldsymbol{v}_{\mathcal{V}}\right) &\coloneqq \frac{1}{2}\left(\boldsymbol{v}_{i+1,j+1} + \boldsymbol{v}_{i+1,j}\right) \\
L^{\mathcal{V}}_{\bar{I}_{i+1/2,j+1}}\left(\boldsymbol{v}_{\mathcal{V}}\right) &\coloneqq \frac{1}{2}\left(\boldsymbol{v}_{i,j+1} + \boldsymbol{v}_{i+1,j+1}\right) \quad .
\end{aligned}$$

With these definitions the discrete gradient $G^{\bar{\mathcal{V}}}_{\mathcal{V}}$ is defined by

$$\begin{aligned}
G^{\bar{\mathcal{V}}}_{V_{i,j}}\left(p_{\bar{\mathcal{V}}}\right) &= \begin{pmatrix} \dfrac{L^{\bar{\mathcal{V}}}_{I_{i+1/2,j}}(p_{\bar{\mathcal{V}}}) - L^{\bar{\mathcal{V}}}_{I_{i-1/2,j}}(p_{\bar{\mathcal{V}}})}{\delta x} \\ \dfrac{L^{\bar{\mathcal{V}}}_{I_{i,j+1/2}}(p_{\bar{\mathcal{V}}}) - L^{\bar{\mathcal{V}}}_{I_{i,j-1/2}}(p_{\bar{\mathcal{V}}})}{\delta y} \end{pmatrix} \\
&= \begin{pmatrix} \dfrac{p_{i+1/2,j+1/2} - p_{i-1/2,j+1/2} + p_{i+1/2,j-1/2} - p_{i-1/2,j-1/2}}{2\delta x} \\ \dfrac{p_{i+1/2,j+1/2} - p_{i+1/2,j-1/2} + p_{i-1/2,j+1/2} - p_{i-1/2,j-1/2}}{2\delta y} \end{pmatrix} \quad .
\end{aligned} \tag{B.1}$$

The divergence $D^{\mathcal{V}}_{\bar{\mathcal{V}}}$ is

$$D^{\mathcal{V}}_{\bar{V}_{i+1/2,j+1/2}}(\boldsymbol{v}_{\mathcal{V}}) = \frac{L^{\mathcal{V}}_{\bar{I}_{i+1,j+1/2}}(u_{\mathcal{V}}) - L^{\mathcal{V}}_{\bar{I}_{i,j+1/2}}(u_{\mathcal{V}})}{\delta x} + \frac{L^{\mathcal{V}}_{\bar{I}_{i+1/2,j+1}}(v_{\mathcal{V}}) - L^{\mathcal{V}}_{\bar{I}_{i+1/2,j}}(v_{\mathcal{V}})}{\delta y}$$
$$= \frac{u_{i+1,j+1} - u_{i,j+1} + u_{i+1,j} - u_{i,j}}{2\delta x} + \frac{v_{i+1,j+1} - v_{i+1,j} + v_{i,j+1} - v_{i,j}}{2\delta y} \ .$$

With the above definitions $D^{\mathcal{V}}_{\bar{\mathcal{V}}}(G^{\bar{\mathcal{V}}}_{\mathcal{V}}(p_{\bar{\mathcal{V}}}))$ is the standard 9-points Laplacian

$$D^{\mathcal{V}}_{\bar{V}_{i+1/2,j+1/2}}\left(G^{\bar{\mathcal{V}}}_{\mathcal{V}}(p_{\bar{\mathcal{V}}})\right) = \frac{1}{4}\frac{\delta x^2 + \delta y^2}{\delta x^2 \delta y^2} a_{i+1/2,j+1/2} - \frac{1}{2}\frac{\delta x^2 - \delta y^2}{\delta x^2 \delta y^2} b_{i+1/2,j+1/2} \ ,$$

where

$$\begin{aligned} a_{i+1/2,j+1/2} &\coloneqq p_{i+3/2,j+3/2} + p_{i-1/2,j+3/2} + p_{i-1/2,j-1/2} + p_{i+3/2,j-1/2} - 4p_{i+1/2,j+1/2} \\ b_{i+1/2,j+1/2} &\coloneqq p_{i+3/2,j+1/2} - p_{i+1/2,j+3/2} + p_{i-1/2,j+1/2} - p_{i+1/2,j-1/2} \quad . \end{aligned}$$

For $\delta x = \delta y$, the second term on the right hand side of the discrete Laplacian disappears and the stencil of $D^{\mathcal{V}}_{\bar{\mathcal{V}}}(G^{\bar{\mathcal{V}}}_{\mathcal{V}}(p_{\bar{\mathcal{V}}}))$ reduces to a 5-points diagonal stencil (cf. Figure 3.2).

Appendix C

The New Projection

C.1 Basis functions for the scalar trial space

An element of $\mathcal{H}^h$ can be written as

$$p(x,y) = \sum_{\bar{V}\in\bar{\mathcal{V}}} p_{\bar{V}}\varphi_{\bar{V}}(x,y)$$

in which $\varphi_{\bar{V}}$ is a basis for this discrete finite element space given by (cf. Figure 3.4)

$$\varphi_{\bar{V}_{i+1/2,j+1/2}} = \begin{cases} (x-x_{i-1/2})(y-y_{j-1/2}) \\ \qquad\qquad \text{for } (x,y)\in [x_{i-1/2},x_{i+1/2}[\times[y_{j-1/2},y_{j+1/2}[\ , \\ (y-y_{j-1/2}) - (x-x_{i-1/2})(y-y_{j-1/2}) \\ \qquad\qquad \text{for } (x,y)\in [x_{i+1/2},x_{i+3/2}[\times[y_{j-1/2},y_{j+1/2}[\ , \\ (x-x_{i-1/2}) - (x-x_{i-1/2})(y-y_{j-1/2}) \\ \qquad\qquad \text{for } (x,y)\in [x_{i-1/2},x_{i+1/2}[\times[y_{j+1/2},y_{j+3/2}[\ , \\ 1-(x-x_{i-1/2}) - (y-y_{j-1/2}) + (x-x_{i-1/2})(y-y_{j-1/2}) \\ \qquad\qquad \text{for } (x,y)\in [x_{i+1/2},x_{i+3/2}[\times[y_{j+1/2},y_{j+3/2}[\ , \\ 0 \qquad\qquad \text{elsewhere.} \end{cases}$$

C.2 Discretization of the new projection

With the same notation as for the original method, the discretization of the operators in the new projection is given in the following. Only, the case for the second projection is given. The operators for the first projection are derived by shifting the indices by one half. Let us define

$$\begin{aligned} p_{x,i,j} &\coloneqq \frac{1}{\delta x}\left(p_{i+1/2,j+1/2} - p_{i-1/2,j+1/2} + p_{i+1/2,j-1/2} - p_{i-1/2,j-1/2}\right) \\ p_{y,i,j} &\coloneqq \frac{1}{\delta y}\left(p_{i+1/2,j+1/2} - p_{i+1/2,j-1/2} + p_{i-1/2,j+1/2} - p_{i-1/2,j-1/2}\right) \\ p_{xy,i,j} &\coloneqq \frac{1}{\delta x\,\delta y}\left(p_{i+1/2,j+1/2} - p_{i-1/2,j+1/2} - p_{i+1/2,j-1/2} + p_{i-1/2,j-1/2}\right) \quad . \end{aligned}$$

The discrete gradient $\mathsf{G}_{\mathcal{V}}^{\bar{\mathcal{V}}}$ is then given by

$$\mathsf{G}_{V_{i,j}}^{\bar{\mathcal{V}}}\left(p_{\bar{\mathcal{V}}}\right) = \begin{pmatrix} p_{x,i,j} \\ \\ p_{y,i,j} \end{pmatrix} + \begin{pmatrix} y - y_j \\ \\ x - x_i \end{pmatrix} p_{xy,i,j} \quad .$$

The divergence $\mathsf{D}_{\bar{\mathcal{V}}}^{\mathcal{V}}$ is defined by

$$\begin{aligned} \mathsf{D}_{\bar{V}_{i+1/2,j+1/2}}^{\mathcal{V}}\left(\boldsymbol{v}_{\mathcal{V}}\right) = {} & \frac{1}{2\,\delta x}\left(u_{i+1,j+1} - u_{i,j+1} + u_{i+1,j} - u_{i,j}\right) + \\ & \frac{\delta y}{8\,\delta x}\left(-u_{y,i+1,j+1} + u_{y,i,j+1} + u_{y,i+1,j} - u_{y,i,j}\right) + \\ & \frac{1}{2\,\delta y}\left(v_{i+1,j+1} - v_{i+1,j} + v_{i,j+1} - v_{i,j}\right) + \\ & \frac{\delta x}{8\,\delta y}\left(-v_{x,i+1,j+1} + v_{x,i+1,j} + v_{x,i,j+1} - v_{x,i,j}\right) \quad . \end{aligned} \tag{C.1}$$

With the above definitions $\mathsf{D}^{\mathcal{V}}_{\bar{\mathcal{V}}}(\mathsf{G}^{\bar{\mathcal{V}}}_{\mathcal{V}}(p_{\bar{\mathcal{V}}}))$ is the 9-points Laplacian proposed by Süli [1991] (cf. Figure 3.5):

$$\begin{aligned}\mathsf{L}^{\bar{\mathcal{V}}}_{\bar{V}_{i+1/2,j+1/2}}(p_{\bar{\mathcal{V}}}) &= \mathsf{D}^{\mathcal{V}}_{\bar{V}_{i+1/2,j+1/2}}\Big(\mathsf{G}^{\bar{\mathcal{V}}}_{\mathcal{V}}(p_{\bar{\mathcal{V}}})\Big) \\ &= \frac{1}{8}\left(\triangle_{xx,i+1/2,j+3/2}(p_{\bar{\mathcal{V}}}) + 6\triangle_{xx,i+1/2,j+1/2}(p_{\bar{\mathcal{V}}}) + \triangle_{xx,i+1/2,j-1/2}(p_{\bar{\mathcal{V}}})\right) + \\ &\quad \frac{1}{8}\left(\triangle_{yy,i+3/2,j+1/2}(p_{\bar{\mathcal{V}}}) + 6\triangle_{yy,i+1/2,j+1/2}(p_{\bar{\mathcal{V}}}) + \triangle_{yy,i-1/2,j+1/2}(p_{\bar{\mathcal{V}}})\right)\end{aligned}$$

with

$$\begin{aligned}\triangle_{xx,i+1/2,j+1/2}(p_{\bar{\mathcal{V}}}) &\coloneqq \frac{1}{\delta x^2}\left(p_{i+3/2,j+1/2} - 2p_{i+1/2,j+1/2} + p_{i-1/2,j+1/2}\right) \\ \triangle_{yy,i+1/2,j+1/2}(p_{\bar{\mathcal{V}}}) &\coloneqq \frac{1}{\delta y^2}\left(p_{i+1/2,j+3/2} - 2p_{i+1/2,j+1/2} + p_{i+1/2,j-1/2}\right) \quad .\end{aligned}$$

Bibliography

Almgren, A. S., Bell, J. B., Colella, P., Howell, L. H., and Welcome, M. L. [1998]. A Conservative Adaptive Projection Method for the Variable Density Incompressible Navier-Stokes Equations. *Journal of Computational Physics*, 142 (1): pp. 1–46. 72

Almgren, A. S., Bell, J. B., and Crutchfield, W. Y. [2000]. Approximate Projection Methods: Part I. Inviscid Analysis. *SIAM Journal on Scientific Computing*, 22 (4): pp. 1139–1159. 41, 82

Almgren, A. S., Bell, J. B., and Szymczak, W. G. [1996]. A Numerical Method for the Incompressible Navier-Stokes Equations Based on an Approximate Projection. *SIAM Journal on Scientific Computing*, 17 (2): pp. 358–369. 23, 37, 42

Arnold, D. N., Brezzi, F., Cockburn, B., and Marini, D. [1998]. Discontinuous Galerkin Methods for Elliptic Problems. In B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., *Discontinuous Galerkin Methods*, volume 11 of *Lecture Notes in Computational Science and Engineering*, pp. 89–101. Springer, New York. 84

Babuška, I. [1971]. Error-Bounds for Finite Element Method. *Numerische Mathematik*, 16: pp. 322–333. 49, 55

Barenblatt, G. I. [1996]. *Scaling, self-similarity, and intermediate asymptotics*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, New York, Melbourne. 4

Bernardi, C., Canuto, C., and Maday, Y. [1988]. Generalized Inf-Sup conditions for the Chebyshev spectral approximation of the Stokes problem. *SIAM Journal on Numerical Analysis*, 25 (6): pp. 1237–1271. 58

Braess, D. [2003]. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer, Berlin, 3rd edition. 50, 52, 56, 67, 68

Brenner, S. C. and Scott, L. R. [1994]. *The mathematical theory of finite element methods*, volume 15 of *Texts in applied mathematics*. Springer, New York. 50

Brezzi, F. [1974]. On the existence, uniqueness and approximation of saddle point problems arising from Lagrangian multipliers. *RAIRO Analyse numérique*, 8: pp. 129–151. 49, 55

Brezzi, F., Douglas Jr., J., Fortin, M., and Marini, L. D. [1987]. Efficient Rectangular Mixed Finite Elements in Two and Three Space Variables. *Mathematical Modelling and Numerical Analysis*, 21 (4): pp. 581–604. 67

Brezzi, F. and Fortin, M. [1991]. *Mixed and Hybrid Finite Element Methods*, volume 15 of *Spinger Series in Computational Mathematics*. Springer, New York. 50

Causin, P. [2002]. *Mixed-hybrid Galerkin and Petrov-Galerkin Finite Element Formulations in Fluid Mechanics*. PhD thesis, Università degli Studi di Milano. 53, 63, 84

Cockburn, B. [1999]. Discontinuous Galerkin methods for convection-dominated problems. In T. Barth and H. Deconink, eds., *High-Order Methods for Computational Physics*, volume 9 of *Lecture Notes in Computational Science and Engineering*, pp. 69–224. Springer, New York. 85

Courant, R., Friedrichs, K. O., and Lewy, H. [1928]. Über die partiellen Differenzengleichungen der mathematischen Physik. *Mathematische Annalen*, 100: pp. 32–74. 24, 28

Geratz, K. J. [1997]. *Erweiterung eines Godunov-Typ-Verfahrens für zwei-dimensionale kompressible Strömungen auf die Fälle kleiner und verschwindender Machzahl*. PhD thesis, Rheinisch-Westfälische Technische Hochschule Aachen. 84

Girault, V. and Raviart, P.-A. [1986]. *Finite Element Methods for Navier-Stokes Equations*, volume 5 of *Spinger Series in Computational Mathematics*. Springer, Berlin. 51, 65

Gresho, P. M. and Chan, S. T. [1990]. On the theory of semi-implicit projection methods for viscous incompressible flow and its implementation via a finite element method that also introduces a nearly consistent mass matrix. Part 2: Implementation. *International Journal for Numerical Methods in Fluids*, 11: pp. 621–659. 75

Harlow, F. H. and Welch, J. E. [1965]. Numerical Calculation of Time-Dependent Viscous Incompressible Flow of Fluid with Free Surface. *The Physics of Fluids*, 8 (12): pp. 2182–2189. 30

Kevorkian, J. and Cole, J. D. [1996]. *Multiple Scale and Singular Perturbation Methods*, volume 114 of *Applied Mathematical Sciences*. Springer, New York. 9

Klein, R. [1995]. Semi-Implicit Extension of a Godunov-Type Scheme Based on Low Mach Number Asymptotics I: One-Dimensional Flow. *Journal of Computational Physics*, 121: pp. 213–237. 9

Klein, R. and Geratz, K. J. [2004]. A Semi-Implicit All Mach Number Godunov-Type Scheme for Compressible Flows. *Mathematical Modelling and Numerical Analysis*. Submitted. 84

Klein, R. and Vater, S. [2003]. *Mathematische Modellierung in der Klimaforschung*. Freie Universität Berlin, Fachbereich Mathematik und Informatik. Vorlesungsskript. 4, 12

Le Maître, O., Levin, J., Iskandarani, M., and Knio, O. M. [2001]. A Multiscale Pressure Splitting of the Shallow-Water Equations I. Formulation and 1D Tests. *Journal of Computational Physics*, 166 (1): pp. 116–151. 1, 84

van Leer, B. [1979]. Towards the ultimate conservative difference scheme V. A second-order sequel to Godunov's method. *Journal of Computational Physics*, 32 (1): pp. 101–136. 28

LeVeque, R. J. [2002]. *Finite Volume Methods for Hyperbolic Problems*, volume 31 of *Cambridge Texts in Applied Mathematics*. Cambridge University Press, Cambridge. 1, 24, 44

Majda, A. J. [2003]. *Introduction to PDEs and Waves for the Atmosphere and Ocean*, volume 9 of *Courant Lecture Notes in Mathematics*. AMS, New York. 1, 2, 3

Meister, A. [1997]. Zur mathematischen Fundierung einer Mehrskalenanalyse der Euler-Gleichungen. Computational Fluid Dynamics and Data Analysis 1 Reihe F, Hamburger Beiträge zur Angewandten Mathematik. 9, 11, 12, 13

Micheletti, S. and Sacco, R. [2001]. Dual-Primal Mixed Finite Elements for Elliptic Problems. *Numerical Methods for Partial Differential Equations*, 17 (2): pp. 137–151. 83

Minion, M. L. [1996]. A projection method for locally refined grids. *Journal of Computational Physics*, 127 (1): pp. 158–178. 72

Munz, C.-D., Roller, S., Klein, R., and Geratz, K. J. [2003]. The extension of incompressible flow solvers to the weakly compressible regime. *Computers & Fluids*, 32 (2): pp. 173–196. 84

Nicolaïdes, R. A. [1982]. Existence, uniqueness and approximation for generalized saddle point problems. *SIAM Journal on Numerical Analysis*, 19 (2): pp. 349–357. i, iv, 58, 83

Osher, S. [1985]. Convergence of generalized MUSCL schemes. *SIAM Journal on Numerical Analysis*, 22 (5): pp. 947–961. 28

Pedlosky, J. [1987]. *Geophysical Fluid Dynamics*. Springer, New York, 2nd edition. 3, 5

Quarteroni, A. and Valli, A. [1997]. *Numerical Approximation of Patial Differential Equations*, volume 23 of *Springer series in computational mathematics*. Springer, Berlin, Heidelberg, New York. 52

Raviart, P.-A. and Thomas, J.-M. [1977]. A Mixed Finite Element Method for 2nd Order Elliptic Problems. In I. Galligani and E. Magenes, eds., *Mathematical Aspects of Finite Element Methods*, volume 606 of *Lecture Notes in Mathematics*, pp. 292–315. Springer. 67

Roberts, J. E. and Thomas, J.-M. [1991]. Mixed and hybrid methods. In P. G. Ciarlet and J. L. Lions, eds., *Finite Element Methods (Part 1)*, volume II of *Handbook of Numerical Analysis*, pp. 523–639. Elsevier Science Publishers B.V. (North-Holland), Amsterdam. 50, 66

Schneider, T., Botta, N., Geratz, K. J., and Klein, R. [1999]. Extension of Finite

Volume Compressible Flow Solvers to Multi-dimensional, Variable Density Zero Mach Number Flows. *Journal of Computational Physics*, 155: pp. 248–286. 7, 23, 28, 29, 62, 77

Schneider, W. [1978]. *Mathematische Methoden der Strömungsmechanik*. Vieweg & Sohn Verlagsgesellschaft, Braunschweig. 9, 11

Schulz-Rinne, C. W. [1993]. *The Riemann problem for two-dimensional gas dynamics and new limiters for high-order schemes*. PhD thesis, Eidgenössische Technische Fachhochschule (ETH) Zürich. Diss. ETH No. 10297. 76

Süli, E. [1991]. Convergence of finite volume schemes for Poisson's equation on nonuniform meshes. *SIAM Journal on Numerical Analysis*, 28 (5): pp. 1419–1430. 36, 95

Thomas, J.-M. and Trujillo, D. [1999]. Mixed finite volume schemes. *International Journal for Numerical Methods in Engineering*, 46 (9): pp. 1351–1366. 64

van der Vorst, H. A. [1992]. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 13 (2): pp. 631–644. 72

Werner, D. [2000]. *Funktionalanalysis*. Springer, Berlin, Heidelberg, New York, 3rd edition. 50