Asymptotic Preserving Finite Volume Schemes for the Singularly-perturbed Shallow Water Equations with Source Terms

Hamed Zakerzadeh

Asymptotic Preserving Finite Volume Schemes for the Singularly-perturbed Shallow Water Equations with Source Terms

Von Der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

M.Sc. **Hamed Zakerzadeh**

aus Tehran, Iran

Berichter: *Univ.-Prof. Dr. rer. nat.* **Sebastian Noelle**
Univ.-Prof. Dr. rer. nat. **Martin Frank**
Assoc. Prof. Ph.D. **Emmanuel Audusse**

Tag der mündlichen Prüfung: 28.08.2017

Abstract

The *shallow water equations* are of quite an importance for modelling oceanic flows as a simple approximation of the *water wave equations*, which describe the gravity-driven free surface flows, when the fluid is incompressible, homogeneous, inviscid, and the pressure is only hydrostatic, owing to the *shallowness assumption*, that is the horizontal length scale is much larger than the vertical one. For the shallow water equations, the *Froude number* characterises the dominance of advective modes compared to gravity (acoustic) modes as the ratio of the bulk velocity to the speed of gravity waves. For the large-scale oceanic phenomena, the Froude number is often small; so, the gravity waves are too fast to contribute to the bulk motion, *i.e.*, they do not affect the solution of the large-scale macroscopic model. For a time-explicit numerical treatment, though, one should devise a method to tackle these fast waves to avoid high computational costs as they restrict the time step through the Courant–Friedrichs–Lewy (CFL) condition. The approach considered throughout this manuscript is to decompose the system into slow and fast parts and to employ an implicit-explicit (IMEX) strategy, *i.e.*, to treat the fast part implicitly and the slow part explicitly.

In addition to the efficiency problem attached to this singularly-perturbed system, one should be careful about the limiting scheme, *i.e.*, if the scheme provides a consistent and stable approximation of the zero-Froude system (lake equations). Even if the convergence to the limit can be shown for the continuous model, preserving such a convergence for the discrete (numerical) model, along with stability and consistency, is by no means trivial and should be carefully analysed. This motivates adopting the framework of *asymptotic preserving (AP) schemes* introduced by [Jin, *SIAM J. Sci. Comp.* 21(2) (1999), pp. 441–454], with the Froude number as the scaling singular parameter. AP schemes are defined as schemes mimicking such a convergence to the limit for the discrete model, *e.g.*, in virtue of uniform consistency and stability.

In this manuscript, we consider two IMEX flux-splitting finite volume schemes for the shallow water equations with uniform consistency and stability w.r.t. the Froude number: the *Lagrange-projection* IMEX scheme and the *reference solution* IMEX scheme. The LP-IMEX scheme is a Godunov-type scheme, which decomposes the system into the acoustic and the transport systems, and employs a Lagrangian formulation for the former. Unfortunately, it is involved in some inherent accuracy issues especially in multiple dimensions, which need to be taken care of; so, we investigate it only for the one-dimensional system. The primary focus would be on the RS-IMEX scheme, which decomposes the solution into the (asymptotic) reference solution and a perturbation around it in order to split the system. We study the RS-IMEX scheme in one and two space dimensions with the bottom topography, and finally, with the additional Coriolis force. For both of these schemes, we present a (rigorous) asymptotic analysis to justify the uniform consistency and stability of the scheme w.r.t. the Froude number, and to corroborate the AP property. We also test the quality of the solutions computed by the RS-IMEX scheme in several numerical examples, particularly for the low-Froude regime.

Acknowledgments

"Wenn es gilt fürs Vaterland, treu die Klingen dann zur Hand,
und heraus mit mut'gem Sang, wär es auch zum letzten Gang!"

– Franz von Kobell *(Bursche heraus!)*

This thesis gives an account of the work carried out during the years 2014–2017 at the *Institut für Geometrie und Praktische Mathematik (IGPM)* in *RWTH Aachen Universtiy*, and has been financially supported by the university through *Graduiertenförderung nach Richtlinien zur Förderung des wissenschaftlichen Nachwuchses (RFwN)* scholarship.

Firstly, I would like to express my sincere gratitude to my thesis advisor, Sebastian Noelle, for all his support and motivation during this period, also for his patience, generosity, and tolerance. I also would like to thank the rest of my doctoral committee: Emmanuel Audusse, Martin Frank, Holger Rauhut and Maria Westdickenberg for their insightful comments and encouragement. My sincere thanks also go to Rupert Klein for sharing his insights and opinions, and for the motivation he gave me by letting me know how much I do not know! I would also like to thank Mária Lukáčová for our great discussion in Mainz, her fruitful comments, and her hospitality.

More specifically, I would like to gratefully thank Christophe Chalons and Mathieu Girardin for giving the motivation and for sharing their insights with me regarding the material of Chapter 2. For Chapter 3, I would like to thank Rupert Klein for helping me in understanding [Kle95], and Jochen Schütz for a helpful discussion on [SN14]. I also gratefully acknowledge an illuminating conversation with Charlotte Perrin, my dear friend and colleague. For chapter 4, my sincere thanks go to Negin Bagherpour—my dear old friend and professor—and my dear brother Mohammad for helpful discussions and ideas regarding some proofs in Section 4.3.2 and Appendix 4.A.

I am indebted to Laurent Gosse and his excellent book [Gos13] for all cute quotes I borrowed from it and for historical details. I would also like to thank Frank Knoben for his superb work as the system administrator, and the Hochschulwachen of the Hauptgebäude for all holidays I disturbed them asking to enter the building (even once on Neujahrstag around 8 a.m.!). I also really appreciate the helps of Sebastian, Charlotte, and Mohammad in reviewing parts of this thesis to avoid several blunders and countless typos. During my stay in Aachen, I have been privileged to have adorable friends around to chill out with, and to spend time together; I would like to thank them all for all they have given to me.

Last but not least, I would like to thank my family: my parents, Nasrin and Mehdi, to whom I owe an eternal gratitude for supporting me all along my life to get where I am standing now, and for their unconditional endless care and love; my sister, Rana, for all her care and advice during the tough days; and my twin brother, Mohammad, who paved all the way with me, all ups and downs, for almost 29 years (including those months before getting born!). Of course and unfortunately, I am not able to express how they mean for me in words, throughout my doctoral studies, in particular, and my life, in general.

Contents

Chapter 1

Introduction

"With my two algorithms one can solve all problems—without error, if God will."

– Khwarizmi, *Algebra (circa 800 AD)*

Throughout this chapter, we present the basics we need for the rest of this manuscript. Firstly, we briefly review hyperbolic balance laws and introduce asymptotic preserving schemes as well as the shallow water equations. Then, we discuss state of the art in designing asymptotic preserving schemes for the shallow water equations, which is followed by the overview of the remaining chapters, describing the contributions of this manuscript.

Contents

1.1 A short introduction to hyperbolic systems

A system of conservation laws considers *conservation* of a quantity in a specific d-dimensional domain, $\Omega \subset \mathbb{R}^d$, and writes

$$\begin{cases} \partial_t \boldsymbol{U}(t,\boldsymbol{x}) + \operatorname{div}_{\boldsymbol{x}} \boldsymbol{F}(\boldsymbol{U}) = \boldsymbol{0}, \\ \boldsymbol{U}(0,\boldsymbol{x}) = \boldsymbol{U}_0(\boldsymbol{x}), \end{cases} \tag{1.1}$$

where $(t,\boldsymbol{x}) \in [0,\infty)\times\Omega$, $\boldsymbol{U}(t,\boldsymbol{x}) \in \mathbb{R}^q$ is the vector of q conservative variables, $\boldsymbol{F} = [\boldsymbol{f}_1,\dots,\boldsymbol{f}_d] \in \mathbb{R}^{q\times d}$ is a smooth flux function, and $\boldsymbol{U}_0(\boldsymbol{x})$ is the initial condition of the system. We set the domain Ω to be a d-dimensional torus $\mathbb{T}^d$ to avoid any issue which may arise from the boundaries. We assume, hereinafter, that the system is *hyperbolic*, *i.e.*, for all directions $\boldsymbol{n} = (n_1,\dots,n_d)^T$

in $\mathbb{R}^d$, the flux Jacobian matrix $A_{\boldsymbol{n}} := \sum_{i=1}^{d} \partial_{\boldsymbol{U}} \boldsymbol{f}_i n_i$ has q real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_q$ with linearly-independent eigenvectors $\{r_k\}_{k=1}^{q}$. It means that $A_{\boldsymbol{n}}$ (denoted more simply as $A := \partial_{\boldsymbol{U}} \boldsymbol{F}$) is diagonalisable, *i.e.*, it can be transformed into a diagonal matrix as $A = R\Lambda R^{-1}$, where Λ is a diagonal matrix—which consists of eigenvalues of A— and R is the matrix of right eigenvectors. We define the *wave family* (λ_k, r_k) to be *genuinely non-linear* if $\lambda_k'(\boldsymbol{U}) \cdot r_k(\boldsymbol{U}) \neq 0$ for all $\boldsymbol{U}$, and to be *linearly degenerate* if $\lambda_k'(\boldsymbol{U}) \cdot r_k(\boldsymbol{U}) = 0$ for all $\boldsymbol{U}$. For one-dimensional (1d) systems, the notion of *Riemann invariants* can be very useful in decoupling the system, *cf.* Chapter 2. It is defined as the variable w_j whose gradient is normal to the right eigenvector r_j, *i.e.*, $\partial_{\boldsymbol{U}} w_j \cdot r_j = 0$. This implies that w_j is constant along r_j; so, the system can be transformed into a diagonal form, as a system of advection of the Riemann invariants $\partial_t w_j + \lambda_j \partial_x w_j = 0$.

The existence, uniqueness, and regularity of the solution of (1.1) is a long-standing question, and only some partial answers are available for some specific systems. It is well-known that, in general, and even with smooth initial data, the system does not possess a continuous solution after a finite time, as the so-called *shock waves* appear and the solution gets discontinuous. So, it is a common practice to weaken the notion of the solution to the *weak* or *distributional* solutions, as defined below.

Definition 1.1.1. *[Weak (distributional) solution] A function* $\boldsymbol{U} \in [L^\infty_{loc}([0,+\infty) \times \Omega)]^q$ *is a weak solution of* (1.1) *with the initial data* $\boldsymbol{U}_0 \in [L^\infty_{loc}(\Omega)]^q$ *if*

$$\int_0^\infty \int_\Omega (\boldsymbol{U} \cdot \partial_t \varphi + \boldsymbol{F} \cdot \nabla_{\boldsymbol{x}} \boldsymbol{\varphi}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t + \int_\Omega \boldsymbol{U}_0 \cdot \boldsymbol{\varphi}(0, \boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0, \qquad \forall \boldsymbol{\varphi} \in [C_0^\infty([0,+\infty) \times \Omega)]^q .$$

It can be shown that (piece-wise smooth) weak solutions satisfy the so-called *Rankine–Hugoniot* jump condition, which governs the evolution of a discontinuity, like shocks. However, the R–K condition is not enough to determine the weak solution uniquely and one can find infinitely many weak solutions satisfying the jump condition. To recover the uniqueness, the notion of *entropy solutions* is employed [Daf10], which are solutions satisfying a relevant entropy inequality

$$\partial_t \eta(\boldsymbol{U}) + \mathrm{div}_{\boldsymbol{x}} \boldsymbol{Q}(\boldsymbol{U}) \leq 0, \qquad \boldsymbol{Q} := [q_1, \ldots, q_d], \tag{1.2}$$

in the sense of distributions, and for *all* entropy pairs $(\eta, \boldsymbol{Q})$, *i.e.*, all pairs of functions $(\eta, \boldsymbol{Q})$ such that η is a scalar and (strictly) convex, and $(\partial_{\boldsymbol{U}} \eta)^T A = (\partial_{\boldsymbol{U}} \boldsymbol{Q})^T$. Although the entropy criterium works well for $d = 1$, it has been shown recently that entropy solutions are not unique for $d > 1$; see [CK16] and the references therein like [DLS10, Ell06].

One can also weaken the solution to the so-called *measure-valued* solution which is in fact a Young measure, rather than an integrable function; see [DiP85, DM87, NMRR96]. For further details about this, in particular, and about the theory of hyperbolic conservation laws, in general, we refer the reader to the monograph [Daf10].

Example 1.1.2 (Isentropic Euler equations)**.** *As a well-known example of hyperbolic conservation laws* (1.1)*, one can name the 2d isentropic Euler equations, which write*

$$\begin{aligned} &\partial_t \varrho + \mathrm{div}_{\boldsymbol{x}}(\varrho \boldsymbol{u}) = 0, \\ &\partial_t (\varrho \boldsymbol{u}) + \mathrm{div}_{\boldsymbol{x}} \left(\varrho \boldsymbol{u} \otimes \boldsymbol{u} + p(\varrho) \mathbb{I}_2\right) = \boldsymbol{0}, \end{aligned} \tag{1.3}$$

where $\varrho > 0$ *is the density of the fluid,* $\boldsymbol{u}$ *is the two-dimensional velocity vector,* $\otimes$ *denotes the Kronecker product,* $p(\varrho) := \kappa \varrho^\gamma$ *(with* $\kappa > 0$ *and* $\gamma > 1$*) is the isentropic pressure law, and* $\mathbb{I}_2$ *is the* 2×2 *identity matrix. It is easy to confirm that the system is hyperbolic, i.e., it has real*

eigenvalues $\boldsymbol{u}\cdot\boldsymbol{n}$ *and* $\boldsymbol{u}\cdot\boldsymbol{n}\pm\sqrt{p'(\varrho)}$*, and a complete set of eigenvectors. The entropy function can be chosen as the total energy of the solution* ϱE*, where the total energy density is written as* $E=\mathcal{E}+\frac{u^2}{2}$*, and* $\mathcal{E}(\varrho):=\frac{\kappa}{\gamma-1}\varrho^{\gamma-1}$ *is the internal energy density, cf. [LW07]). One can show that* ϱE *is strictly convex w.r.t. the conservative variables* $\boldsymbol{U}=(\varrho,\varrho\boldsymbol{u})^T$*. The entropy flux can be shown to be* $Q=(\varrho E+p)\boldsymbol{u}$*; so, the entropy inequality* (1.2) *for this system writes*

$$\partial_t(\varrho E)+\operatorname{div}_{\boldsymbol{x}}\left((\varrho E+p)\boldsymbol{u}\right)\leq 0. \tag{1.4}$$

Having all these issues, even in defining a suitable notion of the solution for the system (1.1), it should not be surprising that finding analytical solutions for (1.1) is out of reach for general non-linear systems. This motivates the use of numerical methods to approximate the solution of (1.1). In order to ensure the quality of the computed solution, schemes should preserve some important properties from the continuous system like the entropy stability; see, *e.g.*, [Tad87, Tad03, ZF16]. Despite the negative uniqueness result of [CK16], on the one hand, the entropy stability can still be used to prove the non-linear stability of solutions. On the other hand, it has been shown recently in [Svä15] that entropy solutions are relevant, at least, for *numerically computed solutions*, by proving the convergence of the Lax–Friedrichs scheme to the entropy solution of the full Euler equations for $d=3$; see also [Svä16].

There is indeed considerable literature dedicated to the numerical approximation of conservation laws for which we refer to [CSJT98]. Here, we only focus on the finite volume (FV) method because of its simplicity, its ability to deal with complex geometries and the inherent conservation properties; see [LeV02, EGH00].

1.1.1 Hyperbolic balance laws

Hyperbolic balance laws are conservation laws with source terms (see [Bou04]), *i.e.*,

$$\partial_t\boldsymbol{U}(t,\boldsymbol{x})+\operatorname{div}_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{U},t,\boldsymbol{x})=\boldsymbol{S}(\boldsymbol{U},t,\boldsymbol{x}), \tag{1.5}$$

where $\boldsymbol{S}\in\mathbb{R}^q$ is a smooth function. The notion of weak solutions can be defined similarly as for conservation laws (1.1); see [Gos13, Bou04].

An important feature of (1.5) lies in the competition between the flux and the source term during the time evolution, leading to the so-called *steady states*, which are often obtained after a long time and solve $\operatorname{div}_{\boldsymbol{x}}\boldsymbol{F}=\boldsymbol{S}$. It is important to note that obtaining these steady-state solutions, numerically, is far from being trivial. It was first noticed in [BV94, GL96] that a naïve treatment of the source term pollutes the numerical solution by some increasing-in-time oscillations, which require using very fine grids to be suppressed. The reason lied in a small unbalance (of the order of grid size) between the discretisations of the flux and the source term, which increased in time and deteriorated the solution. Although there is vast literature dedicated to the so-called *well-balanced (WB) schemes* as a remedy [Bou04, Gos13], a panacea for general systems is out of reach, and WB schemes should be designed almost individually for each case. Moreover, preserving non-stationary equilibrium states is very demanding and only some (partial) results are available like [NXS07, MDBCF16] for the *moving equilibrium*, [LMNK07, CDKLM14, CLP08, BLSZ04, AKNV11, ADDMHP15] for the *quasi-geostrophic equilibrium*, and [BLMY16, BKLL04, TKK16, KM14] for the *hydrostatic equilibrium*.

1.2 Asymptotic preserving schemes

Singular limits of balance laws (or more generally PDEs), characterised by the singular scaling parameter $\varepsilon \in (0,1]$ approaching zero, may present severe difficulties to be treated either in analysis or numerics. The main issue is that the type of the equations changes in the limit [Mas07]. As an example, consider the isentropic Euler system (1.3). It is prevalent in the literature to make the system non-dimensionalised, to see the effects of different physics; *cf.* [MYO90, Lan13]. Here, we apply the standard non-dimensionalisation and (analogous to [DT11]) and define dimensionless variables as $\hat{t} := t/t_\circ$, $\hat{\boldsymbol{x}} := \boldsymbol{x}/L_\circ$, $\hat{\varrho} := \varrho/\varrho_\circ$, $\hat{\boldsymbol{u}} := \boldsymbol{u}/u_\circ$, $\hat{p} := p/p_\circ$, where the subscript $\circ$ stands for characteristic values, $t_\circ := L_\circ/u_\circ$, and $p_\circ := \varrho_\circ c_\circ^2/\gamma$ with $c_\circ := \sqrt{\gamma p_\circ/\varrho_\circ}$ as the characteristic sound speed. With these definitions and after dropping hats, the dimensionless isentropic Euler equations write

$$\begin{aligned} &\partial_t \varrho_\varepsilon + \operatorname{div}_{\boldsymbol{x}}(\varrho_\varepsilon \boldsymbol{u}_\varepsilon) = 0, \\ &\partial_t(\varrho_\varepsilon \boldsymbol{u}_\varepsilon) + \operatorname{div}_{\boldsymbol{x}}\left(\varrho_\varepsilon \boldsymbol{u}_\varepsilon \otimes \boldsymbol{u}_\varepsilon + \frac{1}{\varepsilon^2} p(\varrho_\varepsilon)\mathbb{I}_d\right) = \mathbf{0}, \end{aligned} \tag{1.6}$$

where $\varepsilon := \sqrt{\gamma} Ma$ and the Mach number Ma is defined as the ratio of the characteristic bulk velocity to the characteristic sound speed, *i.e.*, $Ma := u_\circ/c_\circ$. For the dimensionless system (1.6), Ma plays the role of a singular scaling parameter since as $Ma \to 0$, the sound speed $\sqrt{p'(\varrho)}/\varepsilon$ goes to the infinity and the PDE changes to be hyperbolic-elliptic, in the so-called *incompressible limit*:

$$\begin{aligned} &\varrho = const., \qquad \operatorname{div}_{\boldsymbol{x}} \boldsymbol{u} = 0, \\ &\partial_t \boldsymbol{u} + \boldsymbol{u} \cdot \nabla_{\boldsymbol{x}} \boldsymbol{u} + \frac{1}{\varrho} \nabla_{\boldsymbol{x}} \pi = \mathbf{0}, \end{aligned} \tag{1.7}$$

where π is an auxiliary pressure satisfying the divergence constraint as a Lagrange multiplier. Proving the convergence of the solution of the compressible Euler equations (1.6) to the incompressible system (1.7) is very demanding; we refer the reader to consult [KM81, KM82, Dan05, Mas07, Sch05] to review the existing results.

Tackling such singular problems numerically is also complicated as they introduce a "*stiff*" system for which finding an efficient and stable numerical approximation is a longstanding challenge in numerical analysis. In the context of conservation laws, stiffness may be defined as the simultaneous occurrence of eigenvalues (of the flux Jacobian) of different orders of magnitude. The key example is the weakly compressible Euler system (1.6) (with $Ma \sim \varepsilon \ll 1$), which is stiff due to very fast acoustic waves. This stiffness makes the Courant–Friedrichs–Lewy (CFL) condition to restricts the time step non-uniformly with ε such that it should tend to zero, *i.e.*, $\Delta t \lesssim \varepsilon\, \Delta x$. Such a restriction leads to very small time steps, thus a substantial computational cost. Generally speaking, numerical schemes also lose their accuracy in the limit for under-resolved grids, due to the spurious numerical diffusion they generate; for in-depth discussions and remedies see [Del10, DOR10, GV99, GM04, OSB$^+$16, Rie11, RB09b, Rie10]. Similar problems can also happen for other equations like the kinetic equations, where the Knudsen number (the non-dimensionalised mean free path) approaches zero either in the *fluid* or *diffusive* limit, as the source terms is stiff and restricts the time step for an explicit treatment [DP14].

Note that there are several established methods for the limit models *per se*, *e.g.*, there are several working schemes for the incompressible Euler and Navier–Stokes equations; see [DR06] for instance. The crucial question here is that if it is possible to find a scheme working well

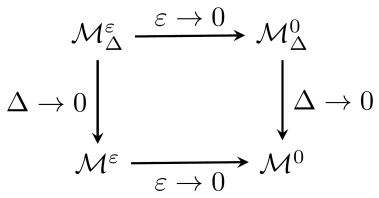


Figure 1.1: Illustration of asymptotic preserving schemes.

regardless of the (scaling) singular parameter ε. One classical way of handling this issue is the *domain decomposition*, which considers the limit system for the regions where ε is small enough, while for other regions a standard solver would be employed. The bottleneck for this strategy is the *coupling condition* between these two regions which is by no means trivial, is *ad hoc* and for which the implementation details are of fundamental importance; see [DP14]. One should also consider that the error incorporated into the final solution due to these issues may deteriorate the accuracy of the scheme itself.

A more recent and unified approach was initiated by [LMM87, LM89] for steady neutron transport in the diffusive regime, followed up by [JL91, JL93, JL96, Kla98, JPT98] for systems with a stiff relaxation or in diffusive limits, and finally resulted in the so-called *asymptotic preserving (AP) schemes* [Jin95, Jin99].[1] The main motivation for AP schemes is to use a uniform grid for all $\varepsilon > 0$, and at the same time to capture the macroscopic behaviour of the system in the limit without any need to resolve microscopic effects, which requires using extremely fine grids in time and space. It, in fact, provides an automatic seamless transition between different scales [Jin10]. For this approach, we assume that the (suitable notion of) *solution* of the PDE with the singular parameter ε converges to the *solution* of the limit PDE as $\varepsilon \to 0$, and aim to show that the counterpart of such a convergence exists at the discrete level. Figure 1.1 illustrates this definition; $\mathcal{M}^{\varepsilon}$ stands for a continuous physical model with the (singular) parameter $\varepsilon \in (0, 1]$, and $\mathcal{M}^{\varepsilon}_{\Delta}$ is a discrete-level model, which provides a consistent discretisation of $\mathcal{M}^{\varepsilon}$. If $\mathcal{M}^{\varepsilon}_{\Delta}$ is a *suitable* and *efficient* scheme for $\mathcal{M}^{\varepsilon}$ uniformly in ε, the scheme is called to be AP; see [Jin10, FR13]. We can define an AP scheme more precisely as follows.

Definition 1.2.1. *[AP schemes] A scheme is called to be AP, provided that it*

(i) gives a consistent discretisation of $\mathcal{M}^{\varepsilon}$ for all $\varepsilon \in (0, 1]$, in particular for the limit problem $\mathcal{M}^{0}$.

(ii) is efficient uniformly in ε, e.g., the CFL condition is ε-uniform and the implicit step can be solved efficiently for all ε.

(iii) it is stable in some suitable sense, uniformly in ε.

For brevity, we call these properties, respectively, Asymptotic Consistency (AC), Asymptotic Efficiency (AEf), and Asymptotic Stability (AS).

Remark 1.2.2. *(i) It is also prevalent in the literature to define the asymptotic stability as the stability of the limit scheme $\mathcal{M}^{0}_{\Delta}$, cf. [Jin10, Gie15]. Also, sometimes, the uniformity of the CFL condition is classified as the asymptotic stability rather than asymptotic efficiency.*

[1] As mentioned by [Gos13], AP schemes have also roots in Soviet Union with the name *Asymptotic Integration Method*; see [Vas94, LE88, Il'69].

(ii) In the sequel and in order to eliminate spurious initial layers (see [Mas07]), we almost always consider well-prepared initial data, which are consistent with the limit $\varepsilon \to 0$, cf. [Mas07, MS01, KLN91, Gre97]. However, ill-prepared initial data have been used in some of the numerical examples. We also refer to Appendix 4.B for the only analysis for ill-prepared initial conditions, in this manuscript.

(iii) As mentioned in [Jin10], the asymptotic consistency suggests that the solution belongs to a manifold, which is driven to the limit manifold as $\varepsilon \to 0$, up to some discretisation error.

(iv) AEf implies the ε-uniform well-posedness of the scheme and in particular the implicit step, which can be translated as having a good condition number if the implicit step is linear, i.e., when it requires solving a linear system of equations. Such an issue can be handled using the classical pre-conditioning techniques as in [Bis15]. Moreover and very recently, the authors in [FN16] have addressed this point more fundamentally for some toy models related to Vlasov–Maxwell equations.

(v) It is important to distinguish between AP schemes and multi-scale schemes. AP schemes deal, in principle, with one and the same model $\mathcal{M}^\varepsilon$ and numerical scheme $\mathcal{M}^\varepsilon_\Delta$ for the whole region between microscopic and macroscopic models. However, generally speaking, a multi-scale scheme such as the heterogeneous multi-scale method (HMM) [WEL+07] can employ different physical models and numerical methods at different scales.

(vi) Interestingly enough, one may translate AP property (for systems with stiff source terms) as "well-balancing with stiffness" since the steady solution can be understood as the long-time limit, by the rescaling $t \mapsto \varepsilon t$ for $\varepsilon \to 0$ [Jin10]; such a point has been elaborated in [Gos13, GT04, GT02, GT03, Gos11, CCG+10].

Achieving asymptotic efficiency regarding the time step restriction, AP schemes often take advantage of the implicit-explicit (IMEX) strategy[2], *i.e.*, to split the flux Jacobian and the source term into stiff and non-stiff parts, and to treat the stiff part implicitly in time and the non-stiff one explicitly in time. IMEX schemes are L_2-stable as long as each step is so, as shown in [HJL12]. Employing an implicit strategy is definitely necessary for stiff terms to find schemes with an ε-uniform time step restriction, but not sufficient at all for the asymptotic stability; see for example [ADG89], where it is shown that even if both split parts are stable in terms of the CFL condition, the resulting scheme can be unconditionally unstable in the L_2-norm. In addition to these IMEX or semi-implicit schemes, there are several works, like [BT16, CDV17], which are devoted to explicit schemes; they define the AP property without considering the uniformity of the time step w.r.t. ε. Of course, such a scheme requires using restrictive time steps; nonetheless, this cost can be justified in the sense that although IMEX schemes allow for larger time steps, they usually should handle a non-linear system of equations, namely by the Newton–Raphson method, with a huge computational cost. Also, the excessive diffusion of the implicit part deteriorates the accuracy and quality of the numerical approximation unless the grid is fine enough or high-order schemes have been employed.

One can also think of fully-implicit schemes like finite volume schemes [GHMN17], mixed finite element-finite volume schemes [FLMN+16], and space-time dG schemes [HM14] (see also [ZM16] for its modified variant without the streamline diffusion). Fully-implicit schemes have

[2] IMEX methods are very well-known for ODEs; see for instance the classic textbook [HW96]). The reader can also consult [ARS97, BR09, CJR97, PR05] for more details about the use of IMEX methods in constructing AP schemes for stiff systems.

the advantage of being *unconditionally stable*, though, they are diffusive and should deal with a non-linear system of equations, which could be truly expensive in terms of the computational cost.

The AP property has been studied extensively for the kinetic equations (see [Jin10, HJL16] for a review), hyperbolic conservation/balance laws [BLMY16, Bis15, BALMN14, NBA$^+$14, DLV17, CDK12, DT11, HJL12, DMTB15], plasma equations [DDN$^+$10, CDV16, FR16, DD16] as well as several other systems. Note that asymptotic consistency proofs in the literature are often formal, based on the asymptotic (Poincaré) expansion. There are few works though, which concern rigorous proofs like [EDMS17b, EDMS17a, JLQX14, FR13, FR16, Bis15, BLMY17]. Furthermore, there have been recently some interests in employing the *relative entropy/energy method* (see [CMS13, GHMN16]) as a tool to measure the difference between the (discrete) solutions of $\mathcal{M}^\varepsilon_\Delta$ and $\mathcal{M}^0$ like [FLMN$^+$16, GHMN17, Fis15] for the compressible Navier–Stokes equations and [BBCM16] for the p-system with damping. Also, there are only a few results regarding the asymptotic stability, either for conservation laws or kinetic equations, like [GJL99, JLQX14, LM08, DLV17, KFJ16, Gie15, Zak17a, BLMY17] using the von Neumann stability analysis, energy methods or entropy stability.

It is worth mentioning that before the acronym AP to appear in the literature, there have been an abundance of studies dedicated to weakly compressible flows. One approach is to improve the low-Mach behaviour of compressible methods, *e.g.*, using the *pre-conditioning* methodologies proposed by [Cho67, Tur87, TFVL93]. They deal with pre-conditioning the system and its numerical dissipation to obtain convergence and accuracy of schemes for the low-Mach regime [DR06, Sect. 9.3]; see also more recent works such as [BEK$^+$16, Del10, DOR10]. Nonetheless, pre-conditioning methods are not appropriate for temporal accuracy of the method [WSW02] or if regions of low and high Mach numbers co-exist. They may also suffer from *very* restrictive time step restrictions, *i.e.*, $\Delta t \lesssim \varepsilon^2 \Delta x$; see [BM05b, BM05a, Del10]. Furthermore, [Kle95] initiated the multiple pressure variables (MPV) approach, treating different orders in the asymptotic expansion of the pressure function in a clever way (*cf.* Remark 3.5.4); see [HP94, KM95] and [MRKG03, PM05, MDR07, Vat13] for further discussions and some extensions. On the other hand, one may extend the incompressible methods, like the *pressure-correction scheme*, for compressible flows; one can name [DLP93, KP96, MD01] for collocated grids, and [BW98, CG84, WSW02, vdHVW03] for staggered grids. An important example is the extension of the celebrated marker-and-cell (MAC) scheme [HW65] to compressible regimes as done in [HA68, HA71], which led to several recent contributions [GGHL08, GHMN17, GHK$^+$11, GHKL15, HKL12, HLN13c, HKL13, HLN13d, HLN13a, HLN13b, HKL14, HLS17]. This second class of schemes may suffer from a non-conservative formulation, which hampers the accurate computation of shock speeds [vdHVW03]. We refer to [PTA12, DR06, KBS$^+$01] for a review of these classical approaches.

Throughout this manuscript, we are aiming to present the AP methodology for the shallow water equations—which we plan to detail now—when some singular parameters are present in the system.

1.3 Shallow water equations

The shallow water equations (SWE) or the Saint-Venant system [BdSV71] (as it is known to the French scientific community) is a reduced 2d model obtained from the 3d incompressible Euler equations for a homogeneous fluid under a gravitational force, used commonly for modelling free surface flows like in ducts, rivers, and oceans. The shallow water model cannot consider thermodynamic effects and *stratification* (variations in density) stemming from the temperature gradients or salinity of the water. The fluid is bounded by a solid wall from below (the so-called *topography*; see Figure 1.2), which implies that the normal component of the velocity should be zero at the bottom. As the surface is *free*, one applies the *kinematic boundary condition* on the top of the fluid region; see [Ped13, Chap. 3].

With these assumptions, the model can be derived by averaging the system in the vertical axis, the direction in which a constant gravitational force is applied. The basic ingredient required for the averaging process is the *shallowness assumption*, *i.e.*, the vertical characteristic length L_v in much smaller than the horizontal one L_h (see Figure 1.2). This assumption is practical, namely for oceanic flows, where $L_h \sim 10^2 - 10^3\,\mathrm{km}$, $L_v \sim 1 - 5\,\mathrm{km}$, and $\delta := L_h/L_v \sim 0.001$-$0.01$. With the shallowness assumption and the incompressibility, one can formally derive the so-called *hydrostatic approximation* which means that the pressure (to leading order) is only hydrostatic, which is the pressure exerted only because of the weight of the water column. One can consult [Ped13, Chap. 3] and [Lio96, Sect. 4.6]) for some formal results and [Lan13] for rigorous justifications of this derivation.

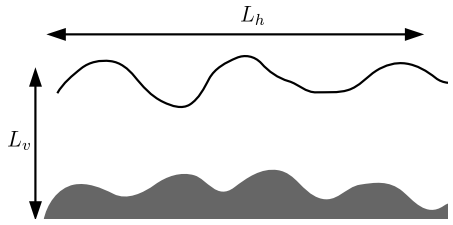


Figure 1.2: Horizontal versus vertical length scales.

Note that the SWE are of quite an importance for modelling oceanic flows as a simplified model for depth-averaged incompressible free surface flows, which give the so-called *water wave equations*; see the monograph [Lan13] for more detailed discussions. This simplicity is advantageous in terms of computational cost; but, restricts the validity of the model, particularly for near-shore *wave shoaling*. Since the SWE are derived by ignoring the higher order terms, $o(\delta)$, in the water wave equations, one may guess that keeping more terms may amend such issues. This, in fact, leads to more involved models such as the *Green–Naghdi* model (*cf.* [Lan13]), which results in some issues like being dispersive, so, troublesome for *wave-breaking*; see [PDZ$^+$14, ZKD$^+$14, DM15, DM16, LM15] for some remedies.

In the utmost generality, and for the domain $\Omega \subset \mathbb{R}^2$ lying in the (x, y) plane, the SWE can be written as

$$\begin{aligned} &\partial_t h + \mathrm{div}_{\boldsymbol{x}}(h\boldsymbol{u}) = 0, \\ &\partial_t(h\boldsymbol{u}) + \mathrm{div}_{\boldsymbol{x}}\left(h\boldsymbol{u}\otimes\boldsymbol{u} + \frac{gh^2}{2}\mathbb{I}_2\right) = \boldsymbol{S}, \end{aligned} \tag{1.8}$$

where h is the height of the water column (above the bottom), $\boldsymbol{u} = (u_1, u_2)$ is the 2d velocity vector, g is the gravity acceleration, and $\boldsymbol{S}$ is a general source term.

Remark 1.3.1. *Physically speaking, the shallow water system is quite different from the isentropic Euler equations* (1.3)*, as it is incompressible and depth-averaged with the hydrostatic pressure. However, these systems look like each other, mathematically; the difference is only due to the pressure term.*

The source term can stem from the bottom (or air) friction, acceleration or deceleration of the flow due to bottom topography or Coriolis forces. For the purpose of this manuscript, we ignore friction terms (and refer to [MD16, DMTB15] and the references therein) and only consider bottom topography and the Coriolis force. This brings us to the 2d *rotating shallow water equations* (RSWE):

$$\begin{aligned} &\partial_t h + \operatorname{div}_{\boldsymbol{x}}(h\boldsymbol{u}) = 0, \\ &\partial_t(h\boldsymbol{u}) + \operatorname{div}_{\boldsymbol{x}}\left(h\boldsymbol{u}\otimes\boldsymbol{u} + \frac{gh^2}{2}\mathbb{I}_2\right) = -gh\nabla_{\boldsymbol{x}}\eta^b - fh\boldsymbol{u}^\perp, \end{aligned} \tag{1.9}$$

where η^b is the bottom function, $\boldsymbol{u}^\perp = (-u_2, u_1)$ is the *perpendicular velocity* and f is the *Coriolis parameter*. We assume that the $f = f_0$ is constant *(zero-plane approximation)*; however, the *beta-plane approximation* $f = f_0 + \beta y$ (y is northward) is also common. Here, the earth's rotation is realised only through the Coriolis force. Let us pronounce that we have neglected the effects of spherical (ellipsoidal) shape of the earth globe as well as the centrifugal forces. We refer to [Ped13, Chap. 3] and [Lio96, Sect. 4.6] for a justification of this system.

Historically speaking, one should mention that the fictitious force introduced in 1835 by Coriolis [Cor35] was to add a missing component to the known centrifugal force in rotating frames, and had nothing to do, in particular, with the meteorology. In fact, before the Coriolis' era, it was known by Hadley in 1735 that the earth's rotation deflects the air currents, and the notion of the Coriolis force came into the field later [Per98]. We refer to [Fos13, Per98] for more details about these historical backgrounds, and to [GSR07, CDGG06, Ped13] for detailed mathematical and physical discussions about the RSWE. Note that for the most of this manuscript, but in Chapter 6, we ignore the Coriolis force, *i.e.*, we set $f = 0$.

For the zero-Coriolis case, and similar to the isentropic Euler system, one can non-dimensionalise the SWE as:

$$\begin{aligned} &\partial_t h + \operatorname{div}_{\boldsymbol{x}}(h\boldsymbol{u}) = 0, \\ &\partial_t(h\boldsymbol{u}) + \operatorname{div}_{\boldsymbol{x}}\left(h\boldsymbol{u}\otimes\boldsymbol{u} + \frac{h^2}{2Fr^2}\mathbb{I}_2\right) = -\frac{h}{Fr^2}\nabla_{\boldsymbol{x}}\eta^b, \end{aligned} \tag{1.10}$$

where $Fr := u_\circ/\sqrt{gH_\circ}$ is the Froude number defined as the ratio of the characteristic bulk velocity to the characteristic speed of *gravity waves*, which are analogous to acoustic waves for the isentropic Euler system. Comparing (1.10) with (1.6) indicates that for the low-Froude regime $Fr \ll 1$, the similar stiffness problems would arise.

1.4 Overview of the manuscript

As an interesting example, we investigate AP schemes for the SWE throughout this manuscript. Apart from studies on the isentropic Euler equations, the literature of AP schemes devoted solely to the SWE (with a source term) is not mature enough. One can find, generally, two types of results, for a system with large friction (with a hyperbolic to parabolic degeneracy) or for the zero-Froude limit (with a hyperbolic to hyperbolic-elliptic degeneracy), when both cases share a similar difficulty of restrictive time steps. For the former, it is essential to design the numerical diffusion appropriately, which is not trivial for unstructured grids; see for instance [Fra12, Bla16] and the references therein. We are not going to handle this case here; instead, the focus is on the latter case, the low-velocity or low-Froude regime, and to study how one should treat the stiffness it arises. For this case, there are some existing IMEX schemes in the literature such as [HJL12, DT11] which are shown to be AP (formally) and perform well in practice. Nonetheless, they are difficult to be analysed rigorously as the implicit step is non-linear.

On the other hand, there are IMEX schemes with a linear implicit step, *e.g.*, [CNPT10, CGK13, CGK16] which employ Riemann invariants of the Euler equations (based on a relaxation approximation) to obtain this linearity. Moreover, a series of papers [BALMN14, BLMY16, Bis15] uses a *linearly-implicit approach* (see Chapter 4) with a linearisation around the equilibrium state to split the shallow water system such that the implicit step is linear. We put the primary emphasis of the manuscript on generalising these two approaches. In this sense, the results of this manuscript can be compared to aforementioned references.

The remainder of this manuscript is organised as follows. In Chapter 2 and motivated by [CNPT10, CGK13, CGK16], we analyse the Lagrange–projection scheme and prove its asymptotic preserving, for the low-Mach isentropic Euler equations as well as the low-Froude SWE. The extension of this approach to two-dimensional systems is an active field of research, and it is open for now if one can preserve the AP property. One can consider this chapter almost independently from the others, as in Chapter 3, we initiate another scheme, the so-called *reference solution implicit-explicit (RS-IMEX)* schemes, by analysing the modified equation of a rather general IMEX scheme. The study of the modified equation, albeit is very formal and limited, motivates the RS-IMEX scheme. This scheme is the generalisation of the method introduced in [BALMN14, Bis15]. In the next chapters, Chapters 4 and 5, we use the RS-IMEX machinery for the SWE in one- and two-dimensional cases, respectively, and prove the AP property, both in analysis and numerics. Then, in Chapter 6, we add the Coriolis force and again investigate the asymptotic preserving of the RS-IMEX scheme. We conclude the manuscript with some remarks and perspectives.

Chapter 2

The Lagrange-projection scheme for the low-Froude shallow water equations

"It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."

– Sherlock Holmes, *A Scandal in Bohemia (1891)*

In this chapter, we show that the Lagrange-projection implicit-explicit scheme applied to the one-dimensional isentropic Euler equations, as in [CNPT10], is asymptotic preserving regarding the Mach number. This consistency analysis has been carried out formally and rigorously. Moreover, we prove the positivity preserving and entropy admissibility of the scheme, under some Mach-uniform restrictions. The analysis is similar to what has been presented in the original paper, but with the emphasis on the uniformity regarding the Mach number, which is crucial for a scheme to be useful in the low-Mach regime. We, then, perform a similar analysis for the one-dimensional shallow water equations with topography and obtain similar stability and consistency results. The contents of this chapter are based on [Zak17a].

Contents

2.1 Introduction

The arbitrary Lagrangian-Eulerian (ALE) approach is a classic one in mechanics, trying to benefit from the Eulerian and Lagrangian formulations, simultaneously; see [DHPRF04] for a nice introduction. Recently, Coquel *et al.* in [CNPT10] utilised this approach to split the waves of Euler-like systems, in a very natural way, to fast acoustic waves and slow advection waves. Inspired by this approach, there have been several works like [CGK13, CGK16, CMV15] which investigate the so-called *Lagrange-projection scheme* for the Euler-like system with a large friction [CGK13], or when the Stokes [CMV15] or the Mach number [CGK16] is small. Such studies have been successful in finding some rigorous stability results, *e.g.*, positivity of the density and the discrete energy inequality. Moreover, in [CGK16], the Lagrange-projection scheme has been analysed for the two-dimensional Euler equations to construct an all-Mach scheme (the scheme with the Mach-uniform order of consistency), where the focus was on the accuracy problems in the low-Mach regime for Godunov-type schemes (of which the Lagrange-projection scheme is a member), and to cure them by a careful look at the truncation error. In fact, it has been shown in [CGK16] that the truncation error of the two-dimensional Lagrange-projection scheme blows up in the low-Mach regime, *i.e.*, it behaves as $\mathcal{O}(\frac{\Delta x}{Ma})$, where Ma stands for the Mach number. The authors of [CGK16] could show that the truncation error can be made uniform regarding the Mach number for a particular modification of the scheme, namely by multiplying the dissipation involved in the discretisation of the pressure terms by an $\mathcal{O}(Ma)$ term. Although this is a promising step, it is not clear if this uniform accuracy in terms of the truncation error is equivalent to the asymptotic consistency, due to the lack of convergence analysis of the scheme. So, a crucial point is to analyse the Lagrange-projection scheme for the asymptotic (incompressible) limit of the isentropic Euler equations, *e.g.*, to confirm uniformity of the results of [CNPT10] w.r.t. the Mach number, and to confirm the asymptotic consistency and stability of the scheme. We wish to mention that it is well-known that Godunov-type schemes show no accuracy problem for low-Mach one-dimensional problems as long as the initial condition is well-prepared, as in Definition 2.3.2; see [Del10, Rie11, DOR10, RB09b, CGK16] for more details. This accuracy of Godunov-type schemes motivates the present chapter as we focus only on the one-dimensional case with a well-prepared initial datum. The extension to multi-dimensional cases is not trivial, *cf.* [CGK16, DJOR16].

In this chapter, we study the issue of consistency and stability of the IMEX Lagrange-projection scheme, or the so-called LP-IMEX scheme as has been proposed in [CGK13], in the incompressible limit of the isentropic Euler equations. In particular, we show that the stability conditions in [CNPT10] are uniform in the Mach number provided that the initial condition is well-prepared. So, all the stability properties in [CNPT10] hold without any restriction regarding the Mach number. Also, we show that the solution is asymptotically consistent for well-prepared initial data (see Theorem 2.3.3). The study has also been extended to the one-dimensional shallow water equations with topography, where the source term presents an additional difficulty in proving asymptotic consistency and well-balancing. For this system and very recently, Chalons *et al.* in [CKKS16] investigated the Lagrange-projection framework with particular attention to the well-balancing and the validity of the entropy inequality. Also, note that we prove Mach-uniform bounds for the (implicit) solution in Sections 2.3.2 and 2.4.2, which justify the asymptotic expansions used throughout the chapter. Indeed, these estimates imply convergence of a subsequence of the discrete computed solution to the incompressible limit, for fixed grids, as the Mach number tends to zero, and in virtue of the Bolzano–Weierstrass theorem and a norm equivalence argument (see Appendix 2.B).

The chapter is organised as follows. In Section 2.2 we introduce the splitting along with a brief introduction to the ALE formalism and relaxation schemes. Then, in Section 2.3, we introduce the IMEX Lagrange-projection scheme with a specific relaxation approximation and discuss the numerical analysis of the scheme. We prove the formal asymptotic consistency, positivity preserving, stability and entropy stability, all under a non-restrictive CFL condition. Then we show that the formal asymptotic consistency is, in fact, rigorous. In Section 2.4 we show similar results for the shallow water equations with a non-flat bottom topography. Appendix 2.B provides some results about the implications of entropy stability for the stability of the solution in the incompressible limit.

2.2 Lagrange-projection idea in the continuous level

For the isentropic Euler equations, one natural way to split the system (waves) is the splitting into acoustic and transport systems (waves). Then, the Lagrange-projection scheme [GR96] consists of solving Riemann problems for the acoustic system in the Lagrangian formulation and then projecting the computed solution onto the fixed Eulerian grid (which is equivalent to the transport system). In this way, the scheme handles Riemann problems in the Lagrangian coordinates, which is easier than in the Eulerian ones, and takes advantage of using a fixed grid; see [GR96, Chapter III, Section 2.5]. It is in this regard that the Lagrange-projection scheme can be understood in the framework of the ALE approach (see [CNPT10]), in which the equations are rewritten in the *referential* coordinates χ which are necessarily neither spatial (Eulerian) x nor material (Lagrangian) X. The referential frame has a relative velocity seen from the spatial frame, which is chosen arbitrarily. Note that the Lagrange-projection scheme is a special case of ALE, in which the relative velocity is chosen such that after each step, the domain remains as the fixed Eulerian one; see [CNPT10, Section 3.3] for more details.

Now, consider the system of the isentropic Euler equations in $[0, +\infty) \times \Omega$, where $\Omega = \mathbb{T}$ is a one-dimensional torus (*cf.* (1.3)):

$$\begin{aligned} &\partial_t \varrho + \partial_x(\varrho u) = 0, \\ &\partial_t(\varrho u) + \partial_x\left(\varrho u^2 + p(\varrho)\right) = 0, \end{aligned} \tag{2.1}$$

with given $\varrho_0(x) := \varrho(0, x)$ and $u_0(x) := u(0, x)$, respectively as the initial density and velocity. $p(\varrho) := \kappa\varrho^\gamma$, with $\kappa > 0$ and $\gamma > 1$, is the isentropic pressure law. As an entropy function, we choose the total energy of the solution ϱE, which can be shown to be strictly convex w.r.t. the conservative variables. The total energy density is written as $E = \mathcal{E} + \frac{u^2}{2}$, where $\mathcal{E}(\varrho) := \frac{\kappa}{\gamma-1}\varrho^{\gamma-1}$ is the internal energy density (see [LW07]). For later use and by denoting τ as the specific volume (the reciprocal of ϱ), we should mention that the internal energy function $\mathcal{E}$ fulfils the *Weyl's assumptions* [CNPT10, Wey49]:

$$\mathcal{E} > 0, \qquad \partial_\tau \mathcal{E} = -p < 0, \qquad \partial_{\tau\tau}\mathcal{E} > 0, \qquad \partial_{\tau\tau\tau}\mathcal{E} < 0. \tag{2.2}$$

Remark 2.2.1. *We stick to this general isentropic pressure function $p(\varrho) = \kappa\varrho^\gamma$ except in Section 2.4, where we pick $\kappa = \frac{1}{2}$ and $\gamma = 2$ to investigate the shallow water equations. We also only consider periodic boundary conditions for the sake of simplicity of the presentation. However, we expect that with a bit of effort and by changing some of the arguments particularly for the asymptotic consistency analysis (see Section 2.3.1 and Section 2.4.1), one can generalise*

the present study to other types of boundary conditions such as open boundary condition; see also [CNPT10] for some interesting results for the case of coupling boundary conditions.

Now, we decompose the original system (2.1) into acoustic and transport sub-systems:

$$\begin{cases} \partial_t \varrho + \varrho \partial_x u = 0, \\ \partial_t(\varrho u) + \varrho u \partial_x u + \partial_x p = 0 \end{cases} \tag{2.3a}$$

$$\begin{cases} \partial_t \varrho + u \partial_x \varrho = 0 \\ \partial_t(\varrho u) + u \partial_x(\varrho u) = 0 \end{cases} \tag{2.3b}$$

and solve them successively. Simply by using the Taylor expansion, it can be seen that this splitting is, in general, (globally) first-order accurate in time. We refer the reader to [HKLR10] for more details about the operator splitting methods. Note that the transport part is merely a transport of conservative variables $(\varrho, \varrho u)$ with the velocity field u.

2.2.1 Lagrange step

In the Lagrangian coordinates, the frame moves with the velocity field. So, what an observer sees is the acoustic part (2.3a). It is not difficult to show that it can also be written as

$$\begin{aligned} \partial_t \tau - \partial_{\mathfrak{m}} u &= 0, \\ \partial_t u + \partial_{\mathfrak{m}} p &= 0, \end{aligned} \tag{2.4}$$

where $\mathrm{d}\mathfrak{m} := \varrho \, \mathrm{d}X$ is the *mass coordinate*, attached to the Lagrangian coordinate X, *cf.* [GR96, Chapter III, Section 2.5]. This is exactly the classical form of the isentropic Euler equations in the Lagrangian framework. To obtain the *non-dimensionalised* equations, we set

$$\hat{t} := \frac{t}{t_\circ}, \quad \hat{x} := \frac{x}{L_\circ}, \quad \hat{\tau} := \tau \varrho_\circ, \quad \hat{u} := \frac{u}{u_\circ}, \quad \hat{p} := \frac{p}{p_\circ}, \quad t_\circ := \frac{L_\circ}{u_\circ}, \quad p_\circ := \varrho_\circ c_\circ^2 / \gamma,$$

where $c_\circ$ is the characteristic sound speed, defined as $c_\circ := \sqrt{\gamma p_\circ / \varrho_\circ}$, and $t_\circ, L_\circ, \varrho_\circ$, $p_\circ$ and $u_\circ$ are characteristic time, length, density, pressure and velocity. Also, we denote the Mach number as the ratio of the characteristic speed to the characteristic sound speed, *i.e.*, $Ma := u_\circ / c_\circ$. Thus, after suppressing hats, the equations become

$$\begin{aligned} \partial_t \tau - \partial_{\mathfrak{m}} u &= 0, \\ \partial_t u + \frac{1}{\varepsilon^2} \partial_{\mathfrak{m}} p &= 0, \end{aligned} \tag{2.5}$$

where $\varepsilon := \sqrt{\gamma} Ma$. From now on and for simplicity, we call ε the Mach number, though it is different from Ma by the factor $\sqrt{\gamma}$, *cf.* [KM82]. Note that the system has two acoustic waves with speeds $\pm\sqrt{-p_\tau}/\varepsilon$.

To solve this system with the aforementioned initial data, we *relax* the system so that all characteristic fields get linearly degenerate, which is easy to solve the Riemann problem for. We actually substitute the source of genuine non-linearity $p(\varrho)$ with the variable π, called relaxation pressure and add another equation for π. This is the heart of so-called *relaxation schemes*; we

refer the reader to [Bou04, Liu87, CLL94, JX95] for more details. Like [CNPT10], we employ the Suliciu relaxation system [Bou04, CGS07], which yields the following dimensionless system:

$$\partial_t \tau - \partial_{\mathfrak{m}} u = 0, \tag{2.6a}$$

$$\partial_t u + \partial_{\mathfrak{m}} \pi_\varepsilon = 0, \tag{2.6b}$$

$$\partial_t \pi_\varepsilon + \alpha_\varepsilon^2 \partial_{\mathfrak{m}} u = \lambda_\varepsilon (p - \pi), \tag{2.6c}$$

with the definitions $\pi_\varepsilon := \frac{\pi}{\varepsilon^2}$, $\alpha_\varepsilon := \frac{a}{\varepsilon}$, and $\lambda_\varepsilon := \frac{\lambda}{\varepsilon^2}$, where a is a constant to be specified and λ is the relaxation parameter.

At least formally, one can observe that in the asymptotic regime $\lambda \to \infty$, π tends to p and the original system will be recovered. Also, one can readily check that the relaxation system only has linearly-degenerate characteristic fields. To use the feature of linear degeneracy, at first, we solve the problem out of equilibrium by setting $\lambda = 0$, and then we project the out-of-equilibrium solution to the equilibrium manifold, *cf.* [CNPT10].

In order to prevent the instabilities to happen for this relaxation system, *i.e.*, to enforce the dissipativity of Chapman–Enskog expansion (see [CLL94, Liu87]), the parameter a must be chosen sufficiently large, according to the so-called *sub-characteristic* or *Whitham stability condition* (see [CC08] for the proof):

$$\alpha_\varepsilon^2 > \frac{\max_\Omega |p_\tau|}{\varepsilon^2}. \tag{2.7}$$

Since the relaxation system with $\lambda = 0$ is strictly hyperbolic with eigenvalues $\pm\alpha_\varepsilon$ and zero, the sub-characteristic condition means that information propagates faster in the relaxation model. Also, linear degeneracy of the fields allows us to analytically solve the Riemann problem when $\lambda = 0$ as one can simply put the relaxation system (2.6a)–(2.6c) into an equivalent diagonal form [CGK13, eq. (12)]:

$$\partial_t \tau - \partial_{\mathfrak{m}} u = 0, \tag{2.8a}$$

$$\partial_t \overrightarrow{w} + \alpha_\varepsilon \partial_{\mathfrak{m}} \overrightarrow{w} = 0, \quad \overrightarrow{w} := \pi_\varepsilon + \alpha_\varepsilon u = \frac{\pi}{\varepsilon^2} + \frac{a}{\varepsilon} u, \tag{2.8b}$$

$$\partial_t \overleftarrow{w} - \alpha_\varepsilon \partial_{\mathfrak{m}} \overleftarrow{w} = 0, \quad \overleftarrow{w} := \pi_\varepsilon - \alpha_\varepsilon u = \frac{\pi}{\varepsilon^2} - \frac{a}{\varepsilon} u. \tag{2.8c}$$

This property justifies by itself the introduction of the proposed relaxation model [CGK16]. Note that $\overrightarrow{w}$ and $\overleftarrow{w}$ are two of Riemann invariants of the relaxation system; the third one is $\mathscr{I} := \pi_\varepsilon + \alpha_\varepsilon^2 \tau$. So, instead of (2.8a) one can use $\partial_t \mathscr{I} = 0$.

Remark 2.2.2. *Naturally-split systems* (2.3a) *and* (2.3b) *are not conservative if they are written in the Eulerian coordinates. As shown in [CNPT10], changing the coordinates to the Lagrangian ones not only helps solving Riemann problems, but also provides a conservative formulation to circumvent the complications stemming from non-conservative products, cf. [DMLM95].*

2.2.2 Projection step

This step is in fact equivalent to remapping the updated solution onto the Eulerian grid so that the referential and spatial (Eulerian) coordinates coincide at the end of each step. Following the

notation in [CNPT10] and with $\phi \in \{\varrho, \varrho u\}$ as a conservative variable, the projection step can be summarised as

$$\partial_t \phi + u \partial_x \phi = 0. \tag{2.9}$$

Like the acoustic part, the transport part (2.3b) or (2.9) can be written in the Lagrangian coordinates, which provides a conservative form; for further details consult [CNPT10]

2.3 LP-IMEX scheme for the isentropic Euler equations

As already mentioned, it is straightforward to solve the Riemann problem for linearly-degenerate systems. In fact, one of the Riemann invariants remains constant along each characteristic line, implying that there is only one set of symmetric scalar linear advection equations to be solved for $\overrightarrow{w}$ and $\overleftarrow{w}$, while $\mathscr{I}$ does not change at all.

At the beginning of the Lagrange (acoustic) step, from temporal step n to some intermediate step $n + 1/2$, the Eulerian and Lagrangian coordinates coincide with each other. Furthermore, the solution of the relaxation system is at equilibrium such that $\pi^n = p^n$. The implicit Lagrange step reads

$$\tau_j^{n+1/2} = \tau_j^n + \frac{\Delta t}{\Delta \mathfrak{m}_j^n} \left(\widetilde{u}_{j+1/2}^{n+1/2} - \widetilde{u}_{j-1/2}^{n+1/2} \right), \tag{2.10a}$$

$$\overrightarrow{w}_j^{n+1/2} = \overrightarrow{w}_j^n - \frac{a \Delta t}{\varepsilon \Delta \mathfrak{m}_j^n} \left(\overrightarrow{w}_j^{n+1/2} - \overrightarrow{w}_{j-1}^{n+1/2} \right), \tag{2.10b}$$

$$\overleftarrow{w}_j^{n+1/2} = \overleftarrow{w}_j^n + \frac{a \Delta t}{\varepsilon \Delta \mathfrak{m}_j^n} \left(\overleftarrow{w}_{j+1}^{n+1/2} - \overleftarrow{w}_j^{n+1/2} \right), \tag{2.10c}$$

where $\Delta \mathfrak{m}_j^n := \varrho_j^n \Delta x$, Δx and Δt are spatial and time steps, and $j \in \{1, 2, \ldots, N\}$ denotes cell indices in the computational domain Ω_N. The interface velocity $\widetilde{u}_{j+1/2}^{n+1/2}$ comes from solving a simple Riemann problem for the relaxation system (2.6a)–(2.6c) with $\lambda = 0$ (see [CNPT10]), and writes

$$\widetilde{u}_{j+1/2}^{n+1/2} := \frac{1}{2} \left(u_j^{n+1/2} + u_{j+1}^{n+1/2} \right) - \frac{1}{2a\varepsilon} \left(\pi_{j+1}^{n+1/2} - \pi_j^{n+1/2} \right). \tag{2.11}$$

Note that there are several (equivalent) variants of the scheme (2.10a)–(2.10c), in different coordinates or with/without using the Riemann invariants; see [CNPT10] for further details.

In the next step, the explicit projection step from the intermediate step $n + 1/2$ to the new temporal step $n + 1$, we map updated values onto the fixed Eulerian grid. There are four cases based on the upwind direction [CGK13, eq. (34)], which can be summarised as

$$\phi_j^{n+1} = \phi_j^{n+1/2} + \frac{\Delta t}{\Delta x} \left[(\widetilde{u}_{j-1/2}^{n+1/2})^+ \phi_{j-1}^{n+1/2} + \left((\widetilde{u}_{j+1/2}^{n+1/2})^- - (\widetilde{u}_{j-1/2}^{n+1/2})^+ \right) \phi_j^{n+1/2} - (\widetilde{u}_{j+1/2}^{n+1/2})^- \phi_{j+1}^{n+1/2} \right], \tag{2.12}$$

with the definitions of the *positive part* $\cdot^+ := \frac{\cdot + |\cdot|}{2}$ and the *negative part* $\cdot^- := \frac{\cdot - |\cdot|}{2}$. Figure 2.1 indicates that how the projection of the solution on the fixed Eulerian grid can be interpreted as

upwinding. $X^{n+1/2}_{j+1/2}$ is the new location of the interface $X^{n}_{j+1/2}$ after the Lagrange step and can be computed by knowing $\widetilde{u}^{n+1/2}_{j+1/2}$. So, the mean value of some quantity ϕ on the fixed grid can be computed based on the direction (sign) of this velocity by finding the contributions of the cell itself as well as its neighbours. For instance in Figure 2.1, there are contributions from the left neighbour and from the cell itself. Adding this projection step to the Lagrange step is what we call the LP-IMEX scheme.

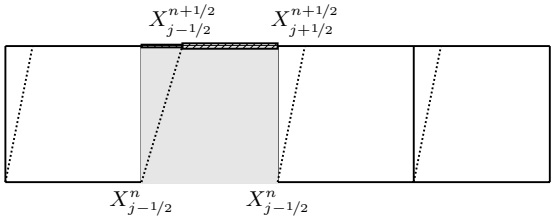


Figure 2.1: The Lagrange update of the grid and the interpretation of the projection step.

2.3.1 Numerical analysis of the LP-IMEX scheme

Considering the LP-IMEX scheme introduced in the previous section, one can obtain some stability results, gathered in Theorem 2.3.3 below. But, firstly and for future reference, let us define the formal incompressible limit of the isentropic Euler equations and the so-called well-prepared initial datum, with the following asymptotic (or Poincaré) expansion for density and velocity

$$\begin{aligned} \varrho(x,t) &= \varrho_{(0)} + \varepsilon\varrho_{(1)} + \varepsilon^2\varrho_{(2)}, \\ u(x,t) &= u_{(0)} + \varepsilon u_{(1)} + \varepsilon^2 u_{(2)}. \end{aligned} \tag{2.13}$$

Definition 2.3.1. *The formal incompressible limit of the isentropic Euler equations* (2.1) *gives the incompressible isentropic Euler equations, and reads (see Appendix 2.A for the formal and [KM82] for the rigorous justification)*

$$\begin{aligned} &\varrho_{(0)}, \varrho_{(1)} = const., \\ &\partial_x u_{(0)} = 0, \\ &\partial_t u_{(0)} + \partial_x \left(u^2_{(0)} + p_{(2)} \right) = 0. \end{aligned}$$

Definition 2.3.2. *For the isentropic Euler equations* (2.1)*, we call the initial data* $(\varrho_{0,\varepsilon}, u_{0,\varepsilon})$ *well-prepared if it holds that (see [KM82, Mei99])*

$$\begin{aligned} \varrho(0,\cdot) = \varrho_{0,\varepsilon} &:= \varrho^0_{(0)} + \varepsilon^2 \varrho^0_{(2),\varepsilon}, \\ u(0,\cdot) = u_{0,\varepsilon} &:= u^0_{(0)} + \varepsilon u^0_{(1),\varepsilon}, \end{aligned} \tag{2.14}$$

where $\varrho_{(0)}$ *and* $u_{(0)}$ *are constant.*

Theorem 2.3.3. *The Lagrange-projection scheme* (2.10a)–(2.10c) *and* (2.12) *with a well-prepared initial datum, satisfies the following properties:*

(i) It can be expressed in the locally conservative form.

(ii) The scheme is AC, i.e., it gives a consistent discretisation of the incompressible Euler equations, as $\varepsilon \to 0$, in terms of Definition 2.3.1.

(iii) Under the ε-uniform CFL constraint (2.23)*, the scheme is positivity preserving, i.e., $\varrho_j^n > 0$ provided that $\varrho_j^0 > 0$ for all $j \in \Omega_N$. Moreover, the density is bounded away from zero, i.e., there exists some $\varrho_{\mathrm{LB}}^n > 0$ such that $\varrho_j^n \geq \varrho_{\mathrm{LB}}^n$ for all $j \in \Omega_N$.*

(iv) Under the CFL constraint (2.23) *and the sub-characteristic condition* (2.33)*, the solution fulfils the local (cell) entropy (energy) inequality, i.e.,*

$$(\varrho E)_j^{n+1} - (\varrho E)_j^n + \frac{\Delta t}{\Delta x}\left[\left(\varrho E\widetilde{u} + \frac{\widetilde{\pi}\widetilde{u}}{\varepsilon^2}\right)_{j+1/2}^{n+1/2} - \left(\varrho E\widetilde{u} + \frac{\widetilde{\pi}\widetilde{u}}{\varepsilon^2}\right)_{j-1/2}^{n+1/2}\right] \leq 0, \tag{2.15}$$

which is consistent with the energy inequality accompanied by the isentropic Euler equations (2.1) $\partial_t(\varrho E) + \partial_x((\varrho E + \frac{p}{\varepsilon^2})u) \leq 0$.

(v) Under the CFL constraint (2.23) *and the sub-characteristic condition* (2.33)*, the computed density, momentum and velocity are stable, i.e., bounded in the ℓ_∞-norm, uniformly in ε.*

We analyse the properties of this scheme in the subsequent subsections. Note that the locally conservative form of the scheme is proved in [CNPT10] and skipped here.

Remark 2.3.4. *Throughout this section and subsequent ones, it is very natural to ask about the order of magnitudes of quantities (in terms of ε). For now, we only do the analysis formally, that is to say, we only take the explicit ε into account and assume that all other quantities are $\mathcal{O}(1)$. In Section 2.3.2, we will justify this assumption.*

2.3.1.1 Asymptotic consistency

At first, we show that the solution is consistent with the incompressible limit in the sense of Definition 2.3.1, *i.e.*, the computed density is constant up to $\mathcal{O}(\varepsilon^2)$, and the calculated velocity is divergence-free (solenoidal) to the leading order. Then, using these results, we prove that the scheme provides a consistent discretisation of the incompressible Euler equation in the limit $\varepsilon \to 0$. Thus, the asymptotic consistency in the sense of Definition 1.2.1 holds.

Considering Definitions 2.3.1 and 2.3.2, we consider a well-prepared solution at step n, *i.e.*,

$$\varrho_j^n = \varrho_{(0)}^n + \varepsilon^2 \varrho_{(2)j}^n,$$
$$u_j^n = u_{(0)}^n + \varepsilon u_{(1)j}^n,$$

where $\varrho_{(0)}^n$ and $u_{(0)}^n$ are constant values. Also, the pressure is at equilibrium, $\pi_j^n \equiv p_j^n$. We aim to show that the scheme (2.10a)–(2.10c) preserves the well-preparedness of the solution at the step n to the intermediate step $n + 1/2$ and then to the next step $n + 1$.

For the Lagrange step, we substitute the asymptotic expansion (2.13) (for τ, u and π) into the scheme and balance terms w.r.t. ε. The $\mathcal{O}(1/\varepsilon)$ terms in the τ-update (2.10a) yield

$$\pi_{(0)j+1}^{n+1/2} - 2\pi_{(0)j}^{n+1/2} + \pi_{(0)j-1}^{n+1/2} = 0.$$

So, $\{\pi_{(0)j}^{n+1/2}\}_{j\in\Omega_N}$ is a linear sequence, thus constant in space due to the periodicity of the domain, *i.e.*, $\pi_{(0)\Delta}^{n+1/2} = \pi_{(0)\Delta}^{n+1/2}$, where Δ stands for a discretised vector, *i.e.*, for all $j \in \Omega_N$.

Since the pressure is not at equilibrium anymore (for step $n+1/2$), π and ϱ are two independent variables and we cannot conclude immediately that the same result holds for the density. But one can establish their relation by combining (2.10a)–(2.10c) to find the update for the relaxation pressure as

$$\varrho_j^n \left(\pi_j^{n+1/2} - \pi_j^n\right) + \frac{a^2\Delta t}{\Delta x}\left(\widetilde{u}_{j+1/2}^{n+1/2} - \widetilde{u}_{j-1/2}^{n+1/2}\right) = 0. \tag{2.16}$$

Then, using the τ-update (2.10a), it yields

$$a^2\left(\tau_j^{n+1/2} - \tau_j^n\right) + \left(\pi_j^{n+1/2} - \pi_j^n\right) = 0. \tag{2.17}$$

It is clear from (2.17) that

$$a^2\big(\varrho_j^{n+1/2} - \varrho_j^n\big) = \varrho_j^n \varrho_j^{n+1/2}\big(\pi_j^{n+1/2} - \pi_j^n\big),$$

which gives that $\varrho_{(0)j}^{n+1/2}(a^2 - \varrho_{(0)}^n(\pi_{(0)}^{n+1/2} - \pi_{(0)}^n)) = a^2\varrho_{(0)}^n$. So, $\varrho_{(0)\Delta}^{n+1/2} = \varrho_{(0)\Delta}^{n+1/2}$, constant in space. Then, due to periodicity and by a spatial summation on (2.10a), it can be obtained that $\varrho_{(0)\Delta}^{n+1/2}$ is constant in time as well, *i.e.*, $\varrho_{(0)\Delta}^{n+1/2} = \varrho_{(0)\Delta}^n$. Also, from the update for the relaxation pressure, (2.16), and again due to periodicity, the numerical fluxes cancel each other out and it turns out that $\pi_{(0)\Delta}^{n+1/2} = \pi_{(0)\Delta}^n$, constant in both time and space.

We then continue with the $\overrightarrow{w}$-update (2.10b):

$$\varrho_j^n\big(\frac{1}{\varepsilon^2}\pi_j^{n+1/2} + \frac{a}{\varepsilon}u_j^{n+1/2}\big) = \varrho_j^n\big(\frac{\pi_j^n}{\varepsilon^2} + \frac{a}{\varepsilon}u_j^n\big) - \frac{a\Delta t}{\varepsilon^2\Delta x}\Big((\pi_j^{n+1/2} - \pi_{j-1}^{n+1/2})/\varepsilon + a\big(u_j^{n+1/2} - u_{j-1}^{n+1/2}\big)\Big).$$

So, balancing $\mathcal{O}(1/\varepsilon^2)$ terms yields

$$\varrho_{(0)}^n\pi_{(0)j}^{n+1/2} = \varrho_{(0)}^n\pi_{(0)}^n - \frac{a\Delta t}{\Delta x}\Big(\pi_{(1)j}^{n+1/2} - \pi_{(1)j-1}^{n+1/2} + a\big(u_{(0)j}^{n+1/2} - u_{(0)j-1}^{n+1/2}\big)\Big),$$

which gives

$$\pi_{(1)j}^{n+1/2} - \pi_{(1)j-1}^{n+1/2} + a\big(u_{(0)j}^{n+1/2} - u_{(0)j-1}^{n+1/2}\big) = 0. \tag{2.18}$$

So, there is the possibility that both $\pi_{(1)\Delta}^{n+1/2}$ and $u_{(0)\Delta}^{n+1/2}$ are constant in space. Showing this, let us balance $\mathcal{O}(1)$ terms in (2.10a):

$$\begin{aligned}\varrho_{(0)j}^n = \varrho_{(0)j}^{n+1/2}&\left(1 + \frac{\Delta t}{2a\Delta x}\Big(a\big(u_{(0)j+1}^{n+1/2} - u_{(0)j-1}^{n+1/2}\big) - \big(\pi_{(1)j-1}^{n+1/2} - 2\pi_{(1)j}^{n+1/2} + \pi_{(1)j+1}^{n+1/2}\big)\Big)\right)\\ &- \frac{\Delta t}{2a\Delta x}\varrho_{(1)j}^{n+1/2}\big(\pi_{(0)j-1}^{n+1/2} - 2\pi_{(0)j}^{n+1/2} + \pi_{(0)j+1}^{n+1/2}\big).\end{aligned}$$

So,

$$a\big(u_{(0)j+1}^{n+1/2} - u_{(0)j-1}^{n+1/2}\big) - \big(\pi_{(1)j-1}^{n+1/2} - 2\pi_{(1)j}^{n+1/2} + \pi_{(1)j+1}^{n+1/2}\big) = 0. \tag{2.19}$$

Combining (2.19) and (2.18) yields that $\pi_{(1)j}^{n+1/2} = \pi_{(1)j+1}^{n+1/2}$ and $u_{(0)j}^{n+1/2} = u_{(0)j+1}^{n+1/2}$. Similar to the analysis of the leading order terms, one can show that $\pi_{(1)\Delta}^{n+1/2}$ and $\varrho_{(1)\Delta}^{n+1/2}$ are constant in time as well as in space. Hence, the solution of the Lagrange step is consistent with the incompressible limit.

For the projection step (2.12), we only show the asymptotic consistency for the case with $\widetilde{u}_{j-1/2}^{n+1/2} < 0$ and $\widetilde{u}_{j+1/2}^{n+1/2} < 0$; other cases can be analysed similarly. For the density, it can be seen that

$$\varrho_j^{n+1} = \varrho_j^{n+1/2} - \frac{\Delta t}{2a\Delta x}\big(\varrho_{j+1}^{n+1/2} - \varrho_j^{n+1/2}\big)\Big(- (\pi_{j+1}^{n+1/2} - \pi_j^{n+1/2})/\varepsilon + a\big(u_{j+1}^{n+1/2} + u_j^{n+1/2}\big)\Big).$$

So, the leading order terms give

$$\begin{aligned}\varrho_{(0)j}^{n+1} = \varrho_{(0)j}^{n+1/2} - \frac{\Delta t}{2a\Delta x}\Big[&- \big(\varrho_{(0)j+1}^{n+1/2} - \varrho_{(0)j}^{n+1/2}\big)\big(\pi_{(1)j+1}^{n+1/2} - \pi_{(1)j}^{n+1/2}\big)\\ &- \big(\varrho_{(1)j+1}^{n+1/2} - \varrho_{(1)j}^{n+1/2}\big)\big(\pi_{(0)j+1}^{n+1/2} - \pi_{(0)j}^{n+1/2}\big)\\ &+ a\big(\varrho_{(0)j+1}^{n+1/2} - \varrho_{(0)j}^{n+1/2}\big)\big(u_{(0)j+1}^{n+1/2} + u_{(0)j}^{n+1/2}\big)\Big],\end{aligned}$$

which implies that the leading order of the computed density is constant, *i.e.*, $\varrho_{(0)\Delta}^{n+1} = \varrho_{(0)\Delta}^{n+1/2} = \varrho_{(0)\Delta}^{n}$. Similarly, one can find that the first-order components are also constant in time and space, *i.e.*, if they do not exist at t_n, there is no $\mathcal{O}(\varepsilon)$ density (or pressure) fluctuation at t_{n+1}:

$$\begin{aligned}\varrho_{(1)j}^{n+1} = \varrho_{(1)j}^{n+1/2} - \frac{\Delta t}{2a\Delta x}\Big[&- \big(\varrho_{(0)j+1}^{n+1/2} - \varrho_{(0)j}^{n+1/2}\big)\big(\pi_{(2)j+1}^{n+1/2} - \pi_{(2)j}^{n+1/2}\big)\\ &- \big(\varrho_{(1)j+1}^{n+1/2} - \varrho_{(1)j}^{n+1/2}\big)\big(\pi_{(1)j+1}^{n+1/2} - \pi_{(1)j}^{n+1/2}\big)\\ &- \big(\varrho_{(2)j+1}^{n+1/2} - \varrho_{(2)j}^{n+1/2}\big)\big(\pi_{(0)j+1}^{n+1/2} - \pi_{(0)j}^{n+1/2}\big)\\ &+ a\big(\varrho_{(0)j+1}^{n+1/2} - \varrho_{(0)j}^{n+1/2}\big)\big(u_{(1)j+1}^{n+1/2} + u_{(1)j}^{n+1/2}\big)\\ &+ a\big(\varrho_{(1)j+1}^{n+1/2} - \varrho_{(1)j}^{n+1/2}\big)\big(u_{(0)j+1}^{n+1/2} + u_{(0)j}^{n+1/2}\big)\Big],\end{aligned}$$

which confirms that $\varrho_{(1)\Delta}^{n+1} = \varrho_{(1)\Delta}^{n+1/2} = \varrho_{(1)\Delta}^{n} = \mathbf{0}$.

To show the *div*-free condition, we consider $\mathcal{O}(1)$ terms of the momentum update in (2.12):

$$\begin{aligned}\varrho_{(0)j}^{n+1}u_{(0)j}^{n+1} = \varrho_{(0)j}^{n+1/2}u_{(0)j}^{n+1/2} - \frac{\Delta t}{2a\Delta x}\Big[&- \big(\varrho_{(0)j+1}^{n+1/2}u_{(0)j+1}^{n+1/2} - \varrho_{(0)j}^{n+1/2}u_{(0)j}^{n+1/2}\big)\big(\pi_{(1)j+1}^{n+1/2} - \pi_{(1)j}^{n+1/2}\big)\\ &- \big(\varrho_{(1)j+1}^{n+1/2}u_{(0)j+1}^{n+1/2} - \varrho_{(1)j}^{n+1/2}u_{(0)j}^{n+1/2}\big)\big(\pi_{(0)j+1}^{n+1/2} - \pi_{(0)j}^{n+1/2}\big)\\ &- \big(\varrho_{(0)j+1}^{n+1/2}u_{(1)j+1}^{n+1/2} - \varrho_{(0)j}^{n+1/2}u_{(1)j}^{n+1/2}\big)\big(\pi_{(0)j+1}^{n+1/2} - \pi_{(0)j}^{n+1/2}\big)\\ &+ a\big(\varrho_{(0)j+1}^{n+1/2}u_{(0)j+1}^{n+1/2} - \varrho_{(0)j}^{n+1/2}u_{(0)j}^{n+1/2}\big)\big(u_{(0)j+1}^{n+1/2} + u_{(0)j}^{n+1/2}\big)\Big].\end{aligned}$$

Thus, $u_{(0)\Delta}^{n+1} = u_{(0)\Delta}^{n+1/2} = u_{(0)\Delta}^{n}$, and the leading order component of the velocity field is constant (solenoidal). Hence, combining the results for the Lagrange and projection steps together, it is obvious that the limit conditions are satisfied.

To prove asymptotic consistency in the sense of Definition 1.2.1, it remains to show the consistency of the discretisation in the limit. The consistency holds for the Lagrange step if the

velocity update

$$\frac{u_j^{n+1/2}-u_j^n}{\Delta t}+\frac{1}{2\varepsilon^2\Delta\mathfrak{m}_j^n}\left(\pi_{j+1}^{n+1/2}-\pi_{j-1}^{n+1/2}\right)-\frac{a/\varepsilon}{2\Delta\mathfrak{m}_j^n}\left(u_{j+1}^{n+1/2}-2u_j^{n+1/2}+u_{j-1}^{n+1/2}\right)=0, \quad (2.20)$$

is a consistent discretisation of $\partial_t u+\frac{1}{\varepsilon^2}\partial_{\mathfrak{m}}\pi=0$ in the limit, when (2.20) gives

$$\frac{u_{(0)j}^{n+1/2}-u_{(0)j}^n}{\Delta t}+\frac{1}{2\Delta\mathfrak{m}_{(0)j}^n}\left(\pi_{(2)j+1}^{n+1/2}-\pi_{(2)j-1}^{n+1/2}\right)-\frac{a}{2\Delta\mathfrak{m}_{(0)j}^n}\left(u_{(1)j+1}^{n+1/2}-2u_{(1)j}^{n+1/2}+u_{(1)j-1}^{n+1/2}\right)=0. \quad (2.21)$$

It is clear that (2.21) is a Rusanov-type scheme applied to $\partial_t u_{(0)}+\partial_{\mathfrak{m}}\pi_{(2)}=0$; so, the Lagrange step is AC.

To show the consistency of the discretisation in the limit for the projection step, we compare (2.9) and (2.12). So, it is sufficient to confirm that $\widetilde{u}_{(0)j+1/2}^{n+1/2}$ is consistent with $u_{(0)}$. This is, in fact, the case, due to the definition of $\widetilde{u}_{j+1/2}^{n+1/2}$ in (2.11) and the asymptotic behaviour of $u_{(0)\Delta}^{n+1/2}$ and $\pi_{(1)\Delta}^{n+1/2}$, namely that $u_{(0)\Delta}^{n+1/2}$ and $\pi_{(1)\Delta}^{n+1/2}$ are constant in space. So, the projection step (2.12) is a consistent discretisation of (2.9) and the scheme is AC in the sense of Definition 1.2.1.

2.3.1.2 Density positivity

In this section, we show that the density is positive under a time step condition which is not restrictive for $\varepsilon \ll 1$. Like [CNPT10, eq. (2.25a)], we define the *local acoustic CFL ratio* $\mu_j := \frac{a\Delta t}{\Delta\mathfrak{m}_j^n}$ and the *local apparent propagation factor* $e_j := \frac{\mu_j/\varepsilon}{1+\mu_j/\varepsilon}$. Then, one can write the Lagrange step (2.10b) as

$$\overrightarrow{w}_j^{n+1/2}=e_j\,\overrightarrow{w}_{j-1}^{n+1/2}+(1-e_j)\,\overrightarrow{w}_j^n.$$

Since $0<e_j<1$ (which can be satisfied uniformly in ε), the updates for $\overrightarrow{w}$ and $\overleftarrow{w}$ are monotone, *i.e.*, no new extremum can be generated. To show it for $\overrightarrow{w}^{n+1/2}$, assume that i is the index of maximum value of $\overrightarrow{w}_j^{n+1/2}$, that is $\overrightarrow{w}_i^{n+1/2}\geq\ \overrightarrow{w}_j^{n+1/2}$ for all $j\in\Omega_N$. So,

$$\overrightarrow{w}_i^{n+1/2}\leq e_i\,\overrightarrow{w}_i^{n+1/2}+(1-e_i)\,\overrightarrow{w}_i^n.$$

Thus, $\overrightarrow{w}_i^{n+1/2}\leq\overrightarrow{w}_i^n$; so, it is bounded from above. The proofs for the lower-bound and $\overleftarrow{w}^{n+1/2}$ are likewise. Hence, defining the upper-bounds $\overrightarrow{M}^n$ and $\overleftarrow{M}^n$ and the lower-bounds $\overrightarrow{m}^n$ and $\overleftarrow{m}^n$ for $\overrightarrow{w}^n$ and $\overleftarrow{w}^n$, one can write for all $j\in\Omega_N$

$$\overrightarrow{m}^n\leq\overrightarrow{w}_j^{n+1/2}\leq\overrightarrow{M}^n,\qquad\overleftarrow{m}^n\leq\overleftarrow{w}_j^{n+1/2}\leq\overleftarrow{M}^n. \quad (2.22)$$

With the bound (2.22) at our disposal, one can show the following theorem.

Theorem 2.3.5. *For some Δt satisfying*

$$\frac{\Delta t}{\Delta x}\leq\frac{2a/\varepsilon}{\left(\overrightarrow{M}^n-\overleftarrow{m}^n\right)^+-\left(\overrightarrow{m}^n-\overleftarrow{M}^n\right)^-}, \quad (2.23)$$

the LP-IMEX scheme preserves the positivity of density provided that $\varrho_j^0>0$ for all $j\in\Omega_N$.

Proof. Along the lines of [CNPT10], for the Lagrange step to satisfy positivity, one gets from the τ-update (2.10a) that

$$\frac{\Delta t}{\Delta x}\left(\widetilde{u}^{n+1/2}_{j-1/2} - \widetilde{u}^{n+1/2}_{j+1/2}\right) < 1, \tag{2.24}$$

which ensures $\varrho_j^{n+1/2} > 0$ for all $j \in \Omega_N$. But on the other hand, Δt should be such that the projection step is a convex combination, which requires

$$\frac{\Delta t}{\Delta x}\left(\left(\widetilde{u}^{n+1/2}_{j-1/2}\right)^+ - \left(\widetilde{u}^{n+1/2}_{j+1/2}\right)^-\right) < 1. \tag{2.25}$$

Between the conditions (2.24) and (2.25), the stronger condition should be chosen, which is (2.25). Then, based on the definition of $\widetilde{u}^{n+1/2}$, we express Δt in terms of $\overrightarrow{M}$, $\overleftarrow{M}$, $\overrightarrow{m}$ and $\overleftarrow{m}$, which concludes the proof. □

The next goal is to show the ε-uniformity of this bound for the time step, *i.e.*, to show that the bound (2.23) does not vanish for $\varepsilon \to 0$. One can pose the following corollary.

Corollary 2.3.6. *For well-prepared initial data, the time step restriction* (2.23) *is ε-uniform.*

Proof. Recall that asymptotic consistency implies that with a well-prepared initial datum and for $\varepsilon \ll 1$, the density (and thus the pressure) has a constant leading order term. So, the differences $\overrightarrow{M}^n - \overleftarrow{m}^n$ and $\overrightarrow{m}^n - \overleftarrow{M}^n$ are not of $\mathcal{O}(1/\varepsilon^2)$ but $\mathcal{O}(1/\varepsilon)$; thus, the CFL constraint (2.23) is uniform in ε. In other words, using the asymptotic expansion for a well-prepared datum gives

$$\begin{aligned}
\overrightarrow{M}^n &\leq \frac{p^n_{(0)}}{\varepsilon^2} + \max_{j\in\Omega_N} p^n_{(2)j} + \frac{a}{\varepsilon}\left(u^n_{(0)} + \varepsilon \max_{j\in\Omega_N}(u^n_{(1)j})\right),\\
\overrightarrow{m}^n &\geq \frac{p^n_{(0)}}{\varepsilon^2} + \min_{j\in\Omega_N} p^n_{(2)j} + \frac{a}{\varepsilon}\left(u^n_{(0)} + \varepsilon \min_{j\in\Omega_N}(u^n_{(1)j})\right),\\
\overleftarrow{M}^n &\leq \frac{p^n_{(0)}}{\varepsilon^2} + \max_{j\in\Omega_N} p^n_{(2)j} - \frac{a}{\varepsilon}\left(u^n_{(0)} + \varepsilon \max_{j\in\Omega_N}(u^n_{(1)j})\right),\\
\overleftarrow{m}^n &\geq \frac{p^n_{(0)}}{\varepsilon^2} + \min_{j\in\Omega_N} p^n_{(2)j} - \frac{a}{\varepsilon}\left(u^n_{(0)} + \varepsilon \min_{j\in\Omega_N}(u^n_{(1)j})\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
\overrightarrow{M}^n - \overleftarrow{m}^n &\leq \frac{a}{\varepsilon}\left(2u^n_{(0)} + \varepsilon\left(\max_{j\in\Omega_N}(u^n_{(0)j}) + \min_{j\in\Omega_N}(u^n_{(1)j})\right)\right) + \left(\max_{j\in\Omega_N} p^n_{(2)j} - \min_{j\in\Omega_N} p^n_{(0)j}\right),\\
\overrightarrow{m}^n - \overleftarrow{M}^n &\leq \frac{a}{\varepsilon}\left(2u^n_{(0)} - \varepsilon\left(\max_{j\in\Omega_N}(u^n_{(1)j}) + \min_{j\in\Omega_N}(u^n_{(1)j})\right)\right) - \left(\max_{j\in\Omega_N} p^n_{(2)j} - \min_{j\in\Omega_N} p^n_{(0)j}\right),
\end{aligned}$$

and one gets

$$\lim_{\varepsilon\to 0}\left[\frac{2a/\varepsilon}{\left(\overrightarrow{M}^n - \overleftarrow{m}^n\right)^+ - \left(\overrightarrow{m}^n - \overleftarrow{M}^n\right)^-}\right] \geq \frac{2a/\varepsilon}{\mathcal{O}(\frac{1}{\varepsilon}) + \mathcal{O}(1)} \geq C. \tag{2.26}$$

Hence, there is an $\mathcal{O}(1)$ constant as the lower-bound of the estimate (2.23), *i.e.*, the condition (2.23) is uniform in ε. □

The following lemma shows that the density is also bounded from below for a finite time.

Lemma 2.3.7. *Under the condition* (2.23)*, the computed density* $\{\varrho_j^{n+1}\}_{j\in\Omega_N}$ *is bounded away from zero in a finite time, where the lower-bound is given by*

$$\varrho_{\mathrm{LB}}^{n+1} := \frac{\min_{j\in\Omega_N}\varrho_j^n}{1+\dfrac{\varepsilon\Delta t}{2a\Delta x}\left[\left(\overrightarrow{M}^n+\overleftarrow{M}^n\right)-(\overrightarrow{m}^n+\overleftarrow{m}^n)\right]}>0. \tag{2.27}$$

Proof. From the τ-update (2.10a) and $\widetilde{u}_{j+1/2}^{n+1/2}=\frac{\varepsilon}{2a}(\overrightarrow{w}_j^{n+1/2}-\overleftarrow{w}_{j+1}^{n+1/2})$, one can get

$$\varrho_j^n=\varrho_j^{n+1/2}\left(1+\frac{\varepsilon\Delta t}{2a\Delta x}\left(\overrightarrow{w}_j^{n+1/2}-\overleftarrow{w}_{j+1}^{n+1/2}-\overrightarrow{w}_{j-1}^{n+1/2}+\overleftarrow{w}_j^{n+1/2}\right)\right).$$

So, to find the minimum value of the computed density, one should determine the maximum value of the rhs. Due to (2.22) and under the condition (2.24), it can be seen that

$$\varrho_j^{n+1/2}\geq\frac{\varrho_j^n}{1+\dfrac{\varepsilon\Delta t}{2a\Delta x}\left[\left(\overrightarrow{M}^n+\overleftarrow{M}^n\right)-(\overrightarrow{m}^n+\overleftarrow{m}^n)\right]}.$$

Thus, since the projection step is a convex combination under the condition (2.23), the lower-bound is obtained as (2.27). □

2.3.1.3 Local energy inequality

We show that the scheme satisfies the energy inequality under an ε-independent time step restriction. For the Lagrange step, based on [CNPT10, Theorem 2.3], we define the entropy function for the symmetric advection problem, (2.8b)–(2.8c), as

$$\eta(\overrightarrow{w},\overleftarrow{w}):=s(\overrightarrow{w})+s(\overleftarrow{w}),\quad s(w):=\frac{\varepsilon^2w^2}{4a^2}.$$

So, it can be rewritten as

$$\eta(\overrightarrow{w},\overleftarrow{w})=\frac{1}{2}\left(u^2+\frac{\pi^2}{\varepsilon^2a^2}\right)=E-\frac{\mathcal{E}}{\varepsilon^2}+\frac{\pi^2}{2a^2\varepsilon^2} \tag{2.28}$$

since after non-dimensionalisation, one gets $E=\frac{\mathcal{E}}{\varepsilon^2}+\frac{u^2}{2}$ where $\mathcal{E}(\varrho)=\frac{\kappa}{\gamma-1}\varrho^{\gamma-1}$. We also define an entropy flux function $q(\overrightarrow{w},\overleftarrow{w})$ as

$$q(\overrightarrow{w},\overleftarrow{w}):=\frac{a}{\varepsilon}\left(s(\overrightarrow{w})-s(\overleftarrow{w})\right)=\frac{\pi u}{\varepsilon^2}. \tag{2.29}$$

Then, the cell entropy inequality reads

$$\eta_j^{n+1/2}-\eta_j^n+\frac{\Delta t}{\Delta\mathfrak{m}_j^n}\left(q_{j+1/2}^{n+1/2}-q_{j-1/2}^{n+1/2}\right)\leq 0. \tag{2.30}$$

Substituting (2.28) and (2.29) into (2.30), one can relate the entropy inequality for the symmetric advection problem to the energy inequality for the acoustic sub-system, *i.e.*,

$$\varrho_j^n\left(E_j^{n+1/2}-E_j^n\right)+\frac{\Delta t}{\Delta x}\left(\left(\frac{\pi u}{\varepsilon^2}\right)_{j+1/2}^{n+1/2}-\left(\frac{\pi u}{\varepsilon^2}\right)_{j-1/2}^{n+1/2}\right)\leq\frac{\varrho_j^n}{\varepsilon^2}\underbrace{\left[\mathcal{E}_j^{n+1/2}-\mathcal{E}_j^n-\frac{\left(\pi_j^{n+1/2}\right)^2-\left(\pi_j^n\right)^2}{2a^2}\right]}_{=:\mathcal{R}_j^{n+1/2}}. \tag{2.31}$$

Then, to prove entropy stability of the scheme, one should show that the entropy residual $\mathcal{R}_j^{n+1/2}$ is non-positive. Considering $\pi_j^n=p_j^n$ and due to (2.17), we rewrite $\mathcal{R}_j^{n+1/2}$ as

$$\begin{aligned}\mathcal{R}_j^{n+1/2}&:=\mathcal{E}_j^{n+1/2}-\mathcal{E}_j^n-\frac{p_j^n}{2a^2}\left(\pi_j^{n+1/2}-p_j^n\right)-\frac{\left(\pi_j^{n+1/2}-p_j^n\right)^2}{2a^2}\\&=\mathcal{E}_j^{n+1/2}-\mathcal{E}_j^n+p_j^n\left(\tau_j^{n+1/2}-\tau_j^n\right)-\frac{a^2}{2}\left(\tau_j^{n+1/2}-\tau_j^n\right)^2.\end{aligned}$$

On the other hand, from a Taylor expansion with an integral remainder, one gets

$$\mathcal{E}_j^{n+1/2}=\mathcal{E}_j^n+\mathcal{E}_\tau|_{x_j,t_n}\left(\tau_j^{n+1/2}-\tau_j^n\right)+\int_{\tau_j^n}^{\tau_j^{n+1/2}}\mathcal{E}_{\tau\tau}(\theta)\left(\tau_j^{n+1/2}-\theta\right)\mathrm{d}\theta.$$

Then, Weyl's assumptions (2.2) and a change of variables in the integral (re-parameterisation) yield that

$$\mathcal{E}_j^{n+1/2}=\mathcal{E}_j^n-p_j^n\left(\tau_j^{n+1/2}-\tau_j^n\right)+\left(\tau_j^{n+1/2}-\tau_j^n\right)^2\int_0^1\mathcal{E}_{\tau\tau}(\tau_j^{n\dagger})(1-\theta)\mathrm{d}\theta,$$

where $\tau_j^{n\dagger}:=\theta\tau_j^{n+1/2}+(1-\theta)\tau_j^n$. So, for the entropy residual to be non-positive, one gets

$$\begin{aligned}\mathcal{R}_j^{n+1/2}&=\left(\tau_j^{n+1/2}-\tau_j^n\right)^2\int_0^1\left(\mathcal{E}_{\tau\tau}(\tau_j^{n\dagger})-a^2\right)(1-\theta)\mathrm{d}\theta\\&=\left(\tau_j^{n+1/2}-\tau_j^n\right)^2\int_0^1\left(-p_\tau(\tau_j^{n\dagger})-a^2\right)(1-\theta)\mathrm{d}\theta\leq 0,\end{aligned} \tag{2.32}$$

and a sufficient condition would be to set the integrand to be negative. Since $p_\tau=-\kappa\gamma\varrho^{1+\gamma}$, it yields

$$a^2\geq\ \kappa\gamma\max_{j\in\Omega_N}\max_{0\leq\theta\leq1}\left(\varrho_j^{n\dagger}\right)^{\gamma+1}=\kappa\gamma\max\left(\|\varrho^{n+1/2}\|_{\ell_\infty}^{\gamma+1},\|\varrho^n\|_{\ell_\infty}^{\gamma+1}\right), \tag{2.33}$$

which satisfies the sub-characteristic condition.

For the projection step, it is clear that due to Jensen's inequality the energy inequality holds:

$$(\varrho E)_j^{n+1}\leq\varrho_j^nE_j^{n+1/2}-\frac{\Delta t}{\Delta x}\left((\varrho E\widetilde{u})_{j+1/2}^{n+1/2}-(\varrho E\widetilde{u})_{j-1/2}^{n+1/2}\right). \tag{2.34}$$

Combining (2.31) and (2.34), we get the energy inequality (2.15), under the ε-uniform time restriction (2.23) and the sub-characteristic condition (2.33).

2.3.1.4 ℓ_∞-stability of the solution

In this section, we prove the stability of the LP-IMEX scheme in the ℓ_∞-norm.

Lemma 2.3.8. *For the well-prepared initial data, the computed density, momentum and velocity are stable in the ℓ_∞-norm, uniformly in ε.*

Proof. We have shown in Appendix 2.B that, for a fixed ε, the entropy stability implies the ℓ_∞-stability provided that the density is shown to be positive. Thus, the density, velocity and so the momentum are stable. For the proof of ε-uniformity of these results see Appendix 2.B. □

2.3.2 Rigorous analysis of asymptotic consistency

The existing asymptotic consistency proofs in the literature are often based on the formal asymptotic expansion as we have already presented in Section 2.3.1, by investigating the method as $\varepsilon \to 0$. The analysis is rather formal as one does not show how the variables change in terms of ε and simply balances the equal powers of ε, assuming implicitly that all the variables are $\mathcal{O}(1)$ (in terms of ε). In this section, we show that it is possible for the LP-IMEX scheme to go further and show asymptotic consistency more rigorously.

The main point is to study the implicit step in order to check how the *unique* updated solution behaves as $\varepsilon \to 0$. Once we show that this solution does not blow up in the limit, one can combine it with the formal analysis (as in Section 2.3.1), and conclude asymptotic consistency rigorously. The approach we present here to justify the formal analysis is close to what has been used in [Bis15], in the context of the finite volume evolution Galerkin (FVEG) scheme [BALMN14]. For future reference, we name this boundedness of the numerical solution for $\varepsilon \to 0$ as "*ε-stability (boundedness)*".

Note that for the scheme written in the form of (2.10a)–(2.10c), $\overrightarrow{w}_\Delta^{n+1/2}$ and $\overleftarrow{w}_\Delta^{n+1/2}$ should be computed implicitly, and then $\tau_\Delta^{n+1/2}$ is obtained explicitly. Now, let us define the matrix $\overrightarrow{J_\varepsilon}$ as the coefficient matrix of size N for the implicit update of $\overrightarrow{w}_\Delta^{n+1/2}$, *i.e.*,

$$\overrightarrow{J_\varepsilon}\,\overrightarrow{w}_\Delta^{n+1/2} = \overrightarrow{y}_\Delta, \qquad \overrightarrow{J_\varepsilon} := \frac{a\Delta t}{\varepsilon}\begin{bmatrix} 1+\dfrac{\varepsilon\Delta\mathfrak{m}_1^n}{a\Delta t} & & & & -1 \\ -1 & 1+\dfrac{\varepsilon\Delta\mathfrak{m}_2^n}{a\Delta t} & & & \\ & \ddots & \ddots & & \\ & & & -1 & 1+\dfrac{\varepsilon\Delta\mathfrak{m}_N^n}{a\Delta t} \end{bmatrix}, \tag{2.35}$$

with vector $\overrightarrow{y}_\Delta$ specified in the case of isentropic Euler equations by (2.10b), as $\overrightarrow{y}_\Delta := \underline{\Delta\mathfrak{m}}^n \odot \overrightarrow{w}_\Delta^n$, where $\underline{\Delta\mathfrak{m}}^n$ is the discretised vector of $\Delta\mathfrak{m}^n$, and $\odot$ denotes the entry-wise or Hadamard product. Likewise, we denote $\overleftarrow{J_\varepsilon}$ and $\overleftarrow{y}_\Delta$. Now, we are in a position to pose the following theorem.

Theorem 2.3.9. *Consider the matrix $\overrightarrow{J_\varepsilon}$ as defined in* (2.35)*. Then*

(i) $\overrightarrow{J_\varepsilon}$ is non-singular for all $\varepsilon > 0$;

(ii) $\lim_{\varepsilon\to 0} \|\overrightarrow{J_\varepsilon}^{-1}\|$ *is bounded for any natural matrix norm.*

Proof. Regarding part (i), it is clear that matrix $\overrightarrow{J_\varepsilon}$ is strictly diagonally dominant (SDD), and it is a classical result that SDD matrices are non-singular; see, *e.g.*, [HJ86, Thm. 6.1.10]. This is enough to show the non-singularity of $\overrightarrow{J_\varepsilon}$, so, to conclude that the solution of the implicit step, $\overrightarrow{w}_\Delta^{n+1/2}$, is unique.

For part (ii), the ∞-norm of the inverse of an SDD matrix $M \in \mathbb{R}^{N\times N}$ can be bounded as [Var75]

$$\|M^{-1}\|_\infty \leq \max_{1\leq i\leq N} \frac{1}{\Delta_i(M)}, \qquad \Delta_i(M) := |M_{ii}| - \sum_{j\neq i} |M_{ij}|. \tag{2.36}$$

For $\overrightarrow{J_\varepsilon}$, one can find that $\Delta_i(\overrightarrow{J_\varepsilon}) = \Delta\mathfrak{m}_i^n > 0$. So, there is an ε-uniform bound for the ∞-norm of matrix inverse. Since N is fixed, all matrix norms are equivalent which concludes the result. □

From Theorem 2.3.9, one can immediately conclude that a unique solution $\overrightarrow{w}^{n+1/2}$ exist for the implicit step, thus for the whole scheme, which has the same order as $\overrightarrow{w}^n$ in terms of ε. Also, by Theorem 2.3.9 and (2.35), one can see that the leading order of $\overrightarrow{w}_\Delta^{n+1/2}$, thus $\pi_{(0)\Delta}^{n+1/2}$, is constant due to a similar result for $\overleftarrow{w}_\Delta^{n+1/2}$. Employing (2.17), one can confirm that $\varrho_\Delta^{n+1/2} = \mathcal{O}(1)$, thus ε-stable. Showing the ε-stability of $u_{(0)\Delta}^{n+1/2}$ needs more work. From (2.35) one can write the update for $u_\Delta^{n+1/2}$ as

$$\left(\Delta\mathfrak{m}_j^n + \frac{a\Delta t}{\varepsilon}\right) u_j^{n+1/2} - \frac{a\Delta t}{2\varepsilon} u_{j-1}^{n+1/2} - \frac{a\Delta t}{2\varepsilon} u_{j+1}^{n+1/2} = u_j^n \Delta\mathfrak{m}_j^n - \frac{\Delta t}{2\varepsilon^2}\left(\pi_{j+1}^{n+1/2} - \pi_{j-1}^{n+1/2}\right), \tag{2.37}$$

which can be recast as a linear system of equations with the companion matrix H_ε defined as

$$H_\varepsilon := \begin{bmatrix} \Delta\mathfrak{m}_1^n + \dfrac{a\Delta t}{\varepsilon} & -\dfrac{a\Delta t}{2\varepsilon} & & & -\dfrac{a\Delta t}{2\varepsilon} \\ -\dfrac{a\Delta t}{2\varepsilon} & \Delta\mathfrak{m}_2^n + \dfrac{a\Delta t}{\varepsilon} & -\dfrac{a\Delta t}{2\varepsilon} & & \\ & \ddots & \ddots & \ddots & \end{bmatrix}. \tag{2.38}$$

Matrix H_ε is SDD; so, it is invertible with a bounded inverse in the limit, as in Theorem 2.3.9. Since we have already proved in Section 2.3.1.1 that $\pi_{(0)\Delta}^{n+1/2}$ and $\pi_{(1)\Delta}^{n+1/2}$ are constant, the rhs of (2.37) is $\mathcal{O}(1)$, and the ε-boundedness of H_ε^{-1} concludes the ε-stability, *i.e.*, $u_{(0)\Delta}^{n+1/2} = \mathcal{O}(1)$.

Remark 2.3.10. *Interestingly enough, one can show that H_ε^{-1} is an almost constant matrix, i.e., it consists of a constant $\mathcal{O}(1)$ part with some $\mathcal{O}(\varepsilon)$ fluctuations. This, combined with periodicity of the domain and being the difference $(\pi_{j+1}^{n+1/2} - \pi_{j-1}^{n+1/2})$ central, implies that the $\mathcal{O}(1/\varepsilon^2)$ term in the rhs vanishes after multiplication by H_ε^{-1} provided that $\pi_{(0)\Delta}^{n+1/2}$ is constant. We refer to Lemma 2.4.5 for the complete proof, where the companion matrix is similar but a bit simpler to be analysed rigorously.*

The boundedness in terms of ε makes the asymptotic consistency analysis in Section 2.3.1 rigorous and shows the behaviour of the quantities in terms of ε. Since the projection step is explicit, its asymptotic consistency can be simply studied as in Section 2.3.1.

Remark 2.3.11. *This approach proves asymptotic consistency rigorously, i.e., the solution moves to the limit as $\varepsilon \to 0$. This is the result that makes the uniformity proofs of the previous sections valid and rigorous, as has been mentioned earlier in Remark 2.3.4.*

To summarise, the scheme is AC and AS, and since $\overrightarrow{J_\varepsilon}$ and $\overleftarrow{J_\varepsilon}$ can be inverted simply due to their structure and that the time step is ε-uniform, it is also AEf. Thus, the scheme is AP in the sense of Definition 1.2.1.

2.4 LP-IMEX scheme for the shallow water equations

In this section and along the same lines as the previous section, we analyse the LP-IMEX scheme applied to the non-dimensionalised shallow water equations (SWE) (1.10) for $d = 1$:

$$\begin{aligned} &\partial_t h + \partial_x m = 0, \\ &\partial_t m + \partial_x \left(\frac{m^2}{h} + \frac{p(h)}{\varepsilon^2} \right) = -\frac{h}{\varepsilon^2} \partial_x \eta^b, \end{aligned} \tag{2.39}$$

where h and $m := hu$ stand respectively for the water height and the momentum and ε denotes the Froude number, defined as the ratio of the characteristic speed to the speed of gravity waves. Also, η^b is the bottom function, and the pressure function is chosen as before but with $\kappa = \frac{1}{2}$ and $\gamma = 2$, *i.e.*, $p(h) = \frac{h^2}{2}$. Note that for this shallow water model to be valid, the bottom slope $\partial_x \eta^b$ should be small enough such that $\tan\theta \approx \theta$ where $\tan\theta$ is the bottom slope; see [BW04, BMCPV03] and references therein. Note also that the energy equation for the shallow water system (2.39) in the *conservative form* writes

$$\partial_t (hE) + \partial_x \left(\big(hE + \frac{p}{\varepsilon^2} + \frac{h\eta^b}{\varepsilon^2}\big) u \right) = 0,$$

with $E = \mathcal{E} + \dfrac{u^2}{2} + \dfrac{\eta^b}{\varepsilon^2}$ and $\mathcal{E} = \dfrac{h}{\varepsilon^2}$, *cf.* [ABB+04] for instance.

We omit the details of the splitting and the numerical scheme, and refer the reader to consult [CGK13, Sect. 3.2.2] and [CMV15]; although the considered source terms are not exactly the same, the structure is similar. Also, [CKKS16] has tailored the scheme to the SWE, very recently. We only need to mention that the transport sub-system is exactly like (2.3b) with the conservative variable $\phi \in \{h, hu\}$ in (2.9), while the acoustic sub-system includes the source term in addition. So, the relaxation system reads as follows, with the same $\alpha_\varepsilon = \frac{a}{\varepsilon}$ as in Section 2.2 for the isentropic Euler equations (see also [CKKS16]):

$$\begin{aligned} &\partial_t \tau - \partial_{\mathfrak{m}} u = 0, \\ &\partial_t \overrightarrow{w} + \alpha_\varepsilon \partial_{\mathfrak{m}} \overrightarrow{w} = -\frac{\alpha_\varepsilon}{\tau \varepsilon^2} \partial_{\mathfrak{m}} \eta^b, \\ &\partial_t \overleftarrow{w} - \alpha_\varepsilon \partial_{\mathfrak{m}} \overleftarrow{w} = \frac{\alpha_\varepsilon}{\tau \varepsilon^2} \partial_{\mathfrak{m}} \eta^b. \end{aligned} \tag{2.40}$$

Using this splitting, one can see that the projection step is like (2.12). Also, motivated by [CGK13, Sect. 5] (see also [CKKS16]), a self-similar solution can be proposed for Riemann

problems using the appropriate notion of *consistency in the integral sense* [HLVL97, Gal02, Gal03], leading to the Lagrange step of the scheme as

$$\tau_j^{n+1/2} = \tau_j^n + \frac{\Delta t}{\Delta \mathfrak{m}_j^n}\left(\widetilde{u}_{j+1/2}^{n+1/2} - \widetilde{u}_{j-1/2}^{n+1/2}\right), \tag{2.41a}$$

$$\overrightarrow{w}_j^{n+1/2} = \overrightarrow{w}_j^n - \frac{a\Delta t}{\varepsilon \Delta \mathfrak{m}_j^n}\left(\overrightarrow{w}_j^{n+1/2} - \overrightarrow{w}_{j-1}^{n+1/2}\right) - \frac{a\Delta t}{\varepsilon^3}\frac{\Delta \mathfrak{m}_{j-1/2}^n}{\Delta \mathfrak{m}_j^n}\eta_{x,j-1/2}^b, \tag{2.41b}$$

$$\overleftarrow{w}_j^{n+1/2} = \overleftarrow{w}_j^n + \frac{a\Delta t}{\varepsilon \Delta \mathfrak{m}_j^n}\left(\overleftarrow{w}_{j+1}^{n+1/2} - \overleftarrow{w}_j^{n+1/2}\right) + \frac{a\Delta t}{\varepsilon^3}\frac{\Delta \mathfrak{m}_{j+1/2}^n}{\Delta \mathfrak{m}_j^n}\eta_{x,j+1/2}^b, \tag{2.41c}$$

where $\Delta \mathfrak{m}_{j+1/2}^n := \frac{\Delta \mathfrak{m}_j^n + \Delta \mathfrak{m}_{j+1}^n}{2}$, $\eta_{x,j+1/2}^b := \frac{\eta_{j+1}^b - \eta_j^b}{\Delta x}$ is the one-sided discretisation of the bottom function, and the interface velocity is defined as

$$\widetilde{u}_{j+1/2}^{n+1/2} := \frac{1}{2}\left(u_j^{n+1/2} + u_{j+1}^{n+1/2}\right) - \frac{1}{2a\varepsilon}\left(\pi_{j+1}^{n+1/2} - \pi_j^{n+1/2}\right) - \frac{1}{2a\varepsilon}\Delta \mathfrak{m}_{j+1/2}^n \eta_{x,j+1/2}^b. \tag{2.42}$$

Notice that this choice of $\eta_{x,j+1/2}^b$ provides the well-balancing for the lake at rest (LaR) equilibrium state which is defined as the set

$$\mathcal{U}_{LaR}^{\Delta} := \left\{ \begin{bmatrix} h_j \\ m_j \end{bmatrix} \,\middle|\, h_j = \eta^s - \eta_j^b, u_j = 0, \forall j \in \Omega_N \right\}, \tag{2.43}$$

with zero velocity and a constant water surface $\eta^s := h + \eta^b$. Note that the failure in satisfying the LaR equilibrium state at the discrete level leads to spurious oscillations.

The basic difference of the LP-IMEX for the SWE (2.41a)–(2.41c) with the LP-IMEX for the isentropic Euler equations (2.10a)–(2.10c) is the source term discretisation in the rhs, which on the one hand, requires refining the arguments for asymptotic consistency (see Section 2.4.1.1 and 2.4.2 below), and on the other, does not allow for the conservative form of the discrete entropy inequality. For the latter, we refer to [CKKS16] where the authors could show the entropy inequality in the *non-conservative* form.

2.4.1 Numerical analysis of the LP-IMEX scheme

Before we proceed with the stability results for the SWE, let us define the formal zero-Froude limit system and the well-prepared initial datum, with the following asymptotic expansion for height and momentum

$$h(x,t) = h_{(0)} + \varepsilon h_{(1)} + \varepsilon^2 h_{(2)},$$
$$m(x,t) = m_{(0)} + \varepsilon m_{(1)} + \varepsilon^2 m_{(2)}.$$

Definition 2.4.1. *The formal zero-Froude limit of the SWE* (2.39) *gives the so-called "lake equations", and reads (see Appendix 2.A as well as [BKL11] for the formal justification)*

$$\eta_{(0)}^s = h_{(0)} + \eta^b = const., \qquad h_{(1)} = const.,$$
$$\partial_x m_{(0)} = 0,$$
$$\partial_t m_{(0)} + \partial_x\left(\frac{m_{(0)}^2}{h_{(0)}} + p_{(2)}\right) = -h_{(2)}\partial_x \eta^b.$$

Definition 2.4.2. *For the 1d SWE* (2.39)*, we call the initial data* $(h_{0,\varepsilon}, m_{0,\varepsilon})$ *well-prepared if it holds that*

$$\begin{aligned} h(0,\cdot) = h_{0,\varepsilon} &:= h^0_{(0)} + \varepsilon^2 h^0_{(2),\varepsilon}, \\ m(0,\cdot) = m_{0,\varepsilon} &:= m^0_{(0)} + \varepsilon m^0_{(1),\varepsilon}, \end{aligned} \tag{2.44}$$

where $h_{(0)}(x) = \eta^s_{(0)} - \eta^b(x)$ *with constant* $\eta^s_{(0)}$ *and* $m_{(0)}$*.*

The following theorem summarises the results of this section.

Theorem 2.4.3. *The LP-IMEX scheme* (2.41a)–(2.41c) *and* (2.12)*, applied to the shallow water equations with a well-prepared initial datum, satisfies the following properties:*

(i) It can be expressed in the locally conservative form for the density.

(ii) The scheme is AC, i.e., it gives a consistent discretisation of the lake equations, as $\varepsilon \to 0$*, in terms of Definition 2.4.1.*

(iii) Under the ε*-uniform CFL constraint* (2.25)*, the scheme is positivity preserving, i.e.,* $h^n_j > 0$ *provided that* $h^0_j > 0$ *for all* $j \in \Omega_N$*.*

(iv) The scheme preserves the lake at rest equilibrium state, i.e., it is well-balanced.

The proof of (i) is clear and skipped. We go through the proof of parts (ii), (iii) and (iv) of the theorem, briefly.

2.4.1.1 Asymptotic consistency

Because of Definition 2.4.1, the argument for the asymptotic consistency should consider two important differences compared to Section 2.3. The SWE with a non-flat bottom topography have a different limit as $\varepsilon \to 0$: rather than the density (or height), the surface elevation is constant. Also, instead of a *div*-free velocity field, the momentum field should be solenoidal. Since the Lagrangian formulation does not consider these two differences into account, it is a bit complicated to check if the scheme drives the solution toward the limit manifold or not; however, one can check the consistency of the discretisation rather readily. Note that as before, we start with a well-prepared initial datum in the sense of Definition 2.4.2.

For the Lagrange step and using the τ-update (2.41a) as well as (2.42), one obtains that

$$\left(\pi^{n+1/2}_{(0)j+1} - \pi^{n+1/2}_{(0)j}\right) + \Delta \mathfrak{m}^n_{(0)j+1/2} \eta^b_{x,j+1/2} = 0. \tag{2.45}$$

To show the consistency of the discretisation in the limit, using (2.45) and (2.42), one can show the consistency of the interface velocity in the limit. Having that, it is straightforward to show that the Lagrange step (2.41a)–(2.41c) is consistent with (2.40) in the limit. For the projection step, the consistency of the interface velocity concludes the argument.

2.4.1.2 Height positivity

The proof is very similar to Section 2.3.1, but compared to that, there is an additional contribution due to the source terms. It is not difficult to show that, due to (2.45), the condition (2.25)—with the interface velocity as defined in (2.42)—can be fulfilled uniformly in ε. So under the non-restrictive CFL condition (2.25), the scheme is positivity preserving.

2.4.1.3 Well-balancing

Note that, one can write the Lagrangian update for $\overrightarrow{w}$ and $\overleftarrow{w}$ (2.41b)–(2.41c) with the same matrices $\overrightarrow{J_\varepsilon}$ and $\overleftarrow{J_\varepsilon}$ as in (2.35). Thus, the scheme is solvable. So, to show that the LP-IMEX scheme preserves the LaR equilibrium state, firstly, we show that the scheme *may* have such a solution, given the initial datum at equilibrium. Then, we argue that since the solution of the scheme is unique, this should be the only case which can happen.

Since the projection step is a convex combination of $\phi_j^{n+1/2}$ under the CFL constraint (2.25), one can confirm that to have a discrete solution in $\mathcal{U}_{LaR}^{\Delta}$, it is sufficient and necessary that $\phi_\Delta^{n+1/2} \in \mathcal{U}_{LaR}^{\Delta}$ and $\widetilde{u}_{j+1/2}^{n+1/2} = 0$ for all $j \in \Omega_N$. Then, we can check if such a state is compatible with the scheme, *i.e.*, if the scheme may have such a solution. It is clear for the projection step; so, let us clarify it for the Lagrange step.

The τ-update is compatible with a zero interface velocity and a steady height. We can obtain the τ–π relation (2.17) for the SWE as well, which clearly shows that given a steady height, the relaxation pressure would be also steady; so, the solution remains at the equilibrium. It only remains to show the compatibility condition for the velocity after the Lagrange update, that is to show $u_\Delta^{n+1/2} = \mathbf{0}$, and also to complete the loop by confirming the compatibility of the zero interface velocity. One can write the update for the velocity as

$$u_j^{n+1/2} = u_j^n - \frac{\Delta t}{2\varepsilon^2 \Delta \mathfrak{m}_j^n} \left(\pi_{j+1}^{n+1/2} - \pi_{j-1}^{n+1/2} + h_{j-1/2}^{n+1/2}(\eta_j^b - \eta_{j-1}^b) + h_{j+1/2}^{n+1/2}(\eta_{j+1}^b - \eta_j^b) \right). \tag{2.46}$$

Since $\pi_\Delta^{n+1/2} = \pi_\Delta^n$ from the arguments above, one can confirm that $u_\Delta^{n+1/2} = \mathbf{0}$ is compatible. Since the velocity is zero and the relaxation pressure is at equilibrium, the definition of the interface velocity (2.42) makes the loop complete.

Up to now, we have shown the compatibility of such a solution at equilibrium. By Theorem 2.3.9, the existence and uniqueness of a solution (which should be the well-balanced solution) is known, thus the well-balancing is concluded. One can also formulate the proof more precisely, like [CKKS16]:

Lemma 2.4.4. *Starting with an initial data on the LaR manifold $\mathcal{U}_{LaR}^{\Delta}$, the LP-IMEX scheme* (2.41a)–(2.41c) *and* (2.12) *preserves the discrete equilibrium.*

Proof. We write the Lagrangian update for $\overrightarrow{w}$ as in (2.35), but with a different vector in the rhs since there is also a contribution from the bottom topography:

$$\left(\frac{a\Delta t}{\varepsilon} + \Delta \mathfrak{m}_j^n \right) \overrightarrow{w}_j^{n+1/2} - \left(\frac{a\Delta t}{\varepsilon} \right) \overrightarrow{w}_{j-1}^{n+1/2} = \Delta \mathfrak{m}_j^n \overrightarrow{w}_j^n - \frac{a\Delta t}{\varepsilon^3} \Delta \mathfrak{m}_{j-1/2}^n \eta_{x,j-1/2}^b. \tag{2.47}$$

Since for all $j \in \Omega_N$, $h_j^n + \eta_j^b$ is constant and $u_j^n = 0$, one can verify that

$$-\frac{a\Delta t}{\varepsilon^3}\Delta\mathfrak{m}_{j-1/2}^n \eta_{x,j-1/2}^b = \frac{a\Delta t}{\varepsilon}\left(\overrightarrow{w}_j^n - \overrightarrow{w}_{j-1}^n\right),$$

and (2.47) can be written as $\overrightarrow{J_\varepsilon}\overrightarrow{w}_\Delta^{n+1/2} = \overrightarrow{J_\varepsilon}\overrightarrow{w}_\Delta^n$, which implies that $\overrightarrow{w}_\Delta^{n+1/2} = \overrightarrow{w}_\Delta^n$ as $\overrightarrow{J_\varepsilon}$ is non-singular. Similarly, $\overleftarrow{w}_\Delta^{n+1/2} = \overleftarrow{w}_\Delta^n$, which implies that $\widetilde{u}_{j+1/2}^{n+1/2} \equiv 0$ and that the Lagrangian step preserves the equilibrium. The proof for the projection step is trivial. □

2.4.2 Rigorous analysis of asymptotic consistency

Given a non-flat bottom topography, Theorem 2.3.9 also applies with the same $\overrightarrow{J_\varepsilon}$ and the rhs vector from (2.47) as

$$\overrightarrow{y}_\Delta = \underline{\Delta\mathfrak{m}}^n \odot \overrightarrow{w}^n - \frac{a\Delta t}{2\varepsilon^3}\underline{\Delta\mathfrak{m}}_-^n \odot \underline{\Delta\eta}_-^b,$$

where $\underline{\Delta\mathfrak{m}}_-^n$ and $\underline{\Delta\eta}_-^b$ are the vectors of $\Delta\mathfrak{m}_{j-1/2}^n$ and $(\eta_j^b - \eta_{j-1}^b)$ respectively. One can see that $\|\overrightarrow{y}_\Delta\| = \mathcal{O}(1/\varepsilon^3)$; so, using the boundedness of $\|(\overrightarrow{J_\varepsilon})^{-1}\|$ is futile to prove the ε-stability of the solution. Instead, one has to study the structure of $(\overrightarrow{J_\varepsilon})^{-1}$, which proposes the following lemma.

Lemma 2.4.5. *Denote* $\overrightarrow{J_\varepsilon}' := \dfrac{\varepsilon}{a\Delta t}\overrightarrow{J_\varepsilon}$. *Then,*

(i) Denote the adjugate matrix of $\overrightarrow{J_\varepsilon}'$ *by* $\mathrm{adj}(\overrightarrow{J_\varepsilon}')$, *and the all-ones matrix of size* N *by* $\mathbf{1}_N$. *Then,* $\mathrm{adj}(\overrightarrow{J_\varepsilon}') = (1 + \mathcal{O}(\varepsilon))\,\mathbf{1}_N$.

(ii) $\det(\overrightarrow{J_\varepsilon}') = \mathcal{O}(\varepsilon)$.

Proof. It is known that the inverse of a circulant matrix is also circulant [Gra06]. So, it is enough if we show that the entries of the first column of $\mathrm{adj}(\overrightarrow{J_\varepsilon}')$ are $1 + \mathcal{O}(\varepsilon)$, which correspond to the first row of the cofactor matrix. We denote $\chi_j := \frac{\varepsilon\Delta\mathfrak{m}_j^n}{a\Delta t}$ and for simplicity of the notation, we assume that $\chi_j = \chi$ is constant; the proof is similar for the non-constant case. For the cofactor matrix, one can see that the entry of the first row and j-th column is

$$\mathrm{cof}(\overrightarrow{J_\varepsilon}')_{1j} = (-1)^{j+1}\det\begin{bmatrix} K_1 & \mathbf{0}_{(j-1)\times(N-j)} \\ \mathbf{0}_{(N-j)\times(j-1)} & K_2 \end{bmatrix},$$

where $\mathbf{0}_{q\times r}$ is the zero matrix of size $q \times r$, and K_1 and K_2 are square matrices of size $(j-1)$ and $(N-j)$, respectively, defined as

$$K_1 := \begin{bmatrix} -1 & 1+\chi & & & \\ & -1 & 1+\chi & & \\ & & \ddots & \ddots & \\ & & & & -1 \end{bmatrix}, \qquad K_2 := \begin{bmatrix} 1+\chi & & & \\ -1 & 1+\chi & & \\ & \ddots & \ddots & \\ & & -1 & 1+\chi \end{bmatrix}.$$

Then, it is clear that

$$\mathrm{cof}(\overrightarrow{J_\varepsilon}')_{1j} = (-1)^{2j}(1+\chi)^{N-j} = (1+\chi)^{N-j},$$

which shows that all the entries of $\mathrm{cof}(\overrightarrow{J_\varepsilon}')$, so $\mathrm{adj}(\overrightarrow{J_\varepsilon}')$, are $1+\mathcal{O}(\varepsilon)$ and concludes part (i). As we mentioned, the proof for the scheme (2.41a)–(2.41c) with non-constant χ_j is similar.

For part (ii), one can compute the determinant directly as $\prod_{j=1}^{N}(1+\chi_j)-1$ which is $\mathcal{O}(\varepsilon)$. □

With $(\overrightarrow{J_\varepsilon}')^{-1} = \mathrm{adj}(\overrightarrow{J_\varepsilon}')/\det(\overrightarrow{J_\varepsilon}')$, Lemma 2.4.5 implies that the implicit operator $(\overrightarrow{J_\varepsilon})^{-1} = \frac{\varepsilon}{a\Delta t}(\overrightarrow{J_\varepsilon}')^{-1}$, is bounded as $\varepsilon \to 0$ and almost constant, up to some deviation of order $\mathcal{O}(\varepsilon)$. Using this and periodic boundary conditions, one can prove the following theorem.

Theorem 2.4.6. *The norm of the updated solution vector $\overrightarrow{w}_\Delta^{n+1/2}$, is of the same order as $\|\overrightarrow{w}_\Delta^n\| = \mathcal{O}(1/\varepsilon^2)$, for periodic or compactly supported bottom topography function.*

Proof. For the statement of the theorem to be true, we should show that the structure of $(\overrightarrow{J_\varepsilon})^{-1}$ implies $\|(\overrightarrow{J_\varepsilon})^{-1}\overrightarrow{y}_\Delta\| = \mathcal{O}(1/\varepsilon^2)$. In other words, it filters the $\mathcal{O}(1/\varepsilon^3)$ part of the vector $\overrightarrow{y}_\Delta$, denoted as $\overrightarrow{y}_\Delta^* := -\frac{a\Delta t}{2\varepsilon^3}\underline{\Delta \mathfrak{m}}_-^n \odot \underline{\Delta \eta}_-^b$. Manipulating $\overrightarrow{y}_\Delta^*$ and since the initial datum is well-prepared, one can find that

$$\begin{aligned} y_j^* &= -\frac{a\Delta t}{2\varepsilon^3}\left(h_j^n + h_{j-1}^n\right)\left(\eta_j^b - \eta_{j-1}^b\right) \\ &= -\frac{a\Delta t}{2\varepsilon^3}\left(2\eta_{(0)j}^{s,n} - \eta_j^b - \eta_{j-1}^b + \mathcal{O}(\varepsilon^2)\right)\left(\eta_j^b - \eta_{j-1}^b\right) \\ &= -\frac{a\Delta t}{2\varepsilon^3}\left(2\eta_{(0)}^{s,n} - \eta_j^b - \eta_{j-1}^b\right)\left(\eta_j^b - \eta_{j-1}^b\right) + \mathcal{O}(1/\varepsilon^2). \end{aligned}$$

Now, it is enough to show that $\overrightarrow{y}^*$ belongs to the kernel of the leading order part of $(\overrightarrow{J_\varepsilon})^{-1}$, which consists of constant vectors. That is to show $\|(\overrightarrow{J_\varepsilon})^{-1}\overrightarrow{y}_\Delta^*\| = \mathcal{O}(1/\varepsilon^2)$, which can be done simply by making a spatial summation and using the boundary condition, *i.e.*,

$$\sum_{j\in\Omega_N} y_j^* = -\frac{a\Delta t}{2\varepsilon^3}\sum_{j\in\Omega_N}\left[2\eta_{(0)}^{s,n}\left(\eta_j^b - \eta_{j-1}^b\right) - \left(\eta_j^{b,2} - \eta_{j-1}^{b,2}\right)\right] = 0.$$

Hence, the $\mathcal{O}(1/\varepsilon^3)$ terms vanish and one is left with the contributions of order $\mathcal{O}(1/\varepsilon^2)$. □

Similar results to Lemma 2.4.5 and Theorem 2.4.6 hold for $\overleftarrow{J_\varepsilon}$ and $\overleftarrow{w}$ respectively. This implies the following corollary.

Corollary 2.4.7. *The asymptotic consistency analysis for the LP-IMEX scheme,* (2.41a)–(2.41c) *and* (2.12) *is rigorous, i.e., the asymptotic expansion is justified.*

Proof. Due to Theorem 2.4.6, the implicit step preserves the order of $\|\overrightarrow{w}_\Delta\|$. It gives $\pi_\Delta^{n+1} = \mathcal{O}(1)$, thus recovers (2.45). As in Section 2.3.2, one can show the boundedness of $\varrho_\Delta^{n+1/2}$ using the τ–π relation (2.17). Also, due to (2.45), the proof of the boundedness of $u_\Delta^{n+1/2}$ is similar to Section 2.3.2 since the coefficient matrix H_ε is exactly like (2.38). This concludes the rigorous asymptotic consistency of the implicit step, and also the whole scheme since the explicit step has already been shown to be asymptotically consistent. □

2.A Formal asymptotic analysis of the shallow water equations

This section is to provide the formal asymptotic analysis for the low-Froude limit of the 1d SWE in the periodic domain Ω (see also [BKL11]). Consider the non-dimensionalised SWE (1.10) for $d = 1$ and with the bottom function η^b (as (2.39)):

$$\begin{aligned} &\partial_t h + \partial_x m = 0, \\ &\partial_t m + \partial_x \left(\frac{m^2}{h} + \frac{p(h)}{\varepsilon^2} \right) = -\frac{h}{\varepsilon^2} \partial_x \eta^b. \end{aligned} \tag{2.48}$$

Then, we substitute the Poincaré expansion for h and m, in terms of the Froude number ε, as

$$\begin{aligned} h(t,x) &= h_{(0)}(t,x) + \varepsilon h_{(1)}(t,x) + \varepsilon^2 h_{(2)}(t,x), \\ m(t,x) &= m_{(0)}(t,x) + \varepsilon m_{(1)}(t,x) + \varepsilon^2 m_{(2)}(t,x), \end{aligned} \tag{2.49}$$

in (2.48), and balance equal powers of ε. $\mathcal{O}(\varepsilon^{-2})$ terms yield $h_{(0)}\, \partial_x (h_{(0)} + b) = 0$; so, the leading order of the water surface (or total height) $\eta^s := h + \eta^b$ is constant in space since $\eta^s_{(0)} := h_{(0)} + \eta^b = \eta^s_{(0)}(t)$. Using this, one can find for the higher order terms that $h_{(0)}\, \partial_x h_{(1)} = 0$, thus $h_{(1)} = h_{(1)}(t)$.

Moreover, the leading order of the continuity equation $\partial_t h_{(0)} + \partial_x m_{(0)} = 0$ gives

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega h_{(0)} \mathrm{d}x = -\int_\Omega \partial_x m_{(0)} \mathrm{d}x = 0,$$

owing to the divergence theorem and the assumption of periodic boundary conditions. Thus, $\partial_t h_{(0)} = 0$ and $\eta^s_{(0)} = \text{const.}$, which give $h_{(0)} = h_{(0)}(x) = \eta^s_{(0)} - \eta^b(x)$ and $m_{(0)} = m_{(0)}(t)$. With similar arguments, one can easily find that $\partial_t h_{(1)} = 0$, so $h_{(1)} = \text{const.}$ and $m_{(1)} = m_{(1)}(t)$. For the evolution of $m_{(0)}$ in time, one gets

$$\partial_t m_{(0)} = -\frac{1}{|\Omega|} \int_\Omega h_{(2)} \partial_x \eta^b \mathrm{d}x = -\frac{1}{|\Omega|} \int_\Omega z_{(2)} \partial_x \eta^b \mathrm{d}x.$$

Thus, the leading order momentum does not evolve in time when the bottom is flat, *i.e.*, $\partial_t m_{(0)} = 0$. Summing up, one can justify Definition 2.4.1 as the formal asymptotic limit of the SWE.

2.B Entropy (energy) stability in the zero-Mach limit

In this section, we discuss the implications of entropy stability for the limit $\varepsilon \to 0$, aiming to show the stability of the solution and, due to the compactness or by the Bolzano–Weierstrass theorem, its strong convergence to a consistent limit. The main objective is to discuss the stability region which entropy stability provides. For this section, we assume the positivity of density and energy inequality for the numerical scheme; so, it is not limited to the LP-IMEX scheme.

Firstly, we recall that positivity and energy inequality give boundedness of the density and velocity, but not directly in the limit $\varepsilon \to 0$. Then, we show this boundedness as $\varepsilon \to 0$, thus the

existence of a converging sub-sequence due to the compactness. We, finally, show that the limit density is the incompressible limit solution. Since this analysis does not use any detail neither from the splitting nor the discretisation, it cannot recover the asymptotic consistency for the limit velocity.

Consider the entropy (energy) function $\mathcal{J} := \varrho E = \frac{1}{2}\frac{(\varrho u)^2}{\varrho} + \frac{\kappa/\varepsilon^2}{\gamma-1}\varrho^\gamma$ and assume a fixed grid of size N defining the discrete domain Ω_N. Having the discrete entropy (energy) inequality, *e.g.*, (2.15), we make a spatial summation to get the *global entropy (energy) inequality*

$$\sum_{j\in\Omega_N} \mathcal{J}_j^{n+1} \leq \sum_{j\in\Omega_N} \mathcal{J}_j^{n} \quad \Longrightarrow \quad \sum_{j\in\Omega_N} \mathcal{J}_j^{n+1} \leq \sum_{j\in\Omega_N} \mathcal{J}_j^{0} \leq C_\varepsilon < \infty. \tag{2.50}$$

If, in addition, we assume positivity, $\mathcal{J}$ is positive and $0 < \mathcal{J}_j^{n+1} \leq C_\varepsilon$ for all $j \in \Omega_N$.

One immediate result, for a fixed ε, is the ℓ_∞-boundedness of discrete solution vectors, *i.e.*, $\varrho_\Delta^{n+1}, u_\Delta^{n+1} \in \ell_\infty$. So, the energy inequality, accompanied by positivity, provides a stability region, which depends on the initial condition as well as ε. The question is how does this stability region change if $\varepsilon \to 0$? If one keeps the grid fixed and considers $\varepsilon \to 0$, the boundedness of the density is rather clear, either due to positivity or due to the boundedness of the energy. But, it is not straightforward to conclude the boundedness of the velocity.

Also, note that the boundedness of the density sequence w.r.t. ε provides strong convergence, that is to say that positive-density solutions, by virtue of compactness, have a converging sub-sequence of vectors $\{\varrho_\Delta^{\varepsilon_k,n}\}_{k\in\mathbb{N}}$ for any step n, where $\varepsilon_k \to 0$ is a sequence approaching the incompressible limit. This sub-sequence converges strongly to *some limit* $\varrho_\Delta^{\varepsilon_\infty,n}$, *i.e.*,

$$\lim_{k\to\infty} \|\varrho_\Delta^{\varepsilon_k,n} - \varrho_\Delta^{\varepsilon_\infty,n}\| = 0.$$

However, it is not clear if the limit is in the space of incompressible solutions. In the following lemma, we show that the computed density converges to its incompressible limit, with the help of energy inequality. We, then, show that the same assumptions are not enough to prove that the computed velocity is *div*-free. Nonetheless, the boundedness of the velocity sequence, so its convergence to some limit, can be obtained.

Lemma 2.B.1. *Consider the sequence* $\{(\varrho_\Delta^{\varepsilon_k,n}, u_\Delta^{\varepsilon_k,n})\}_{k\in\mathbb{N}}$*, accompanied by a well-prepared initial datum, as the discrete solution of the isentropic Euler equations* (2.1)*. Assume that the solution sequence satisfies the density positivity and the energy inequality. Then, the sequence is bounded in* ℓ_∞ *as* $\varepsilon \to 0$ *such that the density sub-sequence approaches the incompressible limit* $\varrho_\Delta^{0,n}$ *with the rate of* $\mathcal{O}(\varepsilon)$*.*

Remark 2.B.2. *The rigorous analysis [KM82] and the formal one based on asymptotic expansions lead us to expect the convergence rate of the density to be of* $\mathcal{O}(\varepsilon^2)$*. So, the convergence rate of Lemma 2.B.1 is not optimal. We see that exactly due to this issue, the asymptotic consistency of the velocity cannot be obtained by these assumptions.*

Proof. Consider a well-prepared initial density as $\varrho_i^{\varepsilon,0} := \varrho^{0,0} + \delta_i^{\varepsilon,0}$ with $\delta_i^{\varepsilon,0} = \mathcal{O}(\varepsilon^2)$ for all $i \in \Omega_N$. We, also, write the density at step n as $\varrho_i^{\varepsilon,n} := \varrho^{0,0} + \delta_i^{\varepsilon,n}$. Mass conservation and positivity imply that $\|\varrho_\Delta^{\varepsilon,n}\|_{\ell_1} = \|\varrho_\Delta^{\varepsilon,0}\|_{\ell_1}$; so, one can simply get

$$\sum_{i\in\Omega_N} \delta_i^{\varepsilon,n} = \sum_{i\in\Omega_N} \delta_i^{\varepsilon,0} = N\mathcal{O}(\varepsilon^2). \tag{2.51}$$

It seems that, in the limit, the density oscillates around a constant state. But this does not imply convergence since the perturbations have no sign. By the global energy inequality (as in (2.50)) one can see that

$$\sum_{i\in\Omega_N}\left(\varrho^{0,0}+\delta_i^{\varepsilon,n}\right)^2\leq\sum_{i\in\Omega_N}\left(\varrho^{0,0}+\delta_i^{\varepsilon,0}\right)^2+C_0\varepsilon^2,\qquad C_0:=\sum_{i\in\Omega_N}\left(\varrho^{0,0}+\delta_i^{\varepsilon,0}\right)\left(u^{0,0}+\mu_i^{\varepsilon,0}\right)^2,\tag{2.52}$$

where $\mu_i^{\varepsilon,0}=\mathcal{O}(\varepsilon)$ to fulfil the well-preparedness of initial datum. Then, combining (2.51) and (2.52) yields $\|\delta_\Delta^{\varepsilon,n}\|_{\ell_2}^2=\|\delta_\Delta^{0,n}\|_{\ell_2}^2+C_0\varepsilon^2=\mathcal{O}(\varepsilon^2)$, which shows that each component converges to the incompressible limit (at least) with the $\mathcal{O}(\varepsilon)$ rate, though, this rate is slower than expected. Furthermore, by straightforward calculations, one can show that

$$\|\varrho_\Delta^{\varepsilon,n}\|_{\ell_2}^2-\left\|\varrho_\Delta^{\varepsilon,0}\right\|_{\ell_2}^2=\|\delta_\Delta^{\varepsilon,n}\|_{\ell_2}^2+\mathcal{O}(\varepsilon^4)=\mathcal{O}(\varepsilon^2).$$

Then, by the complete energy inequality, not (2.52), one can obtain

$$\left\|\left(\varrho u^2\right)_\Delta^{\varepsilon,n}\right\|_{\ell_1}-\left\|\left(\varrho u^2\right)_\Delta^{\varepsilon,0}\right\|_{\ell_1}\leq\mathcal{O}(1).$$

Thus, $\|(\varrho u^2)_\Delta^{\varepsilon,n}\|_{\ell_1}$ is bounded, and since the density uniformly converges to its incompressible limit, the velocity is bounded as well. □

Remark 2.B.3. *With the additional assumption of $\|\delta_\Delta^{\varepsilon,n}\|_{\ell_2}=\mathcal{O}(\varepsilon^2)$, one can show that $\|\mu_\Delta^{\varepsilon,n}\|_{\ell_2}=\mathcal{O}(\varepsilon)$. Thus, the asymptotic consistency would be obtained completely.*

A similar analysis can be done for the case of shallow water equations. Here we state the main result and sketch of the proof.

Lemma 2.B.4. *Consider the sequence $\{(h_\Delta^{\varepsilon_k,n},u_\Delta^{\varepsilon_k,n})\}_{k\in\mathbb{N}}$, accompanied by a well-prepared initial datum, as the discrete solution of the shallow water equations* (2.39)*. Assume that the solution sequence satisfies the height positivity and the energy inequality. Then, the sequence is bounded in ℓ_∞ as $\varepsilon\to 0$ such that the height sub-sequence approaches the zero-Froude limit $h_\Delta^{0,n}$ with the rate of $\mathcal{O}(\varepsilon)$.*

Proof. We consider the well-prepared initial datum as $h_i^{\varepsilon,0}:=\eta^s-\eta_i^b+\delta_i^{\varepsilon,0}$ with $\delta_i^{\varepsilon,0}=\mathcal{O}(\varepsilon)$ for all $i\in\Omega_N$. Then, we write the height at step n as $h_i^{\varepsilon,n}:=\eta^s-\eta_i^b+\delta_i^{\varepsilon,n}$. Using mass conservation, one can find $\sum_i\delta_i^{\varepsilon,n}=\sum_i\delta_i^{\varepsilon,o}$ and from the global energy inequality

$$\begin{aligned}\sum_{i\in\Omega_N}\left[\left(\eta^s-\eta_i^b+\delta_i^{\varepsilon,n}\right)^2+2\eta_i^b\left(\eta^s-\eta_i^b+\delta_i^{\varepsilon,n}\right)\right]\leq&\sum_{i\in\Omega_N}\left[\left(\eta^s-\eta_i^b+\delta_i^{\varepsilon,0}\right)^2+2\eta_i^b\left(\eta^s-\eta_i^b+\delta_i^{\varepsilon,0}\right)\right]\\&+\sum_{i\in\Omega_N}(h^{0,0}+\delta_i^{\varepsilon,0})(u^{0,0}+\mu_i^{\varepsilon,0})^2\varepsilon^2.\end{aligned}$$

Analogous arguments as for Lemma 2.B.1 conclude the boundedness of the velocity sequence. □

Chapter 3

The modified equation analysis

> *"Modified equations have been a commonly used tool in the study of difference schemes. Because of the lack of any theoretical foundation, this use has been accompanied by constant difficulties and results derived from modified equations have sometimes been regarded with apprehension. As a result, a situation arises where authors either disregard entirely the technique or have an unjustified faith in its scope."*
>
> – Griffiths and Sanz-Serna, *On the scope of the method of modified equations (1986)*

In this chapter and motivated by [SN14], we investigate the modified equation as a heuristic tool for the stability analysis of implicit-explicit flux-splitting schemes for stiff systems of conservation laws with the singular parameter ε. With the help of the (truncated) modified equation and inspired by [MP85], we derived some criteria for the stability of symmetric and non-symmetric flux-splittings, and apply the latter to several well-known splittings. We prove that—for the isentropic Euler equations—the Degond–Tang splitting [DT11] and the Haack–Jin–Liu splitting [HJL12], and—for the shallow water equations—the RS-IMEX splitting [Zak16a] are stable in the sense of [MP85]. We also discuss an example of the splitting of the full Euler equations [Kle95, NBA+14]. In fact, the validity of the whole analysis is based on a crucial assumption—which has not been elaborated in [SN14]—that is the truncation of higher order terms of the full modified equation (in terms of discretisation parameters Δx and Δt) is justified. Here, we find a sufficient condition to justify this truncation, which is somewhat restrictive as it requires the initial datum to be almost constant, i.e., to possess only very long waves. This chapter is based heavily on [ZN17].

Contents

3.1 Introduction

For IMEX schemes, one splits the system into stiff and non-stiff parts w.r.t. the parameter ε, then, treats the stiff one implicitly in time while an explicit method is used for non-stiff part. Mainly due to the non-linearity of the system, the way of splitting the system into stiff and non-stiff parts is not unique and believed to be of crucial importance for stability; see [HJL12, DT11] for two recent examples for the isentropic Euler system. In [SN14], the authors tried to present a unified approach in determining the stability of IMEX splittings using the modified equation analysis. In this section, we first review the framework introduced in [SN14]. We consider the linear system of hyperbolic conservation laws in the (one-dimensional) domain $\Omega \subset \mathbb{R}$

$$\partial_t \boldsymbol{U} + \partial_x \boldsymbol{F}(\boldsymbol{U}) = \boldsymbol{0}, \qquad \boldsymbol{F}(\boldsymbol{U}) = A\boldsymbol{U}, \tag{3.1}$$

where $\boldsymbol{U} : [0,\infty) \times \Omega \to \mathbb{R}^q$ are conservative variables and $A \in \mathbb{R}^{q\times q}$ is a constant matrix depending only on the parameter ε, which is real diagonalisable with eigenvalues of $\lambda_1 \geq \ldots \geq \lambda_q$. We assume the initial condition $\boldsymbol{U}(0,x) = \boldsymbol{U}_0(x)$ to be well-prepared so that the time derivatives of the solution, $\partial_t^k \boldsymbol{U}$, are bounded uniformly in ε for $k \in \mathbb{N}$, *cf.* [MS01, KLN91, Gre97]. Now, we decompose the matrix A into stiff and non-stiff parts in an admissible way, as defined below:

Definition 3.1.1. *[Admissible splitting [SN14]] The splitting $A = \widetilde{A} + \widehat{A}$, with "stiff" $\widetilde{A}$ and "non-stiff" $\widehat{A}$, is called to be admissible provided that*

(i) each $\widetilde{A}$ and $\widehat{A}$ induces a hyperbolic system, i.e., they have real eigenvalues and a complete set of eigenvectors;

(ii) the eigenvalues of $\widehat{A}$ are bounded independently of ε, e.g., $\mathcal{O}(1)$, and at least one of the eigenvalues of $\widetilde{A}$ is $\mathcal{O}(\frac{1}{\varepsilon})$.

As in [SN14], we choose a Rusanov-type scheme for both stiff and non-stiff parts, in the computational domain Ω_N with N cells of size $\Delta x := \frac{|\Omega|}{N}$. Also, the time step is denoted by Δt. Such an IMEX scheme can be written either in the *un-split form*

$$\boldsymbol{U}_j^{n+1} = \boldsymbol{U}_j^n - \frac{\Delta t}{\Delta x}\left(\widetilde{\boldsymbol{F}}_{j+1/2}^{n+1} - \widetilde{\boldsymbol{F}}_{j-1/2}^{n+1} + \widehat{\boldsymbol{F}}_{j+1/2}^{n} - \widehat{\boldsymbol{F}}_{j-1/2}^{n}\right),$$

or the *split form* with an explicit step (from the temporal step n to some intermediate step $n+1/2$) and an implicit step (from the intermediate step $n+1/2$ to the new temporal step $n+1$)

$$\begin{cases} \boldsymbol{U}_j^{n+1/2} &= \boldsymbol{U}_j^n - \frac{\Delta t}{\Delta x}\left(\widehat{\boldsymbol{F}}_{j+1/2}^{n} - \widehat{\boldsymbol{F}}_{j-1/2}^{n}\right), \\ \boldsymbol{U}_j^{n+1} &= \boldsymbol{U}_j^{n+1/2} - \frac{\Delta t}{\Delta x}\left(\widetilde{\boldsymbol{F}}_{j+1/2}^{n+1} - \widetilde{\boldsymbol{F}}_{j-1/2}^{n+1}\right), \end{cases}$$

where the numerical fluxes are defined as

$$\begin{aligned} \widetilde{\boldsymbol{F}}_{j+1/2}^{n+1} &:= \frac{1}{2}\widetilde{A}\left(\boldsymbol{U}_{j+1}^{n+1} + \boldsymbol{U}_j^{n+1}\right) - \frac{\widetilde{\alpha}}{2}\left(\boldsymbol{U}_{j+1}^{n+1} - \boldsymbol{U}_j^{n+1}\right), \\ \widehat{\boldsymbol{F}}_{j+1/2}^{n} &:= \frac{1}{2}\widehat{A}\left(\boldsymbol{U}_{j+1}^{n} + \boldsymbol{U}_j^{n}\right) - \frac{\widehat{\alpha}}{2}\left(\boldsymbol{U}_{j+1}^{n} - \boldsymbol{U}_j^{n}\right), \end{aligned}$$

with $\mathcal{O}(1)$ numerical diffusion coefficients $\widetilde{\alpha}$ and $\widehat{\alpha}$ for stiff and non-stiff parts, respectively. Then, the (truncated) modified equation (see [SN14, eq. (10)]) reads

$$\partial_t \boldsymbol{U} + A\partial_x \boldsymbol{U} = \frac{\Delta t}{2}\Big(\frac{\alpha \Delta x}{\Delta t}\mathbb{I}_q - \widehat{A}^2 + \widetilde{A}^2 + [\widetilde{A}, \widehat{A}]\Big)\partial_x^2 \boldsymbol{U}, \tag{3.2}$$

where $\alpha := \widetilde{\alpha} + \widehat{\alpha}$ and $[\widetilde{A}, \widehat{A}] := \widetilde{A}\widehat{A} - \widehat{A}\widetilde{A}$ is the commutator of the stiff and non-stiff Jacobians.

Assuming that $\boldsymbol{U} \in [L_2(\Omega)]^q$, one can apply the Fourier transform to (3.2), which yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\widehat{\boldsymbol{U}} + \left(-i\xi A - \frac{\xi^2 \Delta t}{2}\Big(\frac{\alpha \Delta x}{\Delta t}\mathbb{I}_q - \widehat{A}^2 + \widetilde{A}^2 + [\widetilde{A}, \widehat{A}]\Big)\right)\widehat{\boldsymbol{U}} = 0,$$

with ξ denoting the frequency variable. This gives the following convenient stability result.

Lemma and Definition 3.1.2 (Corollary 1, [SN14])**.** *The modified equation* (3.2) *is L_2-stable if the "frequency matrix"*

$$\mathcal{P}(\xi) := -iA\xi - \xi^2 D_\nu, \quad D_\nu := \frac{\Delta t}{2}\Big(\frac{\alpha \Delta x}{\Delta t}\mathbb{I}_q - \widehat{A}^2 + \widetilde{A}^2 + [\widetilde{A}, \widehat{A}]\Big), \tag{3.3}$$

is stable, i.e., it only has eigenvalues with negative real parts. In this case, we say that the IMEX splitting satisfies condition $(\mathcal{A})$*.*

Remark 3.1.3. *(i) Throughout this chapter, whenever we talk about stability, it means stability in the sense of condition* $(\mathcal{A})$*, unless explicitly stated otherwise.*

(ii) The modified equation (3.2) *is derived formally by truncating Taylor expansions in space and time. We conjecture that a rigorous justification will have to rely on a "low-frequency assumption" such as*

$$\|\xi^k A^k\| = \mathcal{O}(1) \quad \textit{for } \ k = 3, 4, \ldots, \tag{3.4}$$

together with a suitable CFL condition. We will discuss this conjecture in further details in Section 3.2.

(iii) Recalling a famous result of Gel'fand [Gel59], if one considers a convection-diffusion system of equations like (3.3)*, the well-posedness requires the viscosity matrix (in this case D_ν) to be parabolic, i.e., its eigenvalues should have positive real parts. So, in general, parabolicity is a necessary condition for stability. Since in [Gel59] this necessity has been justified only for very high-frequency modes, it cannot be applied in the context of this chapter as the frequencies are small due to a low-frequency assumption. Hence, the parabolicity is not a necessary condition anymore.*

Unfortunately, without any additional structural assumption, obtaining a general stability condition for the matrix $\mathcal{P}$ is very delicate. For example, in [SN14] the authors introduce a characteristic splitting, for which the Jacobians are simultaneously diagonalisable, hence, the commutator $[\widetilde{A}, \widehat{A}]$ vanishes. This immediately provides stability of the modified equation; see Remark 3.4.4 below. Here, we take another approach as, in Section 3.3, we study the eigenvalues of $\mathcal{P}$ assuming symmetry of the splitting and relate it to the strict stability *in the sense of Majda–Pego* [MP85]. Also, in Section 3.4, given a general background state, we study Fourier symbols for linearised modified equations of non-symmetric flux-splittings. Note that our analysis applies to a general background state and any frequency variable for which the modified equation is valid, while the previous work [SN14] evaluated the Fourier symbols numerically using fixed background states and frequencies. Finally, in Section 3.5, we confirm that the modified equations obtained from the splittings in [DT11, HJL12] for the isentropic Euler equations, as well as the RS-IMEX splitting for the shallow water system (see Chapter 4) are stable in the sense of Majda–Pego. We also study Klein's auxiliary splitting [Kle95], and discover a small instability region for the example of two colliding pulses [Kle95, NBA$^+$14], for a moderate CFL number. This seems to give a hint at the numerical difficulties observed in [NBA$^+$14]. Before all these discussions, let us justify the truncation of the modified equation as appeared in (3.2).

3.2 On the validity of the truncated modified equation

It has been proposed since [Hir68, WH74] (see also [VR99, GSS86] for the literature review) that the modified equation can give an interesting intuition about the behaviour of numerical schemes, that is if the solution of the modified equation is stable, the computed solution is stable as well. But, there are some difficulties incorporated with modified equations: they are non-unique [GSS86] and have in fact an infinite number of terms, making a series which is usually not convergent and only valid in the asymptotic limit $(\Delta t, \Delta x) \to 0$. So, for the modified equation to be practical, one should be able to justify the use of its first few terms; see [WH74, MM98] for some interesting discussions on this issue.[1] Here, in addition to the discretisation parameters Δt and Δx, there is the scaling parameter $\varepsilon \ll \Delta t, \Delta x$, which requires justifying the truncation of the modified equation to be rethought; we show that the low-frequency assumption (3.4) achieve this goal for the linear system (3.1). We do the analysis for the linear advection equation. Then, we comment briefly on linear systems.

With the help of [PTA12, Chap. 4] and [VR99, Eq. (8)], we consider the modified equation of some explicit numerical scheme for the (stiff) linear advection equation $\partial_t u + \frac{a}{\varepsilon}\partial_x u = 0$ as

$$\partial_t u + \frac{a}{\varepsilon}\partial_x u = \sum_{\ell=2}^{\infty} \varepsilon^{-\ell} \Delta x^{\ell-1} c_\ell \partial_x^\ell u, \tag{3.5}$$

where $c_\ell < \infty$ are some known coefficients, independent of ε and Δx, and $\Delta t \sim \Delta x$ for the sake of simplicity of presentation. Then, the Fourier transform of the modified equation (3.5) reads

$$\frac{\mathrm{d}}{\mathrm{d}t}\widehat{u} + \frac{i\xi a}{\varepsilon}\widehat{u} = \sum_{\ell=2}^{\infty} i^\ell \varepsilon^{-\ell} \Delta x^{\ell-1} c_\ell \xi^\ell \widehat{u}.$$

Now, assuming that $\xi = \mathcal{O}(\varepsilon)$, which comes from the low-frequency assumption (3.4), only the first few terms of the rhs will formally define its sign (as $\Delta x \ll 1$), so the growth or decay of $\widehat{u}$ in time. Thus, the truncation is formally valid and (3.2) is justified.

One can confirm (see [VR99, Eq. (8)]) that in (3.5) the leading order terms are like $(\frac{a}{\varepsilon})^\ell$. For the case of linear systems, one expects to get similar equations but with powers of A as A^ℓ; this motivates the following conjecture.

Conjecture 3.2.1. *For a linear system with the flux vector $\boldsymbol{F}(\boldsymbol{U}) = A\boldsymbol{U}$, the boundedness of $A^\ell \partial_x^\ell \boldsymbol{U}$ justifies the truncation of the modified equation of an explicit/implicit method. So, in the Fourier space, $\|\xi^\ell A^\ell\| = \mathcal{O}(1)$ should hold.*

This is precisely the low-frequency assumption in Remark 3.1.3. Note that assuming an admissible splitting $A = \widehat{A} + \widetilde{A}$ such that $\|\widehat{A}\|, \|\widetilde{A}\| \lesssim \|A\|$, the analysis for an IMEX scheme can also be done with this low-frequency assumption.

Remark 3.2.2. *Although one cannot truncate the modified equation for high frequencies, it has been shown in [WH74] that it is possible to remedy this issue for linear scalar equations by establishing a relation between the coefficients of the modified equation and the von Neumann stability analysis (which provides a complete stability argument).*

[1] Interestingly enough and as mentioned in [GSS86], there is an analogy between the modified equation and the backward error analysis of Wilkinson, that is to say, that one can use a fixed number of terms of the modified equation and ask for the backward error; see [CJ16] for an illustrative example.

Example 3.2.3. *Consider the non-dimensionalised isentropic Euler equations as* (1.6)*. So, the flux Jacobian (denoted by A) and its power write*

$$A = \begin{bmatrix} 0 & 1 \\ -u^2 + \dfrac{p'(\varrho)}{\varepsilon^2} & 2u \end{bmatrix}, \qquad A^2 = \begin{bmatrix} \dfrac{p'(\varrho) - \varepsilon^2 u^2}{\varepsilon^2} & 2u \\ \dfrac{2u(p'(\varrho) - \varepsilon^2 u^2)}{\varepsilon^2} & \dfrac{3\varepsilon^2 u^2 + p'(\varrho)}{\varepsilon^2} \end{bmatrix}. \tag{3.6}$$

It is straightforward to check that A^k has an entry (so the norm) of $\mathcal{O}(\varepsilon^{-2\lfloor (k+1)/2 \rfloor})$. Thus, assuming $\xi \sim \varepsilon^{\sigma_k}$ for each k, the condition $\|\xi^k A^k\| = \mathcal{O}(1)$ gives σ_k as

$$\sigma_k = \frac{\lfloor \frac{k+1}{2} \rfloor}{k/2}, \tag{3.7}$$

which has been computed in Table 3.1. Because $\sigma := \min_{k>2} \sigma_k = \frac{4}{3}$ and we aim to ignore terms for $k > 2$, the condition $\xi \sim \varepsilon^{4/3}$ justifies truncation of the modified equation. Note that for the full Euler equations, the procedure is similar, which comes to the same σ. Note also that this analysis only suggests a sufficient condition; so, the existence of a smaller σ cannot be excluded by the analysis.

k	1	2	3	4	5	even k	odd k	∞
σ_k	2	1	$\frac{4}{3}$	1	$\frac{6}{5}$	1	$\frac{k+1}{k}$	1

Table 3.1: The values of σ_k from (3.7).

3.3 Stability of symmetric splittings

In this section, we assume that A, $\widehat{A}$ and $\widetilde{A}$ are symmetric, and point out the stability of such a splitting in Corollary 3.3.3. Then, by introducing the notion of strict stability in the sense of Majda–Pego [MP85], we generalise condition ($\mathcal{A}$) (for "*linear*" systems) to "*linearised*" systems. This notion gives a more general stability result for symmetrisable systems (see Theorem 3.3.6). Non-symmetric splittings are treated in Section 3.4.

For any symmetric matrix A, a symmetric splitting is always possible, *e.g.*, if one chooses $\widehat{A} = \operatorname{diag}(A|_{\varepsilon=1})$. For any symmetric splitting, the commutator is a skew-Hermitian matrix, therefore

$$\mathcal{P}(\xi) = -\Bigg[\underbrace{iA\xi + \xi^2 \frac{\Delta t}{2}[\widetilde{A}, \widehat{A}]}_{=:\mathscr{A}} + \underbrace{\xi^2 \frac{\Delta t}{2}\Big(\frac{\alpha \Delta x}{\Delta t}\mathbb{I}_q - \widehat{A}^2 + \widetilde{A}^2\Big)}_{=:\mathscr{H}}\Bigg], \tag{3.8}$$

where $\mathscr{A}$ and $\mathscr{H}$ are skew-Hermitian and Hermitian, respectively. One may conjecture that the eigenvalues of $\mathscr{H}$ would be positive. The following lemma verifies this conjecture.

Lemma 3.3.1. *The Hermitian matrix $\mathscr{H}$ is positive-definite under a non-restrictive CFL condition, independently of ε.*

Proof. One way to conclude the lemma is to use the *eigenvalue stability inequality* (see [Tao12, eq. 1.64]), which states that for two Hermitian matrices L and M of size q, the following holds

$$|\lambda_k(L+M) - \lambda_k(L)| \leq \|M\|_{\text{op}}, \qquad k = 1, \dots, q, \tag{3.9}$$

where the *operator norm* is defined as $\|M\|_{\text{op}} := \max\big(|\lambda_1(M)|, |\lambda_q(M)|\big)$. So, if one puts $L = \widetilde{A}^2$ and $M = -\widehat{A}^2$ in (3.9), it yields

$$-c < \lambda_k(\widetilde{A}^2) - \|\widehat{A}\|_{\text{op}}^2 \leq \lambda_k(-\widehat{A}^2 + \widetilde{A}^2) \leq \lambda_k(\widetilde{A}^2) + \|\widehat{A}\|_{\text{op}}^2,$$

with $c \geq 0$. Due to the order of magnitude of eigenvalues, c can be chosen to be positive and $\mathcal{O}(1)$, namely $c > \|\widehat{A}\|_{\text{op}}^2$, which implies the time step restriction $\Delta t < \alpha \Delta x / \|\widehat{A}\|_{\text{op}}^2$. This CFL condition shifts the eigenvalues to the right (by $\frac{\alpha\Delta x}{\Delta t}\mathbb{I}_q$), so that the eigenvalues of $\mathscr{H}$ are positive.

Another way to conclude the same result is to use the sub-additivity of the numerical range (see Section 3.4): to show that the numerical range of $\widetilde{A}^2$ is positive, and to put the numerical range of $\frac{\alpha\Delta x}{\Delta t}\mathbb{I}_q - \widehat{A}^2$ in the right half-plane under some CFL condition. □

Given these properties of $\mathscr{A}$ and $\mathscr{H}$, there is a sum of a Hermitian and a skew-Hermitian matrix in (3.8), and one can use the Bendixon's theorem in [Hir02] (see [Ben02] for the original work which is limited to real matrices), which shows that given a Hermitian matrix with stable eigenvalues in the left half-plane and a skew-Hermitian matrix, the sum will have stable eigenvalues, *i.e.*, the eigenvalues have negative real parts. To recall, we restate the theorem from [Hir02]; see also [Bro30] for a nice review.

Theorem 3.3.2 (Theorem II, [Hir02])**.** *Consider the matrix $M \in \mathbb{K}^{q\times q}$ with $\mathbb{K} = \mathbb{C}$ or $\mathbb{R}$, where $\lambda_k\big(\mathcal{H}(M)\big) = p_k \in \mathbb{R}$ for $k = 1, \dots, q$ and $\mathcal{H}$ stands for the Hermitian part. Then, the following bounds hold*

$$\min_k p_k \leq \Re\big[\lambda_k(M)\big] \leq \max_k p_k. \tag{3.10}$$

From Lemma 3.3.1 and Theorem 3.3.2, one can conclude immediately the following corollary.

Corollary 3.3.3. *Under a non-restrictive CFL condition, an admissible symmetric splittings is stable, i.e., it satisfies condition ($\mathcal{A}$).*

Remark 3.3.4. *One could also use an energy estimate to show that for the hyperbolic-parabolic system* (3.2) *with a symmetric matrix A, the positive-definiteness of the viscosity matrix D_ν is necessary and sufficient for L_2-stability.*

In order to generalise Lemma 3.1.2 for systems which are linearised around an arbitrary state $\boldsymbol{U}_0$, we introduce the notion of *strict stability* in the sense of Majda–Pego [MP85] below.

Definition 3.3.5 ([MP85])**.** *For the non-linear system $\partial_t \boldsymbol{U} + \partial_x \boldsymbol{F}(\boldsymbol{U}) = \partial_x\left(D_\nu \partial_x \boldsymbol{U}\right)$, the viscosity matrix D_ν is strictly stable at $\boldsymbol{U}_0$ if and only if there exists a $\delta > 0$ such that the eigenvalues $\lambda_k(\xi)$ of the matrix $\mathcal{P}(\xi) := -\boldsymbol{F}'(\boldsymbol{U}_0) i\xi - \xi^2 D_\nu(\boldsymbol{U}_0)$ satisfy the following algebraic condition*

$$\Re\big[\lambda_k(\xi)\big] \leq -\delta|\xi|^2, \quad \text{for all } \xi \in \mathbb{R}. \tag{3.11}$$

This definition also provides some non-linear stability results; see [MP85]. Note that Definition 3.3.5 refers to a given state (arbitrary, but fixed) $\boldsymbol{U}_0$, around which the system is linearised. To

keep the notation simpler and when there is no confusion, we suppress the dependence on $\boldsymbol{U}_0$. Using this framework, one can also find the generalisation of the stability of symmetric splittings at $\boldsymbol{U}_0$, as in [Moc80, MP85]:

Theorem 3.3.6. *Consider the modified equation in Fourier space* (3.3) *and let* $\mathcal{M}(\boldsymbol{U}_0)$ *be a real symmetric positive-definite matrix, symmetrising* $A(\boldsymbol{U}_0)$ *from the left, i.e.,* $(\mathcal{M}A)\,|_{\boldsymbol{U}_0}$ *is symmetric. Then, if* $(\mathcal{M}D_\nu)\,|_{\boldsymbol{U}_0}$ *is positive-definite, the frequency matrix* (3.3)*, so the modified equation* (3.2)*, are strictly stable the modified equation in Fourier space* (3.3) *is strictly stable at* $\boldsymbol{U}_0$*, i.e., there exists a* $\delta > 0$ *such that* $\Re\big[\lambda_k(\mathcal{P}(\xi))\big] \leq -\delta|\xi|^2$.

It is clear that for symmetric splittings, the identity matrix can play the role of the symmetrising matrix $\mathcal{M}$ and Theorem 3.3.6 is reduced to the arguments we have presented above, leading to Corollary 3.3.3.

3.3.1 Extension to two-dimensional systems

One can easily extend the stability analysis to two-dimensional symmetric systems on Cartesian grids. With a similar procedure as [SN14], the modified equation can be obtained as

$$\begin{aligned}\partial_t \boldsymbol{U} + A_1\partial_x \boldsymbol{U} + A_2\partial_y \boldsymbol{U} = \frac{\Delta t}{2}\Big[&\Big(\frac{\alpha_1\Delta x}{\Delta t}\mathbb{I}_q - \widehat{A}_1^2 + \widetilde{A}_1^2 + [\widetilde{A}_1, \widehat{A}_1]\Big)\partial_x^2 \boldsymbol{U}\\ &+ \Big(- (A_1A_2 + A_2A_1) + 2\widetilde{A}_1A_2 + 2\widetilde{A}_2A_1\Big)\partial_{xy}^2 \boldsymbol{U}\\ &+ \Big(\frac{\alpha_2\Delta y}{\Delta t}\mathbb{I}_q - \widehat{A}_2^2 + \widetilde{A}_2^2 + [\widetilde{A}_2, \widehat{A}_2]\Big)\partial_y^2 \boldsymbol{U}\Big].\end{aligned}$$

Then, using two-dimensional Fourier transform, one can find the frequency matrix as

$$\begin{aligned}\mathcal{P}(\xi_1,\xi_2) := -i\xi_1A_1 - i\xi_2A_2 - \frac{\Delta t}{2}\Big[&(\xi_1^2\frac{\alpha_1\Delta x}{\Delta t} + \xi_2^2\frac{\alpha_2\Delta y}{\Delta t})\mathbb{I}_q\\ &+ \xi_1\xi_2[\widetilde{A}_1, \widehat{A}_2] + \xi_1\xi_2[\widetilde{A}_2, \widehat{A}_1]\\ &+ \big(\xi_1\widetilde{A}_1 + \xi_2\widetilde{A}_2\big)^2 - \big(\xi_1\widehat{A}_1 + \xi_2\widehat{A}_2\big)^2\Big].\end{aligned}$$

Since the commutator of two Hermitian matrices is skew-Hermitian, the Hermitian part of $\mathcal{P}(\xi_1,\xi_2)$ writes

$$\mathcal{H}(\mathcal{P})(\xi_1,\xi_2) = -\frac{\Delta t}{2}\Big[(\xi_1^2\frac{\alpha_1\Delta x}{\Delta t} + \xi_2^2\frac{\alpha_2\Delta y}{\Delta t})\mathbb{I}_q + \big(\xi_1\widetilde{A}_1 + \xi_2\widetilde{A}_2\big)^2 - \big(\xi_1\widehat{A}_1 + \xi_2\widehat{A}_2\big)^2\Big],$$

which, like the one dimensional case, can be shown to have positive eigenvalues if the splitting is admissible, *e.g.*, if $\xi_1\widehat{A}_1 + \xi_2\widehat{A}_2$ have $\mathcal{O}(1)$ eigenvalues. So, Lemma 3.3.1 concludes the stability of the frequency matrix.

3.4 Stability of non-symmetric splittings

In this section, we study the stability of the frequency matrix $\mathcal{P}$ without the symmetry assumption so that the commutator contributes to the real parts of the eigenvalues of $\mathcal{P}$ and makes the

analysis more involved. Note that, due to the linearity of the system (3.1), it is always possible to rewrite it as $\partial_t \boldsymbol{V} + B\partial_x \boldsymbol{V} = \boldsymbol{0}$ with a symmetric B—either by the characteristic form (with a diagonal B) or the symmetric decomposition [Fro10, TZ59] (with a general symmetric B—and to use the analysis presented in the preceding section. However, as our main interest in such a linear analysis is to extend basic ideas to linearised systems, we should also consider the general non-symmetric system.

Let us denote the spectrum of $\mathcal{P}$ as $\lambda(\mathcal{P})$. Then, by the theorem of spectral inclusion [GR97, Theorem 1.2-1], this spectrum (and in particular its convex hull) is contained in the closure of the numerical range of $\mathcal{P}$. In other words, $\mathrm{Conv}\,(\lambda(\mathcal{P})) \subseteq W(\mathcal{P})$, where the numerical range $W(\mathcal{P})$ is defined as $W(\mathcal{P}) := \{\langle \boldsymbol{v}, \mathcal{P}\boldsymbol{v}\rangle, \boldsymbol{v} \in \mathbb{C}^q, \|\boldsymbol{v}\|_{\ell_2} = 1\}$. In fact, the real part of the numerical range of $\mathcal{P}$ is bounded by the spectrum of its Hermitian part, *i.e.*,

$$\Re\left[W\left(\mathcal{P}\right)\right] = -\mathrm{Conv}\left(\lambda\big(\xi\mathcal{H}(iA) + \xi^2\mathcal{H}(D_\nu)\big)\right).$$

The eigenvalue stability inequality gives the upper-bound of the numerical range as

$$\Re\left[W(\mathcal{P})\right] \leq -\Big(\xi^2\lambda_q(\mathcal{H}(D_\nu)) - \xi\|\mathcal{H}(iA)\|_{\mathrm{op}}\Big), \tag{3.12}$$

where $\lambda_q(\mathcal{H}(D_\nu))$ denotes the smallest eigenvalue. So, for the stability of $\mathcal{P}$, it is sufficient to set the numerical range to be in the left half-plane. Thus, as a subsidiary result, for symmetric systems $\mathcal{H}(iA)$ vanishes and one can conclude immediately from (3.12) that the positive-definiteness of D_ν implies strict stability with $\delta = \lambda_q(\mathcal{H}(D_\nu))$, the smallest eigenvalues of $\mathcal{H}(D_\nu)$. For a non-symmetric A, although the positive-definiteness of D_ν does not necessarily imply condition $(\mathcal{A})$, we suggest the stability by a modified version of positivity in Theorem 3.4.1 below; this is the same result as [MP85, Thm. 2.1].

Theorem 3.4.1. *For a hyperbolic system* (3.1) *with* $A \in \mathbb{R}^{q\times q}$*, and with eigenvector matrix* R*, the positivity of* $\widetilde{D}_\nu := R^{-1}D_\nu R$ *is sufficient for the stability in terms of condition* $(\mathcal{A})$*.*

Proof. By construction, A is hyperbolic and can be diagonalised as $A = R\Lambda R^{-1}$, where R is the matrix of eigenvectors. Substituting this into the definition of $\mathcal{P}$ yields

$$\mathcal{P}(\xi) = -i\xi A - \xi^2 D_\nu = R\Big(-i\xi\Lambda - \xi^2\widetilde{D}_\nu\Big)R^{-1},$$

where $\widetilde{D}_\nu := R^{-1}D_\nu R$. Since similarity transformations do not change the spectrum, we instead study the eigenvalues of $\widetilde{\mathcal{P}}(\xi)$ defined as $\widetilde{\mathcal{P}}(\xi) := -i\xi\Lambda - \xi^2\widetilde{D}_\nu$.

One can decompose $\widetilde{D}_\nu$ as the sum of Hermitian and skew-Hermitian matrices, *i.e.*, $\mathcal{H}(\widetilde{D}_\nu) + \mathcal{A}(\widetilde{D}_\nu)$. From positivity $-\xi^2\mathcal{H}(\widetilde{D}_\nu)$ is stable and by Theorem 3.3.2, the addition of skew-Hermitian $-\xi^2\mathcal{A}(\widetilde{D}_\nu)$ and $-i\xi\Lambda$ cannot destabilise the stable Hermitian matrix $-\xi^2\mathcal{H}(\widetilde{D}_\nu)$. So $\mathcal{P}$ is stable. □

Note that $\widehat{A}^2 + \widetilde{A}^2 + [\widetilde{A}, \widehat{A}] = (\widetilde{A} - \widehat{A})(\widetilde{A} + \widehat{A})$. Also, the term $\frac{\alpha\Delta x}{\Delta t}\mathbb{I}_q$ in $\widetilde{D}_\nu$ only leads to a shift in the eigenvalues. So, we only need to study $\widetilde{D}'_\nu := R^{-1}(\widetilde{A} - \widehat{A})(\widetilde{A} + \widehat{A})R$ (instead of $\widetilde{D}_\nu$) as claimed by the following lemma. The proof is straightforward and skipped here.

Lemma 3.4.2. *Consider the linear hyperbolic system* (3.1)*. Let R to be the eigenvector matrix of A, and suppose that the splitting $A = \widehat{A} + \widetilde{A}$ is admissible in the sense of Definition 3.1.1. If there exists a lower-bound $\underline{\lambda}_{\mathcal{H}(\widetilde{D}'_\nu)}$ for the eigenvalues of the Hermitian part of $\widetilde{D}'_\nu := R^{-1}(\widetilde{A} - \widehat{A})(\widetilde{A}+\widehat{A})R$, such that $\underline{\lambda}_{\mathcal{H}(\widetilde{D}'_\nu)} = \mathcal{O}(1)$, the splitting is strictly stable in the sense of Majda–Pego.*

Remark 3.4.3. *Compared to the full frequency matrix $\mathcal{P}(\xi)$ used in Lemma 3.1.2, Lemma 3.4.2 is a convenient simplification since $\mathcal{H}(\widetilde{D}'_\nu)$ does not depend on ξ. Such a dependence requires choosing a "distinguished limit" for ξ–ε relation, due to the low-frequency assumption. This will play a crucial role in the example in Section 3.5.4.*

One can go one step further and use the structure of the viscosity matrix, which is known for the modified equation (3.2) unlike most of the literature in the hyperbolic-parabolic systems, *cf.* [MP85, Gel59, CS70]. Because of the hyperbolicity assumption, A and $\widetilde{A}$ are real diagonalisable, *i.e.*,

$$A = R\Lambda R^{-1}, \qquad \widetilde{A} = \widetilde{R}\widetilde{\Lambda}\widetilde{R}^{-1},$$

where R and $\widetilde{R}$ are matrices of eigenvectors and $\widetilde{R} = RQ_{R\to\widetilde{R}}$, where $Q_{R\to\widetilde{R}}$ stands for the change of basis matrix. Substituting these into the definition of $\widetilde{\mathcal{P}}$ gives

$$\widetilde{\mathcal{P}}(\xi) = -i\xi\Lambda - \xi^2 \underbrace{\frac{\Delta t}{2}\Big[\frac{\alpha\Delta x}{\Delta t}\mathbb{I}_q - \Lambda^2 + 2Q_{R\to\widetilde{R}}\widetilde{\Lambda}Q^{-1}_{R\to\widetilde{R}}\Lambda\Big]}_{=\widetilde{D}_\nu}. \tag{3.13}$$

This form of $\widetilde{D}_\nu$ reveals more of its structure. As one can see, for the positivity of $\widetilde{D}_\nu$ the role of $Q_{R\to\widetilde{R}}$ is crucial. For example for the admissible characteristic splitting, $Q_{R\to\widetilde{R}} = \mathbb{I}_q$, and the positivity (and stability) is clear since the components of $\widetilde{D}_\nu$ are diagonal. The equation (3.13) suggests that the splittings whose eigenspaces are close to each other such that $Q_{R\to\widetilde{R}}$ is close to the identity matrix are more likely to be stable, This, essentially, matches the results of [LeV07, Sect. D.7], that if the eigenspaces of the split matrices coincide, the power-boundedness of each explicit and implicit operators is enough for the stability of the whole scheme. We will come back to this form later on, in Chapter 4, when we analyse a flux-splitting scheme for a non-linear hyperbolic system.

Remark 3.4.4. *(i) The IMEX scheme based on the characteristic splitting is uniformly stable not only in the sense that its modified equation is stable (as discussed in [SN14]), but also in the ℓ_2-norm. This is because for such a splitting, one can decouple the system into q scalar equations $\partial_t w_k + \lambda_k \partial_x w_k = 0$ for $k = 1, \dots, q$; then, by the von Neumann stability analysis, both explicit and implicit steps can be shown to be ℓ_2-stable, respectively, under an appropriate (and ε-uniform) CFL condition, and unconditionally (see [Tra09, Sect 3.3.4]).*

(ii) In the light of [HJL12, Lemma 3.1], the stability of each step is clearly enough for the stability of the whole scheme; however, it is far from being necessary in most cases, and often not practical to be fulfilled. For instance, notice that the example in [SN14, Sect. 7] does not have stable steps. One could confirm numerically that for both stable and unstable settings (with $\varepsilon_1 = 0.1$ and $\varepsilon_2 = 0.01$ respectively) the implicit operator $\widetilde{\mathcal{S}}$ is power-bounded while the explicit operator $\widehat{\mathcal{S}}$ is not. Nonetheless, their multiplication $\widetilde{\mathcal{S}}\widehat{\mathcal{S}}$ makes one case stable and the other one unstable. For further details about the stability of the difference equations, the reader can consult [LeV07, Appendix D] and [Tre96, Chap. 4].

(iii) One may conjecture that if the commutator is $\mathcal{O}(1)$, the viscosity matrix is parabolic under a suitable and non-restrictive choice of CFL condition, using the continuity of eigenvalues [Ost66, Appendix K] [Mar49, Chap. 1]. But on one hand, the constant of that continuity grows as $\varepsilon \to 0$, and on the other, it is not even clear if the parabolicity is a relevant condition to be used for low-frequency modes, as mentioned earlier in Remark 3.1.3. Since the smallness of the commutator provides the heuristic for the development of the RS-IMEX scheme in Chapter 4, it is interesting to investigate such a question in more depth.

3.5 Applications

In this section, we show that Lemma 3.4.2 provides the linearised stability at any given state $\boldsymbol{U}_0$ of several splittings used in practice, namely the splitting of Haack–Jin–Liu [HJL12] (abbreviated as HJL hereinafter), Degond–Tang [DT11] (DT hereinafter) and the RS-IMEX splitting. We also discuss the numerical instability which has been reported in [NBA+14] for Klein's auxiliary splitting of the Euler equations [Kle95]. Let us recall that our analysis is based on the modified equation (3.2), hence on the IMEX Euler time integration accompanied by Rusanov-type numerical fluxes.

3.5.1 Haack–Jin–Liu splitting

Consider the (linearised) Jacobian matrix A for the isentropic Euler equations as in (3.6). The HJL splitting [HJL12] decomposes A as

$$\widehat{A} = \begin{bmatrix} 0 & \beta \\ -u^2 + \dfrac{p'(\varrho) - a(t)}{\varepsilon^2} & 2u \end{bmatrix}, \qquad \widehat{\lambda} = u \pm \sqrt{(1-\beta)u^2 + \frac{\beta\big(p'(\varrho) - a(t)\big)}{\varepsilon^2}},$$

$$\widetilde{A} = \begin{bmatrix} 0 & 1-\beta \\ \frac{1}{\varepsilon^2}a(t) & 0 \end{bmatrix}, \qquad \widetilde{\lambda} = \pm\frac{\sqrt{a(t)(1-\beta)}}{\varepsilon},$$

where ϱ, u, and $p(\varrho) = \kappa\varrho^\gamma$ are the density, velocity, and pressure. $\beta \in [0,1]$ is a parameter to be chosen (note that it is called α in [HJL12]) and $a(t) := \min_x p'$. With these settings, the splitting is admissible in the sense of Definition 3.1.1. For further details see [HJL12].

Assume that the system has been linearised around an arbitrary state $\boldsymbol{U}_0 = (\varrho_0, u_0)^T$. Then, in the light of Lemma 3.4.2, we have to study the positivity of $\widetilde{D}'_\nu$. With the aid of Maple™, one can get

$$\begin{aligned}\lim_{\varepsilon\to 0}\Big(\varepsilon^2\lambda^{1,2}_{\mathcal{H}(\widetilde{D}'_\nu)}\Big) &= \lim_{\varepsilon\to 0}\Big[\varepsilon^2(\beta-2)u_0^2 + \Big(a_0 - \beta p'_0 \pm \big((\beta-1)p'_0 + a_0\big)\Big)\Big] \\ &= \lim_{\varepsilon\to 0}\Big[\varepsilon^2(\beta-2)u_0^2 + \Big(a_0 \pm (-p'_0 + a_0) + \beta(-p'_0 \pm p'_0)\Big)\Big].\end{aligned}$$

Owing to the formal analysis for $\varepsilon \ll 1$, the asymptotic expansion gives $p'_0 - a_0 = \mathcal{O}(\varepsilon^2)$, so

$$\lim_{\varepsilon\to 0}\Big(\varepsilon^2\lambda^{1,2}_{\mathcal{H}(\widetilde{D}'_\nu)}\Big) = a_0, (1-2\beta)a_0.$$

Since $a_0 > 0$, both eigenvalues are non-negative in the limit and one can find the lower-bound $\underline{\lambda}_{\mathcal{H}(\widetilde{D}'_\nu)} = \mathcal{O}(1)$, provided we set $\beta \leq 1/2$. So, when $\beta \leq 1/2$, Lemma 3.4.2 implies the strictly stability of the scheme, in the sense of Majda–Pego, under a non-restrictive CFL condition. Note that for the numerical experiments of [HJL12], β is chosen in this stable region and often of $\mathcal{O}(\varepsilon^2)$.

3.5.2 Degond–Tang splitting

In [DT11] and for the isentropic Euler equations with the pressure function $p(\varrho) = \kappa\varrho^\gamma$ (like the HJL splitting example), the following splitting has been proposed for A in (3.6):

$$\widehat{A} = \begin{bmatrix} 0 & 0 \\ -u^2 + \theta p'(\varrho) & 2u \end{bmatrix}, \qquad \widehat{\lambda} = 0,\, 2u,$$
$$\widetilde{A} = \begin{bmatrix} 0 & 1 \\ \dfrac{1-\theta\varepsilon^2}{\varepsilon^2} p'(\varrho) & 0 \end{bmatrix}, \qquad \widetilde{\lambda} = \pm\frac{\sqrt{(1-\theta\varepsilon^2)\, p'(\varrho)}}{\varepsilon},$$

where θ is an *ad hoc* parameter to be chosen between 0 and $1/\varepsilon^2$. Note that it is discussed in [DT11, CDK12, Tan12] that taking $\theta = \mathcal{O}(1)$ leads to the AP property; so, we assume θ to be $\mathcal{O}(1)$. Then, one can clearly confirm that this splitting is admissible in the sense of Definition 3.1.1.

As for the HJL splitting and with $\boldsymbol{U}_0 = (\varrho_0, u_0)^T$, we study the positivity of $\widetilde{D}'_\nu$. With the aid of Maple™, one gets

$$\lim_{\varepsilon\to 0}\left(\varepsilon^2 \lambda^{1,2}_{\mathcal{H}(\widetilde{D}'_\nu)}\right) = \lim_{\varepsilon\to 0}\left[-\varepsilon^2\left(\theta + 2u_0^2\right) p'_0 + p'_0 \pm \mathcal{O}(\varepsilon^2)\right] = p'_0 > 0.$$

Thus, both eigenvalues are positive in the limit, and due to Lemma 3.4.2, the scheme is strictly stable in the sense of Majda–Pego under a non-restrictive CFL condition. Note that this stability does not depend on the choice of θ.

3.5.3 RS-IMEX splitting

RS-IMEX splitting will be introduced in Chapter 4, but to keep the discussion of the modified equation analysis integrated, we discuss its stability in the section. Here, we consider the RS-IMEX splitting for the shallow water equations with a flat bottom topography, in the form $\partial_t \boldsymbol{U} + \partial_x \boldsymbol{F}(\boldsymbol{U}) = \boldsymbol{0}$ with a different formulation from (1.8) (see the equation (4.8)):

$$\boldsymbol{U} = \begin{bmatrix} z \\ m \end{bmatrix}, \qquad \boldsymbol{F} = \begin{bmatrix} m \\ \dfrac{m^2}{h} + \dfrac{z^2 - 2zb}{2\varepsilon^2} \end{bmatrix}, \qquad A = \begin{bmatrix} 0 & 1 \\ -\dfrac{m^2}{(z-b)^2} + \dfrac{z-b}{\varepsilon^2} & \dfrac{2m}{z-b} \end{bmatrix}, \tag{3.14}$$

with the same notation as [BALMN14], *cf.* Chapter 4: z denotes the perturbation of the height from the mean height, $h + b = z$, $m := (z-b)u$ is the momentum and b is the depth function, which is negative and constant. For a scaled version of (3.14) and based on the formal asymptotic analysis in Appendix 2.A, we define the scaled perturbation $\boldsymbol{V} := [v_1, v_2]^T$ (from a constant state)

such that $v_1 := z/\varepsilon^2$ and $v_2 := m$; then, the RS-IMEX splitting with the lake at rest reference solution gives the following flux splitting, *cf.* Chapter 4:

$$\widehat{A} = \begin{bmatrix} 0 & 0 \\ \varepsilon^2 v_1 - \dfrac{\varepsilon^2 v_2^2}{(\varepsilon^2 v_1 - b)^2} & \dfrac{2v_2}{\varepsilon^2 v_1 - b} \end{bmatrix}, \qquad \widehat{\lambda} = 0,\ \frac{2v_2}{\varepsilon^2 v_1 - b},$$
$$\widetilde{A} = \begin{bmatrix} 0 & 1/\varepsilon^2 \\ -b & 0 \end{bmatrix}, \qquad \widetilde{\lambda} = \pm\frac{\sqrt{-b}}{\varepsilon^2}.$$

So, it can be concluded that this splitting is admissible in the sense of Definition 3.1.1.

As for the HJL splitting and with $\boldsymbol{V}_0 = (v_{1,0}, v_{2,0})^T$, one can obtain that

$$\lim_{\varepsilon\to 0}\left(\varepsilon^2 \lambda^{1,2}_{\mathcal{H}(\widetilde{D}'_\nu)}\right) = \lim_{\varepsilon\to 0} \frac{-b^5 + \mathcal{O}(\varepsilon^2)}{(\varepsilon^2 v_{1,0} - b)^4} = -b > 0, \tag{3.15}$$

since $b < 0$. Hence, using Lemma 3.4.2 and similar to the HJL splitting example, the splitting is strictly stable, in the sense of Majda–Pego, under a non-restrictive CFL condition. Note that the leading orders $\varepsilon^2\lambda^{1,2}_{\mathcal{H}(\widetilde{D}'_\nu)}$ are the same for the HJL (with $\beta = \mathcal{O}(\varepsilon^2)$), DT and RS-IMEX splittings.

Remark 3.5.1. *It would be interesting to extend the stability result to equations with a varying bottom topography. Nonetheless, it is not clear how to linearise the Jacobian matrices $\widetilde{A}$ and $\widehat{A}$ (by freezing b), and, simultaneously, the source term (by freezing b_x). Thus, it is more difficult to understand the linearisation error, hence the validity of the stability analysis.*

Example 3.5.2. *In addition to the previous analysis of the modified equation in the low-Mach and the low-Froude number limits ($\varepsilon \ll 1$), we now study $\lambda^{1,2}_{\mathcal{H}(\widetilde{D}'_\nu)}$ for all $\varepsilon \in (0, 1]$. To have the "same" settings for all these three splittings, we consider the pressure law $p(\varrho) = \varrho^2/2$ for the HJL and DT splittings so that it coincides with the pressure function of the shallow water equations. We also choose $(\varrho_0, u_0) = (1, 1)$ for HJL and DT splittings, and $(v_{1,0}, b, v_{2,0}) = (0, -1, 1)$ for the RS-IMEX splitting. With these settings, all the systems are the same and can be compared to each other. We also set the ad hoc parameters of HJL and DT splittings as the typical values, $\beta = \varepsilon^2$ and $\theta = 1$. Figure 3.1 shows that $\lambda^{1,2}_{\mathcal{H}(\widetilde{D}'_\nu)}$ are bounded from below. Indeed, $\lambda^1_{\mathcal{H}(\widetilde{D}'_\nu)}$ is always positive, while $\lambda^2_{\mathcal{H}(\widetilde{D}'_\nu)}$ is positive in the left of the kink—around $\varepsilon \in (0.4, 0.6)$— and negative in the right, but uniformly bounded. Thus, owing to Lemma 3.4.2, all these splittings are asymptotically stable. Note that the plots are hardly distinguishable for small ε.*

3.5.4 Klein's auxiliary splitting

In his influential paper [Kle95], Klein introduced two flux-splittings for the full Euler equations:

$$\boldsymbol{U} = \begin{bmatrix} \varrho \\ \varrho u \\ \varrho E \end{bmatrix}, \quad \boldsymbol{F} = \begin{bmatrix} \varrho u \\ \varrho u^2 + \frac{1}{\varepsilon^2} p \\ (\varrho E + p)\, u \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1 & 0 \\ \frac{\gamma-3}{2}u^2 & (3-\gamma)u & \frac{\gamma-1}{\varepsilon^2} \\ -Hu + \frac{(\gamma-1)\varepsilon^2}{2}u^3 & H - \varepsilon^2(\gamma-1)u^2 & \gamma u \end{bmatrix}, \tag{3.16}$$

where the total energy ϱE satisfies the dimensionless equation of state $\varrho E = \frac{p}{\gamma-1} + \frac{\varepsilon^2}{2}\varrho u^2$, and $H := E + \frac{p}{\varrho}$ stands for total enthalpy.

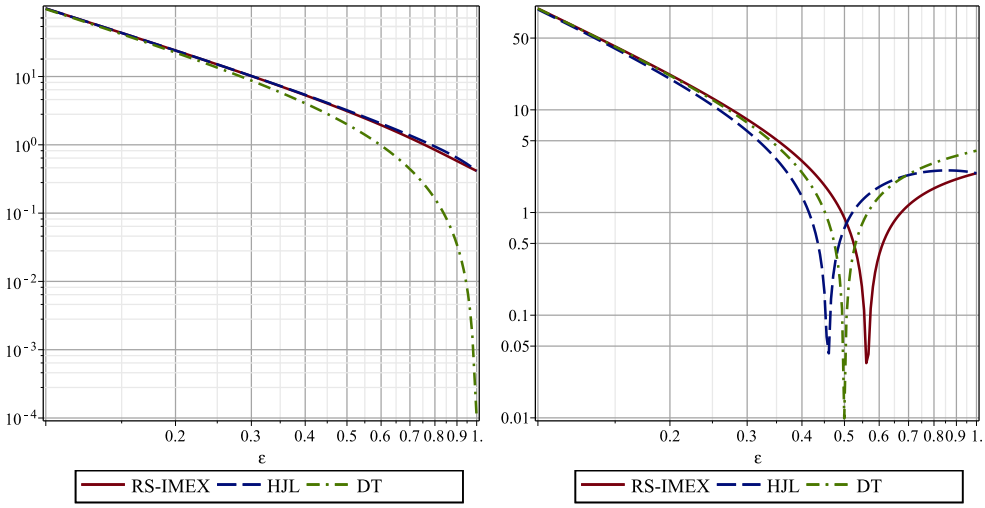


Figure 3.1: $|\lambda^1_{\mathcal{H}(\widetilde{D'_\nu})}|$ (left) and $|\lambda^2_{\mathcal{H}(\widetilde{D'_\nu})}|$ (right) for RS-IMEX, HJL and DT splittings w.r.t. ε.

The main splitting introduces two sub-systems, called system (I) and (II), given by [Kle95, eqs. (3.1)–(3.2)]. In the second splitting, the system (I) is replaced by the so-called *auxiliary* system (I*), which is given by [Kle95, eq. (3.8)]. In this section, we analyse the stability of a flux-splitting IMEX scheme, which uses Klein's auxiliary splitting as a building block (*cf.* [NBA$^+$14]).

Here, the background state for the linearisation is $\boldsymbol{U}_0 = (\varrho_0, \varrho_0 u_0, \varrho_0 E_0)^T$. Following the derivation in [NBA$^+$14], the auxiliary splitting is given by (for $1 < \gamma \leq \frac{5}{3}$)

$$\widetilde{A} = (1-\varepsilon^2)\begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{2}(\gamma-1)u^2 & -(\gamma-1)u & \frac{\gamma-1}{\varepsilon^2} \\ -u\frac{p-p_{\inf}}{\varrho} + \frac{\gamma-1}{2}\varepsilon^2 u^3 & \frac{p-p_{\inf}}{\varrho} - \varepsilon^2(\gamma-1)u^2 & (\gamma-1)u \end{bmatrix},$$

$$\widehat{A} = \begin{bmatrix} 0 & 1 & 0 \\ \Big(\frac{(\gamma-1)\varepsilon^2}{2} - 1\Big)u^2 & \big(2-(\gamma-1)\varepsilon^2\big)\,u & \gamma-1 \\ \widehat{A}_{31} & \widehat{A}_{32} & \widehat{A}_{33} \end{bmatrix},$$

$$\widehat{A}_{31} := -u\left[\big(1+\varepsilon^2(\gamma-1)\big)E - 2\varepsilon^4(\gamma-1)u^2 + (1-\varepsilon^2)\frac{p_{\inf}}{\varrho}\right] + \frac{(\gamma-1)\varepsilon^4}{2}u^3,$$

$$\widehat{A}_{32} := E + \varepsilon^2(\gamma-1)\left(E - \frac{\varepsilon^2}{2}u^2\right) + (1-\varepsilon^2)\frac{p_{\inf}}{\varrho} - (\gamma-1)\varepsilon^4 u^2,$$

$$\widehat{A}_{33} := \big(1+\varepsilon^2(\gamma-1)\big)\,u.$$

The choice of the parameter $p_{\inf} := \min_x p(t,x)$ guarantees the hyperbolicity of split systems, whose eigenvalues read

$$\widehat{\lambda} = u, u \pm c^*, \qquad c^* := \sqrt{\frac{p+(\gamma-1)\Pi}{\varrho}}, \quad \Pi := (1-\varepsilon^2)p_{\inf} + \varepsilon^2 p,$$

$$\widetilde{\lambda} = 0,\ \pm\frac{(1-\varepsilon^2)}{\varepsilon}\sqrt{\frac{(\gamma-1)(p-p_{\text{inf}})}{\varrho}}.$$

So, the splitting is admissible in the sense of Definition 3.1.1 (see [NBA+14]).

Our attempts to compute the eigenvalues of $\mathcal{H}(\widetilde{D}'_\nu)$ for this case, with Maple™, failed. Thus, we study the full frequency matrix $\mathcal{P}(\xi)$. As mentioned in Remark 3.4.3, another complication comes on the scene, which is the relation between ξ and ε through the low-frequency assumption; this makes the limit of the eigenvalues *path-dependent*. Finding the limit, we examine paths I and III in the following (see Figure 3.2). Note that path I is not consistent with the low-frequency assumption (3.4) as it requires $\xi \sim \varepsilon^\epsilon$ for $\epsilon \ll 1$. Path III implies $\xi \sim \varepsilon^{1+\epsilon}$ for some positive ϵ. With a suitable choice of ϵ, this path is consistent with the low-frequency assumption. In the following, we consider the case of $\epsilon = 1/3$ which fulfils the low-frequency assumption for the Euler system as explained in Section 3.2. Path II in Figure 3.2 denotes the intermediate regime $\xi \sim \varepsilon$, and is not consistent with our restrictive low-frequency assumption but has been considered in [Kle95, NBA+14].[2]

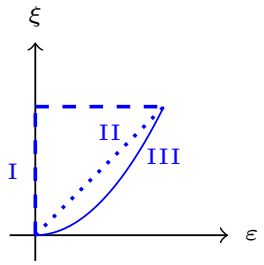


Figure 3.2: Different distinguished limits of ε and ξ.

Path I: Although this path does not match the low-frequency assumption, it is worth being analysed as it is related to the analysis in [SN14]. We, first, compute the $\varepsilon \to 0$ limit, which can be obtained using Maple™ as

$$\begin{aligned}
\lim_{\varepsilon\to 0} \Re\left[\varepsilon^2 \lambda_{\text{I}}^{1,2}\left(\mathcal{P}(\xi)\right)\right] &= \lim_{\varepsilon\to 0} \frac{(\gamma-1)\xi^2\Delta t}{2\varrho_0}\Big[-\varrho_0 E_0(\gamma-1) + p_{\text{inf}} \pm \left(\varrho_0 E_0 + p_{\text{inf}}\right)\Big] \\
&= \lim_{\varepsilon\to 0} \frac{(\gamma-1)\xi^2\Delta t}{2\varrho_0} \begin{cases} 2p_{\text{inf}} + (2-\gamma)\varrho_0 E_0 \\ -\gamma\varrho_0 E_0 \end{cases} \\
\Re\left[\lambda_{\text{I}}^3\left(\mathcal{P}(\xi)\right)\right] &= \frac{\xi^2\Delta t}{2}\left(u_0^2 - \frac{\alpha\Delta x}{\Delta t}\right).
\end{aligned}$$

So, the third non-stiff eigenvalue can be controlled by an ε-uniform CFL condition. Regarding the first two stiff eigenvalues, since $2p_{\text{inf}} + (2-\gamma)\varrho_0 E_0 > 0$ for $\gamma < 2$, the corresponding acoustic eigenvalue is positive in the limit; therefore, the scheme is unstable. Note that the path should get completed by $\xi \to 0$ limit, which does not change the sign of the eigenvalues; thus, the instability result holds as computed in [SN14, Sect. 7].

[2] Note that the choice of $\xi \sim \varepsilon^{4/3}$ gives only one sufficient condition, but it is not necessary; one may obtain less restrictive conditions on the frequencies by more careful analysis.

Path III: Unfortunately for this path, $\lambda_{\mathrm{III}}^{1,2}$ are very complicated while $\lambda_{\mathrm{III}}^{3}$ is similar to the previous case and can be controlled to be negative uniformly in ε. So, we have to study the stability by an example, which is the example of two colliding pulses [Kle95], for which the ε-dependency of the time step has been shown [NBA$^+$14].

Example 3.5.3. *The domain $[-L, L)$ is set to be periodic (like [Kle95, NBA$^+$14]) with $L := 2/\varepsilon^{4/3}$, for $\gamma = 1.4$. The initial data are*

$$\varrho(0,x) = \varrho_{(0)} + \frac{\varepsilon}{2}\varrho_{(1)}\left(1 - \cos\left(\frac{2\pi x}{L}\right)\right), \quad \varrho_{(0)} = 0.955, \quad \varrho_{(1)} = 2, \tag{3.17a}$$

$$p(0,x) = p_{(0)} + \frac{\varepsilon}{2}p_{(1)}\left(1 - \cos\left(\frac{2\pi x}{L}\right)\right), \quad p_{(0)} = 1, \quad p_{(1)} = 2\gamma, \tag{3.17b}$$

$$u(0,x) = \frac{1}{2}u_{(0)}\mathrm{sign}(x)\left(1 - \cos\left(\frac{2\pi x}{L}\right)\right), \quad u_{(0)} = 2\sqrt{\gamma}. \tag{3.17c}$$

As explained in Section 3.2, the frequency should be small, i.e., $\xi \sim \varepsilon^{\frac{4}{3}}$. So, with this initial condition, the small frequency assumption does hold. Now, in order to apply the Majda–Pego stability framework, we linearise around

$$\begin{aligned}
\varrho_0 &= \varrho_{(0)} + \frac{\varepsilon}{2}\varrho_{(1)} = 0.955 + \varepsilon,\\
p_0 &= p_{(0)} + \frac{\varepsilon}{2}p_{(1)} = 1 + \varepsilon\gamma,\\
u_0 &= \sqrt{\gamma},
\end{aligned}$$

and $p_{\mathrm{inf}} = 1$. Note that we have replaced $1 - \cos\left(\frac{2\pi x}{L}\right)$ by its mean value 1. The numerical diffusion and the grid parameters are chosen as in [NBA$^+$14]:

$$\alpha = \sqrt{\frac{\gamma p_0}{\varrho_0}} + \max_x\left(u(0,x)\right), \quad \Delta x = 0.05, \quad \Delta t = \frac{\mathrm{CFL}}{\max_x\left(u(0,x)\right)}\Delta x.$$

As the domain is periodic, it only provides a countable set of frequencies as $\xi = \frac{k\pi}{|\Omega|}$ for $k \in \{1, 2, \ldots, k_{max}\}$, where k_{max} determines the largest Fourier mode that the mesh can carry, and $k_{max} \leq \frac{\pi}{\Delta x}$ [MM98]. We compute the real parts of the eigenvalues of the frequency matrix $\mathcal{P}$ of the modified equation numerically, for the most stable (and non-trivial) mode corresponding to $k = 1$. Figure 3.3 displays $\Re(\lambda_{\mathcal{P}}^1)$, the possibly unstable eigenvalue, for different CFL numbers. The figures are zooms in ε. The figures reveal a small instability region near $\varepsilon \in (0.02, 0.06)$ and for $\mathrm{CFL} = 0.45$. This seems to correspond closely to some of the numerical experiments in [NBA$^+$14], where the CFL number needed to be reduced when changing ε from 0.1 to 0.05. Note, however, that the lack of uniform stability in [NBA$^+$14] is much stronger than the one Figure 3.3 suggests since in [NBA$^+$14] the CFL needed to decrease linearly with the Mach number, while in Figure 3.3, $\Re(\lambda_{\mathcal{P}}^1) \leq 0$ uniformly in ε, for the fixed $\mathrm{CFL} = 0.02$. This discrepancy may possibly be due to a fundamental difference between the Fourier analysis in the present chapter and the real computation in [NBA$^+$14] as, based on Lemma 3.1.2, Figure 3.3 studies a single Fourier mode while, due to the $\mathrm{sign}(x)$ function in (3.17c), the initial data for the velocity contain a superposition of all Fourier modes, which may trigger instabilities not explained by the present analysis. Note that the size of the domain is a bit larger for this example compared to [NBA$^+$14] but it does not affect the results.

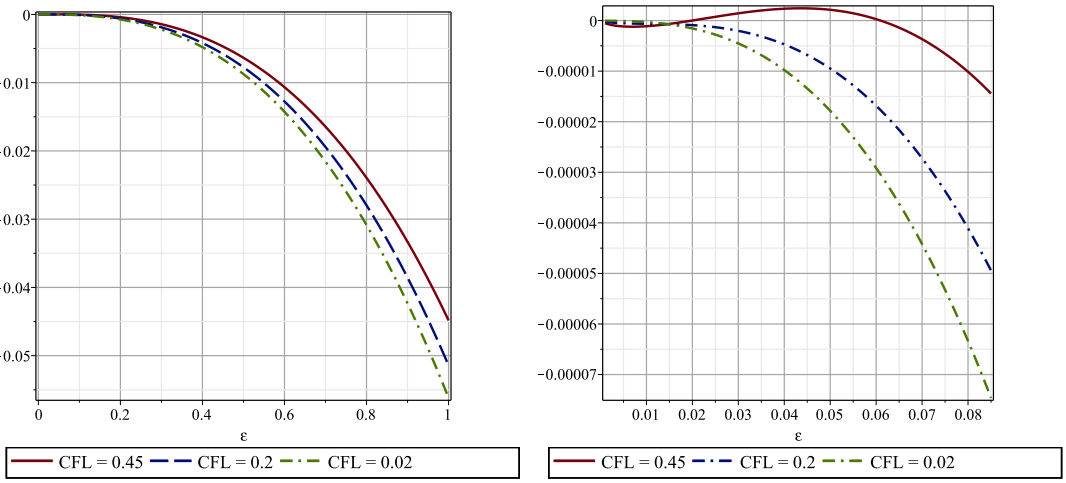


Figure 3.3: $\Re(\lambda^1_{\mathcal{P}})$ for Klein's auxiliary splitting w.r.t. ε, in different regions of ε and for the Fourier mode corresponds to $k = \frac{\pi}{4}\varepsilon^{4/3}$.

Remark 3.5.4. *It is important to point out some differences between the algorithms in [NBA+14] and [Kle95]. Klein develops his approach using the more complex setting of multiple space variables and multiple pressures. Algorithmically, he "combines explicit predictor steps for long wave linear acoustics or global compression with a single implicit scalar Poisson-type corrector scheme" [Kle95, p.3]. Thus, our stability analysis has no direct implication for the scheme proposed in [Kle95]. Rather, it should be seen as a comment to [NBA+14].*

Chapter 4

The RS-IMEX scheme for the 1d shallow water equations

"Die Mathematiker sind eine Art Franzosen: Redet man zu ihnen, so übersetzen sie es in ihre Sprache, und dann ist es alsobald ganz etwas anders."

– Goethe, *Maximen und Reflektionen*

In this chapter, motivated by the modified equation analysis in Chapter 3, we introduce the so-called reference solution implicit-explicit scheme for singularly-perturbed systems of balance laws. RS-IMEX scheme's bottom-line is to use the Taylor expansion of the flux function and the source term around a reference solution (typically the asymptotic limit or an equilibrium solution) to decompose the flux and the source into stiff and non-stiff parts so that the resulting IMEX scheme is asymptotic preserving (AP) w.r.t. the singular parameter ε tending to zero. We prove the asymptotic consistency, asymptotic stability, solvability and well-balancing of the scheme for the case of the one-dimensional shallow water equations when the singular parameter is the Froude number. We will also study several test cases to illustrate the quality of the computed solutions and to confirm the analysis. This chapter is heavily based on [Zak16a].

Contents

4.1 Introduction

In [NBA$^+$14], the authors applied a flux-splitting scheme to the full Euler equations, which uses a variant of Klein's auxiliary splitting [Kle95]. The scheme required an ε-dependent time step for stability. Motivated by this, [SN14] began a stability analysis of the modified equation of linear systems in Fourier variables, and suggested that the commutator of the stiff and non-stiff flux Jacobian matrices may be important for the stability as it is roughly $\mathcal{O}(1/\varepsilon)$; see also [ZN17] for a generalisation of the analysis. That study leads to the main idea of the RS-IMEX scheme whose rigorous asymptotic analysis is the core topic of this chapter.

The fundamental idea is the linearisation around an *(asymptotic) reference solution* such that the resulting modified equation is stable; see Chapter 3. In fact, using the asymptotic reference solution gives a small commutator, which provides a heuristic argument for the stability of the modified equation as mentioned in Remark 3.4.4; see also Remark 4.2.2 below for a discussion on this. In addition to this manuscript, the RS-IMEX scheme has been studied in the works of our collaborators and has been shown to be quantitatively well-behaved in practice. For instance in [SK16], the applicability of the scheme has been illustrated for the stiff Van der Pol equation; also in [KSSN16, KS17], the uniform accuracy and formal asymptotic consistency of the scheme have been studied for the (high-order) RS-IMEX scheme applied to the isentropic Euler equations.

In the present chapter, we restrict our attention to the rigorous AP analysis for the case of 1d shallow water equations (SWE), *i.e.*, asymptotic consistency, asymptotic stability and convergence to the limit for fixed grids (see Remark 4.3.11). These make a solid background for next chapters which extend the scheme to the multi-dimensional SWE with different source terms. Note that broadly speaking, the splitting developed in [BALMN14, Bis15] can be considered as a particular example of the RS-IMEX scheme, with the zero reference solution; see Remark 4.3.5 for more details. We would also like to mention that a somewhat similar idea to the RS-IMEX scheme has been used in [FJ10] (as the so-called *penalisation method* [HJL16]) for the kinetic equations with a low Knudsen number, where the authors split the collision operator using the linearisation around the Maxwellian distribution. Also, one can see similarities between the RS-IMEX scheme and the multiple pressure variables (MPV) approach [KM95]; we will elaborate on this point in Chapter 5.

The remainder of this chapter is organised as follows. In Section 4.2 we present a short introduction to the RS-IMEX scheme, which follows in Section 4.3 and Section 4.4 with the rigorous AP analysis (consistency and stability) of the RS-IMEX scheme for the 1d SWE, with the lake at rest and the zero-Froude limit reference solutions. We will see that although the reference solution is rather simple, the rigorous analysis is not too straightforward. Section 4.5 provides some numerical evidence to confirm the AP analysis and test the quality of the solutions.

4.2 RS-IMEX splitting for hyperbolic systems of balance laws

The goal of this section is to provide an introduction to the RS-IMEX scheme to be applied to the SWE in Section 4.3. Consider the hyperbolic system of balance laws in the d-dimensional domain

$\Omega \subset \mathbb{R}^d$, depending on the singular parameter $\varepsilon \in (0,1]$ (*e.g.*, the Froude or Mach number):

$$\partial_t \boldsymbol{U}(t,\boldsymbol{x};\varepsilon) + \operatorname{div}_{\boldsymbol{x}} \boldsymbol{F}(\boldsymbol{U},t,\boldsymbol{x};\varepsilon) = \boldsymbol{S}(\boldsymbol{U},t,\boldsymbol{x};\varepsilon), \tag{4.1}$$

where $\boldsymbol{U} : [0,+\infty) \times \Omega \times (0,1] \to \mathbb{R}^q$ is the vector of unknowns, $\boldsymbol{F} : \mathbb{R}^q \times [0,+\infty) \times \Omega \times (0,1] \to \mathbb{R}^{q\times d}$ is the flux matrix (in d space dimensions), *i.e.*, $\boldsymbol{F} = [\boldsymbol{f}_1, \dots, \boldsymbol{f}_d]$ with $\boldsymbol{f}_k \in \mathbb{R}^q$, and $\boldsymbol{S} : \mathbb{R}^q \times [0,+\infty) \times \Omega \to \mathbb{R}^q$ is the source term, *e.g.*, due to the gravitational force, Coriolis force, or bottom friction. Note that we often suppress the dependence of $\boldsymbol{U}$, $\boldsymbol{F}$ and $\boldsymbol{S}$ on ε. Ω is chosen to be periodic (a torus), *i.e.*, $\Omega = \mathbb{T}^d$ for the sake of simplicity. To have a hyperbolic system, we also assume that $\boldsymbol{F}$ has a real diagonalisable Jacobian $\boldsymbol{F}' := \partial_{\boldsymbol{U}} \boldsymbol{F}$.

Let us consider the given ε-independent function $\overline{\boldsymbol{U}}$ as the *reference solution*:

$$\overline{\boldsymbol{U}} : [0,+\infty) \times \mathbb{R}^d \to \mathbb{R}^q, \quad (t,\boldsymbol{x}) \mapsto \overline{\boldsymbol{U}}(t,\boldsymbol{x}). \tag{4.2}$$

The ε-independence assumption can be relaxed; but, we stick to it here. Typically, $\overline{\boldsymbol{U}}$ is a steady state solution of the balance law, or the solution of the asymptotic limit equation, derived from (4.1) as $\varepsilon \to 0$, *e.g.*, it can be the lake at rest state for the SWE or the incompressible limit for the Euler equations.

Given the reference solution, we split the solution $\boldsymbol{U}$ of the balance law (4.1) into the reference solution $\overline{\boldsymbol{U}}$ and a perturbation $\boldsymbol{U}_{pert}$, that is $\boldsymbol{U}(t,\boldsymbol{x};\varepsilon) = \overline{\boldsymbol{U}}(t,\boldsymbol{x};\varepsilon) + \boldsymbol{U}_{pert}(t,\boldsymbol{x};\varepsilon)$. We aim to design an algorithm for the perturbation $\boldsymbol{U}_{pert}$ which is asymptotically stable and consistent. The algorithm uses the IMEX approach, and the CFL number for the explicit part should be ε-uniform. Achieving this goal, we split the flux and source terms using the Taylor expansion (linearisation) around $\overline{\boldsymbol{U}}$:

$$\boldsymbol{F}(\boldsymbol{U}) = \boldsymbol{F}(\overline{\boldsymbol{U}}) + \boldsymbol{F}'(\overline{\boldsymbol{U}})\boldsymbol{U}_{pert} + \big(\boldsymbol{F}(\boldsymbol{U}) - \boldsymbol{F}(\overline{\boldsymbol{U}}) - \boldsymbol{F}'(\overline{\boldsymbol{U}})\boldsymbol{U}_{pert}\big) =: \overline{\boldsymbol{F}} + \widetilde{\boldsymbol{F}} + \widehat{\boldsymbol{F}}, \tag{4.3a}$$

$$\boldsymbol{S}(\boldsymbol{U}) = \boldsymbol{S}(\overline{\boldsymbol{U}}) + \boldsymbol{S}'(\overline{\boldsymbol{U}})\boldsymbol{U}_{pert} + \big(\boldsymbol{S}(\boldsymbol{U}) - \boldsymbol{S}(\overline{\boldsymbol{U}}) - \boldsymbol{S}'(\overline{\boldsymbol{U}})\boldsymbol{U}_{pert}\big) =: \overline{\boldsymbol{S}} + \widetilde{\boldsymbol{S}} + \widehat{\boldsymbol{S}}. \tag{4.3b}$$

Note that the stiff part of the splitting is linear by construction, which is very advantageous in terms of computational cost, compared to splittings with non-linear stiff parts like [HJL12, DT11]. Hence, there is no need for solving non-linear systems, *e.g.*, by the Newton-Raphson iteration method. The idea of such a linearisation goes back to [Ros63] (see also [HW96, Chap IV.7]) for ODEs (the so-called linearly-implicit methods) and has been used later in [BALMN14] for the SWE, motivated by [GR10]. So, in a sense, the RS-IMEX splitting is a linearly-implicit method with a general linearisation state.

It may be useful to scale the components of the perturbation by a suitable scaling in order to work with $\mathcal{O}(1)$ terms in the analysis of the scheme. Later on in Section 4.3, we see that an appropriate choice of the scaling matrix, not only makes the analysis more illustrative (see Remark 4.3.10) but also may affect the numerical solution (see Remark 4.4.2). For this reason, we introduce the diagonal matrix $D := \operatorname{diag}(\varepsilon^{d_j})$ with $j = 1, \dots, q$, and define the *scaled* (preconditioned) perturbation $\boldsymbol{V}(t,\boldsymbol{x})$ as $\boldsymbol{V} := D^{-1}\boldsymbol{U}_{pert}$ and denote the corresponding *scaled* flux and source terms by $\boldsymbol{G} := D^{-1}\boldsymbol{F}$ and $\boldsymbol{Z} := D^{-1}\boldsymbol{S}$. So, with $\overline{\boldsymbol{G}}, \widetilde{\boldsymbol{G}}, \widehat{\boldsymbol{G}}, \overline{\boldsymbol{Z}}, \widetilde{\boldsymbol{Z}}$ and $\widehat{\boldsymbol{Z}}$ defined analogously as for $\boldsymbol{F}$ and $\boldsymbol{S}$, the splittings (4.3a) and (4.3b) can be rewritten:

$$\boldsymbol{G} = \overline{\boldsymbol{G}} + \widetilde{\boldsymbol{G}} + \widehat{\boldsymbol{G}}, \qquad \boldsymbol{Z} = \overline{\boldsymbol{Z}} + \widetilde{\boldsymbol{Z}} + \widehat{\boldsymbol{Z}}.$$

Remark 4.2.1. *It is really important to note that the eigenvalues of* $\widetilde{\boldsymbol{F}}' := \partial_{\boldsymbol{U}_{pert}} \widetilde{\boldsymbol{F}}$ *and* $\widehat{\boldsymbol{F}}' := \partial_{\boldsymbol{U}_{pert}} \widehat{\boldsymbol{F}}$ *are exactly the same as the eigenvalues of* $\widetilde{\boldsymbol{G}}' := \partial_{\boldsymbol{V}} \widetilde{\boldsymbol{G}}$ *and* $\widehat{\boldsymbol{G}}' := \partial_{\boldsymbol{V}} \widehat{\boldsymbol{G}}$*, respectively.*

This is because these matrices can be transformed into each other by a similarity transformation with D. So, the scaling does not change the eigenvalues, thus the admissibility of the splitting.

Then, one is left with the following system for the perturbation $\boldsymbol{V} = (v_1, \dots, v_q)^T$:

$$\partial_t \boldsymbol{V} + \operatorname{div}_{\boldsymbol{x}} \left(\widetilde{\boldsymbol{G}}(\overline{\boldsymbol{U}}, \boldsymbol{V}) + \widehat{\boldsymbol{G}}(\overline{\boldsymbol{U}}, \boldsymbol{V}) \right) = \widetilde{\boldsymbol{Z}}(\overline{\boldsymbol{U}}, \boldsymbol{V}) + \widehat{\boldsymbol{Z}}(\overline{\boldsymbol{U}}, \boldsymbol{V}) - \overline{\boldsymbol{T}}(\overline{\boldsymbol{U}}), \tag{4.4}$$

where $\overline{\boldsymbol{T}}(\overline{\boldsymbol{U}})$ is the (*a priori*-known) *scaled residual of the reference solution* and reads

$$\overline{\boldsymbol{T}}(\overline{\boldsymbol{U}}) := D^{-1} \partial_t \overline{\boldsymbol{U}} + \operatorname{div}_{\boldsymbol{x}} \overline{\boldsymbol{G}}(\overline{\boldsymbol{U}}) - \overline{\boldsymbol{Z}}(\overline{\boldsymbol{U}}). \tag{4.5}$$

Remark 4.2.2. *One can confirm that the non-stiff Jacobian is $\widehat{\boldsymbol{G}}' = \boldsymbol{G}'(\boldsymbol{U}) - \boldsymbol{G}'(\overline{\boldsymbol{U}})$ while the stiff one reads $\widetilde{\boldsymbol{G}}' = \boldsymbol{G}'(\overline{\boldsymbol{U}})$. So, the commutator can be obtained as*

$$\left[\widehat{\boldsymbol{G}}', \widetilde{\boldsymbol{G}}'\right] := \widehat{\boldsymbol{G}}' \widetilde{\boldsymbol{G}}' - \widetilde{\boldsymbol{G}}' \widehat{\boldsymbol{G}}' = \left[\boldsymbol{G}'(\boldsymbol{U}), \boldsymbol{G}'(\overline{\boldsymbol{U}})\right].$$

By choosing the reference solution as the asymptotic limit, $\|\boldsymbol{U} - \overline{\boldsymbol{U}}\| \ll 1$, which suggests that the commutator can be small, and supports the stability of the modified equation, cf. Chapter 3 and [ZN17]. Moreover, in the light of the form (3.13)*, one can deduce that the eigenvectors of $\boldsymbol{G}'(\boldsymbol{U})$ and $\boldsymbol{G}'(\overline{\boldsymbol{U}})$ are asymptotically close to each other. So, $Q_{R \to \widetilde{R}}$ approaches the identity matrix, and* (3.13) *confirms the stability of the modified equation.*

Nonetheless, it does not mean that only the asymptotic reference solution provides stability. In fact, the similarity in the structure of the eigenvectors of $\boldsymbol{G}'(\boldsymbol{U})$ and $\boldsymbol{G}'(\overline{\boldsymbol{U}})$—which does not depend on $\|\boldsymbol{U} - \overline{\boldsymbol{U}}\|$—motivates the conjecture that it is the linearly-implicit strategy which makes the modified equation stable, neither using the asymptotic reference solution nor smallness of the commutator. We are not able to prove or disprove that; but, at least, we discuss an example in Section 4.3 showing that even with a non-asymptotic reference solution, which gives a large commutator, $Q_{R \to \widetilde{R}}$ tends to the identity matrix and the modified equation is stable. Although the reference solution may not affect the stability of the scheme, we will show in analysis and practice that the choice of the reference solution does matter for the quality of the computed solution (see Remark 4.4.4 below).

Defining $\boldsymbol{R} := -\operatorname{div}_{\boldsymbol{x}} \boldsymbol{G} + \boldsymbol{Z}$ (with analogous definitions for $\overline{\boldsymbol{R}}$, $\widetilde{\boldsymbol{R}}$ and $\widehat{\boldsymbol{R}}$), one can reformulate (4.4) as

$$\partial_t \boldsymbol{V} = -\overline{\boldsymbol{T}} + \widetilde{\boldsymbol{R}} + \widehat{\boldsymbol{R}}, \tag{4.6}$$

which is a balance law for the scaled perturbation $\boldsymbol{V}$. Note that using (4.4) and (4.6) is not indispensable for the numerical scheme, but it is suitable, notably, for the asymptotic consistency analysis. Note also that $\overline{\boldsymbol{T}} \equiv 0$ if and only if the reference solution $\overline{\boldsymbol{U}}$ satisfies the original system (4.1).

4.2.1 Numerical scheme

The Jacobian $\widetilde{\boldsymbol{F}}'$ in (4.3a) (and $\widetilde{\boldsymbol{G}}'$ due to Remark 4.2.1) has stiff eigenvalues. So, to solve (4.6) numerically, we treat the stiff part $\widetilde{\boldsymbol{R}}$ implicitly in time to avoid restrictive time steps, by using

the implicit Euler time integration. The term $\widehat{\boldsymbol{R}}$ is *expected* to be non-stiff;[1] so, the explicit Euler scheme is employed. Note that in the sequel, we limit ourselves to first-order schemes. The residual $\overline{\boldsymbol{T}}$ is computed independently, *e.g.*, by an appropriate incompressible solver for the Euler system with the incompressible reference solution. Thus, we can define the RS-IMEX scheme as follows, where Δt is the time step, $D_t\phi(t,\boldsymbol{x}) := \frac{\phi(t+\Delta t,\boldsymbol{x})-\phi(t,\boldsymbol{x})}{\Delta t}$, and the subscript Δ stands for a choice of spatial discretisation.

Definition 4.2.3. *Given the reference solution $\overline{\boldsymbol{U}}$, the fully-discrete RS-IMEX scheme for* (4.6) *is given by*

$$D_t\boldsymbol{V}_\Delta^n = -\overline{\boldsymbol{T}}_\Delta^{n+1} + \widetilde{\boldsymbol{R}}_\Delta^{n+1} + \widehat{\boldsymbol{R}}_\Delta^n. \tag{4.7}$$

For the spatial discretisation of the flux, a Rusanov-type numerical flux will be used, which is defined as $f_{i+1/2} := \frac{f(u_i)+f(u_{i+1})}{2} - \frac{\alpha_{i+1/2}}{2}(u_{i+1}-u_i)$, in one space dimension and for the scalar flux $f(u)$ at the interface $i+1/2$, where i denotes the cell index. The numerical diffusion coefficient α is originally chosen such that $\alpha_{i+1/2} \geq \max_{u\in[u_i,u_{i+1}]}|f'(u)|$, which means that for the stiff sub-system, which would be treated implicitly, the numerical diffusion coefficient would be substantial. As the implicit schemes are diffusive inherently, we choose α for the stiff sub-system rather arbitrary, *e.g.*, by computing it for $\varepsilon = 1$ such that $\widetilde{\alpha}, \widehat{\alpha} = \mathcal{O}(1)$. Later on and for numerical examples, we also define corresponding coefficients $0 \leq c_{\widetilde{\alpha}}, c_{\widehat{\alpha}} \leq 1$, and $\widetilde{\alpha}$ and $\widehat{\alpha}$ would be multiplied by them so that only by tuning these coefficients we can control the numerical diffusion in practice (see Section 4.5 as well as next chapters). The extension of this numerical flux to systems and in multi-dimensions is obvious. Note that the source term should be discretised appropriately so that the scheme preserves the equilibrium (well-balancing). We will see in Section 4.3 that the central discretisation is appropriate for the SWE with topography.

In fact for the RS-IMEX scheme, two systems should be solved, one for the reference solution and the other for the scaled perturbation. With a given reference state at step n, one finds the discretised scaled perturbation $\boldsymbol{V}_\Delta^{n+1}$, while the reference state $\overline{\boldsymbol{U}}_\Delta^{n+1}$ may evolve over time and should be computed independently. At the end of each step, the solution can be computed as $\overline{\boldsymbol{U}}_\Delta^{n+1} + D\boldsymbol{V}_\Delta^{n+1}$. The RS-IMEX procedure has been summarised in Algorithm 1.

Algorithm 1 RS-IMEX scheme

1: Get $\overline{\boldsymbol{U}}_\Delta^n$ and $\boldsymbol{V}_\Delta^n$.
2: **Reference step:** Find the updated reference state $\overline{\boldsymbol{U}}_\Delta^{n+1}$.
3: **Explicit step:** Solve $D_t\boldsymbol{V}_\Delta^n = \widehat{\boldsymbol{R}}_\Delta^n$ to find $\boldsymbol{V}_\Delta^{n+1/2}$.
4: **Implicit step:** Solve $D_t\boldsymbol{V}_\Delta^{n+1/2} = -\overline{\boldsymbol{T}}_\Delta^{n+1} + \widetilde{\boldsymbol{R}}_\Delta^{n+1}$ to find the updated perturbation $\boldsymbol{V}_\Delta^{n+1}$.
5: Find the updated solution as $\boldsymbol{U}_\Delta^{n+1} = \overline{\boldsymbol{U}}_\Delta^{n+1} + D\boldsymbol{V}_\Delta^{n+1}$.
6: Continue with step 2.

Remark 4.2.4. *The RS-IMEX scheme introduced here is a bit different from [KSSN16, SK16] mainly in two aspects. Firstly, in those series of papers, the reference solution is computed implicitly as the limit of the singularly-perturbed system while here we employ off-the-shelf methods for this purpose. Moreover, here we use the reformulation* (4.6) *with the scaled perturbation, which makes the analysis easier and more illustrative as we will see later on.*

[1] In general, we do not know if the system is non-stiff or not. But this can be shown for the systems we are dealing with in practice like the shallow water or Euler equations.

4.3 Shallow water equations with the lake at rest reference solution

In this section and as an example of the RS-IMEX scheme for the system (4.1), we follow the procedure described in Section 4.2 to derive the scheme for the 1d SWE with topography and the LaR reference solution. Then, in Section 4.4, we use the zero-Froude limit reference solution.

The non-dimensionalised SWE in one space dimension, using $z = h + b$ (with $b < 0$) and $m = hu$, can be written as in [BALMN14]:

$$\boldsymbol{U} = \begin{bmatrix} z \\ m \end{bmatrix}, \qquad \boldsymbol{F} = \begin{bmatrix} m \\ \dfrac{m^2}{z-b} + \dfrac{z^2 - 2zb}{2\varepsilon^2} \end{bmatrix}, \qquad \boldsymbol{S} = \begin{bmatrix} 0 \\ -\dfrac{z}{\varepsilon^2}\partial_x b \end{bmatrix}. \tag{4.8}$$

In this notation, z is the surface elevation from some chosen constant surface level H_{mean} (namely the mean water level), m is the momentum and b is the water depth measured from H_{mean} with a negative sign (see Figure 4.1). The singular parameter $\varepsilon \in (0, 1]$ is the Froude number; see (1.10).

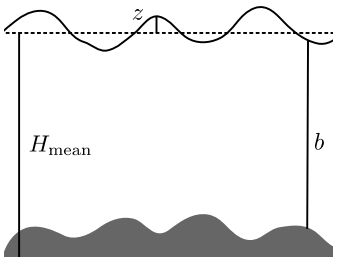


Figure 4.1: Variables used in the shallow water formulation (4.8).

We set the reference state as the LaR equilibrium state, $\overline{\boldsymbol{U}} := (\overline{z}, \overline{m})^T$ with $\overline{z}$ constant in space and $\overline{m} = 0$. Therefore, due to (4.3a)–(4.3b), the splitting reads

$$\overline{\boldsymbol{F}} = \begin{bmatrix} 0 \\ \dfrac{\overline{z}(\overline{z} - 2b)}{2\varepsilon^2} \end{bmatrix}, \qquad \widetilde{\boldsymbol{F}} = \begin{bmatrix} m_{pert} \\ \dfrac{(\overline{z} - b)}{\varepsilon^2} z_{pert} \end{bmatrix}, \qquad \widehat{\boldsymbol{F}} = \begin{bmatrix} 0 \\ \dfrac{m_{pert}^2}{\overline{z} + z_{pert} - b} + \dfrac{z_{pert}^2}{2\varepsilon^2} \end{bmatrix}, \tag{4.9a}$$

$$\overline{\boldsymbol{S}} = \begin{bmatrix} 0 \\ -\dfrac{\overline{z}}{\varepsilon^2}\partial_x b \end{bmatrix}, \qquad \widetilde{\boldsymbol{S}} = \begin{bmatrix} 0 \\ -\dfrac{z_{pert}}{\varepsilon^2}\partial_x b \end{bmatrix}, \qquad \widehat{\boldsymbol{S}} = \boldsymbol{0}. \tag{4.9b}$$

One can see that the Jacobian of $\widetilde{\boldsymbol{F}}$ (w.r.t. $\boldsymbol{U}_{pert}$) has stiff eigenvalues $\widetilde{\lambda} = \mathcal{O}(1/\varepsilon)$, while the eigenvalues of $\widehat{\boldsymbol{F}}'$, denoted by $\widehat{\lambda}$, are non-stiff. More precisely

$$\begin{aligned} \widetilde{\boldsymbol{F}}' &= \begin{bmatrix} 0 & 1 \\ \dfrac{\overline{z} - b}{\varepsilon^2} & 0 \end{bmatrix}, & \widetilde{\lambda} &= \pm\frac{\sqrt{\overline{z} - b}}{\varepsilon}, \\ \widehat{\boldsymbol{F}}' &= \begin{bmatrix} 0 & 0 \\ -u_{pert}^2 + \dfrac{z_{pert}}{\varepsilon^2} & 2u_{pert} \end{bmatrix}, & \widehat{\lambda} &= 0, 2\,u_{pert}, \end{aligned} \tag{4.10}$$

with $u_{pert} := m_{pert}/(\overline{z} + z_{pert} - b)$. Thus, the splitting is admissible in the sense of Definition 3.1.1. Note that the case $\overline{\boldsymbol{U}} = \mathbf{0}$ gives the same splitting as in [BALMN14, Bis15].

Finding the scaling matrix D, we employ the formal asymptotic analysis in Appendix 2.A (see also [KM81, KM82] for the rigorous justification for the flat bottom case), which suggests the formal zero-Froude limit in Definition 4.3.1 below, with the following asymptotic (Poincaré) expansion

$$\begin{aligned} z(t,x) &= z_{(0)} + \varepsilon z_{(1)} + \varepsilon^2 z_{(2)}, \\ m(t,x) &= m_{(0)} + \varepsilon m_{(1)} + \varepsilon^2 m_{(2)}. \end{aligned} \tag{4.11}$$

Definition 4.3.1. *The formal zero-Froude limit of the shallow water system* (4.8) *gives the so-called "lake equations", and reads (see Appendix 2.A as well as [BKL11] for the formal justification)*

$$\begin{aligned} &z_{(0)}, z_{(1)} = const., \\ &\partial_x m_{(0)} = 0, \\ &\partial_t m_{(0)} + \partial_x \left(\frac{m_{(0)}^2}{z_{(0)} - b} + p_{(2)} \right) = -z_{(2)} \partial_x \eta^b. \end{aligned}$$

This suggests the following definition for the well-prepared initial condition for the SWE.

Definition 4.3.2. *For the 1d SWE* (4.8)*, we call the initial data* $(z_{0,\varepsilon}, m_{0,\varepsilon})$ *well-prepared if it holds that*

$$\begin{aligned} z(0,\cdot) = z_{0,\varepsilon} &= z_{(0)}^0 + \varepsilon^2 z_{(2),\varepsilon}^0, \\ m(0,\cdot) = m_{0,\varepsilon} &= m_{(0)}^0 + \varepsilon m_{(1),\varepsilon}^0, \end{aligned} \tag{4.12}$$

where $z_{(0)}$ *and* $m_{(0)}$ *are constant.*

The motivation for scaling the equations was to work with $\mathcal{O}(1)$ quantities. So, due to (4.12), we pick $\overline{z} = z_{(0)}^0$, which implies $D := \mathrm{diag}(\varepsilon^2, 1)$. For simplicity, we stick to this particular choice of $\overline{z}$ throughout this section. Nonetheless, it is rather straightforward to confirm that the asymptotic analysis we are going to present holds for every constant $\overline{z}$, while the choice may affect the numerical diffusion of the scheme, thus the solution.

4.3.1 RS-IMEX scheme

For the scaling matrix $\mathrm{diag}(\varepsilon^2, 1)$, the scaled split perturbation is $\boldsymbol{V} := (v_1, v_2)^T := (z_{pert}/\varepsilon^2, m_{pert})^T$ and the scaled splitting writes

$$\widehat{\boldsymbol{G}} = \begin{bmatrix} 0 \\ \dfrac{v_2^2}{\overline{z} + \varepsilon^2 v_1 - b} + \dfrac{\varepsilon^2}{2} v_1^2 \end{bmatrix}, \qquad \widetilde{\boldsymbol{G}} = \begin{bmatrix} v_2/\varepsilon^2 \\ (\overline{z} - b) v_1 \end{bmatrix}, \tag{4.13a}$$

$$\widehat{\boldsymbol{Z}} = \mathbf{0}, \qquad \widetilde{\boldsymbol{Z}} = \begin{bmatrix} 0 \\ -\partial_x b \, v_1 \end{bmatrix}. \tag{4.13b}$$

Owing to (4.10) and Remark 4.2.1, this splitting is also admissible even with an ill-prepared initial datum (as defined in Appendix 4.B).

Since the LaR equilibrium state is a stationary solution of the system, $\overline{z}$ is constant in time, and the reference solution needs not to be updated. Thus, $\overline{\boldsymbol{T}} \equiv 0$, and one can reformulate the 1d SWE as

$$\partial_t \boldsymbol{V} = -\partial_x \begin{bmatrix} v_2/\varepsilon^2 \\ (\overline{z}-b)v_1 \end{bmatrix} - \partial_x \begin{bmatrix} 0 \\ \dfrac{v_2^2}{\overline{z}+\varepsilon^2 v_1 - b} + \varepsilon^2 \dfrac{v_1^2}{2} \end{bmatrix} + \begin{bmatrix} 0 \\ -\partial_x b v_1 \end{bmatrix}. \tag{4.14}$$

The RS-IMEX scheme approximates this reformulated system as below, written as a two-step scheme:

$$\boldsymbol{V}_i^{n+1/2} = \boldsymbol{V}_i^n - \frac{\Delta t}{\Delta x}\left(\widehat{\boldsymbol{G}}_{i+1/2}^n - \widehat{\boldsymbol{G}}_{i-1/2}^n\right) \qquad \text{(Explicit step)}, \tag{4.15a}$$

$$\boldsymbol{V}_i^{n+1} = \boldsymbol{V}_i^{n+1/2} - \frac{\Delta t}{\Delta x}\left(\widetilde{\boldsymbol{G}}_{i+1/2}^{n+1} - \widetilde{\boldsymbol{G}}_{i-1/2}^{n+1}\right) + \Delta t\, \widetilde{\boldsymbol{Z}}_i^{n+1} \qquad \text{(Implicit step)}, \tag{4.15b}$$

for each cell $i \in \{1, 2, \dots, N\}$ in the computational domain Ω_N with N cells of size Δx, where $\widetilde{\boldsymbol{G}}_{i+1/2}$ and $\widehat{\boldsymbol{G}}_{i+1/2}$ denote the Rusanov flux at cell interfaces as defined in Section 4.2, but for the simplicity of notation with a constant diffusion coefficient α chosen as the maximum value over the domain, with $\widehat{\boldsymbol{G}}$ and $\widetilde{\boldsymbol{G}}$ as in (4.13a), and $\widetilde{\boldsymbol{Z}}_i^n$ is the central discretisation of the source term (4.13b). Denoting ∇_h and Δ_h respectively as the central discretisation of the first and second derivatives, one can rewrite (4.15a)–(4.15b) as

$$\boldsymbol{V}_i^{n+1/2} = \boldsymbol{V}_i^n - \Delta t \nabla_h \begin{bmatrix} 0 \\ \dfrac{v_{2,i}^{n,2}}{\overline{z}+\varepsilon^2 v_{1,i}^n - b_i} + \dfrac{\varepsilon^2}{2} v_{1.i}^{n,2} \end{bmatrix} + \frac{\widehat{\alpha}\Delta x}{2}\Delta t \Delta_h \boldsymbol{V}_i^n, \tag{4.16a}$$

$$\boldsymbol{V}_i^{n+1} = \boldsymbol{V}_i^{n+1/2} - \Delta t \nabla_h \begin{bmatrix} v_{2,i}^{n+1}/\varepsilon^2 \\ (\overline{z}-b_i)\, v_{1,i}^{n+1} \end{bmatrix} + \frac{\widetilde{\alpha}\Delta x}{2}\Delta t \Delta_h \boldsymbol{V}_i^{n+1} - \Delta t \begin{bmatrix} 0 \\ v_{1,i}^{n+1} \nabla_h b_i \end{bmatrix}. \tag{4.16b}$$

Due to Remark 4.2.1, the eigenvalues of $\boldsymbol{F}'$ and $\boldsymbol{G}'$ (and their splittings) are the same; so, the eigenvalues of the non-stiff system are $\mathcal{O}(1)$. Also, note that the reference solution is not close to the solution in the limit. So, this splitting may not give a small commutator needed for the stability of the modified equation. Indeed, the commutator is formally $\mathcal{O}(1/\varepsilon^2)$:

$$\left[\widetilde{\boldsymbol{G}}', \widehat{\boldsymbol{G}}'\right] := \widetilde{\boldsymbol{G}}'\widehat{\boldsymbol{G}}' - \widehat{\boldsymbol{G}}'\widetilde{\boldsymbol{G}}' = \begin{bmatrix} v_1 - \dfrac{v_2^2}{(\overline{z}+\varepsilon^2 v_1 - b)^2} & \dfrac{2v_2/\varepsilon^2}{\overline{z}+\varepsilon^2 v_1 - b} \\ \dfrac{-2(\overline{z}-b)v_2}{\overline{z}+\varepsilon^2 v_1 - b} & -v_1 + \dfrac{v_2^2}{(\overline{z}+\varepsilon^2 v_1 - b)^2} \end{bmatrix}. \tag{4.17}$$

However, as shown in Chapter 3, the modified equation is asymptotically stable (for slow enough Fourier modes). One can confirm that the eigenvector matrices R and $\widetilde{R}$ are very similar in structure so that $Q_{R\to\widetilde{R}}$ is the identity matrix to the leading order, which corroborate the stability of the modified equation by virtue of the form (3.13). Denoting by r_j and $\widetilde{r}_j$ the eigenvectors of the original and the stiff system and $\overline{h} := \overline{z} - b$, due to (4.13a), one obtains

$$r_{1,2} = \begin{bmatrix} \dfrac{\pm 1}{\varepsilon\sqrt{\overline{h}}} + o(\varepsilon^{-1}) \\ 1 \end{bmatrix}, \qquad \widetilde{r}_{1,2} = \begin{bmatrix} \dfrac{\pm 1}{\varepsilon\sqrt{\overline{h}}} + o(\varepsilon^{-1}) \\ 1 \end{bmatrix}, \qquad \lim_{\varepsilon\to 0} Q_{R\to\widetilde{R}} = \mathbb{I}_2. \tag{4.18}$$

So, the modified equation is supposed to be stable despite the very large commutator.

Remark 4.3.3. *Note that* (4.18) *does not expose the structure of* $Q_{R\to\widetilde{R}}$ *completely as* z_{pert} *is assumed to be* $\mathcal{O}(\varepsilon)$*, so vanishes in the limit. A more complete picture can be obtained using the original splitting* (4.9a)*, which results in*

$$r_{1,2} = \begin{bmatrix} \frac{\pm\varepsilon}{\sqrt{h_{(0)}}} + o(\varepsilon) \\ 1 \end{bmatrix}, \qquad \widetilde{r}_{1,2} = \begin{bmatrix} \frac{\pm\varepsilon}{\sqrt{\overline{h}}} + o(\varepsilon) \\ 1 \end{bmatrix}, \qquad \lim_{\varepsilon\to 0} Q_{R\to\widetilde{R}} = \begin{bmatrix} 1+\sqrt{\frac{h_{(0)}}{\overline{h}}} & 1-\sqrt{\frac{h_{(0)}}{\overline{h}}} \\ 1-\sqrt{\frac{h_{(0)}}{\overline{h}}} & 1+\sqrt{\frac{h_{(0)}}{\overline{h}}} \end{bmatrix}.$$

If the $|\overline{h}-h|$ *tends to zero as* $\varepsilon\to 0$*,* (4.18) *is recovered and* $Q_{R\to\widetilde{R}}$ *tends to the identity (up to a scaling).*

4.3.2 Asymptotic analysis of the scheme

We collect the properties of the RS-IMEX scheme in the following theorem.

Theorem 4.3.4. *For the shallow water equations with topography and well-prepared initial data in the sense of Definition 4.3.2, the RS-IMEX scheme* (4.16a)–(4.16b)*, with* (4.13a)–(4.13b)*, a constant* $\widetilde{\alpha}$*, and under an* ε*-uniform time step restriction*

(i) is solvable, i.e., it has a unique solution for all $\varepsilon > 0$*.*

(ii) has an ε*-stable solution, i.e., it is bounded for* $\varepsilon \ll 1$*. So, there is a strongly convergent subsequence of the discrete solutions as* $\varepsilon\to 0$*.*

(iii) is consistent with the asymptotic limit in the fully-discrete settings, i.e., it is asymptotically consistent.

(iv) is asymptotically ℓ_2*-stable for the fixed grid* Δx*, in finite time* $T_f < \infty$ *and with small enough initial data, i.e., there exists a constant* C_{N,T_f} *such that* $\|\boldsymbol{V}_\Delta^n\|_{\ell_2} \le C_{N,T_f}\|\boldsymbol{V}_\Delta^0\|_{\ell_2}$*.*

(v) preserves the lake at rest equilibrium state, i.e., it is well-balanced.

We present the proof of Theorem 4.3.4 in the next sections.

Remark 4.3.5. *As we have already mentioned, the scheme in [BALMN14, Bis15] is a particular example of the RS-IMEX scheme with the zero reference solution. So, one may expect that the analysis in [Bis15] coincides with Theorem 4.3.4. The difference is that the analysis of [Bis15] is basically for the flat bottom case and detailed analysis has been done for various high order reconstructions. By contrast, throughout this chapter, we focus on the first-order schemes in one space dimension and prove asymptotic consistency for a non-flat topography. In Section 4.4, we show that a similar analysis can be used for more general reference solutions.*

4.3.2.1 Solvability of the scheme

Here, we aim to show that there exists a unique solution for the implicit step (so for the scheme) for all $\varepsilon > 0$. At first and for simplicity, we assume the topography b to be constant, which makes the system similar to the isentropic Euler system. Then, we generalise the arguments for the SWE with a varying bottom. To simplify the notation, we $\beta := \frac{\Delta t}{2\Delta x}$.

(i) Constant b Owing to (4.16b), we write the implicit step as $J_\varepsilon \boldsymbol{V}_\Delta^{n+1} = \boldsymbol{V}_\Delta^{n+1/2}$, *i.e.*, the implicit solution operator is J_ε^{-1}. The matrix $J_\varepsilon \in \mathbb{R}^{2N\times 2N}$ writes

$$J_\varepsilon := \begin{bmatrix} P & \dfrac{\beta}{\varepsilon^2}Q \\ \beta\overline{h}Q & P \end{bmatrix}, \tag{4.19}$$

where P and Q are circulant matrices defined as

$$P := \boldsymbol{Circ}\left(1+2\widetilde{\alpha}\beta, -\widetilde{\alpha}\beta, 0, \ldots, 0, -\widetilde{\alpha}\beta\right), \qquad Q := \boldsymbol{Circ}\left(0, 1, 0, \ldots, 0, -1\right).$$

Matrix P is symmetric and strictly diagonally dominant (SDD); so, it has positive real eigenvalues. Matrix Q, as the companion matrix for the central discretisation, is skew-symmetric with eigenvalues on the imaginary axis.

Since P and Q are circulant, they commute [Gra06], and one knows from [Sil00, Thm. 1] (see also [Ber09, Sect. 2.14]) that since all blocks of J_ε commute with each other, the determinant of J_ε can be computed as

$$\det(J_\varepsilon) = \det\left(P^2 - \frac{\overline{h}\beta^2}{\varepsilon^2}Q^2\right).$$

Due to Gerschgorin's circle theorem [HJ86, Chap. 6], the numerical range [HJ91, Chap. 1] of $-\frac{\overline{h}\beta^2}{\varepsilon^2}Q^2$ is non-negative while of P^2 is strictly positive, and both of these parts are symmetric with real eigenvalues. So, using the sub-additivity of numerical range (or the Rayleigh quotient) (*cf.* [HJ91, Chap. 1]), the eigenvalues of the sum cannot be zero. Thus J_ε is not singular, and there exists a unique solution for the scheme.

(ii) Non-constant b For this case, one of the blocks of J_ε is not circulant; the matrix J_ε is written as

$$J_\varepsilon = \begin{bmatrix} P & \dfrac{\beta}{\varepsilon^2}Q \\ \beta R_b & P \end{bmatrix}, \tag{4.20}$$

where R_b is an *almost* circulant matrix such that its i-th row is $(R_b)_i = (b_{i+1}-b_{i-1}, \overline{h}_{i+1}, 0, \ldots, 0, -\overline{h}_{i-1})$, up to a circulation. Note that R_b is circulant only if its arguments are constant for all rows, *i.e.*, if the bottom is flat.

Showing solvability of the scheme for the non-flat bottom case, we can use the fact that since circulant matrices are commutable, they are simultaneously diagonalisable, *i.e.*, any circulant matrix $M \in \mathbb{R}^{N\times N}$ can be diagonalised as $F_N^* M F_N =: \Lambda_M$, where $*$ denotes the conjugate transpose, F_N is a (unique) unitary matrix, which consists of eigenvectors of circulant matrices of size N, and Λ_M is the diagonal matrix of eigenvalues. It is important to mention that F_N does not depend on the entries of M, but only on the size N (see [Gra06]). Using this fact, one can consider the transformed matrix Ξ_ε for showing solvability where

$$\Xi_\varepsilon := \operatorname{diag}\left(F_N^*, F_N^*\right) J_\varepsilon \operatorname{diag}\left(F_N^*, F_N^*\right) = \begin{bmatrix} \Lambda_P & \dfrac{\beta}{\varepsilon^2}\Lambda_Q \\ \beta\overline{h}F_N^* R_b F_N & \Lambda_P \end{bmatrix}.$$

From [Ber09, Fact 2.14.13] and since the blocks Ξ_{11} and Ξ_{12} are commutable, the determinant of Ξ_ε can be written as

$$\det(\Xi_\varepsilon) = \det\left(\Lambda_P^2 - \frac{\overline{h}\beta^2}{\varepsilon^2}\Lambda_Q F_N^* R_b F_N\right).$$

Matrix P is SDD and invertible [HJ86, Thm. 6.1.10]; thus, Λ_P does not have a zero on the diagonal. So, as the matrix $\Lambda_Q F_N^* R_b F_N$ does depend neither on ε nor on β, a suitable choice for β makes Ξ_ε invertible and concludes that J_ε in (4.20) is invertible as well.

4.3.2.2 ε-stability of the solution

In this section, aiming to justify the validity of the formal Poincaré expansion to be used for the formal asymptotic consistency analysis, we prove that the *implicit operator* is bounded in terms of ε. As in Section 2.3.2, we call such a property *ε-stability*. Note that the ε-stability of the implicit operator does not provide ε-stability of the solution *per se*. For that, one also needs the ε-stability of the explicit step at the intermediate time $n+1/2$; see Section 4.3.2.3. Similar to the solvability analysis, we present the proofs for flat and non-flat bottom topographies separately.

(i) Constant b For this case the matrix Ξ_ε can be obtained as

$$\Xi_\varepsilon := \begin{bmatrix} \Lambda_P & \frac{\beta}{\varepsilon^2}\Lambda_Q \\ \beta\overline{h}\Lambda_Q & \Lambda_P \end{bmatrix}. \tag{4.21}$$

Since Q is skew-symmetric, it has only eigenvalues on the imaginary axis, so $\Lambda_Q^* = -\Lambda_Q$. Also, note that $\operatorname{diag}(F_N, F_N)$ is a unitary matrix. Thus, one can bound the norm of J_ε^{-1} as

$$\|J_\varepsilon^{-1}\| \le \|\operatorname{diag}(F_N, F_N)\|\|\operatorname{diag}(F_N^*, F_N^*)\|\|\Xi_\varepsilon^{-1}\| \le \operatorname{cond}(\operatorname{diag}(F_N, F_N))\,\|\Xi_\varepsilon^{-1}\|,$$

for any natural matrix norm. This bound depends on ε only through $\|\Xi_\varepsilon^{-1}\|$; so, we have to show that Ξ_ε^{-1} is uniformly bounded in ε. Before this, let us mention the following lemma for the inverse of partitioned matrices, since we are going to use it several times. This is a classical result; see, *e.g.*, [Ber09, Prop. 2.8.7].

Lemma 4.3.6 (Schur complement)**.** *Consider the partitioned matrix* $M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$. *Then, the inverse of M exists and writes*

$$M^{-1} = \begin{bmatrix} \left(M_{11} - M_{12}M_{22}^{-1}M_{21}\right)^{-1} & -M_{11}^{-1}M_{12}\left(M_{22} - M_{21}M_{11}^{-1}M_{12}\right)^{-1} \\ -M_{22}^{-1}M_{21}\left(M_{11} - M_{12}M_{22}^{-1}M_{21}\right)^{-1} & \left(M_{22} - M_{21}M_{11}^{-1}M_{12}\right)^{-1} \end{bmatrix} \tag{4.22}$$

if all the inverses exist.

Now, we can prove the uniform boundedness of $\|\Xi_\varepsilon^{-1}\|$ in ε.

Lemma 4.3.7. *The inverse of matrix Ξ_ε in* (4.21)*, has a bounded norm for $\varepsilon \to 0$.*

Proof. From Lemma 4.3.6, the inverse of Ξ_ε reads

$$\Xi_\varepsilon^{-1} = \begin{bmatrix} \left(\Lambda_P - \dfrac{\beta^2\overline{h}}{\varepsilon^2}\Lambda_Q^2\Lambda_P^{-1}\right)^{-1} & -\dfrac{\beta}{\varepsilon^2}\Lambda_P^{-1}\Lambda_Q\left(\Lambda_P - \dfrac{\beta^2\overline{h}}{\varepsilon^2}\Lambda_Q^2\Lambda_P^{-1}\right)^{-1} \\ -b\beta\Lambda_P^{-1}\Lambda_Q\left(\Lambda_P - \dfrac{\beta^2\overline{h}}{\varepsilon^2}\Lambda_Q^2\Lambda_P^{-1}\right)^{-1} & \left(\Lambda_P - \dfrac{\beta^2\overline{h}}{\varepsilon^2}\Lambda_Q^2\Lambda_P^{-1}\right)^{-1} \end{bmatrix}.$$

So, one can easily check that each block is bounded, thus is $\|\Xi_\varepsilon^{-1}\|$. □

Remark 4.3.8. *Lemma 4.3.7 concludes that the implicit solution operator J_ε^{-1} is bounded in terms of ε. The immediate result of this ε-stability is that the scaled perturbation $\mathbf{V}_\Delta$ should be $\mathcal{O}(1)$ as long as the explicit step is ε-stable. This result justifies the asymptotic consistency analysis we are going to present in Section 4.3.2.3.*

(ii) Non-constant b For this case, employing the diagonal form of circulant matrices cannot simplify all the blocks of J_ε^{-1} (unlike (4.21)) and the procedure of Lemma 4.3.7 does not seem to be fruitful. Using Lemma 4.3.6 for the inversion of partitioned matrices, one gets (with $\widetilde{\alpha} = 0$ for simplicity)

$$J_\varepsilon^{-1} = \begin{bmatrix} \left(\mathbb{I}_n - \dfrac{\beta^2}{\varepsilon^2}QR_b\right)^{-1} & -\dfrac{\beta}{\varepsilon^2}Q\left(\mathbb{I}_N - \dfrac{\beta^2}{\varepsilon^2}R_bQ\right)^{-1} \\ -\beta R_b\left(\mathbb{I}_N - \dfrac{\beta^2}{\varepsilon^2}QR_b\right)^{-1} & \left(\mathbb{I}_N - \dfrac{\beta^2}{\varepsilon^2}R_bQ\right)^{-1} \end{bmatrix}.$$

As R_b is close to Q, it is plausible to guess that the block $(\mathbb{I}_n - \frac{\beta^2}{\varepsilon^2}QR_b)^{-1}$ is a constant matrix with some $\mathcal{O}(\varepsilon^2)$ fluctuations (see [Zak17a] for further details). However, the fact that the bottom topography is rather general makes the proof difficult. So, we employ an indirect approach, motivated by $\|J_\varepsilon^{-1}\|_{\ell_2} = \sigma_{\min}^{-1}(J_\varepsilon)$ for σ denoting the singular values, and show that the smallest singular value of J_ε does not approach zero in the limit. From Section 4.3.2.1, J_ε is not singular for all $\varepsilon > 0$; so, the singular values are equal to the square root of the eigenvalues of $J_\varepsilon^* J_\varepsilon$. In the following, we prove the non-existence of a vanishing lower-bound for the eigenvalues of $J_\varepsilon^* J_\varepsilon$, which concludes the boundedness of J_ε^{-1}.

Lemma 4.3.9. *For J_ε as in* (4.20)*, there exists a constant C independent of ε, such that* $\lim_{\varepsilon\to 0}\|J_\varepsilon^{-1}\| \leq C$.

Proof. Here, we consider $\widetilde{\alpha} = 0$ to simplify the analysis; however, the analysis for $\widetilde{\alpha} \neq 0$ can be done similarly. Using (4.20), one can write $J_\varepsilon^* J_\varepsilon$ as

$$J_\varepsilon^* J_\varepsilon = \begin{bmatrix} \mathbb{I}_N + \beta^2 R_b^* R_b & \beta\left(\dfrac{Q}{\varepsilon^2} + R_b^*\right) \\ \beta\left(\dfrac{Q}{\varepsilon^2} + R_b^*\right)^* & \mathbb{I}_N + \dfrac{\beta^2}{\varepsilon^4}Q^*Q \end{bmatrix}.$$

Now, consider the vector $\boldsymbol{w} := (\boldsymbol{w}_1, \boldsymbol{w}_2)^T \in \mathbb{C}^{2N}$ living on the unit sphere, *i.e.*, $\|\boldsymbol{w}\|_{\ell_2} = 1$, where both $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are vectors of size N with complex entries. Then, by the definition of numerical range, one gets

$$W(J_\varepsilon^* J_\varepsilon) = \|\beta R_b\boldsymbol{w}_1 + \boldsymbol{w}_2\|_{\ell_2}^2 + \left\|\frac{\beta}{\varepsilon^2}Q\boldsymbol{w}_2 + \boldsymbol{w}_1\right\|_{\ell_2}^2. \tag{4.23}$$

From this, it is clear that if $\boldsymbol{w}_2 \notin \mathcal{N}_Q^{\varepsilon^2} := \{\boldsymbol{w}_2 | \, \|Q\boldsymbol{w}_2\| = \mathcal{O}(\varepsilon^2)\}$, then $\|\frac{\beta}{\varepsilon^2} Q\boldsymbol{w}_2 + \boldsymbol{w}_1\|_{\ell_2}$ goes far from zero when $\varepsilon \to 0$. Otherwise, $\boldsymbol{w}_2 \in \mathcal{N}_Q^{\varepsilon^2}$ and we conclude the result by contradiction, as follows. Assume that $W(J_\varepsilon^* J_\varepsilon)$ approaches zero in the limit; so, from (4.23)

$$\boldsymbol{w}_1 = -\frac{\beta}{\varepsilon^2} Q\boldsymbol{w}_2 + o(1), \tag{4.24a}$$

$$\boldsymbol{w}_2 = -\beta R_b \boldsymbol{w}_1 + o(1). \tag{4.24b}$$

Multiplying (4.24a) by ε^2 yields $\beta Q\boldsymbol{w}_2 = o(\varepsilon^2) - \varepsilon^2 \boldsymbol{w}_1$. For $\varepsilon \to 0$, both terms in the right-hand side have a limit (note that $\boldsymbol{w}_2$ is a bounded function); so, the limit of $\boldsymbol{w}_2$ should lie in the kernel of the central difference operator, *i.e.*, its leading order is *constant* (up to possible checker-board oscillations). That is to say that $\boldsymbol{w}_2 = \boldsymbol{w}_2^{(0)} + \varepsilon^2 \boldsymbol{w}_2^{(2)}$ where $\boldsymbol{w}_2^{(0)}$ consists, in general, of two constants for odd and even indices.

Now, putting $\boldsymbol{w}_1$ from (4.24a) into (4.24b) yields

$$\boldsymbol{w}_2 = \frac{\beta^2}{\varepsilon^2} R_b Q\boldsymbol{w}_2 + o(1),$$

which can be rewritten as $(\mathbb{I}_N - \frac{\beta^2}{\varepsilon^2} R_b Q)\boldsymbol{w}_2 = o(1)$. So, sending to the limit and balancing the leading order terms implies that $\boldsymbol{w}_2^{(0)} = \beta^2 R_b Q \boldsymbol{w}_2^{(2)}$, and it is shown in Appendix 4.A that, due to the periodicity and the structure of Q and R_b, the *constant* $\boldsymbol{w}_2^{(0)}$ can only be zero. So, $\boldsymbol{w}_2$ has a limit and $\boldsymbol{w}_2^{(0)} = \mathbf{0}$.

The equation (4.24b) implies that $R_b \boldsymbol{w}_1 \to \mathbf{0}$. Since the kernel of R_b consists of vectors with a checker-board like structure (see Appendix 4.A), $\boldsymbol{w}_1$ should tend to a *constant*. But from (4.24a), $\boldsymbol{w}_1$ has a difference structure, thus a vanishing mean. As discussed in Appendix 4.A, for a smooth bottom topography, the summation on odd and even indices indicates that, in the leading order, there is not CB structure and $\boldsymbol{w}_1 \to \mathbf{0}$. Hence, $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ is tending to zero, which contradicts the assumption that $\boldsymbol{w}$ lives on the unit sphere. This concludes the lemma. □

Assuming the ε-stability of the explicit step, Lemma 4.3.9 verifies that the scaled perturbation $\boldsymbol{V}_\Delta$ is $\mathcal{O}(1)$, which justifies the formal asymptotic consistency of the next section.

Remark 4.3.10. *So far, one important advantage of the RS-IMEX scheme based on* (4.6) *with a suitable scaling and reference solution has been to enrich Lemma 4.3.7 and Lemma 4.3.9 to conclude the ε-stability of the numerical solution since we directly work with perturbations. Otherwise, one needs to study the structure of J_ε^{-1}, e.g, to show that it extracts a constant part from the solution with some small fluctuations around it; this is generally more difficult.*

Remark 4.3.11. *Note that, by the Bolzano–Weiertrass theorem and a norm equivalence argument, the ε-stability of the solution implies that there exists a sequence $\{\boldsymbol{V}_{\Delta,\varepsilon_k}^{n+1}\}_{k\in\mathbb{N}}$ ($\varepsilon_k \to 0$ as $k \to \infty$) converging strongly to a limit (after extracting a subsequence if necessary). To determine whether this limit is the correct zero-Froude limit will be the topic of Section 4.3.2.3.*

4.3.2.3 Asymptotic consistency

For the RS-IMEX scheme applied to the 1d SWE, the asymptotic consistency requires the leading order of the surface perturbation and the momentum to be constant in space. As we have already proved solvability and ε-stability of the implicit solution operator, the (formal) asymptotic

consistency analysis we aim to present in this section is, in fact, rigorous because the coefficients of the asymptotic expansion are bounded in terms of ε, owing to the ε-stability.

Now, consider the discrete version of the asymptotic expansion (4.11), for all $i \in \Omega_N$ and the temporal step n:

$$z(t_n, x_i) = z_{(0)} + \varepsilon z_{(1)} + \varepsilon^2 z_{(2)}(t_n, x_i), \qquad m(t_n, x_i) = m_{(0)}(t_n) + \varepsilon m_{(1)}(t) + \varepsilon^2 m_{(2)}(t_n, x_i).$$

Since the reference state is the LaR equilibrium state with the scaling matrix $\mathrm{diag}(\varepsilon^2, 1)$, the scaled variables write

$$v_1(t_n, x) = z_{(2)}(t_n, x), \qquad v_2(t_n, x) = m_{(0)}(t_n) + \varepsilon m_{(1)}(t_n) + \varepsilon^2 m_{(2)}(t_n, x). \tag{4.25}$$

Note that (4.25) and ε-stability imply that the scheme provides a consistent discretisation for the leading order of the surface perturbation, simply by construction. So, it remains to determine if the leading order of the momentum is constant in space. Substituting (4.25) into the momentum update of the explicit step (4.16a) yields (with $k = 0, 1$)

$$v^{n+1/2}_{2(k)i} = v^n_{2(k)i} - \frac{\Delta t}{2\Delta x}\frac{v^{n,2}_{2(k)i}}{\overline{h}_{i+1}\overline{h}_{i-1}}(b_{i+1} - b_{i-1}) = v^n_{2(k)c} - \frac{\Delta t}{2\Delta x}\frac{v^{n,2}_{2(k)c}}{\overline{h}_{i+1}\overline{h}_{i-1}}(b_{i+1} - b_{i-1}),$$

where $v^n_{2(k)c}$ is a constant from (4.25). So, the explicit step for the momentum does not introduce an $\mathcal{O}(1/\varepsilon)$ term into the scheme, *i.e.*, $\|\boldsymbol{V}^{n+1/2}_\Delta\| = \mathcal{O}(1)$. Remark 4.3.8 implies that the boundedness of $\boldsymbol{V}^{n+1/2}_\Delta$ leads to the ε-stability of the implicit solution. Thus, from the implicit v_1 update (4.16b), one can (rigorously) conclude that for all $i \in \Omega_N$

$$v^{n+1}_{2(0)i+1} = v^{n+1}_{2(0)i-1}, \qquad v^{n+1}_{2(1)i+1} = v^{n+1}_{2(1)i-1}.$$

So, the updated momentum is *almost* constant, *i.e.*, the discrete divergence operator vanishes in the limit $\nabla_h v^{n+1}_{2,\Delta} = \mathcal{O}(\varepsilon^2)$. Although this is often interpreted as the asymptotic consistency in the literature, it does not imply necessarily that the limit would be obtained. For example, one can confirm that although the discretisation is consistent with the continuous *div*-free condition of the momentum, its null space allows for non-constant sequences, which may lead to the so-called *checker-board oscillations*. Here, we prove that the checker-board phenomenon for the flat bottom case, if happens, is as small as $\mathcal{O}(\varepsilon^2)$. Thus, it does not ruin the numerical solution in the limit. We will illustrate the smallness of checker-board oscillations for the non-flat bottom case by a numerical example, in Section 4.5.1.1.

Lemma 4.3.12. *For the RS-IMEX scheme* (4.16a)–(4.16b) *with a constant* $\widetilde{\alpha}$*, applied to the 1d SWE with flat bottom, the deviations of the computed momentum is* $\mathcal{O}(\varepsilon^2)$*, as* $\varepsilon \to 0$*. In other words, the possible checker-board oscillations for the computed momentum are at most* $\mathcal{O}(\varepsilon^2)$*.*

Proof. The linearity of the implicit step implies that for the differences of the solution $[\![v_{k,i}]\!] := v_{k,i} - v_{k,i-1}$ with $k = 1, 2$, the following holds:

$$J_\varepsilon \begin{bmatrix} [\![\boldsymbol{V}^{n+1}_{1,\Delta}]\!] \\ [\![\boldsymbol{V}^{n+1}_{2,\Delta}]\!] \end{bmatrix} = \begin{bmatrix} [\![\boldsymbol{V}^{n+1/2}_{1,\Delta}]\!] \\ [\![\boldsymbol{V}^{n+1/2}_{2,\Delta}]\!] \end{bmatrix}. \tag{4.26}$$

We will show that the blocks of $K_\varepsilon := J_\varepsilon^{-1}$ behave as

$$\|K_{11}\|, \|K_{12}\|, \|K_{22}\| = \mathcal{O}(1), \qquad \|K_{21}\| = \mathcal{O}(\varepsilon^2). \tag{4.27}$$

Then, owing to (4.25), $\|[\![\boldsymbol{V}_{1,\Delta}^{n+1/2}]\!]\| = \mathcal{O}(1)$ and $\|[\![\boldsymbol{V}_{2,\Delta}^{n+1/2}]\!]\| = \mathcal{O}(\varepsilon^2)$. Combining it with (4.26) yields

$$\left\|[\![\boldsymbol{V}_{2,\Delta}^{n+1}]\!]\right\| = \left\|K_{21}[\![\boldsymbol{V}_{1,\Delta}^{n+1/2}]\!] + K_{22}[\![\boldsymbol{V}_{2,\Delta}^{n+1/2}]\!]\right\| \leq C\left(\left\|[\![\boldsymbol{V}_{2,\Delta}^{n+1/2}]\!]\right\| + \varepsilon^2\left\|[\![\boldsymbol{V}_{1,\Delta}^{n+1/2}]\!]\right\|\right) = \mathcal{O}(\varepsilon^2),$$

which implies that the possible checker-board oscillations are $\mathcal{O}(\varepsilon^2)$.

It only remains to confirm the orders of magnitudes of the blocks in equation (4.27), and in particular, K_{21} and K_{22}. Let us re-write the inverse K_ε as (using Lemma 4.3.6)

$$K_\varepsilon = \begin{bmatrix} \left(P - \dfrac{\beta\overline{h}}{\varepsilon^2}QP^{-1}Q\right)^{-1} & -\dfrac{\beta^2}{\varepsilon^2}P^{-1}Q\left(P - \dfrac{\beta^2\overline{h}}{\varepsilon^2}QP^{-1}Q\right)^{-1} \\ -\beta\overline{h}P^{-1}Q\left(P - \dfrac{\beta^2\overline{h}}{\varepsilon^2}QP^{-1}Q\right)^{-1} & \left(P - \dfrac{\beta^2\overline{h}}{\varepsilon^2}QP^{-1}Q\right)^{-1} \end{bmatrix}. \tag{4.28}$$

Then, it is clear from Lemma 4.3.7 and the structure of K_ε, that $K_{12} = \frac{\beta}{\varepsilon^2\overline{h}}K_{21}$ and $\|K_{12}\| = \mathcal{O}(1)$; so, $\|K_{21}\| = \mathcal{O}(\varepsilon^2)$ and $\|K_{22}\| = \mathcal{O}(1)$, which concludes the proof of Lemma 4.3.12. □

Remark 4.3.13. *When the bottom is non-flat,* (4.26) *does not hold anymore since the momentum equation has contributions of non-constant coefficients terms. However, one can confirm that (assuming* $\widetilde{\alpha} = 0$ *for simplicity)*

$$H_\varepsilon \begin{bmatrix} [\![\boldsymbol{V}_{1,\Delta}^{n+1}]\!] \\ [\![\boldsymbol{V}_{2,\Delta}^{n+1}]\!] \end{bmatrix} = \begin{bmatrix} [\![\boldsymbol{V}_{1,\Delta}^{n+1/2}]\!] \\ [\![\boldsymbol{V}_{2,\Delta}^{n+1/2}]\!] \end{bmatrix}, \qquad H_\varepsilon =: \begin{bmatrix} \mathbb{I}_N & \dfrac{\beta}{\varepsilon^2}Q \\ \beta R_b^\Delta & \mathbb{I}_N \end{bmatrix}. \tag{4.29}$$

The only difference compared to (4.26) *is the block* R_b^Δ *whose non-zero entries of the* i*-th row read* $(\overline{h}_{i+1}, -(b_i - b_{i-1}), -\overline{h}_{i-2})$*, compared to those of* R_b *in* J_ε *which read* $(\overline{h}_{i+1}, (b_{i+1} - b_{i-1}), -\overline{h}_{i-1})$*.*

Assuming that H_ε^{-1} *is* ε*-stable (as it is very similar to* J_ε^{-1}*), one can conclude that* $\|[\![\boldsymbol{V}_{2,\Delta}^{n+1}]\!]\| = \mathcal{O}(1)$*. The matrix* R_b^Δ *is close to* $\overline{h}Q$ *(with* $\mathcal{O}(\Delta x)$ *difference); so, it is plausible to claim that since* $\|[\![\boldsymbol{V}_{2,\Delta}^{n+1/2}]\!]\| = \mathcal{O}(\Delta x)$ *there is an* $\mathcal{O}(\Delta x)$ *deviation from the result of the flat bottom, which gives* $\|[\![\boldsymbol{V}_{2,\Delta}^{n+1}]\!]\| = \mathcal{O}(\varepsilon^2) + \mathcal{O}(\Delta x)$*. Hence, for both cases, one can conclude that the momentum is close to a constant value in the limit.*

To conclude the asymptotic consistency, it is also required to show that the scheme provides a consistent discretisation of $\partial_t m_{(0)}$. Showing that, we consider the limit of the momentum update for each step (with a constant $\widehat{\alpha}$ and $\widetilde{\alpha}$):

$$\text{(Explicit step)} \quad \frac{v_{2(0),i}^{n+1/2} - v_{2(0),i}^{n}}{\Delta t} + \nabla_h\left[\frac{v_{2(0),i}^{2,n}}{\overline{h}_i + \varepsilon^2 v_{1(0),i}^{n}} + \frac{\varepsilon^2}{2}v_{(0)1,i}^{2,n}\right] - \frac{\widehat{\alpha}\Delta x}{2}\Delta_h v_{2(0),i}^{n} = 0. \tag{4.30a}$$

$$\text{(Implicit step)} \quad \frac{v_{2(0),i}^{n+1} - v_{2(0),i}^{n+1/2}}{\Delta t} + \nabla_h\left(\overline{h}_i v_{1(0),i}^{n+1}\right) - \frac{\widetilde{\alpha}\Delta x}{2}\Delta_h v_{2(0),i}^{n} = -v_{1(0),i}^{n+1}\nabla_h b_i. \tag{4.30b}$$

It is clear that (4.30a) and (4.30b) provide consistent discretisations of $\partial_t m_{(0)}$ for both explicit and implicit steps, (4.16a) and (4.16b). Thus, in the light of Lemma 4.3.9 and Lemma 4.3.12 for the ε-stability and smallness of checker-board modes, the scheme is AC.

4.3.2.4 Asymptotic stability

In this section, we discuss the rigorous stability analysis of the RS-IMEX scheme in the ℓ_2-norm, for a fixed grid and in finite time ($T_f < \infty$). Consider the scheme as the following iteration with a constant Δt

$$\boldsymbol{Y}^k = \prod_{i=0}^{s-1} \mathcal{E}_{s-i}\, \boldsymbol{Y}^{k-1}, \qquad k = 0, 1, \ldots, n-1, \quad n = T_f/\Delta t, \tag{4.31}$$

where $\mathcal{E}_i$ for $i = 1, \ldots, s$ are some discrete evolution operators, like explicit and implicit operators for the RS-IMEX scheme.

Motivated by [HJL12, Lemma 3.1] (see also [RM67, Tre96]), one can show that the scheme (4.31) is stable for a finite time and in the ℓ_p-norm, provided that there exist constants c_i independent of Δt such that

$$\|\mathcal{E}_i\|_{\ell_p} \leq 1 + c_i \Delta t, \qquad i = 1, \ldots, s. \tag{4.32}$$

That is to say that $\|\boldsymbol{Y}^n\|_{\ell_p} \leq e^{CT_f} \|\boldsymbol{Y}^0\|_{\ell_p}$ and with the constant C independent of Δt.

Concerning the RS-IMEX scheme, $s = 2$ and $\mathcal{E}_1$ and $\mathcal{E}_2$ denote the explicit and implicit operators, respectively. At first, we consider the implicit step and show that the condition (4.32) holds. Since the explicit step is non-linear, obtaining (4.32) directly is not feasible. Instead, we find a weaker estimate using a discrete Grönwall's inequality. Combining these two results proves the stability.

Stability of the implicit step $\mathcal{E}_2$ As we have mentioned earlier, the implicit operator is J_ε^{-1}. So, one should find some bound of the form $1 + c_2 \Delta t$ for $\|J_\varepsilon^{-1}\|$. Let us assume the norm to be ℓ_2. So, one can write

$$\|\mathcal{E}_2\|_{\ell_2} = \|J_\varepsilon^{-1}\|_{\ell_2} = \frac{1}{\sigma_{\min}(J_\varepsilon)} = \frac{1}{\underline{\omega}^{1/2}(J_\varepsilon^* J_\varepsilon)},$$

where $\underline{\omega}(J_\varepsilon^* J_\varepsilon) := \min |W(J_\varepsilon^* J_\varepsilon)|$. On the other hand, one can conclude from (4.23) that the lower bound of the numerical range $\underline{\omega}(J_\varepsilon^* J_\varepsilon)$ can be written as $1 - \beta c_2''$ with some positive ε-uniform constant c_2''. Defining another constant c_2' such that $\underline{\omega}^{1/2}(J_\varepsilon^* J_\varepsilon) \geq 1 - \beta c_2'$ gives

$$\|\mathcal{E}_2\|_{\ell_2} \leq \frac{1}{1 - \beta c_2'} = \sum_{k=0}^{\infty} (\beta c_2')^k \leq 1 + \beta c_2,$$

due to the Taylor expansion around $\beta = 0$ and with another positive ε-uniform constant c_2. Thus, redefining c_2, $\|\mathcal{E}_2\|_{\ell_2} \leq 1 + c_2 \Delta t$, and the implicit operator is asymptotically stable (in finite time and for a fixed grid).

Stability of the explicit step $\mathcal{E}_1$ To prove the stability of the explicit step is more delicate since it is not linear; consequently c_1 can be obtained but it depends on the solution $\boldsymbol{Y}^k =$

$\left[\boldsymbol{V}^k_{1,\Delta}, \boldsymbol{V}^k_{2,\Delta}\right]^T$. Assuming $\widehat{\alpha} = 0$ for simplicity and from (4.16a), one can bound $\|\mathcal{E}_1 \boldsymbol{Y}^k\|_{\ell_2}$ as

$$\begin{aligned}\|\mathcal{E}_1 \boldsymbol{Y}^k\|_{\ell_2} &\leq \|\boldsymbol{Y}^k\|_{\ell_2} + \frac{2\beta}{h^k_{\min}} \|\langle \boldsymbol{V}^k_{2,\Delta}, \boldsymbol{V}^k_{2,\Delta}\rangle\|_{\ell_2} + \varepsilon^2 \beta \|\langle \boldsymbol{V}^k_{1,\Delta}, \boldsymbol{V}^k_{1,\Delta}\rangle\|_{\ell_2} \\ &\leq \|\boldsymbol{Y}^k\|_{\ell_2} + \beta \left(\frac{2}{h^k_{\min}} + \varepsilon^2\right) \|\boldsymbol{Y}^k\|^2_{\ell_4}, \\ &\leq \left[1 + \beta \left(\frac{2}{h^k_{\min}} + \varepsilon^2\right) \|\boldsymbol{Y}^k\|_{\ell_2}\right] \|\boldsymbol{Y}^k\|_{\ell_2} \end{aligned} \tag{4.33}$$

since for sequence spaces, $\|\boldsymbol{Y}\|_{\ell_q} \leq \|\boldsymbol{Y}\|_{\ell_p}$ for $1 \leq p \leq q$. Here, $h^k_{\min}$ is the lower-bound for the water height at step k, *i.e.*, $h^k_{\min} := \min_{i \in \Omega_N} \left|\overline{z} + \varepsilon^2 v^k_{1,i} - b_i\right|$. Owing to ε-stability, $h^k_{\min}$ is bounded away from zero for a small enough ε. For larger ε, one should add the positivity assumption to conclude the result.

For simplicity, one can rewrite (4.33) as

$$y_{k+1} \leq y_k + \beta_k y_k^2, \qquad y_k := \|\boldsymbol{Y}^k\|_{\ell_2}, \quad \beta_k := \beta \left(\frac{2}{h^k_{\min}} + \varepsilon^2\right). \tag{4.34}$$

The stability of the explicit step means to find an upper-bound for y_k for which we use the following discrete Grönwall's inequality from [WW65].

Theorem 4.3.14 (Thm. 4 [WW65])**.** *Consider the sequences $\{\mu_k\}_{k>0}, \{\nu_k\}_{k>0} \geq 0$ for $k = 0, 1, \ldots$ while $\mu_0 = \nu_0 = 0$. If for the non-negative sequence $\{y_k\}_{k=0,1,\ldots}$ it holds that*

$$y_{k+1} \leq \sigma + \sum_{i=0}^{k} \nu_i y_i + \sum_{i=0}^{k} \mu_i y_i^p, \qquad (\sigma > 0, p \geq 0, p \neq 1),$$

then, by denoting $q := 1 - p$ and $\psi(k) := \prod_{i=0}^{k} (1 + \nu_i)^{-1}$ for $k = 0, 1, \ldots$, the sequence $\{y_k\}_{k \geq 0}$ is bounded as

$$y_{k+1} \leq \frac{1}{\psi(k)} \left(\sigma^q + q \sum_{i=0}^{k} \mu_k \psi^q(i)\right)^{1/q}, \qquad k = 0, 1, \ldots . \tag{4.35}$$

Using Theorem 4.3.14, the following corollary can be obtained.

Corollary 4.3.15. *Given a small enough initial datum, the sequence $\{y_k\}_{k=0,1,\ldots}$ defined in (4.34) is bounded uniformly in ε.*

Proof. One can rewrite (4.34) as

$$y_{k+1} \leq y_0 + \sum_{i=0}^{k} \beta_i y_i^2 = \left(y_0 + \beta_0 y_0^2\right) + \sum_{i=1}^{k} \beta_i y_i^2. \tag{4.36}$$

Comparing to (4.35), we set $\nu_i = 0$, $\mu_0 = 0$, $\mu_{i>0} = \beta_{i>0}$, $p = 1$ and $\sigma = y_0 + \beta_0 y_0^2$. So, $\psi(i) = 1$, and

$$y_{k+1} \leq \left(1/\sigma - \sum_{i=1}^{k} \beta_i\right)^{-1} = \frac{(1 + \beta_0 y_0)\, y_0}{1 - (1 + \beta_0 y_0)\, y_0 \sum_{i=1}^{k} \beta_i}.$$

For y_{k+1} to be bounded, the denominator should be bounded away from zero, which imposes a bound for the norm of the initial condition y_0, *i.e.*,

$$(1 + \beta_0 y_0)\, y_0 \sum_{i=1}^{k} \beta_i < 1. \tag{4.37}$$

So, the norm of solution of the explicit step y_k is bounded under a *"smallness assumption"*. □

Remark 4.3.16. *Note that a simpler version of such bounds can be obtained by induction like [NT92, eq. (3.11)].*

Combining the bounds for explicit and implicit steps, one can bound the norm of the updated solution as

$$y_{k+1} \leq (1 + c_2\Delta t)(1 + c_1\Delta t y_k) y_k, \tag{4.38}$$

where it is assumed that c_1 and c_2 do not change with k, for simplicity. After some straightforward calculations, one gets

$$y_{k+1} \leq (1 + c_2\Delta t)^{k+1} y_0 + c_1\Delta t(1 + c_2\Delta t)^k y_0^2 + \sum_{i=1}^{k} c_1\Delta t(1 + c_2\Delta t)^{k-i} y_i^2.$$

Thus, by picking $\sigma = (1+c_2\Delta t)^{k+1} y_0 + c_1\Delta t(1+c_2\Delta t)^k y_0^2$ and $\nu_i = c_1\Delta t(1+c_2\Delta t)^{k-i}$, Theorem 4.3.14 yields the following stability result for the RS-IMEX scheme.

Theorem 4.3.17. *Given a small enough initial datum and for $\varepsilon \ll 1$, the RS-IMEX scheme* (4.16a)–(4.16b) *is ℓ_2-bounded uniformly in ε and for a finite time.*

Remark 4.3.18. *One can read the smallness condition* (4.37) *as a time step restriction. This condition is restrictive, not in ε, but in terms on the number of grid points. One may circumvent this issue by obtaining some non-linear energy estimates, e.g., as in [Gie15]; this is in the course of investigation.*

Remark 4.3.19. *As we have seen so far, the scheme is AC and AS. Due to Definition 1.2.1, for the scheme to be AP, asymptotic efficiency is also necessary: The CFL condition is ε-uniform (with material velocity), but the condition number of J_ε increases as $\varepsilon \to 0$ (see Remark 4.4.2). Although, literally speaking, the scheme is not AP in the sense of Definition 1.2.1, we call it AP (at least in a weaker sense) since it is AC and AS under a non-restrictive CFL condition.*

4.3.2.5 Well-balancing

To have the LaR equilibrium state at step n, (2.43) implies that $m_i^n = 0$ and z_i^n is constant for all $i \in \Omega_N$. The reference solution is at equilibrium, so is its perturbation, *i.e.*, v_1^n is constant and v_2^n is zero, which implies that $\widehat{\boldsymbol{G}}$ is also constant; so, $\boldsymbol{V}_\Delta^{n+1/2} = \boldsymbol{V}_\Delta^n$. Note that the well-balancing of the explicit step is in fact the consistency of the numerical flux (due to lack of non-stiff source term).

For the implicit step, the central discretisation suffices the compatibility of the equilibrium solution as there is exactly such a term in the difference of Rusanov fluxes. To show this

compatibility, we assume that $v_{2,i}^{n+1}$ is zero and $v_{1,i}^{n+1}$ is constant, which makes the contributions of numerical diffusion to vanish. This compatibility, combined with the unique solvability, suggests that the solution remains stationary, *i.e.*, $\boldsymbol{V}_\Delta^{n+1} = \boldsymbol{V}_\Delta^n$; thus, the scheme is well-balanced. For a more rigorous proof, we write the implicit step as (assuming $\widetilde{\alpha} = 0$ for simplicity)

$$\boldsymbol{V}_{1,\Delta}^{n+1} + \frac{\beta}{\varepsilon^2}\boldsymbol{V}_{2,\Delta}^{n+1} = c\,\mathbf{1}_N, \tag{4.39a}$$

$$\beta R_b \boldsymbol{V}_{1,\Delta}^{n+1} + \boldsymbol{V}_{2,\Delta}^{n+1} = \mathbf{0}_N, \tag{4.39b}$$

with some constant c denoting the constant value for the surface perturbation at $n + 1/2$.

Putting (4.39a) into (4.39b) gives

$$c\beta R_b \mathbf{1}_N + \left(\mathbb{I}_N - \frac{\beta^2}{\varepsilon^2} R_b Q\right) \boldsymbol{V}_{2,\Delta}^{n+1} = \mathbf{0},$$

which implies that $\boldsymbol{V}_{2,\Delta}^{n+1} = \mathbf{0}$ since $\mathbf{1}_N \in \mathcal{N}_{R_b}$ and $(\mathbb{I}_N - \frac{\beta^2}{\varepsilon^2} R_b Q)$ is non-singular (from Section 4.3.2.1). Also, (4.39a) concludes that $\boldsymbol{V}_{1,\Delta}^{n+1}$ is constant, and completes the well-balancing proof.

Remark 4.3.20. *It is important to note that, generally speaking, having the solution at equilibrium does not necessarily imply that the reference solution or its perturbation is at equilibrium. This 1d case with the LaR reference solution is exceptional since the reference solution is constant and stationary.*

4.4 Shallow water equations with the zero-Froude limit reference solution

Here, we consider the SWE as in (4.8) in a periodic domain, and with a flat bottom topography, while the reference solution $\overline{\boldsymbol{U}} = (\overline{z}, \overline{m})^T$ is chosen as the zero-Froude limit solution of (4.8). It can be obtained from Definition 4.3.1 and equation (4.12) that $\overline{z} = z_{(0)}^0$ and $\overline{m} = m_{(0)}^0$, both constant. We have assumed the bottom topography to be flat in order to make the zero-Froude limit stationary (owing to periodic boundary conditions); this makes $\overline{\boldsymbol{T}}$ to vanish and avoids the difficulties stemming from its discretisation in the asymptotic analysis (as will be discussed in Chapter 5). With this reference solution, the splitting can be obtained as

$$\overline{\boldsymbol{F}} := \begin{bmatrix} \overline{m} \\ \dfrac{\overline{m}^2}{\overline{z}-b} + \dfrac{\overline{z}^2 - 2\overline{z}b}{2\varepsilon^2} \end{bmatrix}, \quad \widetilde{\boldsymbol{F}} := \begin{bmatrix} m_{pert} \\ -\dfrac{\overline{m}^2 z_{pert}}{(\overline{z}-b)^2} + \dfrac{\overline{z}-b}{\varepsilon^2} z_{pert} + \dfrac{2\overline{m} m_{pert}}{\overline{z}-b} \end{bmatrix},$$

$$\widehat{\boldsymbol{F}} := \begin{bmatrix} 0 \\ \dfrac{(\overline{m}+m_{pert})^2}{\overline{z}+z_{pert}-b} + \dfrac{z_{pert}^2}{2\varepsilon^2} - \dfrac{\overline{m}^2}{\overline{z}-b} + \dfrac{\overline{m}^2 z_{pert}}{(\overline{z}-b)^2} - \dfrac{2\overline{m} m_{pert}}{\overline{z}-b} \end{bmatrix}.$$

One can check that the splitting is admissible in the sense of Definition 3.1.1; the eigenvalues of $\widetilde{\boldsymbol{F}}'$ are stiff and those of $\widehat{\boldsymbol{F}}'$ are non-stiff:

$$\widetilde{\lambda} = \frac{\overline{m}}{\overline{z}-b} \pm \frac{\sqrt{\overline{z}-b}}{\varepsilon}, \qquad \widehat{\lambda} = 0, 2u_{pert}. \tag{4.40}$$

The asymptotic analysis presented in Appendix 2.A suggests the scaling matrix should be $D := \mathrm{diag}(\varepsilon^2, \varepsilon)$; so, the scaled RS-IMEX splitting reads

$$\begin{aligned}\widetilde{\boldsymbol{G}} &:= \begin{bmatrix} & v_2/\varepsilon \\ -\dfrac{\overline{m}^2 v_1 \varepsilon}{(\overline{z}-b)^2} + \dfrac{(\overline{z}-b)v_1}{\varepsilon} + \dfrac{2\overline{m}v_2}{\overline{z}-b} & \end{bmatrix}, \\ \widehat{\boldsymbol{G}} &:= \begin{bmatrix} 0 & \\ \dfrac{(\overline{m}+\varepsilon v_2)^2}{\varepsilon\,(\overline{z}+\varepsilon^2 v_1 - b)} + \dfrac{\varepsilon v_1^2}{2} - \dfrac{\overline{m}^2}{\varepsilon\,(\overline{z}-b)} + \dfrac{\overline{m}^2 v_1 \varepsilon}{(\overline{z}-b)^2} - \dfrac{2\overline{m}v_2}{\overline{z}-b} & \end{bmatrix}.\end{aligned} \tag{4.41}$$

Due to the well-prepared initial velocity (4.12), the zero-Froude limit reference state makes the wave speeds of the slow system really small, *i.e.*, $\mathcal{O}(\varepsilon)$, as $u_{pert} = \mathcal{O}(\varepsilon)$. Also, $\overline{\boldsymbol{U}}$ is asymptotically close to the solution. Thus, the commutator would be $\mathcal{O}(1)$, *i.e.*,

$$\lim_{\varepsilon\to 0}\left[\widetilde{\boldsymbol{G}}', \widehat{\boldsymbol{G}}'\right] = \begin{bmatrix} v_1 & \dfrac{2v_2}{\overline{z}-b} \\ -2v_2 & -v_1 \end{bmatrix}. \tag{4.42}$$

Similar to the case of the LaR reference solution, the modified equation is stable for this splitting. Also, one can again confirm that the eigenvector matrices R and $\widetilde{R}$ are very similar in structure, which supports the stability of the modified equations. Using (4.41), one obtains

$$\lim_{\varepsilon\to 0}(r_{1,2}) = \lim_{\varepsilon\to 0}(\widetilde{r}_{1,2}) = \begin{bmatrix} \pm 1 \\ \sqrt{\overline{h}} \\ 1 \end{bmatrix} \qquad \Longrightarrow \qquad \lim_{\varepsilon\to 0}\|Q_{R\to\widetilde{R}} - \mathbb{I}_2\| = 0.$$

For this case, the RS-IMEX scheme is defined as in (4.15a)–(4.15b) when $\widehat{\boldsymbol{G}}$ and $\widetilde{\boldsymbol{G}}$ change according to (4.41).

4.4.1 Asymptotic analysis of the scheme

We collect the properties of the RS-IMEX scheme in the following theorem.

Theorem 4.4.1. *For the shallow water equations with a flat bottom and well-prepared initial data in the sense of Definition 4.3.2, the RS-IMEX scheme* (4.15a)–(4.15b)*, with* (4.41) *and a constant* $\widetilde{\alpha}$*,*

(i) is solvable, i.e., it has a unique solution for all $\varepsilon > 0$*.*

(ii) its solution is ε*-stable, i.e., it is bounded for* $\varepsilon \ll 1$*. So, there is convergent subsequence of the discrete solutions as* $\varepsilon \to 0$*.*

(iii) is consistent with the asymptotic limit in the fully-discrete settings, i.e., it is asymptotically consistent.

(iv) is asymptotically ℓ_2*-stable for the fixed grid* Δx*, in finite time* $T_f < \infty$ *and with a small enough initial data, i.e., there exists a constant* C_{N,T_f} *such that* $\|\boldsymbol{V}_\Delta^n\|_{\ell_2} \le C_{N,T_f}\|\boldsymbol{V}_\Delta^0\|_{\ell_2}$*.*

We present the proof of Theorem 4.4.1 in the next sections.

4.4.1.1 Solvability of the scheme

Like Section 4.3.2, it is not difficult to see that J_ε reads

$$J_\varepsilon := \begin{bmatrix} P & \frac{\beta}{\varepsilon}Q \\ \left(-\frac{\overline{m}^2\varepsilon}{\overline{h}^2} + \frac{\overline{h}}{\varepsilon}\right)\beta Q & P + \frac{2\beta\overline{m}}{\overline{h}}Q \end{bmatrix}. \tag{4.43}$$

The blocks of J_ε commute and from [Sil00, Thm. 1] the determinant of J_ε can be computed as

$$\det(J_\varepsilon) = \det\left(\underbrace{P^2 - \frac{\beta^2}{\varepsilon}\left(-\frac{\overline{m}^2\varepsilon}{\overline{h}^2} + \frac{\overline{h}}{\varepsilon}\right)Q^2}_{=:\mathfrak{A}} + \underbrace{\frac{2\beta\overline{m}}{\overline{h}}PQ}_{=:\mathfrak{B}}\right).$$

One can confirm that PQ, so $\mathfrak{B}$, is skew-symmetric and does not change the bounds for the real eigenvalues of the symmetric part $\mathfrak{A}$, owing to the Bendixon's theorem [Ben02, Hir02]. Thus, it remains to show that $\mathfrak{A}$ has only non-zero eigenvalues. Note that the eigenvalues of $P^2 + \frac{\beta^2\overline{m}^2}{\overline{h}^2}Q^2$ can be set positive, by a suitable and ε-uniform choice of β. Using the sub-additivity of the numerical range (spectrum for symmetric matrices), adding $-\frac{\beta^2\overline{h}}{\varepsilon^2}Q^2$ with non-negative eigenvalues makes J_ε non-singular.

4.4.1.2 ε-stability of the solution

Similar to Section 4.3.2, we can find Ξ_ε as

$$\Xi_\varepsilon := \begin{bmatrix} \Lambda_P & \frac{\beta}{\varepsilon}\Lambda_Q \\ \left(-\frac{\overline{m}^2\varepsilon}{\overline{h}} + \frac{\overline{h}}{\varepsilon}\right)\beta\Lambda_Q & \Lambda_P + \frac{2\beta\overline{m}}{\overline{h}}\Lambda_Q \end{bmatrix}.$$

We then can show that Ξ_ε^{-1} has a bounded norm in terms of ε. Due to Lemma 4.3.6, the blocks of Ξ_ε^{-1} read

$$\begin{aligned}
\Xi_{11}^{-1} &= \left(\Lambda_P - \frac{\beta^2}{\varepsilon}\left(-\frac{\overline{m}^2\varepsilon}{\overline{h}} + \frac{\overline{h}}{\varepsilon}\right)\Lambda_Q^2\left(\Lambda_P + \frac{2\beta\overline{m}}{\overline{h}}\Lambda_Q\right)^{-1}\right)^{-1}, \\
\Xi_{12}^{-1} &= -\frac{\beta}{\varepsilon}\Lambda_P^{-1}\Lambda_Q\Xi_{22}^{-1}, \\
\Xi_{21}^{-1} &= -\left(\Lambda_P + \frac{2\beta\overline{m}}{\overline{h}}\Lambda_Q\right)^{-1}\left(-\frac{\overline{m}^2\varepsilon}{\overline{h}^2} + \frac{\overline{h}}{\varepsilon}\right)\beta\Lambda_Q\Xi_{11}^{-1}, \\
\Xi_{22}^{-1} &= \left(\Lambda_P + \frac{2\beta\overline{m}}{\overline{z}-b}\Lambda_Q - \frac{\beta^2}{\varepsilon}\left(-\frac{\overline{m}^2\varepsilon}{\overline{h}^2} + \frac{\overline{h}}{\varepsilon}\right)\Lambda_P^{-1}\Lambda_Q^2\right)^{-1},
\end{aligned}$$

which are all bounded; so, Ξ_ε^{-1} is ε-stable. Assuming the ε-stability of the explicit step (see Section 4.4.1.3), the solution of the implicit step (thus the whole scheme) can be shown to be ε-stable. The ε-stability of the solution implies that the scaled perturbation $\boldsymbol{V}_\Delta$ is $\mathcal{O}(1)$, which justifies the asymptotic consistency analysis we are going to present in the next section.

Remark 4.4.2. *The condition number of J_ε depends on the scaling matrix. For example, one can confirm that using* $\mathrm{diag}(\varepsilon^2, 1)$ *and* $\mathrm{diag}(\varepsilon^2, \varepsilon)$ *makes the condition number to be* $\mathcal{O}(1/\varepsilon^2)$ *and* $\mathcal{O}(1/\varepsilon)$*, respectively. In this sense, the scaling by the diagonal matrix D is the "equilibration of matrices" [GVL12, Sect. 3.5.2] in essence, and may improve the condition number of J_ε; see Table 4.1.*

Table 4.1: Comparison of different scaling for matrix J_ε.

	Scaling by $\mathrm{diag}(\varepsilon^2, 1)$	Scaling by $\mathrm{diag}(\varepsilon^2, \varepsilon)$
J_ε	$\begin{bmatrix} 1 & \mathcal{O}(1/\varepsilon^2) \\ 1 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & \mathcal{O}(1/\varepsilon) \\ \mathcal{O}(1/\varepsilon) & 1 \end{bmatrix}$

4.4.1.3 Asymptotic consistency

We are going to show the asymptotic consistency of the scheme formally. But, as we mentioned before, the analysis is, in fact, rigorous owing to the ε-stability results.

For the explicit step and similar to the case with the LaR reference solution, no $\mathcal{O}(1/\varepsilon)$ contribution is associated with the explicit update since

$$\lim_{\varepsilon \to 0} \left[\frac{(\overline{m} + \varepsilon v_2^n)^2}{\varepsilon\,(\overline{z} + \varepsilon^2 v_1^n - b)} - \frac{\overline{m}^2}{\varepsilon\,(\overline{z} - b)} \right] = \mathcal{O}(1). \tag{4.44}$$

So, it is asymptotically consistent (and ε-stable). This implies that for the implicit step, as shown in the previous section, $\boldsymbol{V}_\Delta^{n+1} = \mathcal{O}(1)$. Balancing $\mathcal{O}(1/\varepsilon)$ terms for the implicit v_1-update implies that $\nabla_h v_{2,\Delta}^{n+1} = \mathcal{O}(\varepsilon)$. These conclude the asymptotic consistency of the scheme.

Remark 4.4.3. *The asymptotic stability analysis for the implicit step is very similar to Section 4.3.2.4. We just wish to stress that for the explicit step, one should use* (4.44) *to find an ε-uniform bound. Hence, one can conclude that the scheme is again AP (in a weaker sense than Definition 1.2.1), i.e., it is AC and AS under a non-restrictive CFL condition while the condition number of J_ε increases as $\varepsilon \to 0$.*

Remark 4.4.4. *Comparing the results of this section to Section 4.3.2, both schemes are AC and AS. As we pointed out in Remark 4.2.2, the modified equation analysis in Chapter 3 suggests that the reference solution does not affect stability of the scheme. However, asymptotically smaller wave speeds for the zero-Froude case (compare* (4.40) *with* (4.10)*) indicate that the choice of reference solution affects the numerical diffusion, so the accuracy. We will illustrate this point in Section 4.5.1.1 for a numerical example.*

4.5 Numerical experiments

In this section, we show that the solutions computed by the RS-IMEX scheme have good quality, comparable to existing schemes. Also, we confirm the AP property (asymptotic consistency and

asymptotic stability) of the scheme, numerically. At first, we consider the flat bottom case in two examples. Then, we continue with a non-flat bottom example.

For all the examples discussed in this section, we put $\widehat{\alpha}$ like in the Lax–Friedrichs scheme as the maximum value of all wave speeds over the whole domain, and $\widetilde{\alpha}$ is likewise but computed for $\varepsilon = 1$ avoiding excessive diffusion. Also, we choose $c_{\widetilde{\alpha}} = c_{\widehat{\alpha}} = 1$. The time step has been computed as $\Delta t := \min(\Delta t_{\text{CFL}}, \Delta t_{\text{Aux}})$ where the CFL time step Δt_{CFL} and the auxiliary time step Δt_{Aux} are defined as

$$\Delta t_{\text{CFL}} := \text{CFL}\,\Delta x / \max_{j\in\Omega_N} \widehat{\alpha}_j, \qquad \Delta t_{\text{Aux}} := \text{CFL}\,\Delta x / \max_{j\in\Omega_N} \widetilde{\alpha}_j|_{\varepsilon=1}.$$

Note that this (non-restrictive) auxiliary time step is only needed to avoid issues when the velocity field is zero.

4.5.1 Shallow water equations with a flat bottom

In this section, we discuss numerical results for the case of SWE with a flat bottom topography. Firstly, we consider a colliding pulses example of [DT11], which has also been discussed in [Bis15]. Then, we discuss another colliding pulses example from [AN12].

4.5.1.1 (i) Colliding pulses

As [DT11, Example 6.1], we consider the following well-prepared initial data in the periodic domain $[0, 1)$:

$$h(0,x) = \mathbf{1}_{[0\le x\le 0.2]\bigcup[0.3<x\le 0.7]\bigcup[0.8<x\le 1]} + (1+\varepsilon^2)\mathbf{1}_{[0.2<x\le 0.3]} + (1-\varepsilon^2)\mathbf{1}_{[0.7<x\le 0.8]},$$
$$m(0,x) = \Big(1-\frac{\varepsilon^2}{2}\Big)\mathbf{1}_{[0\le x\le 0.2]\bigcup[0.8<x\le 1]} + \mathbf{1}_{[0.2<x\le 0.3]\bigcup[0.7<x\le 0.8]} + \Big(1+\frac{\varepsilon^2}{2}\Big)\mathbf{1}_{[0.3<x\le 0.7]},$$

where $\mathbf{1}_\omega$ is the characteristic function in the domain ω, and $H_{\text{ref}} = 1$; so, $\overline{z} = 0$. We also set the final time $T_f = 0.05$ and CFL $= 0.45$. In [DT11, Example 6.1] the pressure function $p(\varrho) = \varrho^2$ has been used; so, we compare the results of the RS-IMEX scheme with [Bis15, Sect. 8.1], where the pressure function is the same as the SWE.

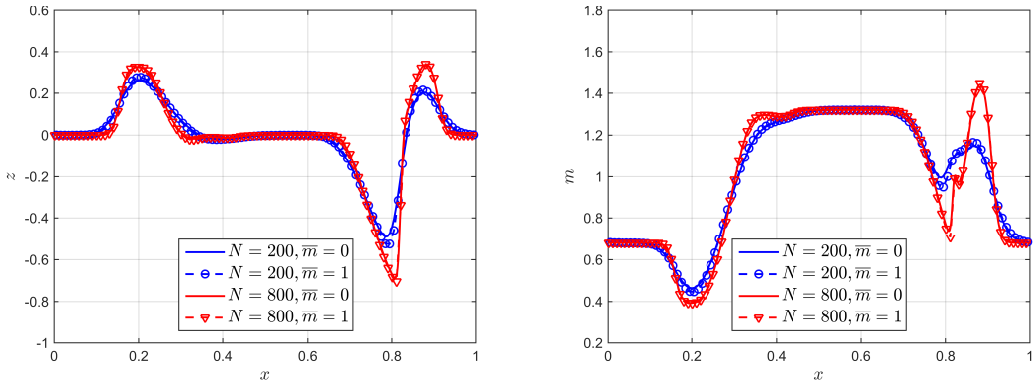
Figures 4.2 and 4.3 show the results of the RS-IMEX scheme with $\overline{m} = 0$ (LaR) and $\overline{m} = 1$ (zero-Froude limit) for $\varepsilon = 0.8$ and $\varepsilon = 0.1$. Compared to [Bis15, Fig. 8.2], it is clear that the computed solutions are well-qualified. Note that for this example, the schemes in [Bis15, Fig. 8.2] uses the same splitting as the RS-IMEX; but, they employ an elliptic approach for the surface perturbation update; see [Bis15] for more details. As Figure 4.2 and Figure 4.3 suggest, the computed surface perturbation z does not change that much with the reference momentum, particularly for $\varepsilon = 0.1$. For the momentum, the $\overline{m} = 1$ case gives slightly more accurate solutions in terms of capturing the extrema. This is due to $\mathcal{O}(\varepsilon^2)$ wave speeds of the non-stiff system (compare (4.40) with (4.10)) which leads to smaller numerical diffusion; this can be clearly seen in Figure 4.3 where the solution is computed on a very fine mesh with $N = 6400$. Note that for $\varepsilon = 0.1$, both schemes cannot capture the details of the waves (micro-structures), which is also the case in [DT11, Bis15].

Figure 4.4 illustrates the experimental order of convergence (EOC) for different ε and $\overline{m} \in \{0, 1\}$, for a normalised version of the error defined as

$$e(\phi_\Delta^{\text{num}}) := \|\phi_\Delta^{\text{num}} - \phi_\Delta^{\text{ref}}\|_{L_1(\Omega_{N_{\text{ref}}})} = \frac{1}{N_{\text{ref}}} \sum_{j \in \Omega_{N_{\text{ref}}}} \left|\phi_j^{\text{num}} - \phi_j^{\text{ref}}\right|. \tag{4.45}$$

where ϕ is the variable of interest (momentum, height, etc.), and ϕ_Δ^{num} and ϕ_Δ^{ref} are respectively the computed solution and the "*reference*" solution computed on a finer mesh with $N = 3200$. The figures shows that the scheme, regardless of the reference solution, has an almost uniform order of convergence for $\varepsilon' in \{0.8, 0.1, 0.05\}$, which coincides with the result of [DT11, Tab. 2]. Both for the surface perturbation and the momentum, the error is normalised by $1/\varepsilon^2$ as the initial data consist of $\mathcal{O}(\varepsilon^2)$ perturbations around zero surface perturbation and around a constant value for the momentum.

Verifying asymptotic consistency and stability, Figure 4.5 shows the solution for a small ε, namely $\varepsilon = 10^{-8}$. It confirms that the solution is close to the limit manifold. That is to say, the surface elevation is almost constant, and the momentum is *div*-free. It also confirms the smallness of the checker-board oscillations. Note that for both cases here (and for all other examples in this chapter), the scheme uses $D = \text{diag}(\varepsilon^2, \varepsilon)$, which makes the condition number of J_ε to be $\mathcal{O}(1/\varepsilon)$ (see Remark 4.4.2). Note also that for the zero-Froude limit reference state, due to $\mathcal{O}(\varepsilon)$ eigenvalues for the non-stiff system as in (4.40), $\Delta t_{\text{CFL}} = \mathcal{O}(1/\varepsilon)$; so, it gets larger as ε decreases. For this example, since there are only $\mathcal{O}(\varepsilon^2)$ deviations of the initial momentum from $\overline{m}$, one expects $\Delta t_{\text{CFL}} = \mathcal{O}(1/\varepsilon^2)$.



(a) Surface perturbation z.

(b) Momentum m.

Figure 4.2: The RS-IMEX solutions for Example (i), with $\varepsilon = 0.8$, CFL $= 0.45$, $T_f = 0.05$, and with two reference states: the LaR and the zero-Froude limit.

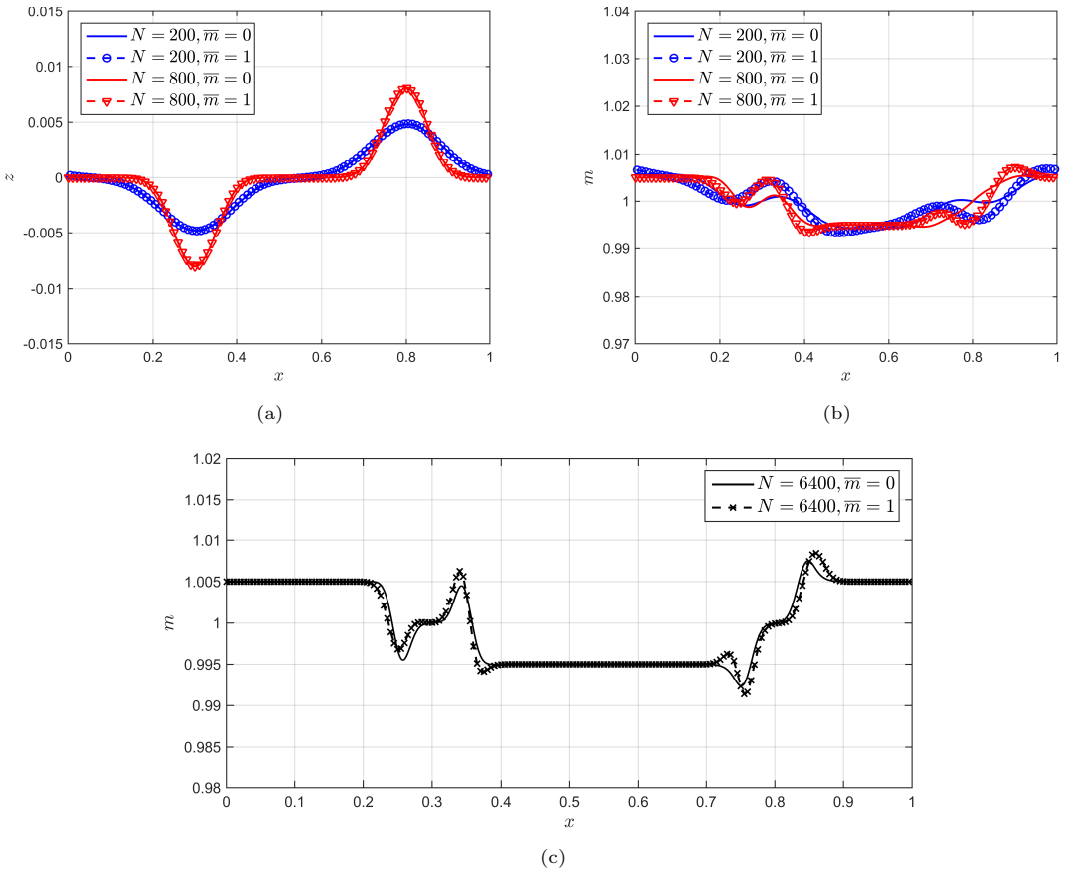


Figure 4.3: (a) and (b): The RS-IMEX solutions for Example (i), with $\varepsilon = 0.1$, CFL $= 0.45$, $T_f = 0.05$, and with two reference states: the LaR and the zero-Froude limit. (c) is like (b) but for a very fine mesh.

4.5.1.2 (ii) Colliding pulses

Consider the following ill-prepared initial data in the periodic domain $[-1, 1)$, as in [AN12] (motivated by [Kle95]):

$$\begin{aligned} h(0,x) &= 0.955 + \frac{\varepsilon}{2}\left(1 - \cos(2\pi x)\right), \\ u(0,x) &= -\text{sign}(x)\sqrt{2}\left(1 - \cos(2\pi x)\right). \end{aligned} \tag{4.46}$$

Figure 4.6 shows the evolution of the water height for the final time $T_f = 0.1$ and $\varepsilon = 0.1$ with $N = 200$, CFL $= 0.45$ and the LaR reference solution. We have also chosen $\overline{z} = -0.045$, *i.e.*, $H_{\text{mean}} = 1$. The figure shows that, comparing to [AN12], the computed solution is accurate. Note that in [AN12], the height is computed by an elliptic approach. Moreover, Figure 4.7 confirms the ε-uniformity of the time step, and stability of the scheme in the ℓ_2-norm, with the growth factor $\mathcal{G}_\phi$, which is defined as $\mathcal{G}_\phi^n := \|\phi_\Delta^n\|_{\ell_2} / \|\phi_\Delta^0\|_{\ell_2}$ for some quantity ϕ. As Figure 4.7 suggests, the scheme is stable uniformly in ε for variables like z, m and u. Note that z should be scaled with ε^2; so, having $\mathcal{G}_z \approx 6$ does not imply that the water height is changing a lot. Also, one can see that as discussed in Appendix 4.B, the scheme moves the solution toward the well-prepared

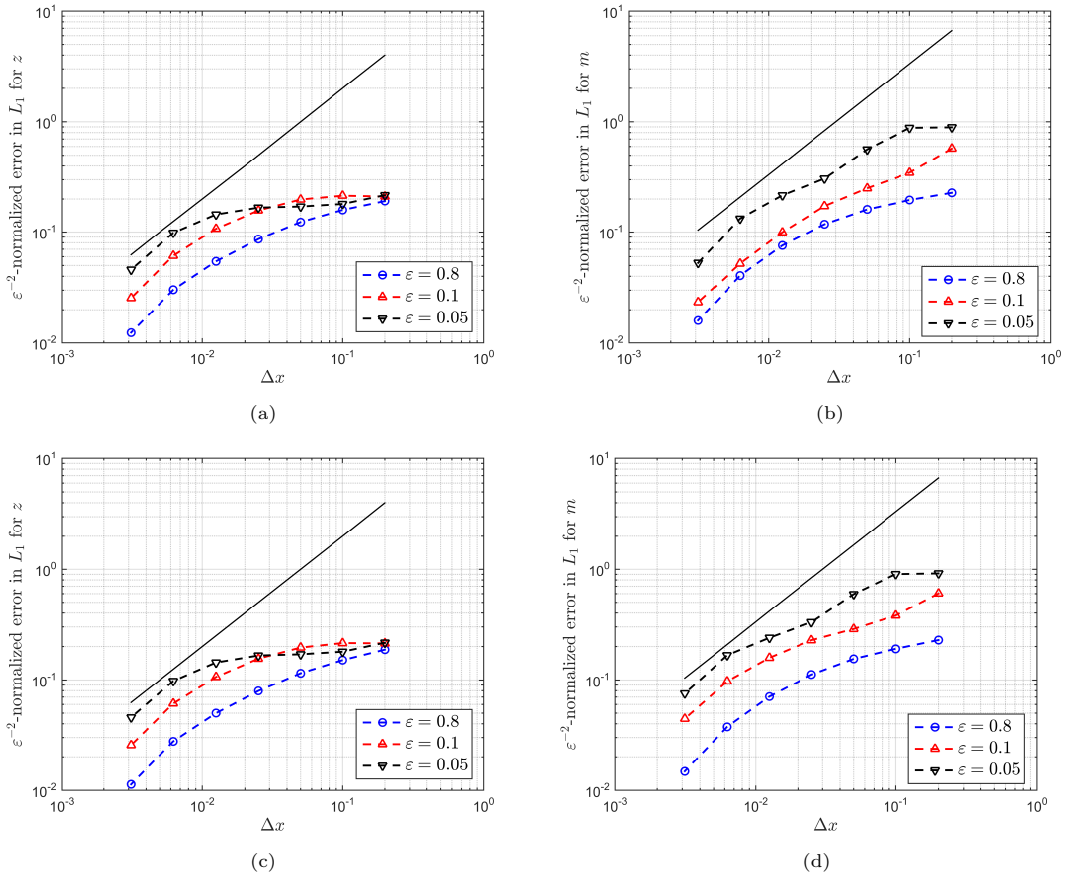


Figure 4.4: The EOC of the RS-IMEX scheme in Example (i), with CFL $= 0.45$, $T_f = 0.05$: (a) and (b) for the LaR ($\overline{m} = 0$) reference state, (c) and (d) for the zero-Froude limit ($\overline{m} = 1$) reference state. The black solid line is the line with slope one.

(limit) manifold. Because the mean value of the momentum is zero, the analysis of Appendix 4.B shows that the scheme makes the momentum $\mathcal{O}(\varepsilon^2)$, which is indicated by a very small $\mathcal{G}_m$ for small ε. Note that after the second step, it is $\widehat{\alpha} = \mathcal{O}(1)$ which dissipates small variations of the solution and gives an almost constant solution at $t = T_f$.

To compare the LaR and the zero-Froude limit reference solutions, for the case (4.46), we keep $\overline{z} = -0.045$ and change the reference momentum to $\overline{m} = \sqrt{2}$ (case ii_b) (which is not the zero-Froude limit). As Figure 4.8 shows, such a choice gives rise to a non-symmetric solution. Since the solution of the PDE does not change regardless of the choice of the reference solution, this issue should stem from the operator splitting, which does not necessarily preserve the structure of the solution. In particular, for this example, this choice of the reference momentum destroys the odd-symmetry of the momentum for each step which cannot be fully compensated due to the splitting error. This non-symmetry is, in fact, a well-known issue for operator splitting schemes; see [DR06, p. 526]. Figure 4.8 confirms this conjecture, as it shows that the solution tends to get symmetric with mesh refinement, *i.e.*, as the operator splitting error gets smaller.

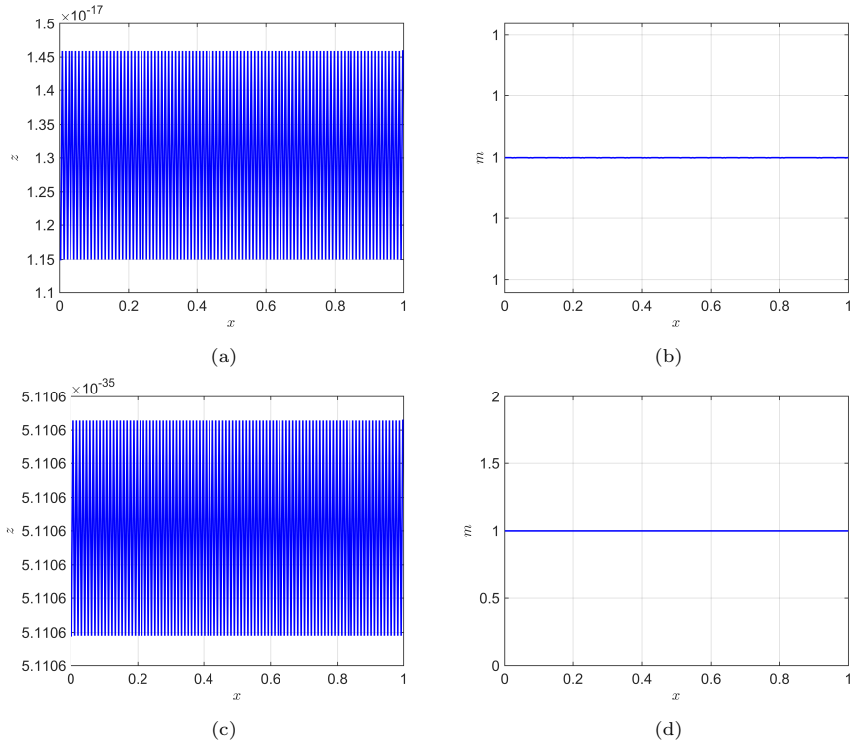


Figure 4.5: Limit of the RS-IMEX solution in Example (i), with $N = 200$, $T_f = 0.05$ and $\varepsilon = 10^{-8}$. (a) and (b) are for the LaR reference solution, (c) and (d) are for the zero-Froude limit reference solution.

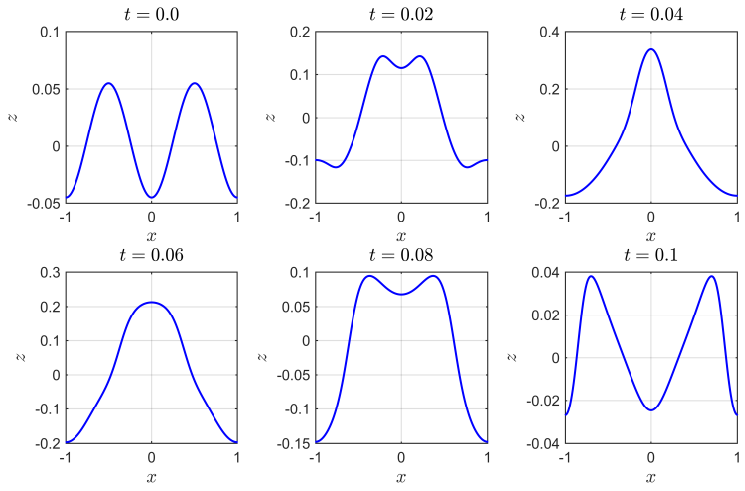
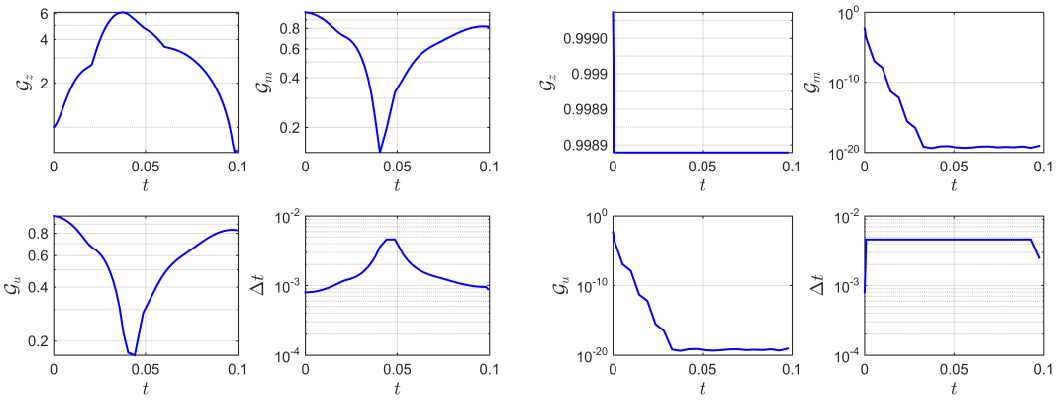


Figure 4.6: Evolution of the surface perturbation for the RS-IMEX solution in Example (ii_a), with $\varepsilon = 0.1$, CFL $= 0.45$, $N = 200$, and the LaR reference solution.



(a) $\varepsilon = 10^{-1}$. (b) $\varepsilon = 10^{-4}$.

Figure 4.7: Growth factor and time step regarding ε, in Example (ii$_a$) with the LaR reference solution.

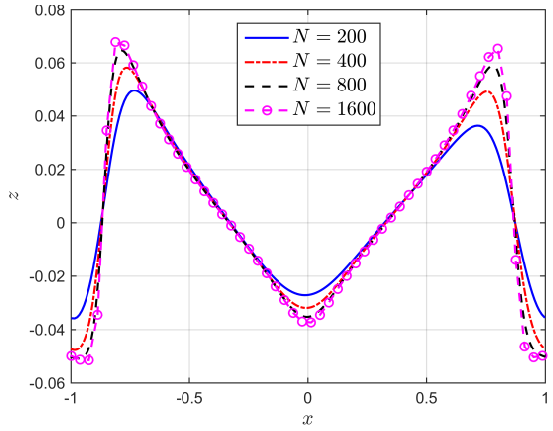


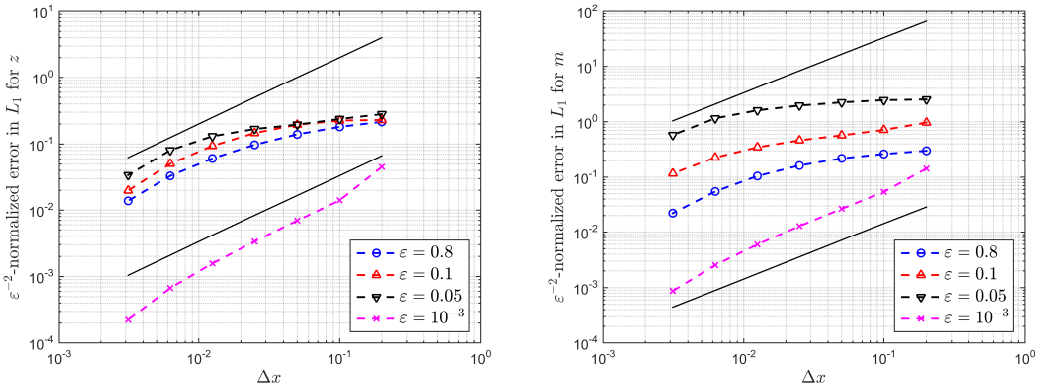
Figure 4.8: Vanishing effect of an unsuitable reference solution in Example (ii$_c$) as $\Delta x \to 0$, for $\varepsilon = 0.1$, $T_f = 0.1$ and $N = 200, 400, 800, 1600$.

4.5.2 Shallow water equations with a non-flat bottom

In this section, we study the result of the RS-IMEX scheme for the non-flat bottom case and confirm the experimental order of convergence for a specific example. Also, we verify the asymptotic consistency of the scheme, numerically. We set the initial condition as in Example (i) but with a non-flat bottom topography $\eta^b(x) = 0.2\sin(3\pi x)$; we denote these settings as Example (iii).

In Figure 4.9, the convergence rate of the scheme has been plotted, which shows the ε-uniform EOC for the scheme. Moreover, Table 4.2, shows the smallness of the checker-board oscillations for v_2. It can be seen that $\|[\![\boldsymbol{V}_{2,\Delta}]\!]\|_{\ell_\infty}$, which indicates the amplitude of possible checker-board oscillations, is of $\mathcal{O}(\varepsilon)$ as $\varepsilon \to 0$, up to some threshold ε where the condition number of J_ε gets very large and affects the solution. This is better than the analysis in Section 4.3.2.3 which

suggests $\mathcal{O}(\varepsilon + \Delta x)$. It can also seen that $\text{cond}_2(J_\varepsilon) = \mathcal{O}(1/\varepsilon)$. The condition number is almost independent of Δx; the refinement can improve the oscillations to some extent (for rather coarse meshes); however, after some point, the amplitude of the oscillations does not change with Δx.



(a) Error in z.

(b) Error in m_1.

Figure 4.9: EOC of the RS-IMEX scheme in Example (iii), with $T_f = 0.05$, CFL $= 0.45$ and the LaR reference solution.

ε	N	$\|[\![\boldsymbol{V}_{2,\Delta}^{n+1}]\!]\|_{\ell_\infty}$	$\text{cond}_2(J_\varepsilon)$	ε	N	$\|[\![\boldsymbol{V}_{2,\Delta}^{n+1}]\!]\|_{\ell_\infty}$	$\text{cond}_2(J_\varepsilon)$
10^{-2}	200	3.68e-05	8.72e+01	10^{-6}	50	1.52e-08	7.98e+05
10^{-3}	200	8.30e-10	8.18e+03	10^{-6}	100	1.75e-11	8.11e+05
10^{-4}	200	5.97e-11	8.17e+03	10^{-6}	200	5.89e-13	8.17e+05
10^{-5}	200	5.83e-12	8.17e+04	10^{-6}	400	1.95e-14	8.21e+05
10^{-6}	200	5.89e-13	8.17e+05	10^{-6}	800	7.73e-14	8.22e+05
10^{-7}	200	6.91e-14	8.17e+06	10^{-6}	1600	2.54e-13	8.23e+05
10^{-8}	200	1.49e-14	8.17e+07				
10^{-9}	200	1.29e-14	8.17e+08				

Table 4.2: Smallness of the checker-board oscillations regarding the refinement in ε and Δx in Example (iii).

4.A On the proof of Lemma 4.3.9

In this section, we complete the proof of Lemma 4.3.9, in particular, we show that the relation $\beta^2 R_b Q \boldsymbol{w}_2^{(2)} = \boldsymbol{w}_2^{(0)}$ implies that $\boldsymbol{w}_2^{(0)}$ can only be zero. We also show that kernel of the matrix R_b includes only vectors with a checker-board like structure (denoted by CB hereinafter), as defined in Lemma 4.A.1 below.

For the non-flat bottom case, the matrix R_b is defined as in (4.20) and $\beta^2 R_b Q \boldsymbol{w}_2^{(2)} = \boldsymbol{w}_2^{(0)}$

gives the following linear system of equations:

$$\beta^2 \begin{bmatrix} \overline{h}_N - \overline{h}_2 & \overline{h}_2 & & & -\overline{h}_N \\ -\overline{h}_1 & \overline{h}_1 - \overline{h}_3 & \overline{h}_3 & & \\ & -\overline{h}_2 & \overline{h}_2 - \overline{h}_4 & \overline{h}_4 & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_{2,2}^{(2)} - \boldsymbol{w}_{2,N}^{(2)} \\ \boldsymbol{w}_{2,3}^{(2)} - \boldsymbol{w}_{2,1}^{(2)} \\ \boldsymbol{w}_{2,4}^{(2)} - \boldsymbol{w}_{2,2}^{(2)} \\ \vdots \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_{2,odd}^{(0)} \\ \boldsymbol{w}_{2,even}^{(0)} \\ \boldsymbol{w}_{2,odd}^{(0)} \\ \vdots \end{bmatrix}. \tag{4.47}$$

We want to characterise the null space of the coefficient matrix and show that the system has a solution only if $\boldsymbol{w}_{2,odd}^{(0)}, \boldsymbol{w}_{2,even}^{(0)} = \mathbf{0}$. One can pose the following lemma.

Lemma 4.A.1. *Consider the linear system of equations $M\boldsymbol{y} = \boldsymbol{c}$ of size N, in accordance with equation* (4.47)*, and with a positive sequence $\{m_k\}_k > 0$, as*

$$\begin{bmatrix} m_N - m_2 & m_2 & & & -m_N \\ -m_1 & m_1 - m_3 & m_3 & & \\ & -m_2 & m_2 - m_4 & m_4 & \\ & & \ddots & \ddots & \ddots \\ m_1 & & & -m_{N-1} & m_{N-1} - m_1 \end{bmatrix} \begin{bmatrix} y_2 - y_N \\ y_3 - y_1 \\ y_4 - y_2 \\ \vdots \\ y_1 - y_{N-1} \end{bmatrix} = \begin{bmatrix} c_o \\ c_e \\ c_o \\ \vdots \\ c_{o/e} \end{bmatrix}. \tag{4.48}$$

Then,

(i) every $\boldsymbol{y} \in \mathcal{N}_M$ is either constant or has a CB-like structure, i.e., $y_k - y_{k-1}$ has different signs for odd and even k's.

(ii) the system (4.47) *is inconsistent, i.e., its solution set is empty, unless $c_i, c_0 = 0$.*

Proof. For part (i), we aim to characterise all vectors $\boldsymbol{y}$ such that $M\boldsymbol{y} = \mathbf{0}$. We rewrite the system as $\widetilde{M}\boldsymbol{\theta} = \mathbf{0}$ by denoting $\theta_j := y_{j+1} - y_j$, where

$$\widetilde{M} := \begin{bmatrix} m_2 & & & m_N \\ m_1 & m_3 & & \\ & \ddots & \ddots & \\ & & m_{N-1} & m_1 \end{bmatrix}. \tag{4.49}$$

One can check that for an odd N, $\det(\widetilde{M}) = \prod_{k=1}^{N} m_k \neq 0$. So, $\widetilde{M}$ is non-singular and has only a trivial null space. That is to say that only the zero vector belongs to its kernel, which corresponds to a constant vector $\boldsymbol{y}$ by definition. If N is even, $\det(\widetilde{M}) = 0$ and one can find a non-zero minor of size $N-1$. So, $\mathcal{N}_M$ is of rank one with the basis $\boldsymbol{\theta}^*$ such that $\theta_k^* = (-1)^k \frac{m_{k-1}}{m_{k+1}} \theta_{k-1}^*$ for $k \in \Omega_N$. The sequence $\{m_k\}_k$ is smooth as it corresponds to the discretisation of a *smooth* bottom function, and it is also positive, *i.e.*, $\{m_k\}_k \geq \min_{k \in \Omega_N} \overline{h}_k > 0$. So, the quotient $m_{k-1}/m_{k+1} \approx 1$ is also smooth for small enough Δx. This implies that, like for $\boldsymbol{\theta}^*$, the components y_k should have a CB-like structure.

Note that for such a $\boldsymbol{\theta}^*$ to belong to the kernel of M, it should fulfil the *compatibility condition* $\sum_j \theta_j = 0$ due to its definition. It is not straightforward to check *a priori* if this relation holds; but fortunately, we only need to confirm the CB-like structure of elements of $\mathcal{N}_M$.

For part (ii), note that the system (4.48) is equivalent to the auxiliary system $\widetilde{M}\boldsymbol{\theta} = \boldsymbol{c}$ as soon as $\boldsymbol{\theta}$ can be written as a difference form. So, it is enough to show the incompatibility of all possible solutions of the auxiliary system.

N is odd If N is odd, we should consider $c_o = c_i = c$. As we already shown, $\widetilde{M}$ is non-singular and the system $\widetilde{M}\boldsymbol{\theta} = \boldsymbol{c}$ has the solution $\boldsymbol{\theta}^*$ such that for $k \in \Omega_N$

$$\begin{cases} m_2\theta_1^* + m_N\theta_N^* = c \\ m_1\theta_1^* + m_3\theta_2^* = c \\ \vdots \\ m_{N-1}\theta_{N-1}^* + m_1\theta_N^* = c \end{cases} \quad \Longrightarrow \quad \theta_{k+1}^* + \frac{m_k}{m_{k+2}}\theta_k^* = \frac{c}{m_{k+2}}. \tag{4.50}$$

Then, making a spatial sum on Ω_N for (4.50) gives

$$\sum_{k\in\Omega_N}\left(1 + \frac{m_k}{m_{k+2}}\right)\theta_k^* = \sum_{k\in\Omega_N}\frac{c}{m_k}.$$

Owing to the smoothness of the sequence $\{m_k\}_k$, one gets $\frac{m_k}{m_{k+2}} \approx 1$, which implies that $\sum\theta_k^* \approx \sum\frac{c}{2m_k}$. This contradicts the compatibility condition $\sum\theta_k^* = 0$, unless $c = 0$.

Denoting $\vartheta_k := \frac{m_k}{m_{k+2}}$ and $\tilde{c}_k := \frac{c}{m_{k+2}}$, a more precise argument can be performed by rewriting (4.50) as

$$\begin{cases} \vartheta_N\theta_N^* + \theta_1^* = \tilde{c}_N \\ \vartheta_1\theta_1^* + \theta_2^* = \tilde{c}_1 \\ \vdots \\ \vartheta_{N-1}\theta_{N-1}^* + \theta_N^* = \tilde{c}_{N-1} \end{cases} \quad \Longrightarrow \quad \begin{cases} \theta_2^* = \tilde{c}_1 - \vartheta_1\theta_1^* \\ \theta_3^* = -\tilde{c}_2 - \vartheta_2\theta_2^* = \tilde{c}_2 - \vartheta_2\tilde{c}_1 + \vartheta_1\vartheta_2\theta_1^* \\ \theta_4^* = \tilde{c}_3 - \vartheta_3\theta_3^* = \tilde{c}_3 - \vartheta_3\theta_2^* + \vartheta_2\vartheta_3\tilde{c}_1 - \vartheta_1\vartheta_2\vartheta_3\theta_1^* \\ \vdots \\ \theta_N^* = (-1)^{N-1}\Big[\displaystyle\prod_{j=1}^{N-1}\vartheta_j\theta_1^* - \sum_{j=1}^{N-1}(-1)^j\tilde{c}_j\prod_{\ell=j+1}^{N-1}\vartheta_\ell\Big] \end{cases}$$

So, one gets two different relations for θ_N^*, which, indeed, should be the same:

$$\theta_N^* = -\vartheta_N^{-1}\theta_1^* + \vartheta_N^{-1}\tilde{c}_N, \qquad \theta_N^* = \prod_{j=1}^{N-1}\vartheta_j\theta_1^* + \sum_{j=1}^{N-1}(-1)^j\tilde{c}_j\prod_{\ell=j+1}^{N-1}\vartheta_\ell.$$

It is not difficult to confirm that $\prod_{j=1}^{N-1}\vartheta_j = \vartheta_N^{-1}$; so

$$\theta_1^* = \frac{\tilde{c}_N}{2} - \frac{\vartheta_N}{2}\sum_{j=1}^{N-1}(-1)^j\tilde{c}_j\prod_{\ell=j+1}^{N-1}\vartheta_\ell. \tag{4.51}$$

In (4.51), since $\{\tilde{c}_k\}_k$ and $\{\nu_k\}_k$ vary smoothly, there are some cancellations for the second term, which suggests that it is of $\mathcal{O}(\Delta x)$. Performing a similar procedure for every $k \in \Omega_N$ implies that the leading order of θ_k^* is $\frac{\tilde{c}_{k-1}}{2}$, which implies that $\sum\theta_k^* \neq 0$, *i.e.*, the solution cannot be compatible, unless $c = 0$.

N is even For this case, the procedure is very similar to the previous one. Assuming the existence of a solution $\boldsymbol{\theta}^*$, one gets

$$\sum_{k\in\Omega_N}\frac{m_k}{m_{k+2}}\theta_k^* = \sum_{(k=2j)\in\Omega_N}\frac{c_e}{m_{k+2}} + \sum_{(k=2j+1)\in\Omega_N}\frac{c_o}{m_{k+2}}, \tag{4.52}$$

which, in general, resembles the previous argument for an odd N. However, for $c_o = -c_e$, the previous argument seems not working as the rhs vanishes. Here, we show that in such a case, for the system to be consistent $c_e = c_e = 0$ should hold which matches the statement of the lemma. Consider the same definition of ϑ_k and $\tilde{c}_k$ as above, with $c_e = -c_o = c$. So,

$$\begin{cases} \vartheta_N \theta_N^* + \theta_1^* = -\tilde{c}_N \\ \vartheta_1 \theta_1^* + \theta_2^* = \tilde{c}_1 \\ \vdots \\ \vartheta_{N-1} \theta_{N-1}^* + \theta_N^* = \tilde{c}_{N-1} \end{cases} \implies \begin{cases} \theta_2^* = \tilde{c}_1 - \vartheta_1 \theta_1^* \\ \theta_3^* = -\tilde{c}_2 - \vartheta_2 \theta_2^* = -\tilde{c}_2 - \vartheta_2 \tilde{c}_1 + \vartheta_1 \vartheta_2 \theta_1^* \\ \theta_4^* = \tilde{c}_3 - \vartheta_3 \theta_3^* = \tilde{c}_3 + \vartheta_3 \theta_2^* + \vartheta_2 \vartheta_3 \tilde{c}_1 - \vartheta_1 \vartheta_2 \vartheta_3 \theta_1^* \\ \vdots \\ \theta_N^* = (-1)^{N-1} \Big[\prod_{j=1}^{N-1} \vartheta_j \theta_1^* - \sum_{j=1}^{N-1} \tilde{c}_j \prod_{\ell=j+1}^{N-1} \vartheta_\ell \Big] \end{cases}$$

One gets two different relations for θ_N^*, which should be the same:

$$\theta_N^* = -\vartheta_N^{-1} \theta_1^* - \vartheta_N^{-1} \tilde{c}_N, \qquad \theta_N^* = -\prod_{j=1}^{N-1} \vartheta_j \theta_1^* + \sum_{j=1}^{N-1} \tilde{c}_j \prod_{\ell=j+1}^{N-1} \vartheta_\ell.$$

Because $\prod_{j=1}^{N-1} \vartheta_j = {\vartheta_N}^{-1}$ and the sign of second terms are different, $\tilde{c}_k = 0$ for all k, *i.e.*, $\boldsymbol{c} = \boldsymbol{0}$. □

Lemma 4.A.1 implies the system (4.47) or (4.48) is only consistent if $\boldsymbol{w}_{2,odd}^{(0)}, \boldsymbol{w}_{2,even}^{(0)} = \boldsymbol{0}$. Also, it confirms that $R_b \boldsymbol{w}_1 \to 0$ if and only if $\boldsymbol{w}_1$ tends to a vector with the CB-like structure. The relation (4.24a) shows that the mean of $\boldsymbol{w}_1^{(0)}$ vanishes for a summation on odd and even indices while, owing to the smoothness of the bottom function and because of this vanishing mean, odd and even entries of $\boldsymbol{w}_1^{(0)}$ should have different signs. This concludes that $\boldsymbol{w}_1^{(0)}$ is the zero vector.

4.B Asymptotic consistency of the RS-IMEX scheme with ill-prepared initial data

Regarding AP schemes for hyperbolic balance laws, the focus is often limited to the well-prepared initial data (Definition 4.3.2). Here, we briefly show that the rigorous asymptotic consistency analysis can also be done for the ill-prepared initial data (*cf.* [FN09, Sect. 4.6]), *i.e.*,

$$\begin{aligned} z_{0,\varepsilon} &= z_{(0)}^0 + \varepsilon z_{(1),\varepsilon}^0, \\ m_{0,\varepsilon} &= m_{(0)}^0 + \varepsilon m_{(1),\varepsilon}^0, \end{aligned} \tag{4.53}$$

where $z_{(0)}^0$ is constant, $z_{(1),\varepsilon}^0 = \mathcal{O}(1)$ and $m_{(0)}^0$ is not solenoidal (constant in 1d).

We consider the LaR reference solution and assume a flat bottom topography. One can check from (4.10) that the splitting is still admissible in the sense of [SN14]. Also, without scaling the perturbation, we pick $\boldsymbol{V} = \boldsymbol{U}_{pert}$.

At first we show the ε-stability of the updated solution to justify the use of asymptotic expansion. From the definition of $\widetilde{\boldsymbol{F}}$ and $\widehat{\boldsymbol{F}}$, one can simply check that the intermediate step solution is

ε-stable as the pressure term $v_1^2/2\varepsilon^2$ is $\mathcal{O}(1)$, owing to (4.53); this implies that $\|\boldsymbol{V}_\Delta^{n+1/2}\| = \mathcal{O}(1)$. Since J_ε^{-1} is ε-stable (with similar arguments as in Section 4.3.2.2), $\|\boldsymbol{V}_\Delta^{n+1}\| = \mathcal{O}(1)$ and the use of asymptotic expansion is justified.

Balancing $\mathcal{O}(1/\varepsilon^2)$ and $\mathcal{O}(1/\varepsilon)$ terms in the implicit momentum update shows that $\nabla_h \left(\overline{h} v_{1,i}^{n+1}\right) = \mathcal{O}(\varepsilon^2)$. This, combined with $\|\boldsymbol{V}_{1,\Delta}^{n+1/2}\| = \mathcal{O}(\varepsilon)$ and the implicit v_1-update, implies that $\nabla_h v_{2,i}^{n+1} = \mathcal{O}(\varepsilon)$. In other words, v_2^{n+1} (similarly v_1^{n+1}) consists of an $\mathcal{O}(1)$ (similarly $\mathcal{O}(\varepsilon)$) constant plus some $\mathcal{O}(\varepsilon)$ (similarly $\mathcal{O}(\varepsilon^2)$) perturbations, where the constant can be shown (by a spatial summation) to be the mean value of the leading order of the initial momentum. Note that for the colliding pulses example 4.5.1.2, these constants are zero. So, after only one step, the solution is moved to the mean value plus small perturbations. Performing a similar procedure for the next step, one obtains $\|\boldsymbol{V}_{1,\Delta}^{n+3/2}\| = \mathcal{O}(\varepsilon^2)$, thus $\nabla_h v_{2,i}^{n+2} = \mathcal{O}(\varepsilon^2)$, which concludes that the solution is completely projected onto the limit manifold, and is moved beyond the initial layer. This gives the correct uniform behaviour for the scheme; see [CJR97] for some discussions on this topic for relaxation systems. Hence, the scheme is AC even with an ill-prepared initial datum in the sense of (4.53).

Chapter 5

The RS-IMEX scheme for the 2d shallow water equations

"Life can only be understood backwards, but it must be lived forwards."

– Kierkegaard (1844)

The present chapter, following Chapter 4, extends the asymptotic analysis of the RS-IMEX scheme for the shallow water equations in two space dimensions. Along the same lines as Chapter 4, we prove the asymptotic preserving property w.r.t. the Froude number, i.e., we prove that the scheme is consistent and stable, uniformly in the Froude number. We discuss the discrete preservation of the equilibrium states and confirm the analytical results by a series of numerical experiments. This chapter is based on [Zak16b].

Contents

5.1 RS-IMEX scheme for the shallow water equations

In this chapter and along the same lines as Chapter 4, we study the RS-IMEX scheme for the 2d SWE with bottom topography. In the current section, the RS-IMEX scheme will be derived for this system, followed by the asymptotic analysis in Section 5.2. Finally, in Section 5.3, several numerical experiments will be presented to confirm the analysis and show the quality of the scheme.

Let us consider the 2d SWE (1.10) in the reformulated form as [BALMN14, Bis15], which is the extension of the 1d SWE (4.8) in Chapter 4:

$$\begin{aligned} &\partial_t z + \operatorname{div}_{\boldsymbol{x}} \boldsymbol{m} = 0, \\ &\partial_t \boldsymbol{m} + \operatorname{div}_{\boldsymbol{x}} \left(\frac{\boldsymbol{m} \otimes \boldsymbol{m}}{z-b} + \frac{z^2 - 2bz}{2\varepsilon^2} \mathbb{I}_2 \right) = -\frac{z}{\varepsilon^2} \nabla_{\boldsymbol{x}} b, \end{aligned} \tag{5.1}$$

with the same notations as in Chapter 4, and in a periodic domain, *i.e.*, $\Omega = \mathbb{T}^2$. However, we use also open domains in two numerical experiments in Sections 5.3.1 and 5.3.2. The system (5.1) converges to the lake equations as $\varepsilon \to 0$:

$$\begin{aligned} &\partial_t \boldsymbol{m} - \operatorname{div}_{\boldsymbol{x}} \left(\frac{\boldsymbol{m} \otimes \boldsymbol{m}}{b} \right) - b \nabla_{\boldsymbol{x}} \pi = \boldsymbol{0}, \\ &\operatorname{div}_{\boldsymbol{x}} \boldsymbol{m} = 0, \end{aligned} \tag{5.2}$$

where the auxiliary pressure (or more precisely, surface perturbation) π acts as the Lagrange multiplier fulfilling the divergence constraint; see Appendix 5.A and Definition 5.2.1 as well as [BKL11]. Note that $-b$ is the zero-Froude limit of the water height, *i.e.*, the water surface is flat in the limit. The aim of this chapter, in particular, is to check if the RS-IMEX scheme provides a consistent and stable approximation of (5.2) in the limit.

Using the general form of hyperbolic balance laws (4.1) for the system (5.1), the choice of conservative variables $\boldsymbol{U} = (z, m_1, m_2)^T$ implies that the flux $\boldsymbol{F}$ and the source term $\boldsymbol{S}$ can be written as

$$\boldsymbol{F} = \begin{bmatrix} m_1 & m_2 \\ \dfrac{m_1^2}{z-b} + \dfrac{z^2 - 2zb}{2\varepsilon^2} & \dfrac{m_1 m_2}{z-b} \\ \dfrac{m_1 m_2}{z-b} & \dfrac{m_2^2}{z-b} + \dfrac{z^2 - 2zb}{2\varepsilon^2} \end{bmatrix}, \quad \boldsymbol{S} = \begin{bmatrix} 0 \\ -z\,\partial_x b/\varepsilon^2 \\ -z\,\partial_y b/\varepsilon^2 \end{bmatrix}.$$

Assuming the reference solution $\overline{\boldsymbol{U}} = (\overline{z}, \overline{m_1}, \overline{m_2})^T$ to be the solution of the lake equations (5.2), and following the RS-IMEX splitting as described in Chapter 4, with the scaling matrix $D = \operatorname{diag}(\varepsilon^2, 1, 1)$, the splitting can be obtained as

$$\overline{\boldsymbol{G}} = \begin{bmatrix} \overline{m_1}/\varepsilon^2 & \overline{m_2}/\varepsilon^2 \\ \dfrac{\overline{m_1}^2}{\overline{z}-b} + \dfrac{\overline{z}^2 - 2\overline{z}b}{2\varepsilon^2} & \dfrac{\overline{m_1 m_2}}{\overline{z}-b} \\ \dfrac{\overline{m_1 m_2}}{\overline{z}-b} & \dfrac{\overline{m_2}^2}{\overline{z}-b} + \dfrac{\overline{z}^2 - 2\overline{z}b}{2\varepsilon^2} \end{bmatrix}, \tag{5.3a}$$

$$\widetilde{\boldsymbol{G}}_1 = \begin{bmatrix} v_2/\varepsilon^2 \\ -\dfrac{\overline{m_1}^2 v_1 \varepsilon^2}{(\overline{z}-b)^2} + \dfrac{2\overline{m_1} v_2}{\overline{z}-b} + (\overline{z}-b) v_1 \\ -\dfrac{\overline{m_1 m_2} v_1 \varepsilon^2}{(\overline{z}-b)^2} + \dfrac{\overline{m_1} v_3}{\overline{z}-b} + \dfrac{\overline{m_2} v_2}{\overline{z}-b} \end{bmatrix}, \quad \widetilde{\boldsymbol{G}}_2 = \begin{bmatrix} v_3/\varepsilon^2 \\ -\dfrac{\overline{m_1 m_2} v_1 \varepsilon^2}{(\overline{z}-b)^2} + \dfrac{\overline{m_1} v_3}{\overline{z}-b} + \dfrac{\overline{m_2} v_2}{\overline{z}-b} \\ -\dfrac{\overline{m_2}^2 v_1 \varepsilon^2}{(\overline{z}-b)^2} + \dfrac{2\overline{m_2} v_3}{\overline{z}-b} + (\overline{z}-b) v_1 \end{bmatrix}, \tag{5.3b}$$

$$\widehat{\boldsymbol{G}}_1 = \begin{bmatrix} \dfrac{m_1^2}{z-b} + \dfrac{z^2-2zb}{2\varepsilon^2} - \dfrac{\overline{m_1}^2}{\overline{z}-b} - \dfrac{\overline{z}^2-2\overline{z}b}{2\varepsilon^2} + \dfrac{\overline{m_1}^2 v_1\varepsilon^2}{(\overline{z}-b)^2} - \dfrac{2\overline{m_1}v_2}{\overline{z}-b} - (\overline{z}-b)v_1 \\ \dfrac{m_1 m_2}{z-b} - \dfrac{\overline{m_1 m_2}}{\overline{z}-b} + \dfrac{\overline{m_1 m_2}v_1\varepsilon^2}{(\overline{z}-b)^2} - \dfrac{\overline{m_1}v_3}{\overline{z}-b} - \dfrac{\overline{m_2}v_2}{\overline{z}-b} \end{bmatrix}, \quad 0$$

$$\widehat{\boldsymbol{G}}_2 = \begin{bmatrix} \dfrac{m_1 m_2}{z-b} - \dfrac{\overline{m_1 m_2}}{\overline{z}-b} + \dfrac{\overline{m_1 m_2}v_1\varepsilon^2}{(\overline{z}-b)^2} - \dfrac{\overline{m_1}v_3}{\overline{z}-b} - \dfrac{\overline{m_2}v_2}{\overline{z}-b} \\ \dfrac{m_2^2}{z-b} + \dfrac{z^2-2zb}{2\varepsilon^2} - \dfrac{\overline{m_2}^2}{\overline{z}-b} - \dfrac{\overline{z}^2-2\overline{z}b}{2\varepsilon^2} + \dfrac{\overline{m_2}^2 v_1\varepsilon^2}{(\overline{z}-b)^2} - \dfrac{2\overline{m_2}v_3}{\overline{z}-b} - (\overline{z}-b)v_1 \end{bmatrix}, \quad 0 \tag{5.3c}$$

$$\overline{\boldsymbol{Z}} = \begin{bmatrix} 0 \\ -\overline{z}\,\partial_x b/\varepsilon^2 \\ -\overline{z}\,\partial_y b/\varepsilon^2 \end{bmatrix}, \qquad \widetilde{\boldsymbol{Z}} = \begin{bmatrix} 0 \\ -v_1\partial_x b \\ -v_1\partial_y b \end{bmatrix}, \qquad \widehat{\boldsymbol{Z}} = \boldsymbol{0}. \tag{5.3d}$$

Unlike Chapter 4, this choice of scaling matrix is not what the formal asymptotic analysis suggests (see Appendix 5.A). However, we will see in Section 5.2.2.2 that this choice is more appropriate for the rigorous asymptotic consistency analysis. One should verify that the Jacobian matrices $\widehat{\boldsymbol{G}}'$ and $\widetilde{\boldsymbol{G}}'$ have a complete set of eigenvectors, and that eigenvalues of $\widehat{\boldsymbol{G}}'$ are non-stiff as there is no $\mathcal{O}(1/\varepsilon)$ term in $\widehat{\boldsymbol{G}}$ after simplification. The hyperbolicity of $\widetilde{\boldsymbol{G}}'$ is trivial by construction. For the Jacobian of the slow system $\widehat{\boldsymbol{G}}'$ (or more precisely $\widehat{\boldsymbol{G}}'_{\boldsymbol{n}}$), the eigenvalues can be obtained as (see [KSSN16] for instance)

$$\widehat{\lambda}_1 = 0, \qquad \widehat{\lambda}_2 = (\boldsymbol{u}-\overline{\boldsymbol{u}})\cdot\boldsymbol{n}, \qquad \widehat{\lambda}_2 = 2(\boldsymbol{u}-\overline{\boldsymbol{u}})\cdot\boldsymbol{n},$$

with the velocity $\boldsymbol{u} := \frac{\boldsymbol{m}}{z-b}$, the reference velocity $\overline{\boldsymbol{u}} := \frac{\overline{\boldsymbol{m}}}{\overline{z}-b}$ and the unit normal vector $\boldsymbol{n}$. So, the splitting is admissible in the sense of Definition 3.1.1. Note that this splitting will be reduced to the splitting in [BALMN14, Bis15] as soon as one picks $\overline{\boldsymbol{m}} = \boldsymbol{0}$.

Based on Algorithm 1, the RS-IMEX scheme can be written in the split form:

$$\boldsymbol{V}_{ij}^{n+1/2} = \boldsymbol{V}_{ij}^{n} - \frac{\Delta t}{\Delta x}\left(\widehat{\boldsymbol{G}}^{n}_{1,i+1/2j} - \widehat{\boldsymbol{G}}^{n}_{1,i-1/2j}\right) - \frac{\Delta t}{\Delta y}\left(\widehat{\boldsymbol{G}}^{n}_{2,ij+1/2} - \widehat{\boldsymbol{G}}^{n}_{2,ij-1/2}\right), \tag{5.4a}$$

$$\begin{aligned}\boldsymbol{V}_{ij}^{n+1} = \boldsymbol{V}_{ij}^{n+1/2} &- \frac{\Delta t}{\Delta x}\left(\widetilde{\boldsymbol{G}}^{n+1}_{1,i+1/2j} - \widetilde{\boldsymbol{G}}^{n+1}_{1,i-1/2j}\right) - \frac{\Delta t}{\Delta y}\left(\widetilde{\boldsymbol{G}}^{n+1}_{2,ij+1/2} - \widetilde{\boldsymbol{G}}^{n+1}_{2,ij-1/2}\right) \\ &+ \Delta t\,\widetilde{\boldsymbol{Z}}_{ij}^{n+1} - \Delta t\,\overline{\boldsymbol{T}}_{ij}^{n+1},\end{aligned} \tag{5.4b}$$

for each cell $(i,j) \in \{1,2,\ldots,N_x\}\times\{1,2,\ldots,N_y\}$ in the computational domain Ω_N, with spatial steps Δx and Δy and the time step Δt, where $\widetilde{\boldsymbol{G}}_{i+1/2j}$ and $\widehat{\boldsymbol{G}}_{i+1/2j}$ are Rusanov fluxes at cell interfaces as defined in Section 4.2.1, with $\widehat{\boldsymbol{G}}$ and $\widetilde{\boldsymbol{G}}$ as in (5.3b)–(5.3c). $\widetilde{\boldsymbol{Z}}_{ij}^{n+1}$ is the central discretisation of the source terms in (5.3d) and $\overline{\boldsymbol{T}}_{ij}^{n+1}$ is the discretisation of the residual of the reference solution, and is computed as

$$\overline{\boldsymbol{T}}_{ij}^{n+1} = D^{-1}\frac{\overline{\boldsymbol{U}}_{ij}^{n+1} - \overline{\boldsymbol{U}}_{ij}^{n}}{\Delta t} + \frac{\overline{\boldsymbol{G}}^{n+1}_{1,i+1/2j} - \overline{\boldsymbol{G}}^{n+1}_{1,i-1/2j}}{\Delta x} + \frac{\overline{\boldsymbol{G}}^{n+1}_{2,ij+1/2} - \overline{\boldsymbol{G}}^{n+1}_{2,ij-1/2}}{\Delta y} - \overline{\boldsymbol{Z}}_{ij}^{n+1}, \tag{5.5}$$

again using the Rusnaov numerical flux but with the numerical diffusion coefficient $\overline{\alpha}$; we will discuss about the role of this diffusion in Appendix 5.B.

By denoting $\nabla_{h,x}$ and $\Delta_{h,x}$, respectively, as the central discretisation of the first and second derivatives in the x-direction, one can rewrite (5.4a)–(5.4b) as

$$\boldsymbol{V}_{ij}^{n+1/2} = \boldsymbol{V}_{ij}^{n} - \Delta t \nabla_{h,x} \widehat{\boldsymbol{G}}_{1,ij}^{n} - \Delta t \nabla_{h,y} \widehat{\boldsymbol{G}}_{2,ij}^{n} + \frac{\widehat{\alpha}_1 \Delta x}{2} \Delta t \Delta_{h,x} \boldsymbol{V}_{ij}^{n} + \frac{\widehat{\alpha}_2 \Delta y}{2} \Delta t \Delta_{h,y} \boldsymbol{V}_{ij}^{n}, \tag{5.6a}$$

$$\begin{aligned} \boldsymbol{V}_{ij}^{n+1} = \boldsymbol{V}_{ij}^{n+1/2} &- \Delta t \nabla_{h,x} \widetilde{\boldsymbol{G}}_{1,ij}^{n+1} - \Delta t \nabla_{h,y} \widetilde{\boldsymbol{G}}_{2,ij}^{n+1} + \frac{\widetilde{\alpha}_1 \Delta x}{2} \Delta t \Delta_{h,x} \boldsymbol{V}_{ij}^{n+1} + \frac{\widetilde{\alpha}_2 \Delta y}{2} \Delta t \Delta_{h,y} \boldsymbol{V}_{ij}^{n+1} \\ &+ \Delta t \widetilde{\boldsymbol{Z}}_{ij}^{n+1} - \Delta t \overline{\boldsymbol{T}}_{ij}^{n+1}. \end{aligned} \tag{5.6b}$$

For the reference solution $\overline{\boldsymbol{U}}$, the surface perturbation $\overline{z}$ is constant (in time and space) and the momentum field $\overline{\boldsymbol{m}}$ is solenoidal. We also pick $\overline{\alpha}_1 = \overline{\alpha}_2 = 0$ for the sake of simplicity. So, one can write $\overline{\boldsymbol{T}}$ block-wise as $\overline{\boldsymbol{T}}_{\Delta}^{n+1} := [\overline{\boldsymbol{T}}_{1,\Delta}^{n+1}, \overline{\boldsymbol{T}}_{2,\Delta}^{n+1}, \overline{\boldsymbol{T}}_{3,\Delta}^{n+1}]^T$ such that

$$\begin{aligned} \overline{\boldsymbol{T}}_{1,ij}^{n+1} &= \left(\nabla_{h,x} \overline{m_1}_{ij}^{n+1} + \nabla_{h,y} \overline{m_2}_{ij}^{n+1}\right)/\varepsilon^2, \\ \overline{\boldsymbol{T}}_{2,ij}^{n+1} &= D_t \overline{m_1}_{ij}^{n} + \nabla_{h,x} \left(\frac{\overline{m_1}_{ij}^{n+1,2}}{\overline{z} - b_{ij}} \right) + \nabla_{h,y} \left(\frac{\overline{m_1}_{ij}^{n+1} \overline{m_2}_{ij}^{n+1}}{\overline{z} - b_{ij}} \right), \\ \overline{\boldsymbol{T}}_{3,ij}^{n+1} &= D_t \overline{m_2}_{ij}^{n} + \nabla_{h,x} \left(\frac{\overline{m_1}_{ij}^{n+1} \overline{m_2}_{ij}^{n+1}}{\overline{z} - b_{ij}} \right) + \nabla_{h,y} \left(\frac{\overline{m_1}_{ij}^{n+1,2}}{\overline{z} - b_{ij}} \right). \end{aligned} \tag{5.7}$$

So far, the scheme for computing the scaled perturbation has been introduced. The remaining point to be clarified is how to solve the equations for the reference solution, which is explained in the next section. Note that from now on and for the sake of simplicity, we assume the same number of grid points in both directions, *i.e.*, $N_x = N_y = N$. Also, we pick $\widehat{\alpha}_1 = \widehat{\alpha}_2$ and $\widetilde{\alpha}_1 = \widetilde{\alpha}_2$.

Remark 5.1.1. *(i) Note that unlike the one-dimensional system in Chapter 4, the reference velocity field is not constant in general; so, the limit reference solution is not the solution of the original system, i.e., the residual $\overline{\boldsymbol{T}}$ does not vanish.*

(ii) In fact, $\overline{\boldsymbol{T}}_{\Delta}$ corresponds to the discretisation of the lake equations, without the second order "incompressible" pressure $p_{(2)}$ or π. So, other terms in (5.6a)–(5.6b) *can be seen as an approximation of that missing pressure term. Formally speaking, in the limit $\varepsilon \to 0$ and $\Delta \to 0$, one expects to recover the limit system; so, $(v_2, v_3) \to (0, 0)$, which leaves the desired pressure term $(\overline{z} - b)v_1$.*

(iii) It is not difficult to verify that (5.3a)–(5.3d) *give similar system as in the multiple pressure variables (MPV) approach, where a system for the perturbations around the incompressible Euler system is obtained, cf. [MDR07, eqs. (30)–(32)] for instance. However, that approach makes use of an implicit method to solve for the perturbations and ignore higher order terms in terms of the Mach number in order to deal with a linearised system.*

(iv) The non-vanishing of $\overline{\boldsymbol{T}}_{\Delta}$ can be seen from the definition of $\overline{\boldsymbol{T}}_{\Delta}$ in (5.7) *as the discrete divergence of the limit momentum field "may" vanish only approximately, and there "may" be a missing term in $\overline{\boldsymbol{T}}_{2,\Delta}$ and $\overline{\boldsymbol{T}}_{3,\Delta}$ regarding the contributions of the incompressible pressure. In fact, with a constant reference surface perturbation, the pressure gradient vanishes completely in $\overline{\boldsymbol{T}}_{2,\Delta}$ and $\overline{\boldsymbol{T}}_{3,\Delta}$ while its $\mathcal{O}(1)$ contribution is present in the lake equations. So, $\|\overline{\boldsymbol{T}}_{2,\Delta}\|$ and $\|\overline{\boldsymbol{T}}_{3,\Delta}\|$ would be generally $\mathcal{O}(1)$. This leads to the fact that starting with an initial datum on the limit manifold, i.e., $\boldsymbol{V}_{\Delta}^{n} = \boldsymbol{0}$, the scheme produces some $\mathcal{O}(\Delta t)$*

disturbances such that $\boldsymbol{V}_\Delta^{n+1} \neq \boldsymbol{0}$. *This is the matter of importance as it indicates that the explicit part of the scheme does not vanish in the limit, i.e.,* $\lim_{\varepsilon\to 0} \nabla \cdot \widehat{\boldsymbol{G}} \neq 0$. *Showing this, we assume that* $\boldsymbol{V}_\Delta^n = \boldsymbol{V}_\Delta^{n+1} = \boldsymbol{0}$ *and that the discrete divergence constraint is fulfilled for the reference solution. So, all terms in the scheme*

$$\boldsymbol{V}_\Delta^{n+1} = \boldsymbol{V}_\Delta^n - \Delta t \overline{\boldsymbol{T}}_\Delta^{n+1} + \Delta t \left(-\nabla_{h,\boldsymbol{x}} \cdot \widetilde{\boldsymbol{G}}_\Delta + \widetilde{\boldsymbol{Z}}_\Delta\right)^{n+1} - \Delta t \left(\nabla_{h,\boldsymbol{x}}^{\cdot} \widehat{\boldsymbol{G}}_\Delta\right)^{n},$$

vanish but $\overline{\boldsymbol{T}}_{2,\Delta}$ *and* $\overline{\boldsymbol{T}}_{3,\Delta}$, *which implies that* $\boldsymbol{V}_\Delta^{n+1} = \boldsymbol{0}$ *cannot hold. Thus, the explicit part may not be zero. We will illustrate this in Section 5.3.4 by a numerical example.*

5.1.1 Solving for the reference solution

In the RS-IMEX procedure, one needs to solve the reference system in time and compute $\overline{\boldsymbol{T}}$. Here, the reference system is chosen as the zero-Froude limit (the lake equations), which is the same as the incompressible limit of the isentropic Euler system if the bottom topography is flat. Because the lake equations are globally well-posed (see [LOT96]), finding a numerical approximation of its solution is justified.

Indeed, there exist several numerical methods for the incompressible Euler or Navier–Stokes equations, see, *e.g.*, [PTA12, DR06]. Here and to solve the lake equations (5.2) numerically, we employ the so-called *projection scheme* (mainly by Chorin [Cho68, Cho69] and Temam [Tem69]) because of its simplicity and applicability to collocated grids. The projection method, in the time-discrete form, for the lake equations (5.2) can be outlined as follows (see [PTA12]):

(i) Update the momentum field *only* due to the advection term $-\mathrm{div}_{\boldsymbol{x}}\left(\frac{\boldsymbol{m}^n \otimes \boldsymbol{m}^n}{b}\right)$, using the (local) Lax–Friedrichs scheme. This leads to the momentum field $\boldsymbol{m}^\star$.

(ii) Now, consider the pressure term and impose the *div*-free condition. Then, solve the elliptic equation for the updated auxiliary surface perturbation π^{n+1}:

$$-\mathrm{div}_{\boldsymbol{x}}\left(b \nabla_{\boldsymbol{x}} \pi^{n+1}\right) = \frac{\mathrm{div}_{\boldsymbol{x}} \boldsymbol{m}^\star}{\Delta t}. \tag{5.8}$$

(iii) Update the momentum field to $\boldsymbol{m}^{n+1}$ with the updated pressure field (using the central scheme):

$$\frac{\boldsymbol{m}^{n+1} - \boldsymbol{m}^\star}{\Delta t} - b \nabla_{\boldsymbol{x}} \pi^{n+1} = 0.$$

The procedure seems straightforward; however, as we impose periodic boundary conditions for the auxiliary pressure π (and z), computing the solution of the elliptic equation (5.8) is not trivial, because the companion matrix for the discretised equation is not invertible (for doubly-periodic domains). This discrete system is, though, solvable under a *solvability condition*, which is, in fact, a no-net-flux condition on the boundaries. For continuous and discretised equations, this condition writes

$$-\Delta_{\boldsymbol{x}} \theta = f \text{ in } \Omega : \qquad \int_\Omega f \mathrm{d}\boldsymbol{x} = 0$$

$$-\Delta_{h,\boldsymbol{x}}\theta_{ij} = f_{ij} \text{ in } \Omega_N : \qquad \sum_{(i,j)\in\Omega_N} f_{ij} = 0$$

Due to this singularity of the coefficient matrix, lots of methods for linear systems of equations (LSE) cannot be employed. Nevertheless, one can consider the following approaches.

Discrete Fourier transform (DFT) For the flat bottom case, the elliptic equation has constant coefficients. So, for a doubly-periodic domain, one can make use of DFT and find the solution very efficiently. Since this is a somewhat standard approach in the context of numerical schemes for elliptic problems, we skip the details here and refer the reader to [VL92].

Parabolic regularisation of the problem For the non-flat bottom case, DFT is of no use as the coefficients are not constant. But, one can introduce a regularised parabolic problem, by adding a time derivative in some pseudo time τ, whose stationary solution gives the solution of the original elliptic one. For the elliptic equation $-\Delta_{\boldsymbol{x}}\theta = f$, the regularised problem is $\partial_\tau\theta - \Delta_{\boldsymbol{x}}\theta = f$, leading to the following non-singular discretised system for the implicit Euler time integration:

$$\left(1 - \Delta\tau\Delta_{h,\boldsymbol{x}}\right)\theta_{ij}^{n+1} = \theta_{ij}^{n} - \Delta\tau f_{ij}.$$

One can even circumvent solving an LSE by using explicit methods, *e.g.*, the explicit Euler method:

$$\theta_{ij}^{n+1} = \left(1 + \Delta\tau\Delta_{h,\boldsymbol{x}}\right)\theta_{ij}^{n} - \Delta\tau f_{ij}.$$

Seeking the stationary solution of this system gives the solution of the original Poisson problem. Note that we impose the steadiness of the solution approximately by requiring the relative temporal change of the solution to be less than 0.01% (from the solution at the previous step).

5.2 Asymptotic analysis of the scheme

Before we proceed with the main theorem of this section, let us fix the definition of the well-prepared initial data, using the following asymptotic (Poincaré) expansion

$$\begin{aligned} z(t,\boldsymbol{x}) &= z_{(0)} + \varepsilon z_{(1)} + \varepsilon^2 z_{(2)}, \\ \boldsymbol{m}(t,\boldsymbol{x}) &= \boldsymbol{m}_{(0)} + \varepsilon \boldsymbol{m}_{(1)} + \varepsilon^2 \boldsymbol{m}_{(2)}. \end{aligned} \tag{5.9}$$

Definition 5.2.1. *The formal zero-Froude limit of the shallow water system* (5.1) *gives the so-called lake equations, and reads (see Appendix 5.A as well as [BKL11] for the formal justification)*

$$\begin{aligned} & z_{(0)}, z_{(1)} = const., \\ & \operatorname{div}_{\boldsymbol{x}} \boldsymbol{m}_{(0)} = 0, \\ & \partial_t \boldsymbol{m}_{(0)} + \operatorname{div}_{\boldsymbol{x}} \left(\frac{\boldsymbol{m}_{(0)} \otimes \boldsymbol{m}_{(0)}}{z_{(0)} - b} \right) + \nabla_{\boldsymbol{x}} p_{(2)} = -z_{(2)} \nabla_{\boldsymbol{x}} \eta^b. \end{aligned}$$

Thus, the well-prepared initial data can be defined as follows.

Definition 5.2.2. *For the 2d SWE* (5.1)*, we call the initial data* $(z_{0,\varepsilon}, \boldsymbol{m}_{0,\varepsilon})$ *well-prepared if it holds that*

$$\begin{aligned} z(0,\cdot) &= z_{0,\varepsilon} = z^0_{(0)} + \varepsilon^2 z^0_{(2),\varepsilon}, \\ \boldsymbol{m}(0,\cdot) &= \boldsymbol{m}_{0,\varepsilon} = \boldsymbol{m}^0_{(0)} + \varepsilon \boldsymbol{m}^0_{(1),\varepsilon}, \end{aligned} \tag{5.10}$$

where $z^0_{(0)}$ *is constant and* $\operatorname{div}_{\boldsymbol{x}} \boldsymbol{m}^0_{(0)} = 0$.

Considering well-preparedness for the initial datum, we pose the main theorem of this chapter.

Theorem 5.2.3. *For the 2d SWE with topography and a well-prepared initial datum in a periodic domain, the RS-IMEX scheme* (5.6a)–(5.6b)*, with* (5.3a)–(5.3d)*, the zero-Froude limit reference solution, a constant* $\widetilde{\alpha}$*, and under an* ε*-uniform time step restriction*

(i) is solvable, i.e., it has a unique solution for all $\varepsilon > 0$.

(ii) is consistent with the asymptotic limit in the fully-discrete settings, i.e., it is asymptotically consistent.

(iii) is asymptotically ℓ_2*-stable for the fixed grid, in finite time* $T_f < \infty$ *and with small enough initial data provided the reference solution is stable, i.e., there exists a constant* C_{N,T_f} *such that* $\|\boldsymbol{V}^n_\Delta\|_{\ell_2} \leq C_{N,T_f} \|\boldsymbol{V}^0_\Delta\|_{\ell_2}$.

(iv) preserves the lake at rest equilibrium state, provided that both $\overline{\boldsymbol{U}}_\Delta$ *and* $\boldsymbol{V}_\Delta$ *are at equilibrium.*

(v) may produce checker-board oscillations for the surface perturbation only as small as $\mathcal{O}(\varepsilon^2)$.

We discuss the proof of Theorem 5.2.3 in the next sections.

5.2.1 Solvability

Assuming $\widetilde{\alpha} = 0$ and $\Delta x = \Delta y$ for simplicity, one can write the coefficient matrix of the implicit step, J_ε, as

$$J_\varepsilon = \begin{bmatrix} \mathbb{I}_{N^2} & \frac{\beta}{\varepsilon^2} J_{12} & \frac{\beta}{\varepsilon^2} J_{13} \\ \beta J_{21} & \mathbb{I}_{N^2} + \beta J_{22} & \beta J_{23} \\ \beta J_{31} & \beta J_{32} & \mathbb{I}_{N^2} + \beta J_{33} \end{bmatrix}, \tag{5.11}$$

with $\beta := \frac{\Delta t}{2\Delta x}$ and where all J_{ij} are $\mathcal{O}(1)$; we can write the blocks more explicitly as

$$\begin{aligned} &J_{12} = Q_x, && && J_{13} = Q_y, && \\ &J_{21} = \operatorname{diag}(Q^b_x) + Q^{\overline{h}}_x - \varepsilon^2 (Q_x^{\overline{u_1}^2} + Q^{\overline{u_1 u_2}}_y), && J_{22} = 2Q^{\overline{u_1}}_x + Q^{\overline{u_2}}_y, && J_{23} = Q^{\overline{u_1}}_y, \\ &J_{31} = \operatorname{diag}(Q^b_y) + Q^{\overline{h}}_y - \varepsilon^2 (Q^{\overline{u_1 u_2}}_x + Q_y^{\overline{u_2}^2}), && J_{32} = Q^{\overline{u_2}}_x, && J_{33} = Q^{\overline{u_1}}_x + 2Q^{\overline{u_2}}_y, \end{aligned} \tag{5.12}$$

where Q^ϕ_x and Q^ϕ_y stand for corresponding matrices of central discretisation of the variable ϕ in each direction and $\operatorname{diag}(Q^b_x)$ and $\operatorname{diag}(Q^b_y)$ are diagonal matrices with central discretisation of b

as entries (like Chapter 2); compare J_ε with [Bis15, eq. (6.99)], which makes clear that the main difference is that $\overline{\boldsymbol{u}} \neq \mathbf{0}$.

So, the matrix J_ε, which is the inverse of the solution operator of the implicit step (5.6b), can be rewritten as $J_\varepsilon := \mathbb{I}_{3N^2} + \beta \Xi_\varepsilon$, where Ξ_ε is a matrix not depending on β. Hence, with a suitable choice of β, none of the eigenvalues of $\beta\Xi_\varepsilon$ is equal to -1, implying that J_ε is non-singular, and the implicit step, so the whole scheme, is solvable. The proof for $\widetilde{\alpha} \neq 0$ is likewise.

5.2.2 Asymptotic consistency

Like Chapter 4, we discuss the asymptotic consistency of the scheme in two ways, rigorously and formally. At first, we investigate the ε-stability of the solution, that is if the scaled perturbation is $\mathcal{O}(1)$. Then, we do the formal asymptotic consistency analysis, which would be rigorous, in virtue of the ε-stability. We assume $\widetilde{\alpha} = 0$ for the sake of simplicity.

5.2.2.1 ε-stability of the implicit step operator

For the ε-stability of the solution, in addition to the formal asymptotic analysis of the explicit step, one needs to show that the solution of the implicit step is ε-stable. Like Chapter 4, we, firstly, show that J_ε has a bounded inverse in terms of ε. But, unlike the 1d case, $\overline{\boldsymbol{T}}_\Delta$ does not necessarily vanish and may change the order of the rhs in the implicit step. More precisely, it can be seen from (5.7) that $\overline{\boldsymbol{T}}_{1,\Delta}$ is $\mathcal{O}(1/\varepsilon^2)$ since the projected velocity field on the grid may not have a zero divergence. This inexact discrete divergence makes $\overline{\boldsymbol{T}}_{1,\Delta}$ to be unbounded in the limit; so, the boundedness of $\lim_{\varepsilon\to 0} J_\varepsilon^{-1}$ is not sufficient to conclude ε-stability of the solution. Hence, we should also analyse the structure of the blocks of J_ε^{-1} in the next step.

After some manipulations, one can confirm that the following holds for the numerical range $W(J_\varepsilon^* J_\varepsilon)$:

$$
\begin{aligned}
W(J_\varepsilon^* J_\varepsilon) = &\left\| \boldsymbol{w}_1 + \frac{\beta}{\varepsilon^2} J_{12}\boldsymbol{w}_2 + \frac{\beta}{\varepsilon^2} J_{13}\boldsymbol{w}_3 \right\|_{\ell_2}^2 + \left\| \boldsymbol{w}_2 + \beta J_{21}\boldsymbol{w}_1 + \beta J_{22}\boldsymbol{w}_2 + \beta J_{23}\boldsymbol{w}_3 \right\|_{\ell_2}^2 \\
&+ \left\| \boldsymbol{w}_3 + \beta J_{31}\boldsymbol{w}_1 + \beta J_{32}\boldsymbol{w}_2 + \beta J_{33}\boldsymbol{w}_3 \right\|_{\ell_2}^2,
\end{aligned}
$$

where $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3 \in \mathbb{C}^N$ and $\|\boldsymbol{w}_1\|_{\ell_2}^2 + \|\boldsymbol{w}_2\|_{\ell_2}^2 + \|\boldsymbol{w}_3\|_{\ell_2}^2 = 1$. Defining $\mathcal{N}_M^{\varepsilon^2} := \left\{ \boldsymbol{w} \,|\, \|M\boldsymbol{w}\| = \mathcal{O}(\varepsilon^2) \right\}$, it is clear that for $\boldsymbol{w}_2 \notin \mathcal{N}_{J_{12}}^{\varepsilon^2}$ or $\boldsymbol{w}_3 \notin \mathcal{N}_{J_{13}}^{\varepsilon^2}$, the numerical range $W(J_\varepsilon^* J_\varepsilon)$ is bounded away from zero. Otherwise, we can conclude the result by contradiction, as follows, by assuming that the bottom function is constant.

If the numerical range $W(J_\varepsilon^* J_\varepsilon)$ approaches zero in the limit, it implies that

$$
\boldsymbol{w}_1 = -\frac{\beta}{\varepsilon^2}(J_{12}\boldsymbol{w}_2 + J_{13}\boldsymbol{w}_3) + o(1), \tag{5.13a}
$$

$$
(\mathbb{I}_{N^2} + \beta J_{22})\,\boldsymbol{w}_2 = -\beta\,(J_{21}\boldsymbol{w}_1 + J_{23}\boldsymbol{w}_3) + o(1), \tag{5.13b}
$$

$$
(\mathbb{I}_{N^2} + \beta J_{33})\,\boldsymbol{w}_3 = -\beta\,(J_{31}\boldsymbol{w}_1 + J_{32}\boldsymbol{w}_2) + o(1). \tag{5.13c}
$$

Then, one can check that in the limit $\varepsilon \to 0$, the matrix J_{21} obtains a skew-symmetric pattern. Furthermore, for the flat bottom case, the structure of J_{21} and J_{12} are the same, up to a scaling

(see equation (5.12)). This implies that in the limit, $J_{21}J_{12}$ (similarly $J_{31}J_{13}$) is *almost* symmetric and diagonally-dominant, up to higher order terms $\mathcal{O}(\varepsilon^2)$. Adding $\mathbb{I}_N^2+\beta J_{22}$ (similarly $\mathbb{I}_N+\beta J_{33}$), with a suitably chosen β, makes the sum *strictly* diagonally dominant (SDD). So, due to the result of [Var75], the matrices $M_2 := (\mathbb{I}_{N^2}+\beta J_{22}-\frac{\beta^2}{\varepsilon^2}J_{21}J_{12})$ and $M_3 := (\mathbb{I}_{N^2}+\beta J_{33}-\frac{\beta^2}{\varepsilon^2}J_{31}J_{13})$ have bounded inverses, *i.e.*,

$$\lim_{\varepsilon\to 0}\left\|\Big(\mathbb{I}_{N^2}+\beta J_{22}-\frac{\beta^2}{\varepsilon^2}J_{21}J_{12}\Big)^{-1}\right\|<\infty,\qquad \lim_{\varepsilon\to 0}\left\|\Big(\mathbb{I}_{N^2}+\beta J_{33}-\frac{\beta^2}{\varepsilon^2}J_{31}J_{13}\Big)^{-1}\right\|<\infty. \quad \text{(O1)}$$

Owing to (O1), and by manipulating (5.13b)–(5.13c), one can justify the following relations

$$\begin{aligned}
&\boldsymbol{w}_2 = C_2\boldsymbol{w}_3+o(1), \qquad C_2 := \Big(\mathbb{I}_{N^2}+\beta J_{22}-\frac{\beta^2}{\varepsilon^2}J_{21}J_{12}\Big)^{-1}\Big(\frac{\beta^2}{\varepsilon^2}J_{21}J_{13}-\beta J_{23}\Big),\\
&\boldsymbol{w}_3 = C_3\boldsymbol{w}_2+o(1), \qquad C_3 := \Big(\mathbb{I}_{N^2}+\beta J_{33}-\frac{\beta^2}{\varepsilon^2}J_{31}J_{13}\Big)^{-1}\Big(\frac{\beta^2}{\varepsilon^2}J_{31}J_{12}-\beta J_{32}\Big),
\end{aligned} \tag{5.14}$$

where the matrices C_2 and C_3 are $\mathcal{O}(1)$ as M_2 and M_3 make the large terms $\frac{\beta^2}{\varepsilon^2}J_{21}J_{13}$ and $\frac{\beta^2}{\varepsilon^2}J_{31}J_{12}$ to vanish; this can be be proved along the similar lines as in Chapter 2.

If, like [Bis15], one additionally assumes $\overline{\boldsymbol{u}}=\boldsymbol{0}$, the blocks of J_ε will get simplified extensively as $J_{22}=J_{33}=J_{23}=J_{32}=\boldsymbol{0}_{N^2}$, and the analysis will be much simpler because the relations (5.13b) and (5.13c) yield

$$\boldsymbol{w}_2 = -\beta\overline{h}J_{12}\boldsymbol{w}_1+o(1), \qquad \boldsymbol{w}_3 = -\beta\overline{h}J_{12}\boldsymbol{w}_1+o(1).$$

Then, the leading order of (5.13a) indicates $J_{12}\boldsymbol{w}_2^{(0)}+J_{13}\boldsymbol{w}_3^{(0)}=\boldsymbol{0}$. So,

$$\big(J_{12}^2+J_{13}^2\big)\boldsymbol{w}_1^{(0)}=0,$$

which gives a classical central discretisation of the Poisson equation, and implies that $\boldsymbol{w}_1^{(0)}$ should lie in the null spaces of J_{12} and J_{13}. Thus $\boldsymbol{w}_2^{(0)}=\boldsymbol{w}_3^{(0)}=\boldsymbol{0}$.

Because $J_{12}\boldsymbol{w}_1^{(0)}=\boldsymbol{0}$ and $J_{13}\boldsymbol{w}_1^{(0)}=\boldsymbol{0}$, the vector $\boldsymbol{w}_1^{(0)}$ is either constant or has a checkerboard (CB) like structure. It is helpful to see $\boldsymbol{w}_1^{(0)}$ as a lexicographically-ordered array of the following matrix:

$$\begin{bmatrix}
\square & \blacktriangle & \square & \blacktriangle & \dots\\
\blacksquare & \triangle & \blacksquare & \triangle & \dots\\
\square & \blacktriangle & \square & \blacktriangle & \dots\\
\blacksquare & \triangle & \blacksquare & \triangle & \dots\\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix} \tag{5.15}$$

This implies that $\boldsymbol{w}_1^{(0)}$ has at most 4 degrees of freedom. On the other hand, $\boldsymbol{w}_1$ has a central difference structure due to (5.13a), so has a vanishing mean. Thus, in the light of (5.15), taking a sum on the leading order of (5.13a) for each type of the entries $(\triangle,\square,\blacktriangle,\blacksquare)$ and using the periodicity imply that the rhs vanishes while the lhs is proportional to $\boldsymbol{w}_1^{(0)}$. Hence $\boldsymbol{w}_1^{(0)}=\boldsymbol{0}$, and $\lim_{\varepsilon\to 0}(\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{w}_3)=(\boldsymbol{0},\boldsymbol{0},\boldsymbol{0})$ which contradicts the assumption that $\|\boldsymbol{w}\|_{\ell_2}=1$ and concludes the ε-stability of the implicit solution *operator* since the numerical range cannot tend to zero.

Remark 5.2.4. *(i) Note that we could show the rigorous proof of ε-stability of the implicit solution operator only for the case of flat bottom topography and with the zero reference solution, which were the assumptions in the asymptotic consistency proof in [Bis15].*

(ii) For the non-flat bottom case or the non-zero reference velocity field, the proof would be much more involved, e.g., the matrices C_2 and C_3 are not bounded anymore with a non-flat bottom topography; this makes the analysis more complicated. In that case, we would rely on the numerical evidence to conclude the result.

(iii) In numerical examples we will study later on in Section 5.3, $\lim_{\varepsilon\to 0}\|J_\varepsilon^{-1}\|$ has been checked to be bounded; so, we have this ε-stability. Based on the numerical evidence we suggest that the result would hold generally.

5.2.2.2 ε-stability of the implicit solution

So far, we have shown the ε-stability of the implicit solution *operator*. However, we need the ε-stability of the implicit step *solution* for the rigorous asymptotic consistency. As mentioned earlier, an obstacle to conclude it immediately from the ε-stability of the implicit solution operator is that the reference solution computed by the projection scheme does not satisfy the *div*-free condition exactly (even for the well-prepared initial data) and enters a truncation error into the rhs vector of the implicit step, *i.e.*, $\|\overline{\boldsymbol{T}}_{1,\Delta}\| = \mathcal{O}(\Delta x/\varepsilon^2)$, which is large for a small ε. So, having $\|J_\varepsilon^{-1}\| = \mathcal{O}(1)$ does not suffice to show that $\|\boldsymbol{V}_\Delta^{n+1}\| = \mathcal{O}(1)$. Hence, proving ε-stability gets more involved and requires making use of the structure of J_ε^{-1}, aiming to show that $\mathcal{O}(1/\varepsilon^2)$ terms in $\overline{\boldsymbol{T}}_{1,\Delta}$ are not present in the updated solution.

Remark 5.2.5. *Note that the issue with $\overline{\boldsymbol{T}}_{1,\Delta}$ has roots in the representation of the reference momentum field in the discrete space, which does not preserve necessarily the solenoidality of the momentum field. Note also that with this choice of d_2 and d_3, there is not such an issue with $\overline{\boldsymbol{T}}_{2,\Delta}$ and $\overline{\boldsymbol{T}}_{3,\Delta}$.*

Finding the structure of $J_\varepsilon^{-1} =: K_\varepsilon$, we use the identity $K_\varepsilon J_\varepsilon = \mathbb{I}_{3N^2}$, which can be written as follows by abusing the notations we have used previously in (5.11) for blocks of J_ε.

$$\begin{bmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{bmatrix} \begin{bmatrix} J_{11} & J_{12}/\varepsilon^2 & J_{13}/\varepsilon^2 \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{bmatrix} = \mathbb{I}_{3N^2},$$

which implies that for $i = 1, 2, 3$

$$K_{i1}J_{12} = \mathcal{O}(\varepsilon^2), \qquad K_{i1}J_{13} = \mathcal{O}(\varepsilon^2). \tag{5.16}$$

In other words, K_{i1} can be decomposed as an $\mathcal{O}(1)$ matrix whose null space contains the columns of J_{12} and J_{13}, plus a matrix which is $\mathcal{O}(\varepsilon^2)$. So, it remains to show that the former does cancel $\mathcal{O}(1/\varepsilon^2)$ terms in $\overline{\boldsymbol{T}}_{1,\Delta}$. Showing this, one needs to determine the null space of K_{i1} using the conditions in (5.16).

For the implicit step (5.6b), blocks J_{12} and J_{13} are companion matrices of central discretisations in x- and y-directions, respectively. The first condition in (5.16) implies that after multiplying each row of K_{i1} by each column of J_{12}, only $\mathcal{O}(\varepsilon^2)$ terms remain. As J_{12} consists of skew-symmetric circulant blocks $\boldsymbol{Circ}\,(0, 1, 0, \ldots, 0, -1)$, the condition (5.16) requires that for each

row, entries of odd and even columns of the first (and analogously other) $N_x \times N_y$ sub-block(s) of K_{i1} are equal to each other up to $\mathcal{O}(\varepsilon^2)$; but the values of these entries may be different for each sub-block and row. A similar argument with J_{13} implies that for each row of K_{i1}, the structure repeats for every other sub-block, *i.e.*, each row has a structure like

$$\square_{odd}\ \blacktriangle_{odd}\ \square_{odd}\ \blacktriangle_{odd}\ \dots \mid \square_{even}\ \blacktriangle_{even}\ \square_{even}\ \blacktriangle_{even}\ \dots \mid \square_{odd}\ \blacktriangle_{odd}\ \square_{odd}\ \blacktriangle_{odd}\ \dots$$

Hence, owing to the periodicity and since $\overline{\boldsymbol{T}}_{1,\Delta}$ consists of central differences, the leading order of $K_{i1}\overline{\boldsymbol{T}}_{1,\Delta}$ vanishes and some $\mathcal{O}(1)$ terms remain. This implies that no $\mathcal{O}(1/\varepsilon^2)$ term stays in the implicit update; thus, the solution of the implicit step is $\mathcal{O}(1)$ and the scheme is ε-stable, thanks to the ε-stability of the explicit step, which is the topic of the next section.

5.2.2.3 Formal asymptotic consistency

Confirming the ε-stability of the explicit step, we assume that $\|\boldsymbol{V}_\Delta^n\| = \mathcal{O}(1)$, which is compatible with the well-prepared initial data (in the sense of Definition 5.2.1), and confirm that $\|\boldsymbol{V}_\Delta^{n+1/2}\| = \mathcal{O}(1)$. Since $\widehat{\boldsymbol{G}}_{1,1} = \widehat{\boldsymbol{G}}_{2,1} = 0$, one can immediately conclude that $\|\boldsymbol{V}_{1,\Delta}^{n+1/2}\| = \mathcal{O}(1)$. For $\boldsymbol{V}_{2,\Delta}^{n+1/2}$ (and similarly $\boldsymbol{V}_{3,\Delta}^{n+1/2}$), one can simply confirm that

$$\lim_{\varepsilon\to 0}\left(\nabla_{h,x}\widehat{\boldsymbol{G}}_{1,2,ij}^n + \nabla_{h,y}\widehat{\boldsymbol{G}}_{2,2,ij}^n\right) = \mathcal{O}(1),$$

such that no $\mathcal{O}(1/\varepsilon^2)$ or $\mathcal{O}(1/\varepsilon)$ contribution would exist in the explicit update, as

$$\lim_{\varepsilon\to 0}\Bigg[\ \nabla_{h,x}\left(\frac{m_1^2}{z-b} + \frac{z^2-2zb}{2\varepsilon^2} - \frac{\overline{m_1}^2}{\overline{z}-b} - \frac{\overline{z}^2-2\overline{z}b}{2\varepsilon^2} + \frac{\overline{m_1}^2 v_1\varepsilon^2}{(\overline{z}-b)^2} - \frac{2\overline{m_1}v_2}{\overline{z}-b} - (\overline{z}-b)v_1\right)_{ij}^n$$
$$+\nabla_{h,y}\left(\frac{m_1m_2}{z-b} - \frac{\overline{m_1m_2}}{\overline{z}-b} + \frac{\overline{m_1m_2}v_1\varepsilon^2}{(\overline{z}-b)^2} - \frac{\overline{m_1}v_3}{\overline{z}-b} - \frac{\overline{m_2}v_2}{\overline{z}-b}\right)_{ij}^n\Bigg] = \mathcal{O}(1).$$

So, the explicit step does not change the leading order of $\boldsymbol{V}_{2,\Delta}^n$ and analogously $\boldsymbol{V}_{3,\Delta}^n$. This completes the ε-stability proof of the previous section.

To confirm asymptotic consistency, we consider the implicit step and show that the limit of the solution is consistent with the limit manifold. From the v_1-update and considering (5.7) and (5.3a)–(5.3c), one can simply check that the momentum field is solenoidal, *i.e.*,

$$\nabla_{h,x}\left(\overline{m_1} + v_2\right)_{ij}^{n+1} + \nabla_{h,y}\left(\overline{m_2} + v_3\right)_{ij}^{n+1} = \mathcal{O}(\varepsilon^2). \tag{5.17}$$

The interesting point is that although the divergence of the reference momentum field is expected to be $\mathcal{O}(\Delta x)$, the solver for the perturbation compensates this issue and makes the divergence to vanish as $\varepsilon \to 0$.

Since proving the consistency of the evolution of the leading order of the momentum is straightforward, the asymptotic consistency of the scheme is concluded, but only up to possible oscillations for the momentum field in the null space of central difference operators $\nabla_{h,x}$ and $\nabla_{h,y}$ which leads to potential checker-board oscillations.

Remark 5.2.6. *The ε-stability of the solution implies immediately that since $\|\boldsymbol{V}_{1,\Delta}^{n+1}\| = \mathcal{O}(1)$, the possible checker-board oscillations for the surface perturbation are $\mathcal{O}(\varepsilon^2)$. This seems to solve the problem in [KSSN16] regarding the checker-board oscillations in a periodic domain. This is why we stated in Remark 4.2.4 that using the reformulated scheme* (4.7) *is more illustrative; in essence, it makes the perturbation variable $\boldsymbol{V}_\Delta$ more accessible for the asymptotic analysis. This result seems to suggest that it is not necessary to add a large diffusion for precluding checker-board oscillations. This is in contradiction with [KSSN16, KS17, Bis15], where adding a large $\mathcal{O}(1/\varepsilon^2)$ numerical stabilisation term to the continuity equation has been proposed, although those schemes are not literally the same as the one analysed here. Note that such a large $\mathcal{O}(\Delta x/\varepsilon^2)$ diffusion makes the scheme excessively diffusive and degrades its accuracy unless one takes $\Delta x \sim \varepsilon^2$ (resolved grid), which is not practical and contradicts the AP property of the scheme [Jin16].*

5.2.3 Asymptotic stability

The proof of the asymptotic stability of the 2d RS-IMEX scheme is very similar to the 1d case, which has been explained in detail in Chapter 4. There are two basic elements required for the proof, ε-stability of the implicit step and a non-linear ε-uniform bound for the explicit one, alongside with the assumption of positivity which is justified in the $\varepsilon \ll 1$ regime. Since the reference solution is not stationary and should be computed in time, one should also obtain some estimate for the solution of the reference solver. For now, we simply assume that the reference solver is stable in a proper sense such that the computed reference solution is bounded in a norm; see [GMS06] and the references therein for further details. This implies that the residual of the reference solution should be bounded as well, *i.e.*, there are constants such that $\|\overline{\boldsymbol{T}}_\Delta^k\| \leq \overline{c}^k$. Thus, very similar to Chapter 4, one can estimate the norm of the computed solution for the step k as (assuming constants not depending on the time for simplicity)

$$y_{k+1} \leq (1 + c_2\Delta t)(1 + c_1\Delta t y_k)y_k + \Delta t(1 + c_2\Delta t)\overline{c}.$$

The bound for the first term in the rhs is exactly like in Chapter 4. The second term can be simply bounded as $\sum_{j=0}^{k} \Delta t(1 + c_2\Delta t)^j \overline{c}$.

Lemma 5.2.7. *Given a small enough initial datum and for $\varepsilon \ll 1$, the RS-IMEX scheme* (5.6a)–(5.6b) *is ℓ_2-stable in finite time, i.e., $\|\boldsymbol{V}_\Delta^n\|_{\ell_2} \leq C_{N,T_f}\|\boldsymbol{V}_\Delta^0\|_{\ell_2}$.*

Proof. Like Chapter 4, the boundedness of J_ε^{-1} confirms that the implicit operator is power-bounded for a finite time $T_f < \infty$. For the explicit step, the proof can be carried out using a discrete Grönwall's inequality [WW65], which would be very similar to Chapter 4. □

5.2.4 Well-balancing

Showing the well-balancing of schemes for the LaR equilibrium state, it is crucial to check if the discretisations of the source terms are consistent with of the pressure flux. For the RS-IMEX scheme (5.6a)–(5.6b), as the reference surface perturbation is a constant value and both of these terms have been discretised by central differences, it is clear that the discretisations are consistent with each other. Nonetheless, this consistency may not be enough in proving well-balancing as the reference momentum is not constant. Clarifying this, note that to prove the well-balancing

of a scheme, one assumes that the solution is initially on the equilibrium manifold and shows that the updated solution is also at equilibrium. So, let us assume that $\boldsymbol{U}_\Delta^n \in \mathcal{U}_{LaR}^\Delta$, defined as

$$\mathcal{U}_{LaR}^\Delta := \left\{ \begin{bmatrix} z_{ij} \\ m_{1,ij} \\ m_{2,ij} \end{bmatrix} \,\middle|\, z_{ij} = \eta^s - b_{ij}, \boldsymbol{m}_{ij} = \boldsymbol{0}, \forall (i,j) \in \Omega_N \right\}, \tag{5.18}$$

for a constant water surface level η^s. Referring back to Algorithm 1, one can see that $\boldsymbol{U}_\Delta^n$ has two parts, and belonging to $\mathcal{U}_{LaR}^\Delta$ enforces a constraint on the sum of $\overline{\boldsymbol{U}}_\Delta^n$ and $\boldsymbol{V}_\Delta^n$. As $\overline{\boldsymbol{U}}_\Delta^n$ is not generally constant, assuming $\boldsymbol{U}_\Delta^n \in \mathcal{U}_{LaR}^\Delta$ does not determine if $\overline{\boldsymbol{U}}_\Delta^n \in \mathcal{U}_{LaR}^\Delta$ or $\boldsymbol{V}_\Delta^n \in \mathcal{U}_{LaR}^\Delta$. So, one should consider a non-zero momentum field in the well-balancing analysis, for both the reference solution and the perturbation. Thus, non-stationary equilibrium states should be taken into account whose discrete preservation is much more complicated to be studied, *cf.* [NXS07]. This is, in fact, the price we are paying for decomposing the solution, which was helpful for the asymptotic consistency analysis.

This issue with the well-balancing analysis does not affect the results of [BALMN14, Bis15] and Chapter 4 since they employ the LaR reference solution which is a constant state. Also, note that the same difficulty would happen even for the flat bottom case.

Keeping the integrity of the chapter, we discuss some observations as well as a remedy regarding well-balancing of the RS-IMEX scheme in Appendix 5.B. Here, we only wish to show that assuming both parts of the solution to be at equilibrium enforces the updated solution to be also at equilibrium, so the well-balancing of the scheme.

Lemma 5.2.8. *For the RS-IMEX scheme* (5.6a)–(5.6b) *in a periodic domain, if* $\overline{\boldsymbol{U}}_\Delta^n, \boldsymbol{V}_\Delta^n \in \mathcal{U}_{LaR}^\Delta$ *then* $\overline{\boldsymbol{U}}_\Delta^{n+1}, \boldsymbol{V}_\Delta^{n+1} \in \mathcal{U}_{LaR}^\Delta$. *So, the scheme is well-balanced regarding the lake at rest equilibrium state.*

Proof. For the reference solution, it is clear that $\boldsymbol{m}_\Delta^\star = \boldsymbol{0}$, so $\Delta_{h,\boldsymbol{x}} \pi_\Delta^{n+1} = \boldsymbol{0}$, which only has constant solutions in a periodic domain. Thus, $\boldsymbol{m}_\Delta^{n+1} = \boldsymbol{m}_\Delta^n = \boldsymbol{0}$, $\overline{\boldsymbol{U}}_\Delta^{n+1} \in \mathcal{U}_{LaR}^\Delta$, and the projection scheme preserves the LaR equilibrium state. For the explicit part of the scheme, as $\overline{z}$ and v_1^n are constant, one is left with $(-zb + \overline{z}b + \varepsilon^2 b v_1)/\varepsilon^2$ in the explicit flux of the momentum equation, which is zero; so, $\boldsymbol{V}_\Delta^{n+1/2} \in \mathcal{U}_{LaR}^\Delta$. For the implicit step, $\overline{\boldsymbol{T}}_\Delta^{n+1} = \boldsymbol{0}$ and the LaR solution is compatible as the only non-zero term it leaves in the stiff flux is the pressure flux $(z-b)v_1$, which is discretised consistently with the source term. Since the scheme is solvable, the LaR solution is necessarily the unique solution of the scheme, *i.e.*, $\boldsymbol{V}_\Delta^{n+1} \in \mathcal{U}_{LaR}^\Delta$. □

5.3 Numerical experiments

In this section, we verify the quality of the computed solutions by the RS-IMEX scheme and confirm the asymptotic analysis of Section 5.2.3 by the help of several numerical examples. At first, we test the performance of the scheme for the $\varepsilon = \mathcal{O}(1)$ regime, in Sections 5.3.1 and 5.3.2. Then, in Sections 5.3.3 and 5.3.4, we consider examples with $\varepsilon \ll 1$ and illustrate the asymptotic preserving property of the scheme. The time step is computed as in Section 4.5, with $c_{\widetilde{\alpha}} = 0$ and $c_{\widehat{\alpha}} = 1$ (like [Bis15]), and with an additional constraint for the time step required for the reference solver.

5.3.1 (i) 2d quasi-stationary states

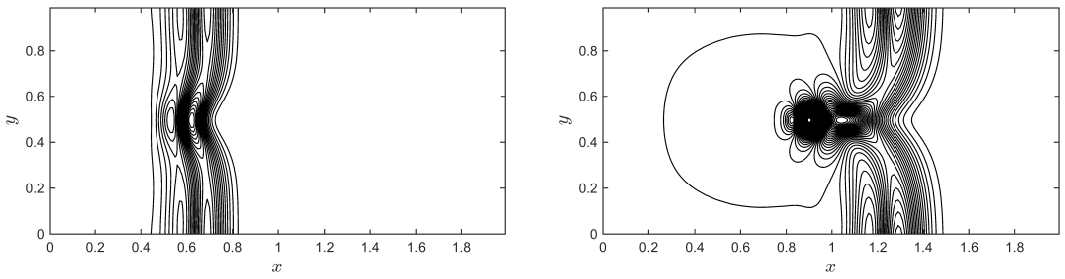
This example is used in [LMNK07] to test the preservation of a stationary steady state and the approximation of small perturbations around it. Consider the domain $[0, 2] \times [0, 1]$ with open boundaries, where the bottom topography is given by

$$\eta^b(x, y) = 0.8 \exp(-5(x - 0.9)^2 - 50(y - 0.5)^2),$$

the mean water level is set as $H_{\text{mean}} = 1$, $\varepsilon = 1$, and the initial data are

$$z(0, x, y) = 0.01\ \mathbf{1}_{[0.05 \le x \le 0.15]},$$
$$u_1(0, x, y) = u_2(0, x, y) = 0.$$

We pick the reference solution as zero and CFL $= 0.45$. In Figure 5.1, we present the contours of the surface perturbation, computed on the 200×100 grid. The initial perturbation propagates without any oscillations until it reaches the hump. As the wave speed is slower over the hump, due to the smaller water height, the initially planar perturbation gets distorted. Also, note that the solution is symmetric in the y-direction. The solution of the RS-IMEX scheme matches very well the existing results like [LMNK07, NPPN06].



(a) $T_f = 0.6$. (b) $T_f = 1.2$.

Figure 5.1: Solution of the RS-IMEX scheme for the 2d quasi-stationary example (i): Surface perturbation computed on the 200×100 grid, with CFL $= 0.45$.

5.3.2 (ii) 2d Riemann problem

Similar to [HJL12] and inspired by the well-known examples of [LL98], we also run a test on a 2d Riemann problem in the domain $[0, 1]^2$ with open boundaries and the following initial conditions in each quadrant of the domain:

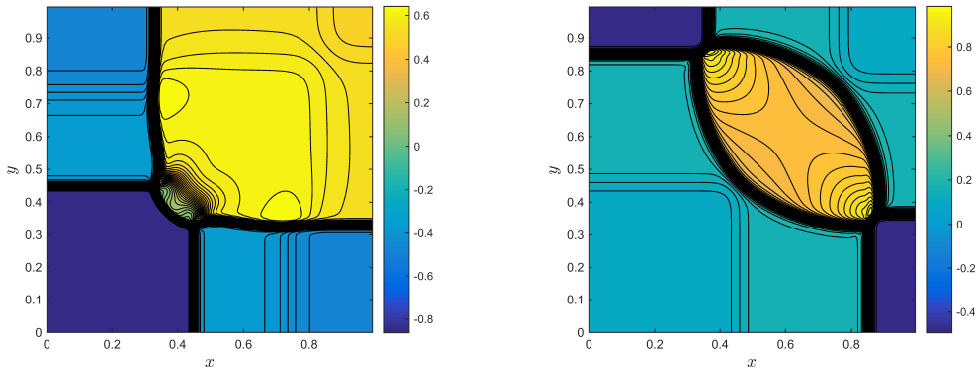
$h(0) = 0.5323$	$h(0) = 1.5$
$u_1(0) = 1.206$	$u_1(0) = 0$
$u_2(0) = 0$	$u_2(0) = 0$
$h(0) = 0.138$	$h(0) = 0.5323$
$u_1(0) = 1.206$	$u_1(0) = 0$
$u_2(0) = 1.206$	$u_2(0) = 1.206$

(Configuration 3)

$h(0) = 0.5065$	$h(0) = 1.1$
$u_1(0) = 0.8939$	$u_1(0) = 0$
$u_2(0) = 0$	$u_2(0) = 0$
$h(0) = 1.1$	$h(0) = 0.5065$
$u_1(0) = 0.8939$	$u_1(0) = 0$
$u_2(0) = 0.8939$	$u_2(0) = 0.8939$

(Configuration 4)

We set $\varepsilon = 1$ and $H_{\text{mean}} = 1$, and choose the reference solution as zero. Figure 5.2 shows the results of these two configurations, computed on the 150×150 grid and with CFL $= 0.45$. As the figures suggest, these configurations result in four shock waves. Also, note that the solutions have the symmetry w.r.t. the diagonal.



(a) Configuration 3 ($T_f = 0.3$).

(b) Configuration 4 ($T_f = 0.25$).

Figure 5.2: Solution of the RS-IMEX scheme for different configurations of the 2d Riemann problem (ii): Surface perturbation computed on the 150×150 grid, with CFL $= 0.45$.

5.3.3 (iii) Periodic flow

This example was used in [DT11] and with a slight difference in [HJL12]. Here, we use the settings of [DT11]; we consider the periodic domain $[0, 1)^2$ with a flat bottom topography $b(x, y) = -1$, and with the following well-prepared initial data (see Figure 5.3):

$$\begin{aligned} z(0, x, y) &= \varepsilon^2 \sin^2(2\pi(x + y)), \\ m_1(0, x, y) &= \sin(2\pi(x - y)) + \varepsilon^2 \sin(2\pi(x + y)), \\ m_2(0, x, y) &= \sin(2\pi(x - y)) + \varepsilon^2 \cos(2\pi(x + y)). \end{aligned}$$

We decompose the initial momentum field and pick the solenoidal leading order part as the initial reference momentum field. The solutions for $\varepsilon \in \{0.8, 0.05\}$ have been plotted in Figure 5.4, which are computed on the same spatial grid as in [DT11] with CFL $= 0.45$. The figures suggest that the solution of the RS-IMEX scheme is comparable to the Degond–Tang method [DT11].

Experimental order of convergence Studying the asymptotic accuracy of the scheme, we check the experimental order of convergence for two different values of ε. The error is computed with the help of a numerical solution on a fine reference grid, $N_{\text{ref},x} = N_{\text{ref},y} = 320$, as the "*reference*" solution; we define the error in the L_1-norm as

$$e(\phi_\Delta^{\text{num}}) := \|\phi_\Delta^{\text{num}} - \phi_\Delta^{\text{ref}}\|_{L_1(\Omega_{N_{\text{ref}}})} = \frac{1}{N_{\text{ref}}^2} \sum_{(i,j)\in\Omega_{N_{\text{ref}}}} \left|\phi_{ij}^{\text{num}} - \phi_{ij}^{\text{ref}}\right|, \tag{5.19}$$

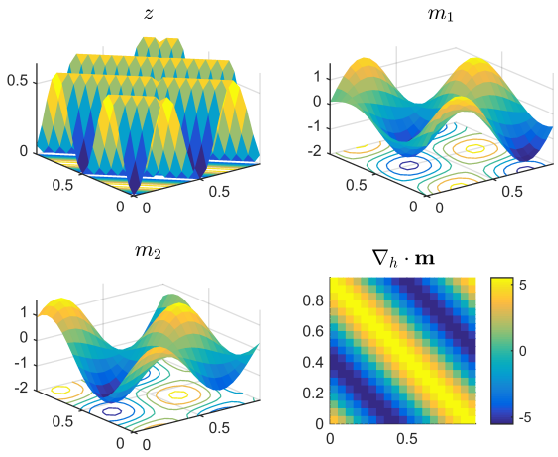
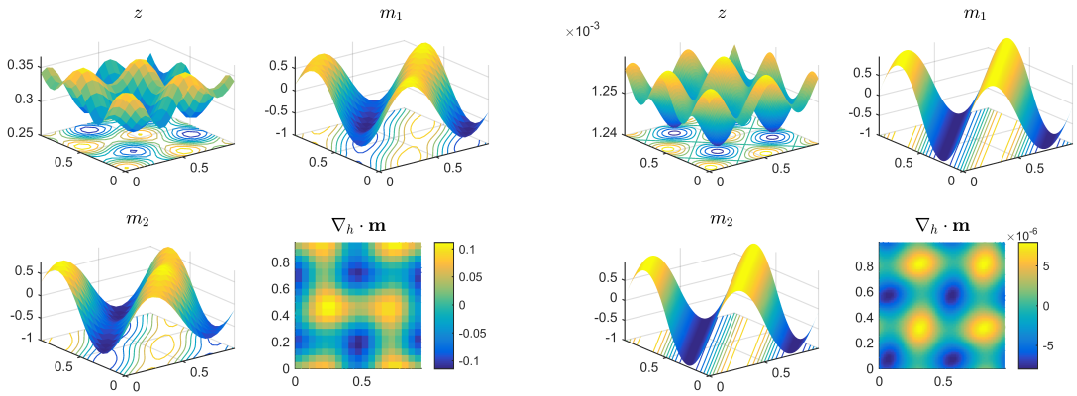


Figure 5.3: Initial condition for the periodic flow example (iii) with $\varepsilon = 0.8$, computed on the 80×80 grid.



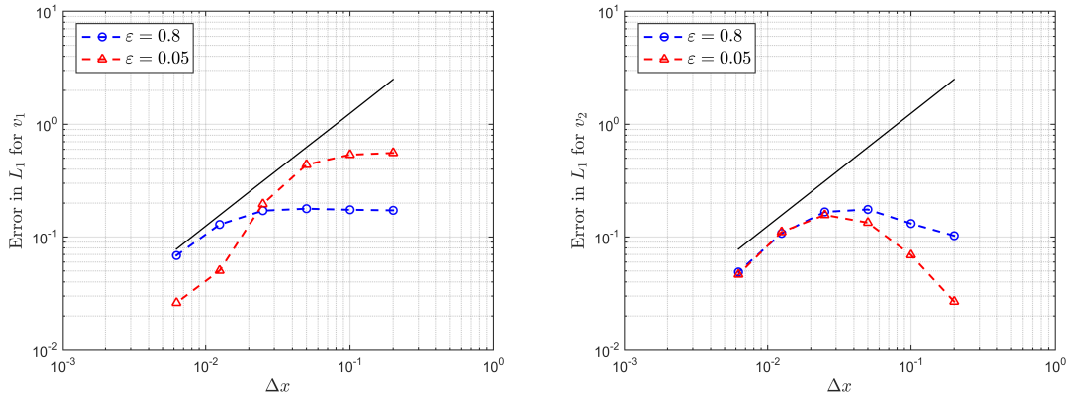
(a) $\varepsilon = 0.8$ on the 20×20 grid.

(b) $\varepsilon = 0.05$ on the 80×80 grid.

Figure 5.4: Solution of the RS-IMEX scheme for the periodic flow example (iii), with CFL $= 0.45$ and $T_f = 1$.

where ϕ is the variable of interest (momentum, height, etc.) and ϕ_Δ^{num} and ϕ_Δ^{ref} are respectively the numerical solution given by the scheme and the "*reference*" solution. Figure 5.5 confirms that the EOC is ε-uniform for the scaled perturbations v_1 and v_2 (similarly for v_3).

Asymptotic preserving property To confirm the asymptotic consistency of the scheme numerically, we obtain the solution for $\varepsilon = 5 \times 10^{-6}$ with $T_f = 0.1$ and on the 80×80 grid. As shown in Figure 5.6, the solution seems to be consistent with the limit, *i.e.*, the divergence is zero up to the machine accuracy, and the water surface is almost constant up to some small oscillations of order $10^{-12} \sim \varepsilon^2$.



(a) EOC for v_1.

(b) EOC for v_2.

Figure 5.5: Experimental order of convergence of the RS-IMEX scheme for the periodic flow example (iii), with CFL = 0.45 and $T_f = 1$.

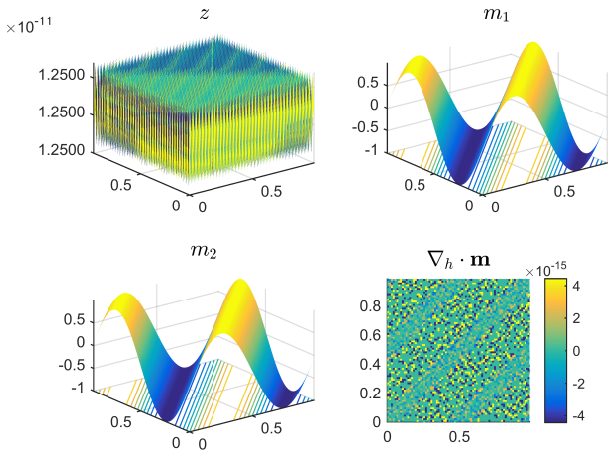


Figure 5.6: Solution of the RS-IMEX scheme for the periodic flow example (iii) with $\varepsilon = 5 \times 10^{-6}$, CFL = 0.45 and $T_f = 1$, computed on the 80×80 grid.

Efficiency of the RS-IMEX scheme A very natural question may arise regarding the efficiency of the scheme; one may argue that the scheme is not efficient since two independent solutions should be computed in time, the reference solution and its perturbation, which doubles the computational cost. It has been explained in [KSSN16] that since the RS-IMEX scheme is more accurate on the same grid (at least compared to the scheme in [HJL12] and for examples in Section 5.3.3 and Section 5.3.4), it compensates that drawback as one can use a coarser mesh.

Moreover, as mentioned in Remark 4.2.4, there is an important difference between the RS-IMEX scheme in [KSSN16] and the RS-IMEX scheme in this manuscript, which is how to find the evolution of the reference solution in time. We use an established efficient method for this purpose (like the projection method) while in [KSSN16] the authors solve the limit system with an implicit step with a huge computational cost. For this reason, the scheme presented here

does not suffer from being costly. This can be confirmed by Table 5.1, which shows that the cost of computing the reference solution is not comparable to the whole CPU time. The bottleneck is, in fact, the implicit solver; however, note that the backslash operator of Matlab™ has been used for this purpose. This operator employs an LU decomposition, which explains the cost of the implicit solver. As the comparison in Table 5.1 shows, employing an iterative method like GMRES (without pre-conditioning) may circumvent this issue, while affecting the robustness of the code since there are several parameters to be tuned *a priori*.

Table 5.1: CPU time comparison (in seconds) for different ε on the fixed 80×80 grid and for $T_f = 1$ and $\Delta t/\Delta x = 0.25$ in the periodic flow example (iii).

	Total	Implicit step LSE solver	Poisson solver for π_Δ^{n+1}
LU	322.833	178.293 (55.2%)	0.200 (0.1%)
GMRES	240.911	41.257 (17.1%)	0.228 (0.1%)

(a) $\varepsilon = 0.8$.

	Total	Implicit step LSE solver	Poisson solver for π_Δ^{n+1}
LU	338.938	194.766 (57.5%)	0.198 (0.1%)
GMRES	241.308	53.318 (22.1%)	0.387 (0.2%)

(b) $\varepsilon = 0.05$.

Consistency of the reference solver One can check that since the initial velocity field is *div*-free and $u_1(0,\cdot,\cdot) = u_2(0,\cdot,\cdot)$, the exact solution for the reference system is steady state and does not evolve over time. Note that since $m_1(0,\cdot,\cdot) = m_2(0,\cdot,\cdot)$, the equation for m_1 writes

$$\partial_t \overline{m_1} + \frac{\overline{m_1}}{\overline{z}-b}\nabla_x \cdot \overline{\boldsymbol{m}} + \underbrace{\frac{\overline{m_1}}{\overline{z}-b}\partial_x \overline{m_1} + \frac{\overline{m_2}}{\overline{z}-b}\partial_y \overline{m_1}}_{\frac{\overline{m_1}}{\overline{z}-b}(\partial_x \overline{m_1} + \partial_y \overline{m_2}) = 0} = 0. \tag{5.20}$$

Thus, $\partial_t \overline{m_1} = 0$, and similarly $\partial_t \overline{m_2} = 0$.

On the other hand, the projection scheme does not preserve this steady state exactly, *i.e.*, the discrete version of (5.20), so $\boldsymbol{m}_\Delta^\star = \boldsymbol{m}_\Delta^0$, does not hold. This is, basically, due to the numerical diffusion required for stability, which has first-order consistency. So, the scheme adds $\mathcal{O}(\Delta t \Delta x)$ disturbances at each step, which leads to preserving the equilibrium approximately (up to $\mathcal{O}(\Delta x)$). Table 5.2 confirms this and shows the difference between the computed reference solution and the exact solution w.r.t. mesh refinement. Note that this error is of the order of truncation errors and does not affect the formal consistency order of the scheme.

Table 5.2: Error in computation of the reference solution for the periodic flow example (iii), with $T_f = 1$ and on different grids.

$N_x \times N_y$	$\|\overline{u}_{1,\Delta}^{projection}(T_f) - \overline{u}_{1,\Delta}^{exact}(T_f)\|_{\ell_\infty}$	Order
20×20	6.55e-1	-
40×40	4.75e-1	0.46
80×80	3.11e-1	0.61
160×160	1.87e-1	0.73

Stability of the reference solver Figure 5.7 indicates the stability of the computed reference velocity field as its ℓ_∞-norm is bounded, where the initial data are set for $\varepsilon = 0.8$. Note that the two lines coincide with each other.

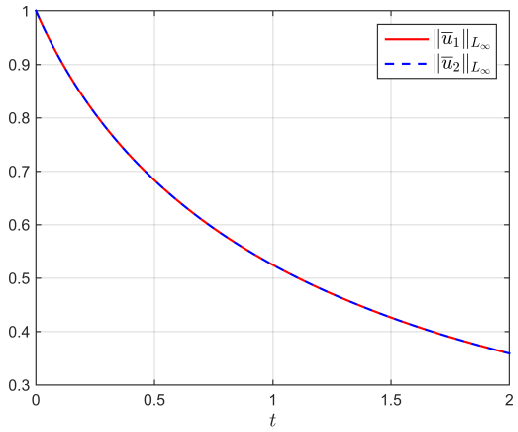


Figure 5.7: Stability of the projection scheme for the periodic flow example (iii): Norm of the reference velocity components versus time for $T_f = 2$, computed on the 40×40 grid, with CFL $= 0.45$.

5.3.4 (iv) Travelling vortex

This is one of the few examples for the SWE whose exact solution is available; see [RB09a]. We consider the well-prepared initial condition (see Figure 5.8) as in [BALMN14] with the periodic domain $[0, 1)^2$:

$$\begin{aligned} z(0, x, y) &= \mathbf{1}_{[r \le \frac{\pi}{\omega}]} \left(\frac{\Gamma\varepsilon}{\omega}\right)^2 (g(\omega r) - g(\pi)), \\ u_1(0, x, y) &= u_0 + \mathbf{1}_{[r \le \frac{\pi}{\omega}]} \Gamma (1 + \cos(\omega r))\, (y_c - y), \\ u_2(0, x, y) &= \mathbf{1}_{[r \le \frac{\pi}{\omega}]} \Gamma (1 + \cos(\omega r))\, (x - x_c), \end{aligned} \tag{iv$_a$}$$

with $b(x, y) = -110$, $u_0 = 0.6$ and

$$\begin{aligned} r &:= \|\boldsymbol{x} - \boldsymbol{x}_c\|, \quad \boldsymbol{x}_c = (0.5, 0.5)^T, \quad \Gamma = 1.4, \quad \omega = 4\pi, \\ g(r) &:= 2\cos r + 2r \sin r + \frac{1}{8}\cos 2r + \frac{r}{4}\sin 2r + \frac{3}{4}r^2. \end{aligned}$$

We pick the initial velocity field as the initial reference velocity $\overline{\boldsymbol{u}}_0$. Figure 5.9 confirms the quality of the solution for $\varepsilon \in \{0.8, 0.01\}$ compared to [Bis15]; it is computed for a short time $T_f = 0.1$ with CFL $= 0.45$, on the 100×100 grid, and with no implicit diffusion, *i.e.*, $c_{\widetilde{\alpha}} = 0$. The figures suggest that the scheme does not preserve the symmetry of the solution, though, the un-symmetry is very small. This is, in fact, a well-known issue for operator splitting schemes; see [DR06, p. 526].

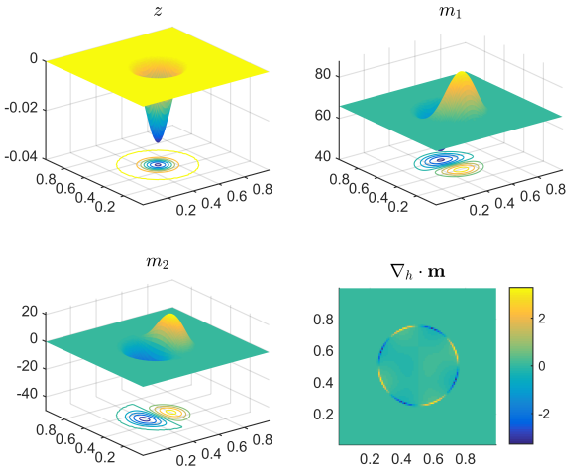
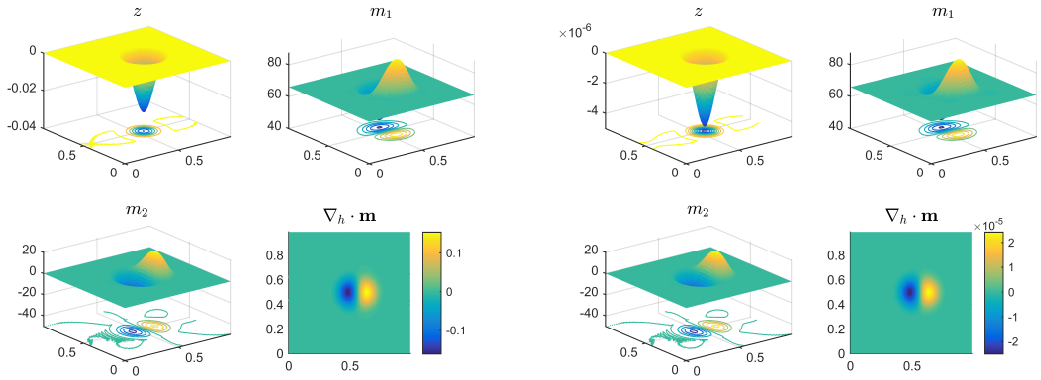


Figure 5.8: Initial condition for the travelling vortex example (iv_a) with $\varepsilon = 0.8$, computed on the 100×100 grid.

With these preliminary results at our disposal, we study the accuracy of the scheme in the next section, which is feasible since the exact solution of this example is available. Then, we investigate the AP property of the scheme numerically.



(a) $\varepsilon = 0.8$. (b) $\varepsilon = 0.01$.

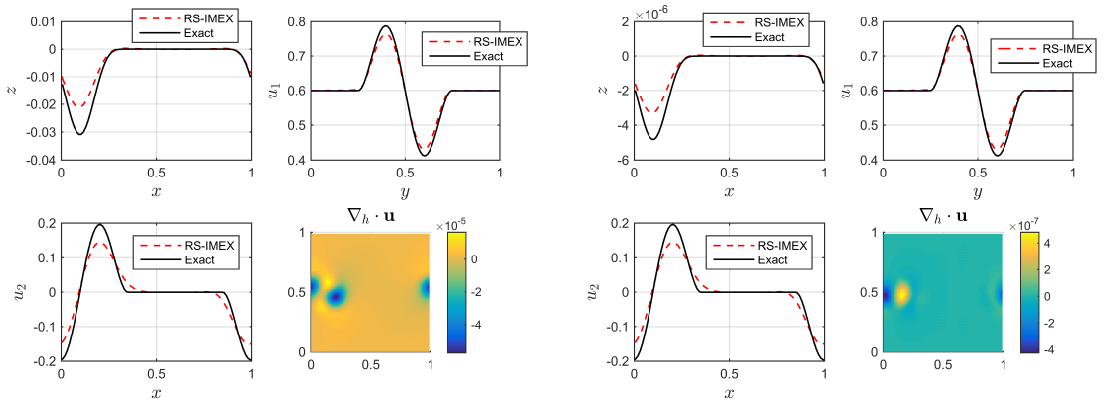
Figure 5.9: Solution of the RS-IMEX scheme for the travelling vortex example (iv_a), on the 100×100 grid and with CFL $= 0.45$ and $T_f = 0.1$.

Experimental order of convergence The exact solution of the travelling vortex example is simply the initial condition advected by u_0 with the period $T_\pi = \frac{5}{3}$, *i.e.*, for $\phi \in \{z, u_1, u_2\}$, it holds that $\phi(t, x, y) = \phi(0, x - u_0 t, y)$. Employing this exact solution, we can find the experimental order of convergence with the error defined as (5.19), but with a different norm and reference solution. Tables 5.3 shows EOC for different ε; it is clear that the order of convergence is not deteriorated for small ε, and it is close to the theoretical one. That is to say that the scheme is uniformly-accurate for all $\varepsilon > 0$. We also illustrate this fact in Figure 5.10, where both exact

and numerical solutions are plotted along the centrelines of the domain, for the 80×80 grid, $T_f = 1$, and with $\varepsilon \in \{0.8, 0.01\}$.

Table 5.3: Experimental order of convergence for the travelling vortex example (iv_a) with $T_f = 1$ and for different ε.

	$\varepsilon = 0.8$				$\varepsilon = 0.01$			
N	e_{z,ℓ_∞}	EOC_{z,ℓ_∞}	e_{u_1,ℓ_∞}	EOC_{u_1,ℓ_∞}	e_{z,ℓ_∞}	EOC_{z,ℓ_∞}	e_{u_1,ℓ_∞}	EOC_{u_1,ℓ_∞}
20	2.61e-2	-	1.04e-1	-	4.08e-6	-	1.04e-1	-
40	2.00e-2	0.38	6.80e-2	0.61	3.12e-6	0.38	6.80e-2	0.61
80	1.23e-2	0.70	3.63e-2	0.91	1.92e-6	0.71	3.63e-2	0.91
160	6.20e-3	0.99	1.65e-3	1.14	9.69e-7	0.99	1.65e-3	1.14
	$\varepsilon = 10^{-4}$				$\varepsilon = 10^{-6}$			
N	e_{z,ℓ_∞}	EOC_{z,ℓ_∞}	e_{u_1,ℓ_∞}	EOC_{u_1,ℓ_∞}	e_{z,ℓ_∞}	EOC_{z,ℓ_∞}	e_{u_1,ℓ_∞}	EOC_{u_1,ℓ_∞}
20	4.08e-10	-	1.04e-1	-	4.08e-14	-	1.04e-1	-
40	3.13e-10	0.38	6.80e-2	0.61	3.13e-14	0.38	6.80e-2	0.61
80	1.92e-10	0.71	3.63e-2	0.91	1.92e-14	0.71	3.63e-2	0.91
160	9.69e-11	0.99	1.65e-3	1.14	9.69e-15	0.99	1.65e-3	1.14



(a) $\varepsilon = 0.8$. (b) $\varepsilon = 0.01$.

Figure 5.10: Error of the RS-IMEX scheme for the travelling vortex example (iv_a) with the incompressible reference solution, on the 80×80 grid and with CFL $= 0.45$ and $T_f = 1$.

Asymptotic preserving property In this part, we aim to confirm the AP property of the scheme, which is to say that the solution is consistent with the limit solution in an appropriate sense and it is stable, uniformly. Figure 5.11 shows the solution of the scheme for a small ε, in particular $\varepsilon = 10^{-6}$. One can see that there is a very good agreement between the result of the RS-IMEX scheme and the exact solution (which is the initial data since $T_f = T_\pi$). It is also clear that there is no checker-board oscillation for the momentum and surface perturbation. Figure 5.12a shows the scaled perturbation $\boldsymbol{V}_{3,\Delta}$; one can see that it grows with time. But as has been shown in Figure 5.12b, it is of order of the scheme, $\mathcal{O}(\Delta x)$, so can be controlled efficiently.

Long-time simulation It is also of interest to check the behaviour of the scheme in a long run. For this purpose, we change the final time to $T_f = 2T_\pi$. Figure 5.13 confirms that for different

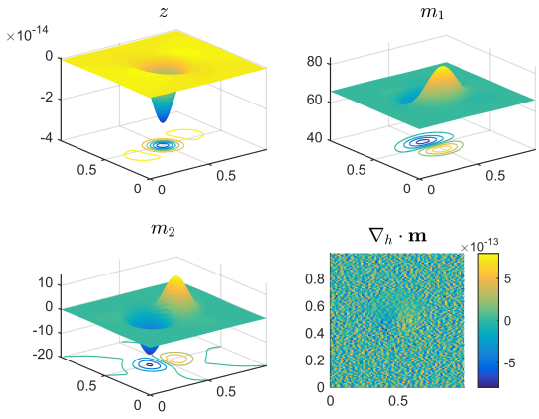
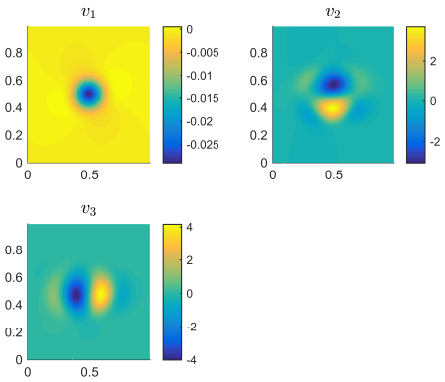
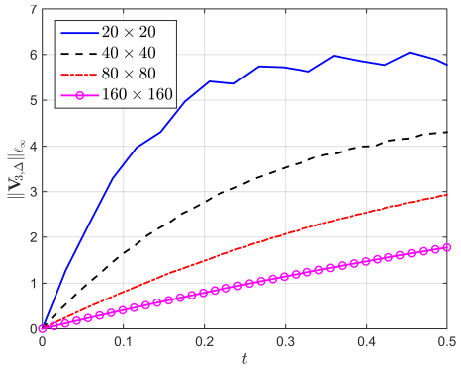


Figure 5.11: Solution of the RS-IMEX scheme for the travelling vortex example (iv), computed on the 100 × 100 grid, with $\varepsilon = 10^{-6}$, CFL = 0.45 and $T_f = T_\pi$.



(a) Perturbation on the 100 × 100 grid for $T_f = T_\pi$.



(b) Time dependence of the norm of the perturbation regarding the mesh refinement.

Figure 5.12: Behaviour of the scaled perturbation for the travelling vortex example (iv$_a$), with CFL = 0.45 and $\varepsilon = 10^{-6}$.

ε, the computed solution does not show any kind of instability; but, it is a bit dissipative, which is expected since the scheme is first-order.

Effects of the reference solution Finding the evolution of the reference solution in time requires additional computational costs, which should be justified. For this example, we show that using the asymptotic reference solution provides better accuracy; thus, it is reasonable to invest in finding a suitable reference solution. Figure 5.14 illustrates the error of the RS-IMEX solution with the zero reference solution. One can clearly observe that, compared to Figure 5.10, the scheme is much more diffusive and less accurate.

Behaviour of the scheme in the limit As explained earlier in Remark 5.1.1, the residual of the reference solution $\overline{\boldsymbol{T}}_\Delta$ does not vanish in general, *e.g.*, owing to the reference momentum

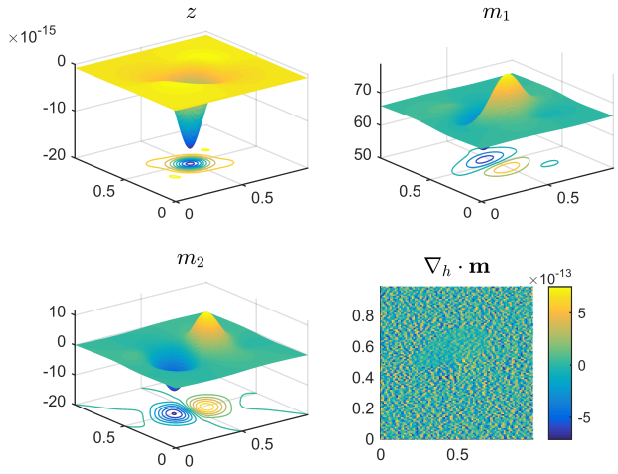
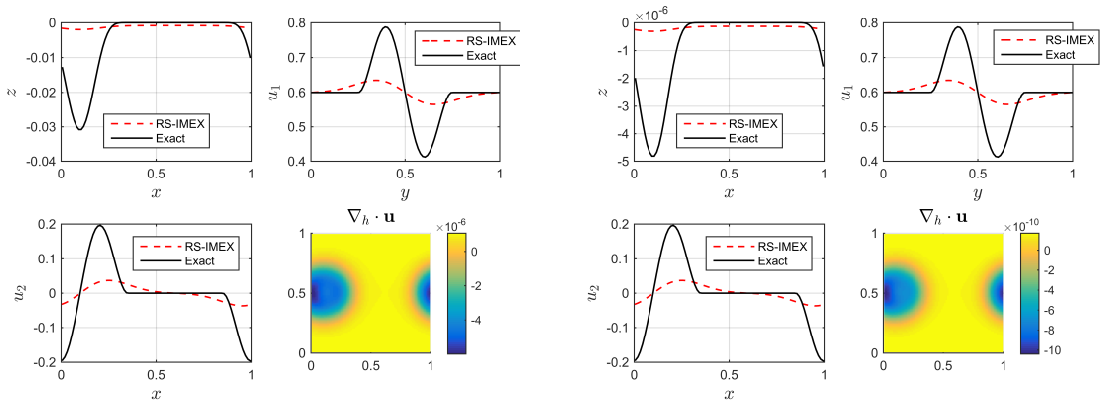


Figure 5.13: Long-time solution for the travelling vortex example (iv_a) by the RS-IMEX scheme, computed on the 100×100 grid for $\varepsilon = 10^{-6}$ with CFL $= 0.45$ and $T_f = 2T_\pi$.



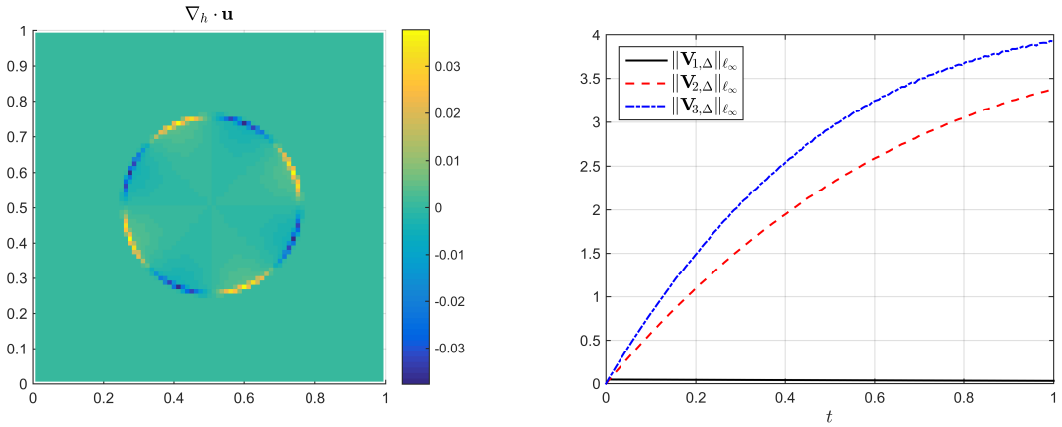
(a) $\varepsilon = 0.8$. (b) $\varepsilon = 0.01$.

Figure 5.14: Error of the RS-IMEX scheme for the travelling vortex example (iv_a) with the zero reference solution $\overline{\boldsymbol{U}} = \mathbf{0}$, on the 80×80 grid, with CFL $= 0.45$ and $T_f = 1$. The results should be compared with Figure 5.10.

field whose discrete divergence may not be translated precisely on the grid or due to the missing incompressible pressure term in $\overline{\boldsymbol{T}}_{2,\Delta}$ and $\overline{\boldsymbol{T}}_{3,\Delta}$. To observe how this issue affects the computed solution, we consider a travelling vortex example with an initial datum on the limit manifold, *i.e.*, with the same initial velocity field as (iv_a), but with $z(0,x,y) = 0$ so that $\boldsymbol{V}_\Delta^0 = \mathbf{0}$. We also pick $\varepsilon = 10^{-6}$, and denote these settings as Example (iv_b). As Figure 5.15a shows, the discrete divergence (as in $\overline{\boldsymbol{T}}_{1,\Delta}$) is only approximately zero, *i.e.*, it is $\mathcal{O}(\Delta x)$, as the velocity field is not smooth enough (continuously differentiable with jumps in the second derivatives), the central difference is only first-order consistent. Also, $\overline{\boldsymbol{T}}_{2,\Delta}, \overline{\boldsymbol{T}}_{3,\Delta} \neq \mathbf{0}$, which can be explained similarly to the periodic flow example as the analytical solution for the reference system is an advection in the x-direction, which can be preserved by the projection scheme only approximately. However, there is an important difference with the periodic flow example since for this case the incompressible

pressure is not zero; so, $\overline{\boldsymbol{T}}_{2,\Delta}, \overline{\boldsymbol{T}}_{3,\Delta} = \mathcal{O}(1)$.

These contributions produce some disturbances, accelerate the scaled perturbation $\boldsymbol{V}_\Delta^n$ and make it to grow (by order $\mathcal{O}(\Delta t \Delta x)$ at each step), as shown in Figure 5.15b. Note that due to the ε-stability results we have proved, the perturbation of the water surface should be $\mathcal{O}(\varepsilon^2)$, which is confirmed by Figure 5.16, where the solution has been compared with the exact one.



(a) (b)

Figure 5.15: (a): Numerical divergence of the initial velocity field for the travelling vortex example (iv$_b$), computed on the 80×80 grid. (b): Time evolution of the norm of the perturbation from the reference solution for the solution computed on the 80×80 grid with $\varepsilon = 10^{-6}$, CFL $= 0.45$ and $T_f = 1$.

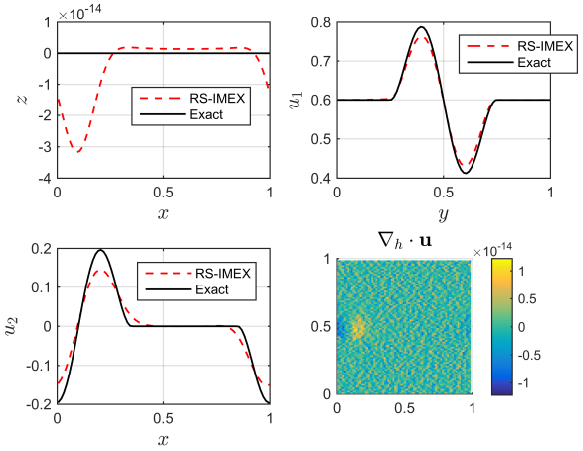


Figure 5.16: Comparison of the computed solution by the RS-IMEX scheme with the exact one in Example (iv$_b$) with $\varepsilon = 10^{-6}$ on the 80×80 grid, and with CFL $= 0.45$ and $T_f = 1$.

To complete the claim of Remark 5.1.1, we should study the contribution of each part of the scheme. Figure 5.17 shows norm of the *update* corresponding to $\|\overline{\boldsymbol{T}}_\Delta\|_{\ell_\infty}$, $\|\nabla_{h,\boldsymbol{x}} \cdot \widehat{\boldsymbol{G}}_\Delta\|_{\ell_\infty}$ and $\|\nabla_{h,\boldsymbol{x}} \cdot \widehat{\boldsymbol{G}}_\Delta\|_{\ell_\infty}$ for the reference, explicit and implicit parts, respectively. It is evident from the figure that in the limit $\varepsilon \to 0$, the contribution of none of these steps can be ignored. Note that

the $\mathcal{O}(1/\varepsilon^2)$ contributions of implicit and reference steps cancel each other, as shown in Section 6.3.2.2, and an $\mathcal{O}(1)$ contribution would remain, like the contribution of the explicit step.

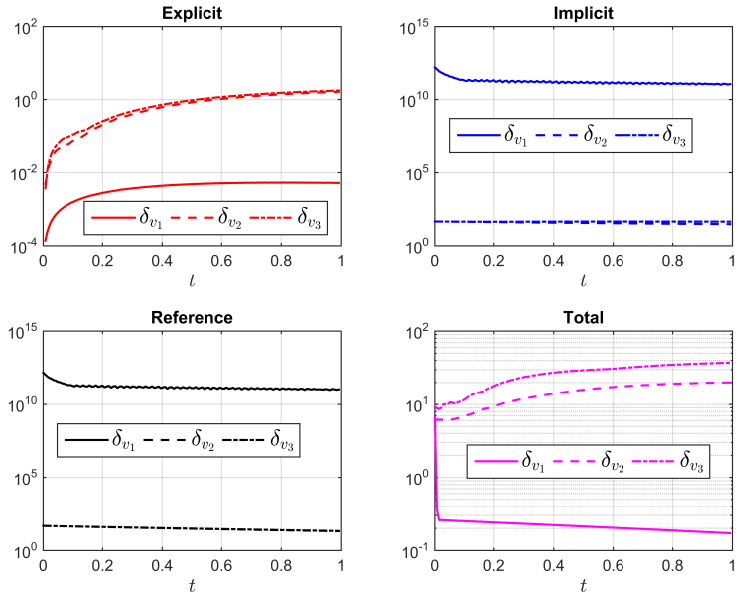


Figure 5.17: Norm of the update for each step as well as the total update of the scheme in Example (iv$_b$), for the solution computed on the 80×80 grid with $\varepsilon = 10^{-6}$, CFL $= 0.45$ and $T_f = 1$.

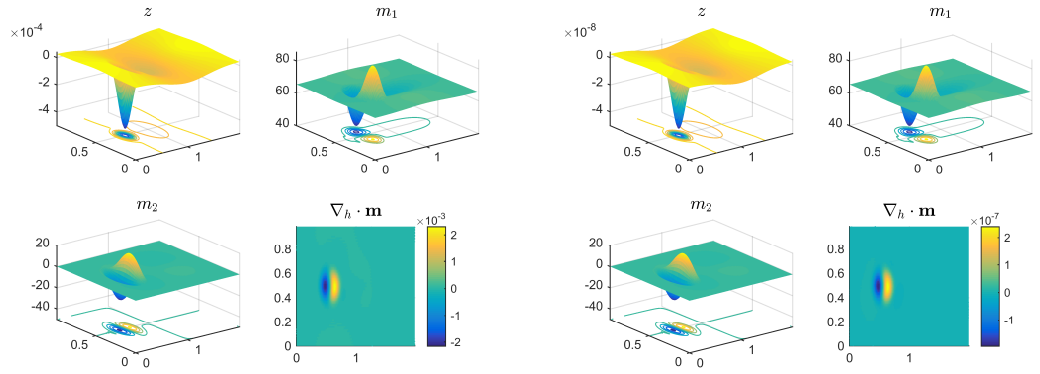
5.3.5 (v) Travelling vortex with topography

In this section, we study the previous travelling vortex example but with a non-flat topography. Like [BALMN14], the bottom topography is given by

$$\eta^b(x, y) = 10 \exp(-5(x-1)^2 - 50(x-0.5)^2)$$

in the periodic domain $[0, 2) \times [0, 1)$. Note that this initial condition is not well-prepared as the initial momentum field is not solenoidal (but the velocity field is). In this case, the exact solution is no longer available; so, one should compare the results of the RS-IMEX scheme with [BALMN14, Bis15], for $T_f = 0.1$, CFL $= 0.3$ and on the 160×80 grid. Note that, due to varying topography, one cannot use DFT and should employ the parabolic regularisation method described in Section 5.1.1. Here, we use the explicit version of the regularisation with the 0.01% tolerance for the steadiness of the solution (in the pseudo-time and using a normalised version of the difference between temporal steps). Figure 5.18 verifies the asymptotic consistency of the scheme as the surface perturbation and divergence of the momentum field are $\mathcal{O}(\varepsilon^2)$.

Regarding asymptotic stability of the scheme, Figure 5.19 confirm that for a small ε, the scaled perturbation remains bounded in time. The solution is computed on the 160×80 grid, with CFL $= 0.3$ and for $T_f = 2$. The computational cost of the reference solver is not that much different from the flat bottom case, as the solution seeks the steady state rather fast.



(a) $\varepsilon = 0.1$. (b) $\varepsilon = 0.001$.

Figure 5.18: Solution the RS-IMEX scheme for Example (v) with different ε on the 160×80 grid, with CFL $= 0.3$ and $T_f = 1$.

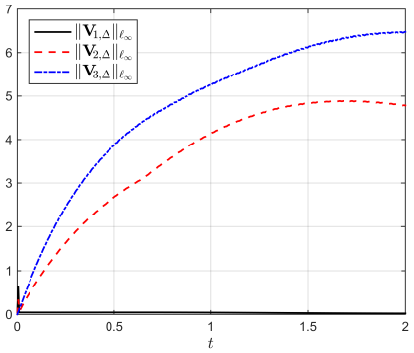


Figure 5.19: Asymptotic stability of the RS-IMEX scheme Example (v): Norm of the perturbation versus time for $\varepsilon = 10^{-6}$ on the 80×160 grid, with CFL $= 0.3$ and $T_f = 2$.

5.A Asymptotic analysis of the shallow water equations

This section is to provide the formal asymptotic analysis for the low-Froude 2d SWE. In the periodic domain Ω, consider the usual formulation of dimensionless SWE (1.10), with η^b as the bottom function and ε as the Froude number:

$$\begin{aligned} &\partial_t h + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{m} = 0, \\ &\partial_t \boldsymbol{m} + \operatorname{div}_{\boldsymbol{x}} \left(\frac{\boldsymbol{m} \otimes \boldsymbol{m}}{h} \right) + \nabla_{\boldsymbol{x}} \left(\frac{h^2}{2\varepsilon^2} \right) = -\frac{h}{\varepsilon^2} \nabla_{\boldsymbol{x}} \eta^b. \end{aligned} \tag{5.21}$$

We use the Poincaré expansion of h and $\boldsymbol{m}$ in terms of ε as

$$\begin{aligned} h(t,\boldsymbol{x}) &= h_{(0)}(t,\boldsymbol{x}) + \varepsilon h_{(1)}(t,\boldsymbol{x}) + \varepsilon^2 h_{(2)}(t,\boldsymbol{x}), \\ \boldsymbol{m}(t,\boldsymbol{x}) &= \boldsymbol{m}_{(0)}(t,\boldsymbol{x}) + \varepsilon m_{(1)}(t,\boldsymbol{x}) + \varepsilon^2 m_{(2)}(t,\boldsymbol{x}). \end{aligned} \tag{5.22}$$

Then, we substitute (5.22) in (5.21), and balance equal powers of ε. So, $\mathcal{O}(\varepsilon^{-2})$ terms yield $h_{(0)} \nabla_{\boldsymbol{x}} (h_{(0)} + b) = \mathbf{0}$, which implies that the leading order of the water surface $\eta^s := h + \eta^b$

is constant in space, *i.e.*, $\eta^s_{(0)} = \eta^s_{(0)}(t)$. Using this, one can find that $h_{(0)}\nabla_{\boldsymbol{x}} h_{(1)} = \mathbf{0}$, so $h_{(1)} = h_{(1)}(t)$.

Moreover, the leading order of the continuity equation $\partial_t h_{(0)} + \mathrm{div}_{\boldsymbol{x}} \boldsymbol{m}_{(0)} = 0$ yields that

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \left(h_{(0)} + \eta^b\right) \mathrm{d}x = -\int_{\partial\Omega} \boldsymbol{m}_{(0)}.\boldsymbol{n}\mathrm{d}s = 0,$$

owing to the divergence theorem and the periodicity of Ω. Thus, $\partial_t h_{(0)} = 0$ and $\eta^s_{(0)} = \text{const.}$, which gives $h_{(0)} = \eta^s_{(0)} - \eta^b(\boldsymbol{x})$. Hence, from the continuity equation, we get the *div*-free condition for the leading order momentum field, *i.e.*, $\mathrm{div}_{\boldsymbol{x}} \boldsymbol{m}_{(0)} = 0$. With similar arguments, one can easily find that $\partial_t h_{(1)} = 0$, so $h_{(1)} = \text{const.}$ and $\mathrm{div}_{\boldsymbol{x}} \boldsymbol{m}_{(1)} = 0$. Summing up all these gives the formal asymptotic limit of the SWE as in Definition 5.2.1.

5.B On the well-balancing of the RS-IMEX scheme

In this section, we discuss an example regarding the well-balancing issue for the RS-IMEX scheme, and based on that, we suggest a remedy for recovering the well-balancing. We limit our focus on cases with a flat bottom topography; similar concerns could be raised for the case of a non-flat bottom but will be skipped here.

Consider the SWE in the periodic domain $[0,1)^2$ with $H_{\text{mean}} = 1$, $\varepsilon = 0.8$, and with the following initial condition, together with the $\overline{\boldsymbol{U}} + D\boldsymbol{V}$ decomposition:

$$\begin{cases} z(0,x,y) = 0 \\ m_1(0,x,y) = 0 \\ m_2(0,x,y) = 0 \end{cases}, \begin{cases} \overline{z}(0,x,y) = 0 \\ \overline{m_1}(0,x,y) = \sin(2\pi(x-y)) \\ \overline{m_2}(0,x,y) = \sin(2\pi(x-y)) \end{cases}, \begin{cases} v_1(0,x,y) = 0 \\ v_2(0,x,y) = -\sin(2\pi(x-y)) \\ v_3(0,x,y) = -\sin(2\pi(x-y)) \end{cases}. \quad \text{(vi)}$$

This setting is peculiar as, for the RS-IMEX scheme with such a zero initial datum, one does not consider a non-zero reference solution. However, this example is only designed to investigate how the well-balancing can be obtained when $\overline{\boldsymbol{U}}^n_\Delta + \boldsymbol{V}^n_\Delta = \mathbf{0}$ when neither the reference part $\overline{\boldsymbol{U}}^n_\Delta$ nor the perturbation $\boldsymbol{V}^n_\Delta$ is at equilibrium. The reference part corresponds to a steady state solution and can be solved by the projection scheme. As explained in Section 5.3.3, the projection scheme cannot preserve such a steady state and adds an $\mathcal{O}(\Delta t \Delta x)$ disturbance at each step. As Figure 5.20 shows, the surface perturbation computed by the RS-IMEX scheme is zero, up to the machine accuracy; but, the calculated velocity field is far from zero (though can be controlled by the mesh refinement). These unbalances do not originate solely from the reference solver; even if with the exact $\overline{\boldsymbol{U}}$, the scheme does not preserve the equilibrium exactly and the result is almost the same as Figure 5.20.

Understanding the reason behind this observation, we write down the whole scheme in one step and pick $\widetilde{\alpha} = 0$ for simplicity. For the v_1-update, we find ($\widehat{\alpha}_1, \widehat{\alpha}_2 = \widehat{\alpha}$):

$$v^{n+1}_{1,ij} = v^n_{1,ij} - \frac{\Delta t}{\varepsilon^2}\nabla_{h,\boldsymbol{x}} \cdot \begin{pmatrix} \overline{m_1} + v_2 \\ \overline{m_2} + v_3 \end{pmatrix}^{n+1}_{ij} + \frac{\widehat{\alpha}\Delta x}{2}\Delta t \Delta_{h,\boldsymbol{x}} v^n_{1,ij}, \tag{5.23}$$

which implies that the zero surface perturbation is compatible with a zero momentum field.

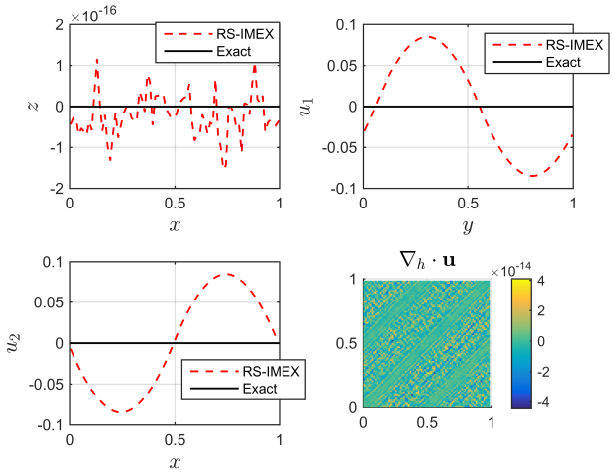


Figure 5.20: Comparison of the RS-IMEX solution with the exact one, with $\varepsilon = 0.8$, $T_f = 0.1$, CFL $= 0.45$ on the 80×80 grid for the stationary well-balancing example (vi).

For the v_2-update (and analogously v_3-) one gets

$$
\begin{aligned}
v_{2,ij}^{n+1} = v_{2,ij}^{n} &- \Delta t \nabla_{h,x} \widehat{\boldsymbol{G}}_{1,2,ij}^{n} - \Delta t \nabla_{h,y} \widehat{\boldsymbol{G}}_{2,2,ij}^{n} + \frac{\widehat{\alpha}\Delta x}{2} \Delta t \Delta_{h,\boldsymbol{x}} v_{2,ij}^{n} \\
&- \Delta t \nabla_{h,x} \widetilde{\boldsymbol{G}}_{1,2,ij}^{n+1} - \Delta t \nabla_{h,y} \widetilde{\boldsymbol{G}}_{2,2,ij}^{n+1} \\
&- \left(\overline{m_1}^{n+1} - \overline{m_1}^{n}\right)_{ij} - \Delta t \nabla_{h,x} \left(\frac{\overline{m_1}^2}{\overline{z} - b} + \frac{\overline{z}^2 - 2\overline{z}b}{2\varepsilon^2} \right)_{ij}^{n+1} - \Delta t \nabla_{h,y} \left(\frac{\overline{m_1 m_2}}{\overline{z} - b} \right)_{ij}^{n+1}.
\end{aligned}
\tag{5.24}
$$

To verify the compatibility of the LaR solution in (5.24), we assume an exact (steady) reference solution, *i.e.*, $\overline{z}_\Delta^n = \overline{z}_\Delta^{n+1} = 0$ and $\overline{\boldsymbol{m}}_\Delta^{n+1} = \overline{\boldsymbol{m}}_\Delta^n$, as well as $\boldsymbol{m}_\Delta^{n+1} = \boldsymbol{m}_\Delta^n = \mathbf{0}$ and $v_{1,\Delta}^{n+1} = const.$. For such a solution, one can verify that the only remaining term in (5.24) is the numerical diffusion for the explicit step $\frac{\widehat{\alpha}\Delta x}{2} \Delta t \Delta_{h,\boldsymbol{x}} v_{2,\Delta,ij}^n$. The problem is that, unlike Chapter 4 for the 1d case, $v_{2,\Delta}^n$ is not constant; so, unless $\widehat{\alpha} = 0$, the numerical diffusion does not vanish. It accelerates the flow and destroys the compatibility with an $\mathcal{O}(\Delta x)$ disturbance. Thus, we can guess that setting $\widehat{\alpha}$ to zero balances the scheme, which has been corroborated by Figure 5.21, where the only change compared to Figure 5.20 is that $\widehat{\alpha} = 0$. Note that setting $\widehat{\alpha} = 0$ furnishes some oscillations which leads to instability; so, it is not favourable.

A more interesting remedy would be to add a diffusion for $\overline{\boldsymbol{T}}_{2,\Delta}$ and $\overline{\boldsymbol{T}}_{3,\Delta}$, *i.e.*, $\overline{\alpha} = \widehat{\alpha}$. Because the sum of the two parts of the momentum should be zero for a LaR solution, this modification balances the numerical diffusion coming from the explicit part and fulfils the compatibility. Figure 5.22 presents two case considering this modification, with the exact or the numerically-computed reference solution. It indicates that the modified scheme can preserve the LaR equilibrium state for this example. Note that such a strategy may not work if the reference solution is not steady state.

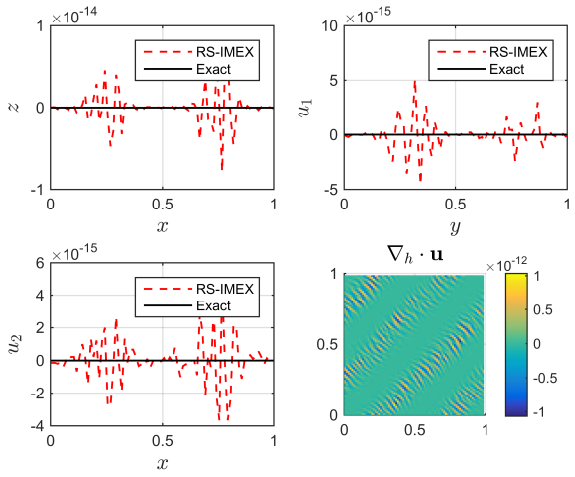
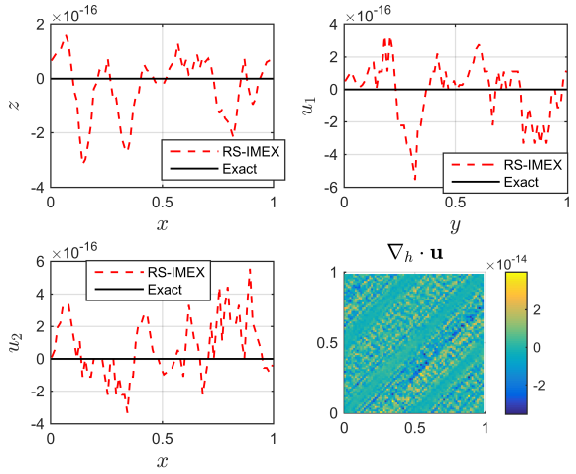
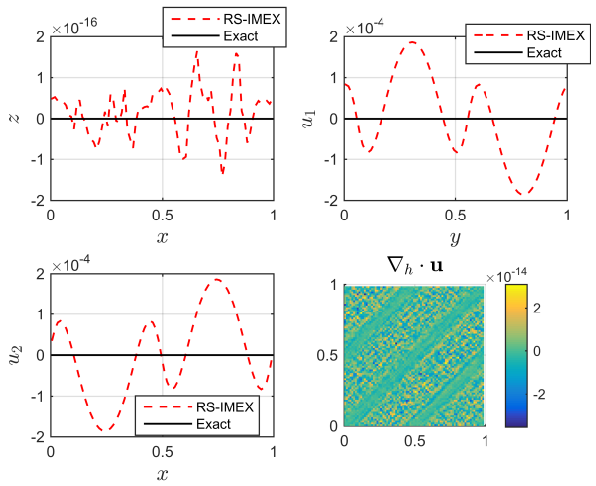


Figure 5.21: Comparison of the RS-IMEX solution with the exact one, with $\varepsilon = 0.8$, $T_f = 0.1$, CFL $= 0.45$, and $\widehat{\alpha} = 0$, computed on the 80×80 grid for the stationary well-balancing example (vi).



(a) Exact $\overline{\boldsymbol{U}}_\Delta$.



(b) $\overline{\boldsymbol{U}}_\Delta$ is computed by the projection scheme.

Figure 5.22: Comparison of the RS-IMEX solution with the exact one, with $\varepsilon = 0.8$, $T_f = 0.1$, CFL $= 0.45$, and $\overline{\alpha} = \widehat{\alpha}$, computed on the 80×80 grid for the stationary well-balancing example (vi).

Chapter 6

The RS-IMEX scheme for the 2d rotating shallow water equations

"How can I help seeing what is in front of my eyes? Two and two are four! Sometimes, Winston. Sometimes they are five. Sometimes they are three. Sometimes they are all of them at once. You must try harder. It is not easy to become sane."

– George Orwell, *1984*

In this chapter, we investigate the applicability of the RS-IMEX scheme, already studied in Chapters 4 and 5 for the shallow water equations, for the "rotating" shallow water equations, where the additional Coriolis force is present in the system. We show the asymptotic consistency of the scheme in the quasi-geostrophic distinguished limit, which is a commonly-adopted characterisation of the rotation-gravity interplay in the ocean modelling. We also test the quality of the scheme by several numerical examples. This chapter is based on [Zak17b].

Contents

6.1 Introduction

Despite the inherent limitations of the shallow water equations, as mentioned in Chapter 1, they are often used to model the oceanic flows; see for instance [Ped13, Maj03]). The large-scale oceanic flows are mainly affected by the earth's rotation, which motivates studying the so-called 2d rotating shallow water equations (RSWE). Consider the domain $\Omega \subset \mathbb{R}^2$ lying in the (x, y)

plane; then, the RSWE write as (1.9):

$$\begin{aligned} &\partial_t h + \operatorname{div}_{\boldsymbol{x}}(h\boldsymbol{u}) = 0, \\ &\partial_t(h\boldsymbol{u}) + \operatorname{div}_{\boldsymbol{x}}\left(h\boldsymbol{u}\otimes\boldsymbol{u} + \frac{gh^2}{2}\mathbb{I}_2\right) = -gh\nabla_{\boldsymbol{x}}\eta^b - fh\boldsymbol{u}^\perp, \end{aligned} \tag{6.1}$$

where h is the water height, η^b is the bottom function, $\boldsymbol{u} = (u_1, u_2)$ is the 2d velocity vector, $\boldsymbol{u}^\perp = (-u_2, u_1)$ is the *perpendicular velocity*, g is the gravity acceleration constant, $\mathbb{I}_2$ is the 2×2 identity matrix and f is the Coriolis parameter assumed to be a constant value (zero-plane approximation). We limit our focus to periodic domains $\Omega = \mathbb{T}^2$ for simplicity.

Back then, when the system (6.1) was introduced, the huge computational cost for a numerical approximation of its solution could not be paid. So, several simplified models have been introduced to mimic the main behaviour of this system. Charney was the first who could simplify this meteorological system in a successful and practical way [Cha48, Cha49]; by *filtering out* noises or fast gravity waves, which do not contribute to the bulk motion of the fluid. So, the model is left with slow *Rossby waves*, resulting in the so-called *barotropic quasi-geostrophic equations*; see [Maj03, MW06, EM96, Ped13, Dur13] for an overview. Accurately enough, one can claim that the modern era of numerical schemes for geophysical flows, in general, and ocean currents, in particular, has been started with this approximation and the ice-breaking paper [CFvN50], which presented a numerical method for it. Recently, thanks to high computing resources, one is able to employ more sophisticated models like the RSWE (6.1) without such simplifying assumptions. Nonetheless, in most practical cases, the fast gravity waves are present and make the system stiff. This stiffness requires using very fine grids or devising schemes covering several scales in time and space at once. Tackling such an issue, we adopt the AP framework to design a scheme to capture the macroscopic behaviour of the system (6.1) for an under-resolved grid.

Due to the additional Coriolis force, an extra parameter will be introduced in the system, the so-called Rossby number *Ro*. To study AP property requires choosing the desired *distinguished limit*, as a characterisation of the relation between *Ro* and *Fr*, with only one scaling parameter ε. Throughout this chapter, we limit our focus to the so-called *quasi-geostrophic limit*, which is a singular limit denoted by $\varepsilon \to 0$. The rationale for this choice is the famous result by Majda [Maj03] that, in this limit, the RSWE converge to the quasi-geostrophic equations or the barotropic vorticity equations, which are the equations derived formally by [Cha48]. This ensures that, at least in the continuous level, there is a convergence for this singular limit. Thus, regarding the famous AP diagram in Figure 1.1, it is justified to look for a scheme to preserve this convergence at the discrete level. It would be of course interesting to consider more general cases with centrifugal forces and viscosity, like those analysed in [FGGVN12, FGN12, FN14a, FN14b]; but, we skip them here.

Like other balances laws, preserving equilibrium states of the system is also crucial for numerical schemes designed for the system (6.1). This can be really difficult; see [ADDMHP15, AKNV11, AKO09, BLSZ04, LMNK07] for some simplified cases. For example in [LMNK07, CDKLM14], the authors devised second-order schemes preserving spacial one-dimensional equilibrium states, the so-called *jets in the rotational frame*, or in [AKNV11], an extension of the *hydrostatic reconstruction* method [BKLL04, ABB$^+$04] has been employed to preserve the *geostrophic balance* between the pressure gradient and the Coriolis force, combined with the technique presented in [VK09] to find the *auxiliary water height*.

In fact, other than [HZLMP11, ADDMHP15], the question of designing AP schemes for the

RSWE has not yet been discussed in the literature and consists the main part of this chapter.[1] It is already shown in Chapters 4 and 5, that the RS-IMEX scheme is well-behaved for the SWE but without the Coriolis force; so, it is of our interest to check if the scheme works well with the additional Coriolis force. This would be the main goal of the present chapter, organised as follows. In Section 6.2, we introduce the RS-IMEX scheme for the RSWE after the reformulation and non-dimensionalisation of system (6.1). Then, in Section 6.3, we present the numerical analysis for the scheme in terms of well-balancing and asymptotic preserving property, followed by a set of numerical examples in Section 6.4.

6.2 RS-IMEX scheme for the rotating shallow water equations

One can deduce from the *Buckingham π-theorem* [Buc15] that there are three different dimensionless groups for this system: the Strouhal number Sr, the Froude number Fr and the Rossby number Ro. But since we consider two height scales, as in [Maj03, Chap. 4], we should also introduce another dimensionless group Θ. The height scales are $H_\circ$ for the mean water level chosen equal to the actual mean water level H_{mean}, and $Z_\circ$ for the surface perturbation from H_{mean} (denoted by z as in Chapters 4 and 5), *i.e.*, $h = H_{\text{mean}} + z - \eta^b$. Defining dimensionless variables as

$$\hat{\boldsymbol{x}} := \frac{\boldsymbol{x}}{L_\circ}, \quad \hat{t} := \frac{t}{t_\circ}, \quad \hat{\boldsymbol{u}} := \frac{\boldsymbol{u}}{u_\circ}, \quad \hat{z} := \frac{z}{Z_\circ}, \quad \hat{\eta}^b := \frac{\eta^b}{Z_\circ}, \quad \hat{h} := \frac{h}{H_\circ},$$

where characteristic states are denoted by the subscript $\circ$, one obtains $\hat{h} = 1 + \Theta(\hat{z} - \hat{\eta}^b)$ and can rewrite (6.1) as (*cf.* [Maj03])

$$\begin{aligned} &Sr\,\partial_{\hat{t}}\left(\Theta\hat{z}\right) + \operatorname{div}_{\hat{\boldsymbol{x}}}\left(\hat{h}\hat{\boldsymbol{u}}\right) = 0, \\ &Sr\,\partial_{\hat{t}}\left(\hat{h}\hat{\boldsymbol{u}}\right) + \operatorname{div}_{\hat{\boldsymbol{x}}}\left(\hat{h}\hat{\boldsymbol{u}}\otimes\hat{\boldsymbol{u}} + \frac{\hat{h}^2}{2Fr^2}\mathbb{I}_2\right) = -\frac{\Theta}{Fr^2}\hat{h}\nabla_{\hat{\boldsymbol{x}}}\hat{\eta}^b - \frac{\hat{h}}{Ro}\boldsymbol{u}^\perp, \end{aligned} \tag{6.2}$$

with the following definitions for Sr, Fr, Ro and Θ:

$$Sr := \frac{L_\circ}{u_\circ t_\circ}, \quad Fr := \frac{u_\circ}{\sqrt{gH_\circ}}, \quad Ro := \frac{u_\circ}{fL_\circ}, \quad \Theta := \frac{Z_\circ}{H_\circ}.$$

It is well-known (see [KVPR10]) that the physical and mathematical properties of the system may depend on the relation between these groups; so, one should select one path (so-called distinguished limit) before going further. Like [VK09], we choose $Sr = 1$, which means that the reference time scale is chosen as the advective time scale. Also, defining $F^{1/2} := fL_\circ/\sqrt{gH_\circ} = \mathcal{O}(1)$, we choose

$$Ro = \varepsilon \ll 1, \quad Fr = F^{1/2}\varepsilon, \quad \Theta = F\varepsilon.$$

This is the so-called *quasi-geostrophic distinguished limit*, *i.e.*, the Rossby and Froude numbers are small; there is an exact balance between the pressure gradient and the Coriolis force; and

[1] [HZLMP11] extends the well-balanced schemes developed in [LMNK07] and presents two large time step methods for the low-Froude regime. [ADDMHP15] performs an asymptotic accuracy analysis for a linear rotating model mimicking the Coriolis force, in the context of [DJOR16].

the variation of the bottom topography and surface perturbation are very mild compared to the height of the water column, owing to $\Theta \sim \varepsilon$, *i.e.*, $\|z\|, \|\nabla_{\boldsymbol{x}}\eta^b\| = \mathcal{O}(\varepsilon)$ (see [Maj03, Ped13]). This limit also requires $Z_\circ = fu_\circ L_\circ/g$. With similar notations as previous chapters, we can rewrite (6.2) as (after suppressing hats):

$$\begin{aligned} &\partial_t z + \frac{1}{\Theta}\mathrm{div}_{\boldsymbol{x}}\boldsymbol{m} = 0, \\ &\partial_t \boldsymbol{m} + \mathrm{div}_{\boldsymbol{x}}\left(\frac{\boldsymbol{m}\otimes\boldsymbol{m}}{\Theta z - b} + \frac{\Theta z^2 - 2bz}{2\varepsilon}\mathbb{I}_2\right) = -\frac{1}{\varepsilon}z\nabla_{\boldsymbol{x}}b - \frac{1}{\varepsilon}\boldsymbol{m}^\perp, \end{aligned} \tag{6.3}$$

where $\boldsymbol{m} := (\Theta z - b)\boldsymbol{u}$ is the momentum vector and b is the dimensionless water depth measured from H_{mean} (scaled by $H_\circ$) with a negative sign, *i.e.*, $1 - \Theta\eta^b = -b$; see Figure 4.1. This implies that the topography's contribution in the rhs of (6.3) is $\mathcal{O}(1)$. It is important to remark that Θz is the surface perturbation; picking $\Theta = 1$ recovers the notation of Chapters 4 and 5.

Remark 6.2.1. *Analogously to Ertel's theorem for the conservation of potential vorticity (PV) (cf. [Ped13, Chap. 2]), one can show that for the original system* (6.1)*, the so-called potential vorticity* $\Pi_s := \frac{f+\zeta}{h}$ *is conserved, i.e.,* $(\partial_t + \boldsymbol{u}\cdot\nabla_x)\,\Pi_s = 0$*, where* ζ *is the magnitude of the vorticity* $\zeta := \|\nabla_x \times \boldsymbol{u}\|$*. For the non-dimensionalised system* (6.3)*,* Π_s *is obtained as*

$$\widehat{\Pi}_s = \frac{1 + \varepsilon\zeta}{\Theta z - b}. \tag{6.4}$$

As mentioned before, Majda showed in [Maj03] that as $\varepsilon \to 0$ the system (6.3) or (6.2) converges to the *quasi-geostrophic equations (QGE)*:

$$\boldsymbol{u}_{(0)} = \nabla_x^\perp z_{(0)}, \tag{6.5a}$$

$$\Delta_{\boldsymbol{x}} z_{(0)} = \zeta_{(0)}, \tag{6.5b}$$

$$\left(\partial_t + \boldsymbol{u}_{(0)}\cdot\nabla_x\right)\left(\zeta_{(0)} - Fz_{(0)} + F\eta^b_{(0)}\right) = 0, \tag{6.5c}$$

where the subscript (0) stands for the leading order term in the Poincaré expansion. Equation (6.5a) means that the solution is at *geostrophic equilibrium* locally in time. It also implies that the surface perturbation $z_{(0)}$ can be read as the stream function ψ, *i.e.*, $\nabla_{\boldsymbol{x}}^\perp\psi = \boldsymbol{u}_{(0)}$; so, the velocity field is solenoidal. Defining ξ as the leading order of $\widehat{\Pi}_s$ for $\varepsilon \ll 1$ (*i.e.*, $\xi := \widehat{\Pi}_{s(0)}$) and using (6.4) imply that equation (6.5c) is the conservation of the (leading order of the) potential vorticity $\xi := \zeta_{(0)} - Fz_{(0)} + Fb_{(0)}$ while the (relative) vorticity $\zeta_{(0)}$ is given by (6.5b). Note that (6.5a)–(6.5c) can be also realised in the usual velocity formulation, instead of this vorticity–stream function formulation, as the $\varepsilon \to 0$ limit of

$$\left(\partial_t + \boldsymbol{u}\cdot\nabla_x\right)\boldsymbol{u} + \varepsilon^{-1}\boldsymbol{u}^\perp + \varepsilon^{-1}\nabla_{\boldsymbol{x}} z = \boldsymbol{0}. \tag{6.6}$$

Remark 6.2.2. *For future reference, we define the geostrophic equilibrium by the notion of "apparent topography" or "auxiliary water depth"; see [Bou04, BLSZ04, AKNV11]. This is to model the effect of the Coriolis force with an auxiliary height so that one can use the same well-balancing methods as for non-rotational systems. Considering the "potentials"* Φ *and* Ψ*, we define "potential energies"* K *and* L *as*

$$\begin{aligned} K &:= g(h + \eta^b - \Phi), \qquad \partial_x\Phi := \frac{f}{g}u_2, \\ L &:= g(h + \eta^b + \Psi), \qquad \partial_y\Psi := \frac{f}{g}u_1. \end{aligned} \tag{6.7}$$

Then, for the solution to be at the geostrophic equilibrium it should hold that

$$\mathrm{div}_{\boldsymbol{x}}\boldsymbol{u} \equiv 0, \qquad \partial_x K \equiv 0, \qquad \partial_y L \equiv 0. \tag{6.8}$$

6.2.1 Numerical scheme

As we already explained in Chapters 4 and 5, the RS-IMEX scheme decomposes the solution $\boldsymbol{U}$ as $\boldsymbol{U} = \overline{\boldsymbol{U}} + \boldsymbol{U}_{pert}$, where $\overline{\boldsymbol{U}}$ is a chosen solution and $\boldsymbol{U}_{pert}$ is the remaining part. For practical as well as analytical reasons (see Chapter 4), one is interested to pick $\overline{\boldsymbol{U}}$ as a solution which is asymptotically close to $\boldsymbol{U}$; so, for the RSWE, we pick the QGE (6.5a)–(6.5c) as the reference system. Then, we use a Taylor expansion around that solution to split the flux and source terms into two parts, linear stiff and non-linear non-stiff parts. Let us rewrite the system (6.3) as

$$\partial_t \boldsymbol{U} + \mathrm{div}_{\boldsymbol{x}} \boldsymbol{F}(\boldsymbol{U}, \boldsymbol{x}) = \boldsymbol{S}^B(\boldsymbol{U}, \boldsymbol{x}) + \boldsymbol{S}^C(\boldsymbol{U}),$$

where $\boldsymbol{U} = (z, m_1, m_2)^T$ and one can identify $\boldsymbol{F}$, $\boldsymbol{S}^B$ and $\boldsymbol{S}^C$ as

$$\boldsymbol{F} = \begin{bmatrix} m_1/\Theta & m_2/\Theta \\ \dfrac{m_1^2}{\Theta z - b} + \dfrac{\Theta z^2 - 2zb}{2\varepsilon} & \dfrac{m_1 m_2}{\Theta z - b} \\ \dfrac{m_1 m_2}{\Theta z - b} & \dfrac{m_2^2}{\Theta z - b} + \dfrac{\Theta z^2 - 2zb}{2\varepsilon} \end{bmatrix}, \quad \boldsymbol{S}^B = \begin{bmatrix} 0 \\ -z\partial_x b/\varepsilon \\ -z\partial_y b/\varepsilon \end{bmatrix}, \quad \boldsymbol{S}^C = \begin{bmatrix} 0 \\ m_2/\varepsilon \\ -m_1/\varepsilon \end{bmatrix}.$$

Assuming the reference solution $\overline{\boldsymbol{U}} = (\overline{z}, \overline{m_1}, \overline{m_2})^T$ to be the solution of the QGE, and following the RS-IMEX splitting as described in Chapter 4, the splitting can be obtained as

$$\overline{\boldsymbol{F}} = \begin{bmatrix} \overline{m_1}/\Theta & \overline{m_2}/\Theta \\ \dfrac{\overline{m_1}^2}{\Theta\overline{z} - b} + \dfrac{\Theta\overline{z}^2 - 2\overline{z}b}{2\varepsilon} & \dfrac{\overline{m_1 m_2}}{\overline{z} - b} \\ \dfrac{\overline{m_1 m_2}}{\Theta\overline{z} - b} & \dfrac{\overline{m_2}^2}{\Theta\overline{z} - b} + \dfrac{\Theta\overline{z}^2 - 2\overline{z}b}{2\varepsilon} \end{bmatrix}, \tag{6.9a}$$

$$\widetilde{\boldsymbol{F}}_1 = \begin{bmatrix} v_2/\Theta \\ -\dfrac{\overline{m_1}^2 v_1 \Theta}{(\Theta\overline{z} - b)^2} + \dfrac{2\overline{m_1} v_2}{\Theta\overline{z} - b} + \dfrac{(\Theta\overline{z} - b)}{\varepsilon} v_1 \\ -\dfrac{\overline{m_1 m_2} v_1 \Theta}{(\Theta\overline{z} - b)^2} + \dfrac{\overline{m_1} v_3}{\Theta\overline{z} - b} + \dfrac{\overline{m_2} v_2}{\Theta\overline{z} - b} \end{bmatrix}, \qquad \widetilde{\boldsymbol{F}}_2 = \begin{bmatrix} v_3/\Theta \\ -\dfrac{\overline{m_1 m_2} v_1 \Theta}{(\Theta\overline{z} - b)^2} + \dfrac{\overline{m_1} v_3}{\Theta\overline{z} - b} + \dfrac{\overline{m_2} v_2}{\Theta\overline{z} - b} \\ -\dfrac{\overline{m_2}^2 v_1 \Theta}{(\Theta\overline{z} - b)^2} + \dfrac{2\overline{m_2} v_3}{\Theta\overline{z} - b} + \dfrac{(\Theta\overline{z} - b)}{\varepsilon} v_1 \end{bmatrix}, \tag{6.9b}$$

$$\widehat{\boldsymbol{F}}_1 = \begin{bmatrix} 0 \\ \dfrac{m_1^2}{\Theta z - b} + \dfrac{\Theta z^2 - 2zb}{2\varepsilon} - \dfrac{\overline{m_1}^2}{\Theta\overline{z} - b} - \dfrac{\Theta\overline{z}^2 - 2\overline{z}b}{2\varepsilon} + \dfrac{\overline{m_1}^2 v_1 \Theta}{(\Theta\overline{z} - b)^2} - \dfrac{2\overline{m_1} v_2}{\Theta\overline{z} - b} - \dfrac{(\Theta\overline{z} - b)}{\varepsilon} v_1 \\ \dfrac{m_1 m_2}{\Theta z - b} - \dfrac{\overline{m_1 m_2}}{\Theta\overline{z} - b} + \dfrac{\overline{m_1 m_2} v_1 \Theta}{(\Theta\overline{z} - b)^2} - \dfrac{\overline{m_1} v_3}{\Theta\overline{z} - b} - \dfrac{\overline{m_2} v_2}{\Theta\overline{z} - b} \end{bmatrix},$$

$$\widehat{\boldsymbol{F}}_2 = \begin{bmatrix} 0 \\ \dfrac{m_1 m_2}{\Theta z - b} - \dfrac{\overline{m_1 m_2}}{\Theta\overline{z} - b} + \dfrac{\overline{m_1 m_2} v_1 \Theta}{(\Theta\overline{z} - b)^2} - \dfrac{\overline{m_1} v_3}{\Theta\overline{z} - b} - \dfrac{\overline{m_2} v_2}{\Theta\overline{z} - b} \\ \dfrac{m_2^2}{\Theta z - b} + \dfrac{\Theta z^2 - 2zb}{2\varepsilon} - \dfrac{\overline{m_2}^2}{\Theta\overline{z} - b} - \dfrac{\Theta\overline{z}^2 - 2\overline{z}b}{2\varepsilon} + \dfrac{\overline{m_2}^2 v_1 \Theta}{(\Theta\overline{z} - b)^2} - \dfrac{2\overline{m_2} v_3}{\Theta\overline{z} - b} - \dfrac{(\Theta\overline{z} - b)}{\varepsilon} v_1 \end{bmatrix}, \tag{6.9c}$$

$$\overline{\boldsymbol{S}}^B = \begin{bmatrix} 0 \\ -\overline{z}\partial_x b/\varepsilon \\ -\overline{z}\partial_y b/\varepsilon \end{bmatrix}, \quad \widetilde{\boldsymbol{S}}^B = \begin{bmatrix} 0 \\ -v_1\partial_x b/\varepsilon \\ -v_1\partial_y b/\varepsilon \end{bmatrix}, \quad \overline{\boldsymbol{S}}^C = \begin{bmatrix} 0 \\ \overline{m_2}/\varepsilon \\ -\overline{m_1}/\varepsilon \end{bmatrix}, \quad \widetilde{\boldsymbol{S}}^C = \begin{bmatrix} 0 \\ v_3/\varepsilon \\ -v_2/\varepsilon \end{bmatrix}, \tag{6.9d}$$

and $\widehat{\boldsymbol{Z}}^B = \widehat{\boldsymbol{Z}}^C = \boldsymbol{0}$. Similar to Chapters 4 and 5, one can verify that the Jacobian matrices $\widehat{\boldsymbol{F}}'$ and $\widetilde{\boldsymbol{F}}'$ have complete sets of eigenvectors, and that the eigenvalues of $\widehat{\boldsymbol{F}}'$ are non-stiff as there is no $\mathcal{O}(1/\varepsilon)$ term in $\widehat{\boldsymbol{F}}$ (note that $\Theta = F\varepsilon$). So, the splitting is admissible in the sense of Definition 3.1.1. One can confirm that, besides the scaling, the only difference of the system in this chapter with the one in Chapter 5 is the additional Coriolis force $\boldsymbol{S}^C$.

Based on RS-IMEX algorithm, the scheme can be written as a two-step explicit-implicit scheme:

$$\begin{aligned} \boldsymbol{V}_{ij}^{n+1/2} = \boldsymbol{V}_{ij}^n &- \frac{\Delta t}{\Delta x}\left(\widehat{\boldsymbol{F}}^n_{1,i+1/2j} - \widehat{\boldsymbol{F}}^n_{1,i-1/2j}\right) - \frac{\Delta t}{\Delta y}\left(\widehat{\boldsymbol{F}}^n_{2,ij+1/2} - \widehat{\boldsymbol{F}}^n_{2,ij-1/2}\right) \\ &+ \Delta t\left(\widehat{\boldsymbol{S}}^B_{ij} + \widehat{\boldsymbol{S}}^C_{ij}\right)^n, \end{aligned} \tag{6.10a}$$

$$\begin{aligned} \boldsymbol{V}_{ij}^{n+1} = \boldsymbol{V}_{ij}^{n+1/2} &- \frac{\Delta t}{\Delta x}\left(\widetilde{\boldsymbol{F}}^{n+1}_{1,i+1/2j} - \widetilde{\boldsymbol{F}}^{n+1}_{1,i-1/2j}\right) - \frac{\Delta t}{\Delta y}\left(\widetilde{\boldsymbol{F}}^{n+1}_{2,ij+1/2} - \widetilde{\boldsymbol{F}}^{n+1}_{2,ij-1/2}\right) \\ &+ \Delta t\left(\widetilde{\boldsymbol{S}}^B_{ij} + \widetilde{\boldsymbol{S}}^C_{ij}\right)^{n+1} - \Delta t\,\overline{\boldsymbol{T}}^{n+1}_{ij}, \end{aligned} \tag{6.10b}$$

for each cell $(i,j) \in \{1,2,\ldots,N_x\} \times \{1,2,\ldots,N_y\}$ in the computational domain Ω_N, with spatial steps Δx and Δy and the time step Δt, where $\widetilde{\boldsymbol{F}}_{i+1/2j}$ and $\widehat{\boldsymbol{F}}_{i+1/2j}$ are Rusanov fluxes at cell interfaces as defined in Section 4.2.1, with $\widehat{\boldsymbol{F}}$ and $\widetilde{\boldsymbol{F}}$ as in (6.9b)–(6.9c). $\widetilde{\boldsymbol{S}}^{n+1}_{ij}$ is the central discretisation of the source terms in (6.9d) and $\overline{\boldsymbol{T}}^{n+1}_{ij}$ is the discretisation of the residual of the reference solution, and is computed as

$$\overline{\boldsymbol{T}}^{n+1}_{ij} = \frac{\overline{\boldsymbol{U}}^{n+1}_{ij} - \overline{\boldsymbol{U}}^n_{ij}}{\Delta t} + \frac{\overline{\boldsymbol{F}}^{n+1}_{1,i+1/2j} - \overline{\boldsymbol{F}}^{n+1}_{1,i-1/2j}}{\Delta x} + \frac{\overline{\boldsymbol{F}}^{n+1}_{2,ij+1/2} - \overline{\boldsymbol{F}}^{n+1}_{2,ij-1/2}}{\Delta y} - \left(\overline{\boldsymbol{S}}^B_{ij} + \overline{\boldsymbol{S}}^C_{ij}\right)^{n+1}. \tag{6.11}$$

By denoting $\nabla_{h,x}$ and $\Delta_{h,x}$ respectively as the central discretisations of the first and second derivatives in the x-direction, one can rewrite (6.10a)–(6.10b) as

$$\boldsymbol{V}_{ij}^{n+1/2} = \boldsymbol{V}_{ij}^n - \Delta t\nabla_{h,x}\widehat{\boldsymbol{F}}^n_{1,ij} - \Delta t\nabla_{h,y}\widehat{\boldsymbol{F}}^n_{2,ij} + \frac{\widehat{\alpha}_1\Delta x}{2}\Delta t\Delta_{h,x}\boldsymbol{V}^n_{ij} + \frac{\widehat{\alpha}_2\Delta x}{2}\Delta t\Delta_{h,y}\boldsymbol{V}^n_{ij}, \tag{6.12a}$$

$$\begin{aligned} \boldsymbol{V}_{ij}^{n+1} = \boldsymbol{V}_{ij}^{n+1/2} &- \Delta t\nabla_{h,x}\widetilde{\boldsymbol{F}}^{n+1}_{1,ij} - \Delta t\nabla_{h,y}\widetilde{\boldsymbol{F}}^{n+1}_{2,ij} + \frac{\widetilde{\alpha}_1\Delta x}{2}\Delta t\Delta_{h,x}\boldsymbol{V}^{n+1}_{ij} + \frac{\widetilde{\alpha}_2\Delta x}{2}\Delta t\Delta_{h,y}\boldsymbol{V}^n_{ij} \\ &+ \Delta t\left(\widetilde{\boldsymbol{S}}^B_{ij} + \widetilde{\boldsymbol{S}}^C_{ij}\right)^{n+1} - \Delta t\,\overline{\boldsymbol{T}}^{n+1}_{ij}. \end{aligned} \tag{6.12b}$$

Assuming $\overline{\alpha}_{1,2} = 0$ and with the reference surface perturbation $\overline{z}$ and the reference momentum

field $\overline{\boldsymbol{m}}$, one can write $\overline{\boldsymbol{T}}$ block-wise as $\overline{\boldsymbol{T}}_{\Delta}^{n+1} := [\overline{\boldsymbol{T}}_{1,\Delta}^{n+1}, \overline{\boldsymbol{T}}_{2,\Delta}^{n+1}, \overline{\boldsymbol{T}}_{3,\Delta}^{n+1}]^T$ such that

$$
\begin{aligned}
\overline{\boldsymbol{T}}_{1,ij}^{n+1} &= D_t \overline{z}_{ij}^{n} + \frac{1}{\Theta}\left(\nabla_{h,x}\overline{m_1}_{ij} + \nabla_{h,x}\overline{m_2}_{ij}\right)^{n+1}, \\
\overline{\boldsymbol{T}}_{2,ij}^{n+1} &= D_t \overline{m_1}_{ij}^{n} + \nabla_{h,x}\left(\frac{\overline{m_1}_{ij}^2}{\Theta\overline{z}_{ij} - b_{ij}}\right)^{n+1} + \nabla_{h,y}\left(\frac{\overline{m_1}_{ij}\overline{m_2}_{ij}}{\Theta\overline{z}_{ij} - b_{ij}}\right)^{n+1} + \frac{1}{2\varepsilon}\nabla_{h,x}\left(\Theta\overline{z}_{ij}^2 - 2b_{ij}\overline{z}_{ij}\right)^{n+1} \\
&\quad + \frac{1}{\varepsilon}\overline{z}_{ij}^{n+1}\nabla_{h,x}b_{ij} - \frac{1}{\varepsilon}\overline{m_2}_{ij}^{n+1}, \\
\overline{\boldsymbol{T}}_{3,ij}^{n+1} &= D_t \overline{m_2}_{ij}^{n} + \nabla_{h,x}\left(\frac{\overline{m_1}_{ij}\overline{m_2}_{ij}}{\Theta\overline{z}_{ij}^{n+1} - b_{ij}}\right)^{n+1} + \nabla_{h,y}\left(\frac{\overline{m_1}_{ij}^2}{\Theta\overline{z}_{ij} - b_{ij}}\right)^{n+1} + \frac{1}{2\varepsilon}\nabla_{h,y}\left(\Theta\overline{z}_{ij}^2 - 2b_{ij}\overline{z}_{ij}\right)^{n+1} \\
&\quad + \frac{1}{\varepsilon}\overline{z}_{ij}^{n+1}\nabla_{h,y}b_{ij} + \frac{1}{\varepsilon}\overline{m_1}_{ij}^{n+1}.
\end{aligned}
\tag{6.13}
$$

So far, the scheme for computing the perturbation $\boldsymbol{V}_{\Delta}$ has been introduced. The remaining point to be clarified is how to solve the equations for the reference solution, which is explained in the next section. Note that from now on and for the sake of simplicity, we assume the same number of grid points in both directions, *i.e.*, $N_x = N_y = N$. Also, we pick $\widehat{\alpha}_1 = \widehat{\alpha}_2$ and $\widetilde{\alpha}_1 = \widetilde{\alpha}_2$.

6.2.2 Solving for the reference solution

As explained before, we consider the solution of QGE as the reference solution. This system is, in fact, very well-known in the meteorology and has been studied numerically since [CFvN50]. Although the system seems to be simple, obtaining stable numerical schemes is very challenging. A very successful idea is due to Arakawa [Ara66]; he showed that using a particular staggered grid, one can obtain non-linear stability for the semi-discrete scheme. Note that, as remarked in [KM05], solving (6.5a)–(6.5c), which are in the vorticity-stream function form, is preferable as it conserves automatically the geostrophy. For the sake of completeness, we review the Arakawa Jacobian method briefly and refer the reader to consult [KAK11] and [Dur13, Sect 3.6].

Here, we implement the Arakawa Jacobian method as in [KAK11], using a predictor-corrector approach. By having ψ^n from the initial height, we obtain ξ^{n+1} using the *Arakawa Jacobian*. Then, we solve for ψ^{n+1}, and redo this procedure to correct the predicted solution. The key ingredient of this method is the Arakawa Jacobian, which provides a recipe for a stable discretisation of (6.5c), rewritten as

$$
\partial_t \xi + J(\psi, \xi) = 0, \qquad J(\psi, \xi) := \frac{\partial \psi}{\partial x}\frac{\partial \xi}{\partial y} - \frac{\partial \psi}{\partial y}\frac{\partial \xi}{\partial x}. \tag{6.14}
$$

A naïve discretisation of (6.14) leads to instability as observed in [Phi59]. Arakawa in [Ara66] introduced a particular discretisation of this Jacobian, denoted here by $J_{Arakawa}$ such that the semi-discrete (discrete in space) scheme preserves the mean kinetic energy, mean PV, mean square PV *(enstrophy)* and the mean wave number.[2] Thus, the schemes provides some sort of non-linear stability; see [Dur13]. Thanks to periodic domains, we neglect the boundary treatment issues detailed in [KAK11, KM05]. Algorithm 2 provides the sketch of the method.

[2] Note that the original work of Arakawa was about the barotropic vorticity equation $(\partial_t + \boldsymbol{u}\cdot\nabla_x)\zeta = 0$. Here, we have ξ instead of ζ.

Algorithm 2 Arakawa method

1: Consider the initial height as the stream function ψ_Δ^n.
2: Compute the initial PV as $\xi_\Delta^n = (\Delta_{h,\boldsymbol{x}} - F)\,\psi_\Delta^n + F\eta_\Delta^b$.
3: Predict $\xi_\Delta^{n+1/2}$ as $\xi_\Delta^{n+1/2} = \xi_\Delta^n - \Delta t J_{Arakawa}(\psi_\Delta^n, \xi_\Delta^n)$.
4: Predict $\psi_\Delta^{n+1/2}$ as $(\Delta_{h,\boldsymbol{x}} - F)\,\psi_\Delta^{n+1/2} = \xi_\Delta^{n+1/2} - F\eta_\Delta^b$
5: Repeat steps 2–4 to correct predicted values and obtain $(\psi_\Delta^{n+1}, \xi_\Delta^{n+1})$:

$$\begin{aligned}\xi_\Delta^{n+1/2\dagger} &= (\Delta_{h,\boldsymbol{x}} - F)\,\psi_\Delta^{n+1/2} + F\eta_\Delta^b\\ \xi_\Delta^{n+1} &= \xi_\Delta^n - \Delta t J_{Arakawa}(\psi_\Delta^{n+1/2}, \xi_\Delta^{n+1/2\dagger})\\ (\Delta_{h,\boldsymbol{x}} - F)\,\psi_\Delta^{n+1} &= \xi_\Delta^{n+1} - F\eta_\Delta^b\end{aligned}$$

6: Continue with step 2 with the initial value $(\psi_\Delta^{n+1}, \xi_\Delta^{n+1})$.

6.3 Asymptotic analysis of the scheme

Before we proceed with the main theorem of this section, let us fix the definition of the well-prepared initial data:

Definition 6.3.1. *For the RSWE* (6.3)*, we call the initial data* $(z_{0,\varepsilon}, \boldsymbol{u}_{0,\varepsilon})$ *well-prepared if it holds that*

$$\begin{aligned}z(0,\cdot) &= z_{0,\varepsilon} = z_{(0)}^0 + \varepsilon z_{(1),\varepsilon}^0,\\ \boldsymbol{u}(0,\cdot) &= \boldsymbol{u}_{0,\varepsilon} = \boldsymbol{u}_{(0)}^0 + \varepsilon \boldsymbol{u}_{(1),\varepsilon}^0,\end{aligned} \tag{6.15}$$

where $(z_{(0)}^0, \boldsymbol{u}_{(0)}^0)$ *is the solution of the QGE, i.e.,* $\boldsymbol{u}_{(0)}^0$ *is solenoidal with the stream function* $z_{(0)}^0$*.*

Theorem 6.3.2. *For the rotating shallow water equations with topography and a well-prepared initial datum in a periodic domain, the RS-IMEX scheme* (6.12a)–(6.12b)*, with* (6.9a)–(6.9d)*, the QGE reference solution, a constant* $\widetilde{\alpha}$*, and under an* ε*-uniform time step restriction*

(i) is solvable, i.e., it has a unique solution for all $\varepsilon > 0$*.*

(ii) is consistent with the asymptotic limit in the fully-discrete settings, i.e., it is asymptotically consistent.

(iii) is asymptotically ℓ_2*-stable for the fixed grid, in finite time* $T_f < \infty$ *and with a small enough initial data provided the reference solution is stable, i.e., there exists a constant* C_{N,T_f} *such that* $\|\boldsymbol{V}_\Delta^n\|_{\ell_2} \leq C_{N,T_f}\|\boldsymbol{V}_\Delta^0\|_{\ell_2}$*.*

(iv) preserves the lake at rest (LaR) equilibrium state, provided that both $\overline{\boldsymbol{U}}_\Delta$ *and* $\boldsymbol{V}_\Delta$ *are at equilibrium.*

(v) may produce checker-board oscillations for the surface perturbation only as small as $\mathcal{O}(\varepsilon)$*.*

We discuss the proof of Theorem 6.3.2 in the next sections.

6.3.1 Solvability

Like Chapter 5, but with an additional source term and a different scaling matrix, and by assuming $\widetilde{\alpha} = 0$ and $\Delta x = \Delta y$, one can write the coefficient matrix of the implicit step, J_ε, as

$$J_\varepsilon = \begin{bmatrix} \mathbb{I}_{N^2} & \frac{\beta}{\varepsilon} J_{12} & \frac{\beta}{\varepsilon} J_{13} \\ \frac{\beta}{\varepsilon} J_{21} & \mathbb{I}_{N^2} + \beta J_{22} & \beta J_{23} - \frac{\Delta t}{\varepsilon} \mathbb{I}_{N^2} \\ \frac{\beta}{\varepsilon} J_{31} & \beta J_{32} + \frac{\Delta t}{\varepsilon} \mathbb{I}_{N^2} & \mathbb{I}_{N^2} + \beta J_{33} \end{bmatrix}, \tag{6.16}$$

where all J_{ij} are of $\mathcal{O}(1)$ and can be defined similarly as in (5.11):

$$\begin{aligned} J_{12} &= F^{-1} Q_x, & J_{13} &= F^{-1} Q_y, \\ J_{21} &= \operatorname{diag}(Q_x^b) + Q_x^{\overline{h}} - \varepsilon \Theta (Q_x^{\overline{u_1}^2} + Q_y^{\overline{u_1 u_2}}), & J_{22} &= 2 Q_x^{\overline{u_1}} + Q_y^{\overline{u_2}}, & J_{23} &= Q_y^{\overline{u_1}}, \\ J_{31} &= \operatorname{diag}(Q_y^b) + Q_y^{\overline{h}} - \varepsilon \Theta (Q_x^{\overline{u_1 u_2}} + Q_y^{\overline{u_2}^2}), & J_{32} &= Q_x^{\overline{u_2}}, & J_{33} &= Q_x^{\overline{u_1}} + 2 Q_y^{\overline{u_2}}, \end{aligned}$$

where Q_x^ϕ and Q_y^ϕ stand for corresponding matrices of central discretisation of ϕ in each direction and $\operatorname{diag}(Q_x^b)$ and $\operatorname{diag}(Q_y^b)$ are diagonal matrices with central discretisation of b as entries (like Chapter 5).

So, the matrix J_ε, which is the inverse of the solution operator of the implicit step (6.12b), can be rewritten as $J_\varepsilon := \mathbb{I}_{3N^2} + \Delta t \, \Xi_\varepsilon$, where Ξ_ε is a matrix not depending on Δt. Hence, with a suitable choice of Δt, none of the eigenvalues of $\Delta t \, \Xi_\varepsilon$ is equal to -1, implying that J_ε is non-singular, and the implicit step, so the whole scheme, is solvable. The proof for $\widetilde{\alpha} \neq 0$ is likewise.

6.3.2 Asymptotic consistency

We discuss the asymptotic consistency of the scheme in two ways, rigorously and formally. At first in Section 6.3.2.1, we investigate the ε-stability, *i.e.*, if the perturbation $\boldsymbol{V}_\Delta$ is $\mathcal{O}(1)$. Then, in Section 6.3.2.2, we perform the formal asymptotic consistency analysis, which turns out to be rigorous, in virtue of ε-stability.

6.3.2.1 ε-stability of the implicit step

For the ε-stability of the solution, one needs to show that the solution of the implicit step is ε-stable, in addition to the formal asymptotic analysis of the explicit step. At first, we show that J_ε has a bounded inverse in terms of ε. Unlike Chapter 5, the rhs in (6.13) will be shown to be $\mathcal{O}(1)$. This concludes the ε-stability of the solution computed by the implicit step, owing to the boundedness of the implicit solution operator.

With the same procedure as in previous chapters, one can confirm that the numerical range of $J_\varepsilon^* J_\varepsilon$ is

$$W(J_\varepsilon^* J_\varepsilon) = \left\| \frac{\beta}{\varepsilon} J_{12} \boldsymbol{w}_2 + \frac{\beta}{\varepsilon} J_{13} \boldsymbol{w}_3 + \boldsymbol{w}_1 \right\|_{\ell_2}^2 + \left\| \frac{\beta}{\varepsilon} J_{21} \boldsymbol{w}_1 - \frac{\Delta t}{\varepsilon} \boldsymbol{w}_3 + \beta J_{23} \boldsymbol{w}_3 + \beta J_{22} \boldsymbol{w}_2 + \boldsymbol{w}_2 \right\|_{\ell_2}^2$$

$$+\left\|\frac{\beta}{\varepsilon}J_{31}\boldsymbol{w}_1+\frac{\Delta t}{\varepsilon}\boldsymbol{w}_2+\beta J_{32}\boldsymbol{w}_2+\beta J_{33}\boldsymbol{w}_3+\boldsymbol{w}_3\right\|_{\ell_2}^2,$$

where $\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{w}_3\in\mathbb{C}^N$ and $\|\boldsymbol{w}_1\|_{\ell_2}^2+\|\boldsymbol{w}_2\|_{\ell_2}^2+\|\boldsymbol{w}_3\|_{\ell_2}^2=1$. Similar to Chapter 5, one can argue by contradiction to show that the numerical range cannot approach zero. Assume that $W(J_\varepsilon^*J_\varepsilon)$ approaches zero in the limit; so,

$$\boldsymbol{w}_1=-\frac{\beta}{\varepsilon}(J_{12}\boldsymbol{w}_2+J_{13}\boldsymbol{w}_3)+o(1), \tag{6.17a}$$

$$(\mathbb{I}_{N^2}+\beta J_{22})\,\boldsymbol{w}_2=-\frac{\beta}{\varepsilon}J_{21}\boldsymbol{w}_1-\left(\beta J_{23}-\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\right)\boldsymbol{w}_3+o(1), \tag{6.17b}$$

$$(\mathbb{I}_{N^2}+\beta J_{33})\,\boldsymbol{w}_3=-\frac{\beta}{\varepsilon}J_{31}\boldsymbol{w}_1-\left(\beta J_{32}+\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\right)\boldsymbol{w}_2+o(1). \tag{6.17c}$$

With a suitable choice of β, one can confirm that $\beta J_{23}-\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}$ and $\beta J_{32}+\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}$ are invertible with a bounded inverse, *i.e.*,

$$\lim_{\varepsilon\to0}\left\|\left(\beta J_{23}-\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\right)^{-1}\right\|=\mathcal{O}(\varepsilon),\qquad\lim_{\varepsilon\to0}\left\|\left(\beta J_{32}+\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\right)^{-1}\right\|=\mathcal{O}(\varepsilon). \tag{O1}$$

So, using (6.17b), (6.17c) and (O1) yields

$$\begin{aligned}\boldsymbol{w}_3&=\left(\beta J_{23}-\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\right)^{-1}\Big(-\frac{\beta}{\varepsilon}J_{21}\boldsymbol{w}_1-\Big(\mathbb{I}_{N^2}+\beta J_{22}\Big)\boldsymbol{w}_2\Big)+o(\varepsilon),\\ \boldsymbol{w}_2&=\left(\beta J_{32}+\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\right)^{-1}\Big(-\frac{\beta}{\varepsilon}J_{31}\boldsymbol{w}_1-\Big(\mathbb{I}_{N^2}+\beta J_{33}\Big)\boldsymbol{w}_3\Big)+o(\varepsilon).\end{aligned} \tag{6.18}$$

Manipulating (6.18), one can find a relation for $\boldsymbol{w}_2$ and $\boldsymbol{w}_3$ in terms of $\boldsymbol{w}_1$:

$$\begin{aligned}P_2'\boldsymbol{w}_2&=P_2''\boldsymbol{w}_1+o(\varepsilon),\\ P_3'\boldsymbol{w}_3&=P_3''\boldsymbol{w}_1+o(\varepsilon),\end{aligned} \tag{6.19}$$

where

$$\begin{aligned}P_2'&:=\mathbb{I}_{N^2}-\Big(\beta J_{32}+\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\Big)^{-1}\Big(\mathbb{I}_{N^2}+\beta J_{33}\Big)\Big(\beta J_{23}-\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\Big)^{-1}\Big(\mathbb{I}_{N^2}+\beta J_{22}\Big),\\ P_2''&:=-\frac{\beta}{\varepsilon}\Big(\beta J_{32}+\frac{\Delta t}{\varepsilon}\Big)^{-1}J_{31}+\frac{\beta}{\varepsilon}\Big(\beta J_{32}+\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\Big)^{-1}\Big(\mathbb{I}_{N^2}+\beta J_{33}\Big)\Big(\beta J_{23}-\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\Big)^{-1}J_{21},\\ P_3'&:=\mathbb{I}_{N^2}-\Big(\beta J_{23}-\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\Big)^{-1}\Big(\mathbb{I}_{N^2}+\beta J_{22}\Big)\Big(\beta J_{32}+\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\Big)^{-1}\Big(\mathbb{I}_{N^2}+\beta J_{33}\Big),\\ P_3''&:=-\frac{\beta}{\varepsilon}\Big(\beta J_{23}-\frac{\Delta t}{\varepsilon}\Big)^{-1}J_{21}+\frac{\beta}{\varepsilon}\Big(\beta J_{23}-\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\Big)^{-1}\Big(\mathbb{I}_{N^2}+\beta J_{22}\Big)\Big(\beta J_{32}+\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\Big)^{-1}J_{31}.\end{aligned}$$

Note that matrices P_2'' and P_3'' are bounded because, using (O1), it is easy to confirm that

$$\lim_{\varepsilon\to0}\left\|\frac{\beta}{\varepsilon}\Big(\beta J_{23}-\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\Big)^{-1}J_{21}\right\|=\mathcal{O}(1),\qquad\lim_{\varepsilon\to0}\left\|\frac{\beta}{\varepsilon}\Big(\beta J_{32}+\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\Big)^{-1}J_{31}\right\|=\mathcal{O}(1). \tag{O2}$$

So, using (O1) and (O2), one gets

$$P_2'=\mathbb{I}_{N^2}-\mathcal{O}(\varepsilon^2),\qquad P_2''=-\frac{\beta}{\varepsilon}\Big(\beta J_{32}+\frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\Big)^{-1}J_{31}+\mathcal{O}(\varepsilon),$$

$$P_3' = \mathbb{I}_{N^2} - \mathcal{O}(\varepsilon^2), \qquad P_3'' = -\frac{\beta}{\varepsilon}\Big(\beta J_{23} - \frac{\Delta t}{\varepsilon}\mathbb{I}_{N^2}\Big)^{-1} J_{21} + \mathcal{O}(\varepsilon).$$

Since P_2' and P_3' are boundedly-invertible, we define $P_2 := (P_2')^{-1}P_2''$ and $P_3 := (P_3')^{-1}P_3''$ to rewrite (6.17a) only in terms of $\boldsymbol{w}_1$ as

$$\Big(\mathbb{I}_{N^2} + \frac{\beta}{\varepsilon}J_{12}P_2 + \frac{\beta}{\varepsilon}J_{13}P_3\Big)\boldsymbol{w}_1 = o(1). \tag{6.21}$$

Since the bottom is almost flat, $\|\nabla_{h,\boldsymbol{x}} b_{ij}\| = \mathcal{O}(\varepsilon)$, the leading order of J_{21} and J_{31} are the same as J_{12} and J_{13} (up to a scaling). Also, $(\frac{\beta\varepsilon}{\Delta t}J_{32} + \mathbb{I}_{N^2})^{-1}$ is like $\mathbb{I}_{N^2} + \mathcal{O}(\frac{\beta\varepsilon}{\Delta t})$. These imply that the leading order terms of (6.21) vanish since $[J_{12}, J_{13}] = \mathbf{0}_{N^2}$.

Balancing $\mathcal{O}(1)$ terms shows that $(\mathbb{I}_{N^2} + \frac{\beta}{\varepsilon}\Xi)\boldsymbol{w}_1^{(0)} = \mathbf{0}$, where Ξ consists of $\mathcal{O}(\varepsilon)$ terms in P_2 and P_3. So, it is plausible to claim that with a suitable choice of β, this matrix is non-singular and $\boldsymbol{w}_1^{(0)} = \mathbf{0}$; our numerical evidence verifies this. Thus, due to (6.19), one finds $\boldsymbol{w}_2^{(0)} = \boldsymbol{w}_3^{(0)} = \mathbf{0}$; thus, $\lim_{\varepsilon\to 0}(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3) = (\mathbf{0}, \mathbf{0}, \mathbf{0})$, which contradicts the assumption that $\|\boldsymbol{w}\|_{\ell_2} = 1$ and concludes the ε-stability of the implicit solution *operator* since the numerical range cannot tend to zero.

However, we need ε-stability of *the solution* for the rigorous asymptotic consistency. For that, one also needs to show that $\|\overline{\boldsymbol{T}}_\Delta\| = \mathcal{O}(1)$, which concludes that the solution of the implicit step is $\mathcal{O}(1)$, owing to the boundedness of J_ε^{-1}. We show that the projection of the reference solution on the discrete grid is consistent to the leading order such that no large term remains in $\overline{\boldsymbol{T}}_\Delta$. Showing that, assume $(\psi_{ij}, \overline{\boldsymbol{u}}_{ij})$ to be an approximate solution of the QGE. By construction, there is a discrete stream function which gives the discrete velocity field (by a central difference), *i.e.*, $\nabla_{h,\boldsymbol{x}}\psi_{ij} \equiv \overline{\boldsymbol{u}}_{ij}^{\perp}$. This implies that $\nabla_{h,\boldsymbol{x}} \cdot \overline{\boldsymbol{u}}_{ij} \equiv 0$. Note that our scaling assumptions mean that the bottom topography is almost flat (with $\mathcal{O}(\varepsilon)$ deviations); thus, $\nabla_{h,\boldsymbol{x}} \cdot \overline{\boldsymbol{m}}_{ij} \equiv 0$ and $\|\overline{\boldsymbol{T}}_{1,\Delta}\| = \mathcal{O}(1)$. For $\overline{\boldsymbol{T}}_{2,\Delta}$ and $\overline{\boldsymbol{T}}_{3,\Delta}$, one can see that $\mathcal{O}(1/\varepsilon)$ terms gives $\nabla_{h,\boldsymbol{x}}\psi_{ij} - \overline{\boldsymbol{u}}_{ij}^{\perp}$, which vanishes by construction; so, $\|\overline{\boldsymbol{T}}_{3,\Delta}\|, \|\overline{\boldsymbol{T}}_{3,\Delta}\| = \mathcal{O}(1)$. Hence, $\mathcal{O}(1/\varepsilon)$ terms in $\overline{\boldsymbol{T}}_\Delta$ vanish in the limit. This concludes the proof of ε-stability, owing to ε-stability of the explicit step, which is topic of the next section.

Remark 6.3.3. *Note that in Chapter 5, the projection scheme has been used for the lake equations, which does not satisfy the div-free condition exactly. So, for that case, the proof of the ε-stability required studying the structure of J_ε^{-1}.*

6.3.2.2 Formal asymptotic consistency

Firstly, we show that the explicit step is ε-stable, *i.e.*, it does not produce large, namely $\mathcal{O}(1/\varepsilon)$, contributions in the explicit update. We assume that $\|\boldsymbol{V}_\Delta^n\| = \mathcal{O}(1)$, which is compatible with the well-prepared initial data, and confirm that $\|\boldsymbol{V}_\Delta^{n+1/2}\| = \mathcal{O}(1)$. Since $\widehat{\boldsymbol{F}}_{1,1} = \widehat{\boldsymbol{F}}_{2,1} = 0$, one can immediately conclude that $\|\boldsymbol{V}_{1,\Delta}^{n+1/2}\| = \mathcal{O}(1)$. For $\boldsymbol{V}_{2,\Delta}^{n+1/2}$ (and similarly $\boldsymbol{V}_{3,\Delta}^{n+1/2}$), one simply gets that as $\varepsilon \to 0$ (note that $\Theta = F\varepsilon$)

$$\nabla_{h,x}\left(\frac{m_1^2}{\Theta z - b} + \frac{\Theta z^2 - 2zb}{2\varepsilon} - \frac{\overline{m_1}^2}{\Theta\overline{z} - b} - \frac{\Theta\overline{z}^2 - 2\overline{z}b}{2\varepsilon} + \frac{\overline{m_1}^2 v_1\Theta}{(\Theta\overline{z} - b)^2} - \frac{2\overline{m_1}v_2}{\Theta\overline{z} - b} - \frac{(\Theta\overline{z} - b)}{\varepsilon}v_1\right)_{ij}^n$$
$$+\nabla_{h,y}\left(\frac{m_1 m_2}{\Theta z - b} - \frac{\overline{m_1 m_2}}{\Theta\overline{z} - b} + \frac{\overline{m_1 m_2}v_1\Theta}{(\Theta\overline{z} - b)^2} - \frac{\overline{m_1}v_3}{\Theta\overline{z} - b} - \frac{\overline{m_2}v_2}{\Theta\overline{z} - b}\right)_{ij}^n = \mathcal{O}(1).$$

So, $\lim_{\varepsilon\to 0}\left(\nabla_{h,x}\widehat{\boldsymbol{F}}^n_{1,2,ij}+\nabla_{h,y}\widehat{\boldsymbol{F}}^n_{2,2,ij}\right)=\mathcal{O}(1)$ and the explicit step does not change the leading order of $\boldsymbol{V}^n_{2,\Delta}$ (and $\boldsymbol{V}^n_{3,\Delta}$). This completes the ε-stability proof of the explicit step.

To complete the asymptotic consistency analysis, we show that the implicit step solution is consistent with the limit manifold. Based on the ε-stability of Section 6.3.2.1, we suppose that $\|\boldsymbol{V}^{n+1}_{\Delta}\|=\mathcal{O}(1)$. From the v_1-update and considering (6.13) and (6.9a)–(6.9c), the momentum field (up to $\mathcal{O}(\varepsilon)$) is solenoidal, *i.e.*,

$$\nabla_{h,x}\left(\overline{m_1}+v_2\right)^{n+1}_{ij}+\nabla_{h,y}\left(\overline{m_2}+v_3\right)^{n+1}_{ij}=\mathcal{O}(\varepsilon). \tag{6.22}$$

Using v_2-update (similarly for v_3), one can balance $\mathcal{O}(1/\varepsilon)$ terms, which gives

$$-\nabla_{h,x}\left(bv_{1(0)}+b\overline{z}\right)^{n+1}_{ij}=-\left(\overline{z}+v_{1(0)}\right)^{n+1}_{ij}\nabla_{h,x}b_{ij}+(\overline{m_2}+v_{3(0)})^{n+1}_{ij}. \tag{6.23}$$

This is a consistent discretisation of (6.5a) since the bottom is almost flat, $\|\nabla_{h,\boldsymbol{x}}b_{ij}\|=\mathcal{O}(\varepsilon)$. In other words, (6.23) implies that

$$\nabla_{h,\boldsymbol{x}}\left(\overline{z}+v_{1(0)}\right)^{n+1}_{ij}=\boldsymbol{u}^{\perp,n+1}_{(0),ij}.$$

Thus, up to $\mathcal{O}(\varepsilon)$, the solution is consistent with the limit manifold. Since the consistency of the evolution of the leading order of the momentum is clear (by equation (6.6)), the asymptotic consistency of the scheme is concluded, but only up to possible checker-board oscillations for the momentum field in the null space of central difference operators $\nabla_{h,x}$ and $\nabla_{h,y}$. Note that the ε-stability of the solution implies immediately that since $\|\boldsymbol{V}^{n+1}_{1,\Delta}\|=\mathcal{O}(1)$, the possible checker-board oscillations for the surface perturbation are $\mathcal{O}(\varepsilon)$ for $\varepsilon\ll 1$.

Remark 6.3.4. *The asymptotic stability of the scheme can be carried out similarly as in Chapter 5 as, assuming the ε-stability of the implicit solution operator, the additional stiff source term does not make any difference for the analysis.*

6.3.3 Well-balancing

It is already discussed in Chapter 5 that well-balancing analysis for the RS-IMEX scheme may be challenging; the reason is that even for relatively simple steady states, the analysis should handle moving equilibria. In fact, by decomposing the solution, we gain more information on the asymptotic behaviour of the solution for $\varepsilon\ll 1$, but at the same time, we lose another part of the information as we only know that the sum of the reference solution and its perturbation is at equilibrium, which gives us no specific knowledge about individual parts. This issue would be more pronounced for rotating shallow water equations with the quasi-geostrophic equilibrium as in this case even the reference surface perturbation is not constant. For this case, $\overline{\boldsymbol{U}}_{\Delta}$ is at equilibrium by construction, *i.e.*, $\overline{\boldsymbol{U}}^n_{\Delta}\in\mathcal{U}^{\Delta}_{GE}$ for all n, where $\mathcal{U}^{\Delta}_{GE}$ is the geostrophic equilibrium manifold defined as

$$\mathcal{U}^{\Delta}_{GE}:=\left\{\begin{bmatrix}z_{ij}\\ m_{1,ij}\\ m_{2,ij}\end{bmatrix}\;\middle|\;\nabla_{h,\boldsymbol{x}}h_{ij}=\boldsymbol{u}^{\perp}_{ij},\forall(i,j)\in\Omega_N\right\}.$$

To prove the well-balancing w.r.t. this equilibrium, one needs to show that $\boldsymbol{V}^{n+1}_{\Delta}\in\mathcal{U}^{\Delta}_{GE}$ provided that $\boldsymbol{V}^{n}_{\Delta}\in\mathcal{U}^{\Delta}_{GE}$, which is not true for the scheme (6.12a)–(6.12b) without any additional well-balancing mechanism.

We conclude the well-balancing discussion with the following lemma, which shows that assuming both split parts of the solution to be at rest equilibrium, the well-balancing of the scheme holds.

Lemma 6.3.5. *For the RS-IMEX scheme* (6.12a)–(6.12b) *in a periodic domain, assume that* $\boldsymbol{U}_\Delta^n \in \mathcal{U}_{LaR}^\Delta$ *(defined as* (5.18)*). If* $\overline{\boldsymbol{U}}_\Delta^n, \boldsymbol{V}_\Delta^n \in \mathcal{U}_{LaR}^\Delta$ *then* $\overline{\boldsymbol{U}}_\Delta^{n+1}, \boldsymbol{V}_\Delta^{n+1} \in \mathcal{U}_{LaR}^\Delta$*. So, the scheme is well-balanced regarding the lake at rest equilibrium state.*

Proof. One can check that the reference solution will be stationary, *i.e.*, $\overline{\boldsymbol{U}}_\Delta^{n+1} \in \mathcal{U}_{LaR}^\Delta$. So, the Coriolis force vanishes and the problem is reduced to the well-balancing for the shallow water equations with topography as studied in Chapter 5. This concludes the proof. □

6.4 Numerical experiments

In this section, we test the quality of the computed solutions by the RS-IMEX scheme and corroborate the AP property with the help of several numerical examples. The time step is computed as in Section 5.3, with an additional constraint for the time step required for the Arakawa method. The choices of $c_{\widetilde{\alpha}}$ and $c_{\widehat{\alpha}}$ are reported for each example. Note that motivated by the well-balancing discussion in Appendix 5.B, we set $\overline{\alpha} = \widehat{\alpha}$ and tune it with $c_{\overline{\alpha}}$, which will be reported for each example.

6.4.1 (i) 1d Rossby adjustment in an open domain

This example is a classical *Rossby adjustment* [Ros38]—the relaxation of an arbitrary initial configuration toward the state of linear geostrophic equilibrium—of an unbalanced jet-shaped momentum in the open domain $[-20, 20]$, and has been investigated by several authors, *e.g.*, [AKO09, BLSZ04, CDKLM14, LMNK07]. The initial datum is a rest state superimposed by a 1d jet (a localised uni-directional velocity distribution):

$$z(0,x) = 0, \qquad u_1(0,x) = 0, \qquad u_2(0,x) = \frac{2(1+\tanh(4x/\ell+2))(1-\tanh(4x/\ell-2))}{(1+\tanh(2))^2},$$

where the maximum zonal velocity is one and the width of the jet $\ell = 2$. Also, $\eta^b(x,y) = 0$, $H_{\text{mean}} = 1$ and $f = g = 1$. Adopting this example in the framework of system (6.3), we should non-dimensionalise (6.1); so, we pick $H_\circ, u_\circ, t_\circ, L_\circ = 1$ and $Z_\circ = f u_\circ L_\circ / g = 1$, which gives $Ro, Fr, F, \Theta, \varepsilon = 1$. Note that this choice for the non-dimensionalisation is only for the sake of simplicity; the physical choice is $L_\circ = \ell$ and $u_\circ = 2$, which gives $Z_\circ, \Theta, F = 4$ and $Fr = 2$. For the RS-IMEX scheme, we pick $c_{\widetilde{\alpha}} = 0$ and $c_{\widehat{\alpha}} = 1$ and the zero reference solution.

The initial jet adjusts a momentum unbalance in a transient phase, leading to the emission of gravity waves out of the jet. The time evolution of the water height is illustrated in Figure 6.1 and matches the aforementioned works quite well. As time evolves, the solution tends to the geostrophic balance as demonstrated in Figure 6.2. The conservation of the potential vorticity is confirmed by Figure 6.3. The initial profile is a bit shifted to the right, but dissipated as well, due to the inherent diffusive behaviour of first-order schemes. As pointed out in [BLSZ04], even for long time simulations, there are still small oscillations around the geostrophic equilibrium

because some modes with the frequencies close to f remain for a longer time in the core of the jet. These very slow propagating waves have almost zero group velocities; see [BLSZ04].

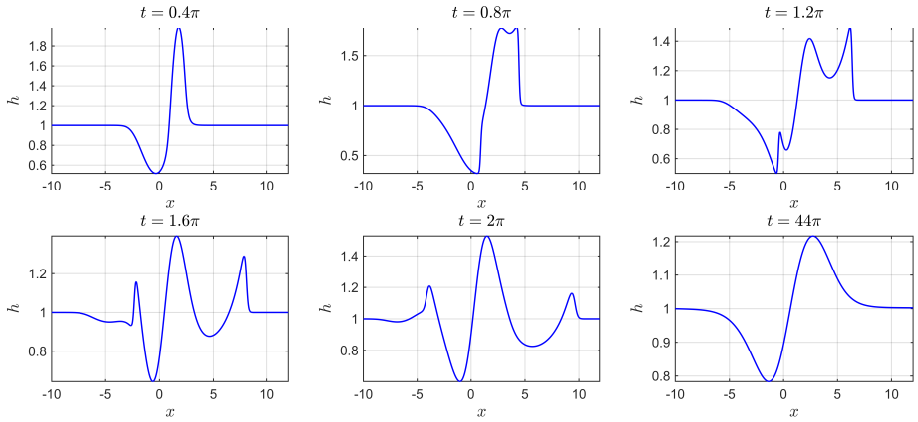
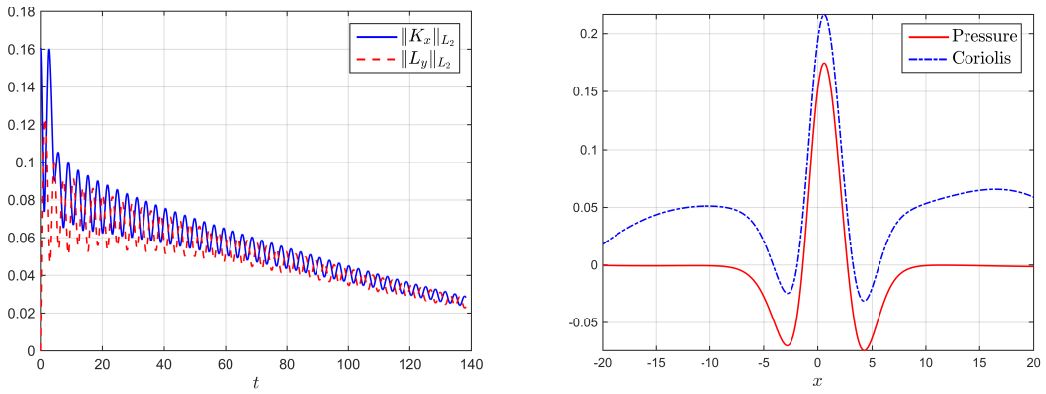


Figure 6.1: Evolution of the water height in Example (i), computed with $N_x = 10000$, CFL $= 0.45$ and for $T_f = 44\pi$.



(a) (b)

Figure 6.2: Adjustment toward the geostrophic equilibrium in Example (i).

6.4.2 (ii) 1d geostrophic steady state

This example is as [CDKLM14, Ex. 1] (see also [CLP08]), in the domain $[-5, 5]$ with open boundaries. The bottom topography is flat $\eta^b(x, y) = 0$, $f = 10$ and $g = 1$, when the flow is initially at the geostrophic equilibrium with

$$K(0, x) = 2, \qquad u_1(0, x) = 0, \qquad u_2(0, x) = \frac{2g}{f} x e^{-x^2}.$$

For non-dimensionalisation, we pick $H_\circ, u_\circ, t_\circ, L_\circ = 1$. Also, $Z_\circ = f u_\circ L_\circ / g = 10$ should be chosen, which gives $Ro = \varepsilon = 0.1$, $Fr = 1$, $F = 100$, $\Theta = 10$. In order to find $h(0, x)$, one should use (6.7) and solve $\Phi_x|_{t=0} = \frac{f}{g} u_2|_{t=0} = 2x e^{-x^2}$, which implies that $\Phi(0, x) = -e^{-x^2}$ and

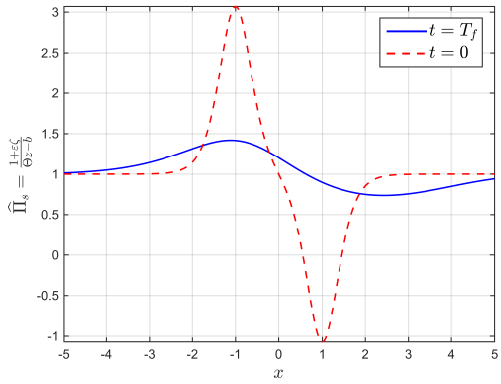
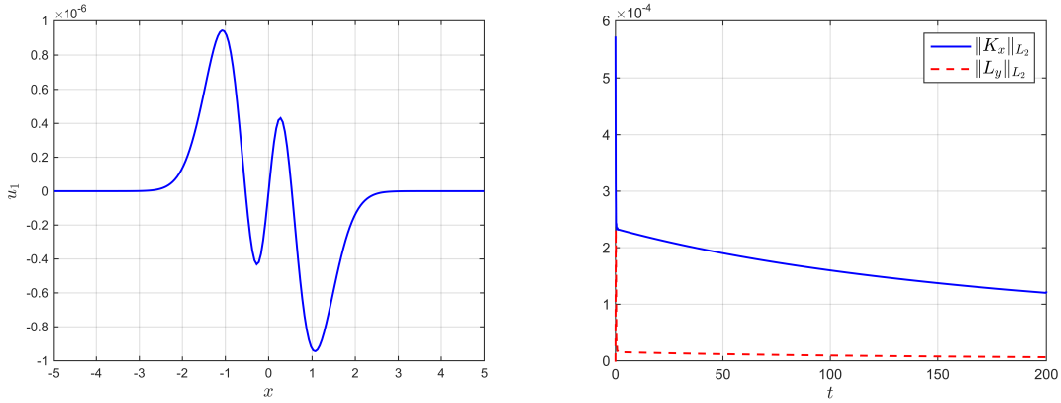


Figure 6.3: Potential vorticity profile (in x-direction) in Example (i) for $T_f = 44\pi$.

$h(0,x) = 2 - e^{-x^2}$ (up to a constant). So, we choose $H_{\text{mean}} = 2$, which is not equal to $H_\circ$. Also, for the RS-IMEX scheme, we pick $c_{\widetilde{\alpha}} = 0$ and $c_{\widehat{\alpha}} = 0.1$ and the zero reference solution.

Figure 6.4 indicates that the scheme preserves the steady state very well, as equilibrium variables u_1 and $\partial_x K$, computed using a uniform grid with 200 cells for $T_f = 200$, are still small. As u_1 is almost zero, one expects to see no advection for the potential vorticity in the x-direction, which is confirmed by Figure 6.5.



(a) (b)

Figure 6.4: Preservation of the equilibrium state in Example (ii) by the RS-IMEX scheme, computed with $N_x = 200$, CFL $= 0.45$ and for $T_f = 200$.

6.4.3 (iii) 1d geostrophic steady state with a periodic bottom

This example is as [CDKLM14, Ex. 2] with $f = g = 1$ and $\eta^b(x) = 1 + \frac{f}{g}\sin(\frac{\pi}{5}x)$ (this is a shift compared to the original example to have a non-negative bottom function). The flow is initially

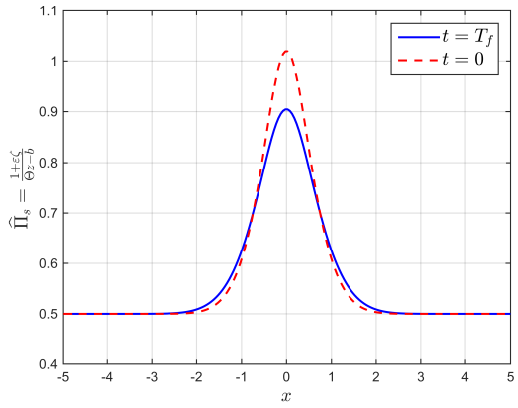


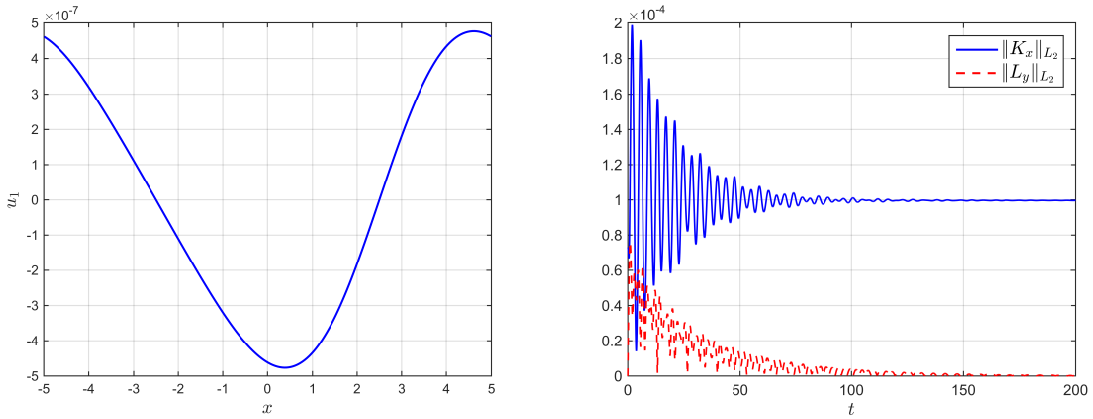
Figure 6.5: Potential vorticity profiles (in x-direction) in Example (ii) for $T_f = 200$.

at the geostrophic equilibrium in the periodic domain $[-5, 5)$ with

$$K(0,x) = 1, \qquad u_1(0,x) = 0, \qquad u_2(0,x) = \frac{\pi}{5}\cos(\frac{\pi}{5}x).$$

For non-dimensionalisation, we pick $H_\circ, u_\circ, t_\circ, L_\circ = 1$. So $Z_\circ = 1$ should be chosen, which gives $Ro, Fr, F, \Theta, \varepsilon = 1$. From the definition of K, one finds $\Phi(0,x) = \sin(\frac{\pi}{5}x) + C_\Phi$ with a constant C_Φ, and $h(0,x) = C_\Phi$. We then choose $C_\phi = 1.1$ and $H_{\text{mean}} = 2.1$, which imply that $z(0,x) = \sin(\frac{\pi}{5}x)$. Also, for the RS-IMEX scheme, we pick $c_{\widetilde{\alpha}} = 0$ and $c_{\overline{\alpha}} = c_{\widehat{\alpha}} = 0.1$ as well as the quasi-geostrophic reference solution with $\overline{z}(0,x) = z(0,x)$.

Similar to Example (i), Figure 6.6 indicates that the scheme preserves this steady state very well, as equilibrium variables u_1 and $\partial_x K$, computed using a uniform grid with 200 cells for $T_f = 200$, are still small. Also, as Figure 6.7 shows and since $u_1 \approx 0$, the potential vorticity has not been advected in the x-direction.



(a) (b)

Figure 6.6: Preservation of the equilibrium state in Example (iii) by the RS-IMEX scheme, computed with $N_x = 200$, CFL $= 0.45$ and for $T_f = 200$.

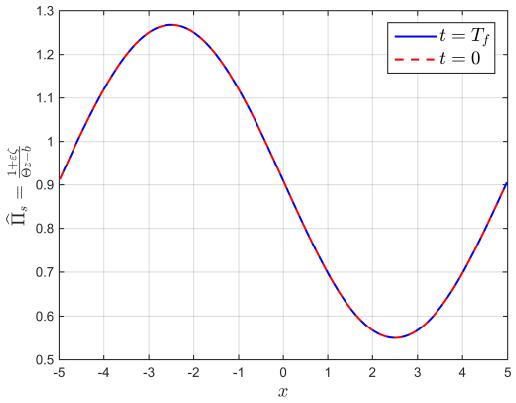


Figure 6.7: Potential vorticity profiles (in x-direction) in Example (iii) for $T_f = 200$.

6.4.4 (iv) 2d geostrophic (Rossby) adjustment

For this example, as in [CDKLM14, Ex. 4], the computational domain is $[-10, 10]^2$ with open boundaries, the bottom topography is flat $\eta^b(x, y) = 0$, $H_{\text{ref}} = 1$ and $f = g = 1$, with the following initial data:

$$z(0, x, y) = \frac{1}{4}\left(1 - \tanh\left(10\sqrt{2.5x^2 + 0.4y^2} - 1\right)\right), \qquad u_1(0, x, y) = u_2(0, x, y) = 0.$$

Like Example (i), one finds $Ro, Fr, F, \Theta, \varepsilon = 1$ with a suitable non-dimensionalisation. For the RS-IMEX scheme, we pick $c_{\widetilde{\alpha}} = 0$, $c_{\widehat{\alpha}} = 1$ and the zero reference solution. To capture the dynamics we set the time step as $\Delta t = 0.2\Delta x$.

The evolution of the water surface for the RS-IMEX scheme on the 400×400 grid is presented in Figure 6.8. The initial perturbation generates two circular shock waves propagating outwards with a clockwise rotating elevation staying behind the shocks. As time evolves, the solution converges to a nontrivial geostrophic steady state, as confirmed by Figure 6.9.

6.4.5 (v) 2d geostrophic jet

In this example, as in [CDKLM14, Ex. 5], we test the RS-IMEX scheme for the general 2d geostrophic jet. The computational domain is $[-10, 10)^2$ with open boundaries, the bottom topography is flat $\eta^b(x, y) = 0$, $H_{\text{mean}} = 1$, and $f = g = 1$, with the following initial data:

$$\begin{aligned}
z(0, x, y) &= \frac{1}{4}\left(1 - \tanh\left(10\sqrt{2.5x^2 + 0.4y^2} - 1\right)\right), \\
u_1(0, x, y) &= \frac{y}{\sqrt{2.5x^2 + 0.4y^2}}\left(1 - \left(\tanh\left(10\sqrt{2.5x^2 + 0.4y^2} - 10\right)\right)^2\right), \\
u_2(0, x, y) &= \frac{-6.25x}{\sqrt{2.5x^2 + 0.4y^2}}\left(1 - \left(\tanh\left(10\sqrt{2.5x^2 + 0.4y^2} - 10\right)\right)^2\right).
\end{aligned}$$

Like Example (iv), one finds $Ro, Fr, F, \Theta, \varepsilon = 1$ with a suitable non-dimensionalisation. For the RS-IMEX scheme, we pick $c_{\widetilde{\alpha}} = 0$, $c_{\widehat{\alpha}} = 1$ and the zero reference solution.

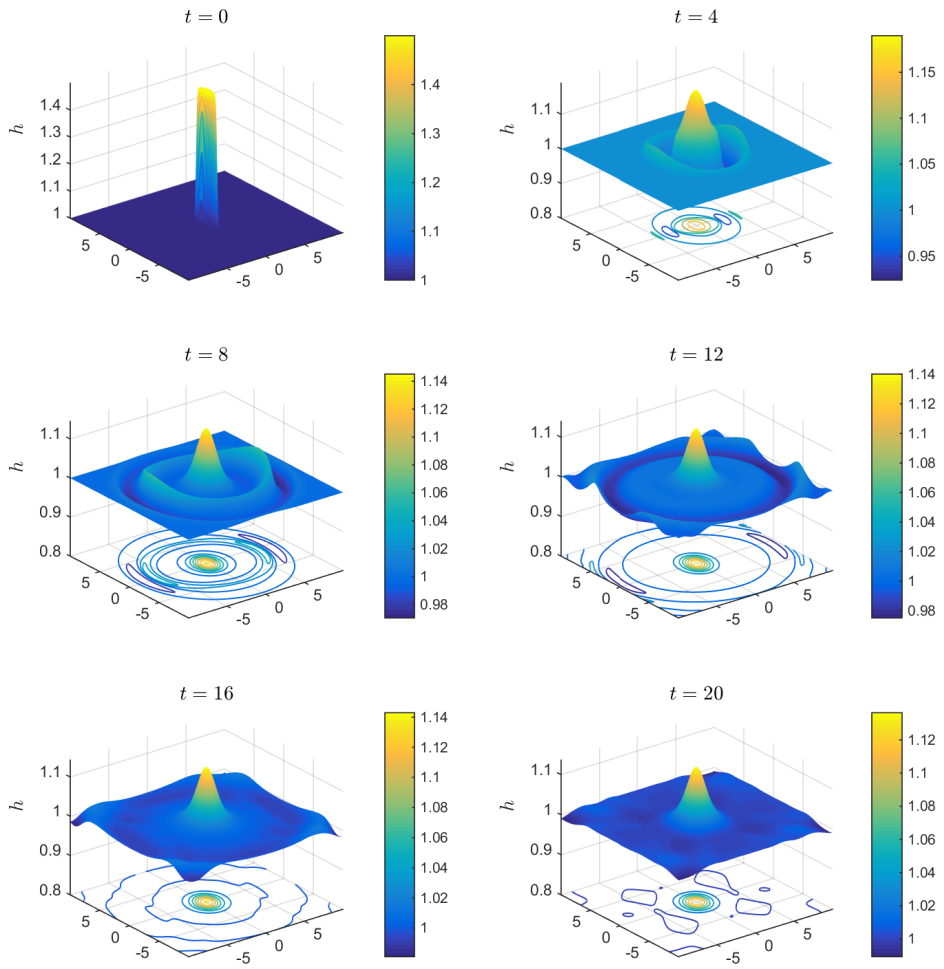


Figure 6.8: Evolution of the water height in Example (iv), computed with $N_x = N_y = 400$, CFL $= 0.45$ and for $T_f = 20$.

In Figure 6.10, we present long-time evolution of $\partial_x K$ and $\partial_y L$, computed using the 50×50 uniform grid. Comparing the results with [CDKLM14], it is evident that the RS-IMEX scheme approximates this 2d geostrophic jets accurately.

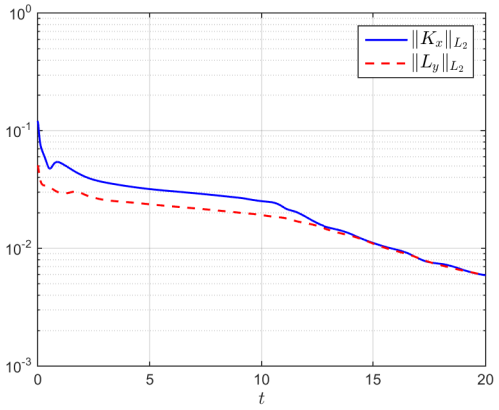


Figure 6.9: Adjustment toward the geostrophic equilibrium in Example (iv), computed with $N_x = N_y = 400$ and $\Delta t = 0.2\Delta x$.

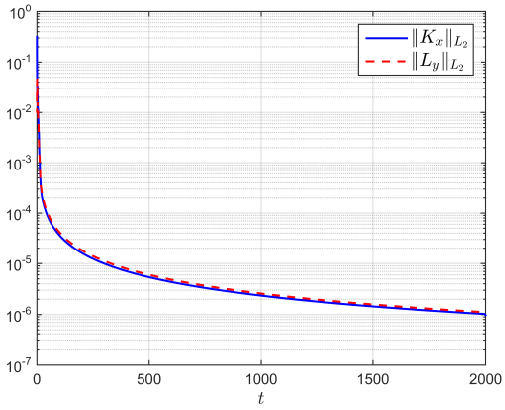


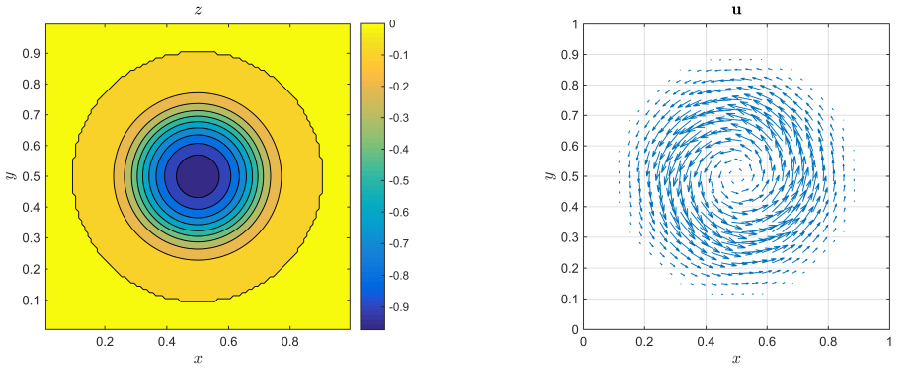
Figure 6.10: Adjustment toward the geostrophic equilibrium in Example (v), computed with $N_x = N_y = 50$ and CFL $= 0.45$.

6.4.6 (vi) 2d stationary vortex

This example considers a 2d stationary (and non quasi-geostrophic) vortex in the periodic domain $[0,1)^2$, as in [AKO09]. The initial data are

$$\boldsymbol{u}_0(r,\theta) = \vartheta_\theta(r)\hat{\theta}, \qquad \vartheta_\theta(r) := 5r\mathbf{1}_{[r<\frac{1}{5}]} + (2-5r)\mathbf{1}_{[\frac{1}{5}\le r<\frac{2}{5}]}, \qquad z_0'(r) = \vartheta_\theta + \frac{\varepsilon}{r}\vartheta_\theta^2,$$

where r is the distance to the vortex centre $(0.5, 0.5)^T$ (see Figure 6.11) and we set $H_{\text{mean}} = 2$. It is not difficult to check that with this choice of the initial condition, the height is stationary and the pressure gradient is in balance with the Coriolis force and the advective terms. So, this vortex is a particular case of *fully non-linear 2d gradient wind equilibrium* [AKO09]. Also, note that $\frac{\varepsilon}{r}\vartheta_\theta^2$ is the contribution of the advective terms in this balance; so, it is $\mathcal{O}(\varepsilon)$ and the balance would be geostrophic as $\varepsilon \to 0$, which implies that the initial data are well-prepared. For the RS-IMEX scheme, we pick CFL $= 0.45$, $c_{\widehat{\alpha}} = 1$, $c_{\widetilde{\alpha}} = 0.1$, $c_{\overline{\alpha}} = 0$ and the quasi-geostrophic reference solution with $\overline{z}_0'(r) = \vartheta_\theta$, which implies $\overline{\boldsymbol{u}}_0(r) = \vartheta_\theta$.

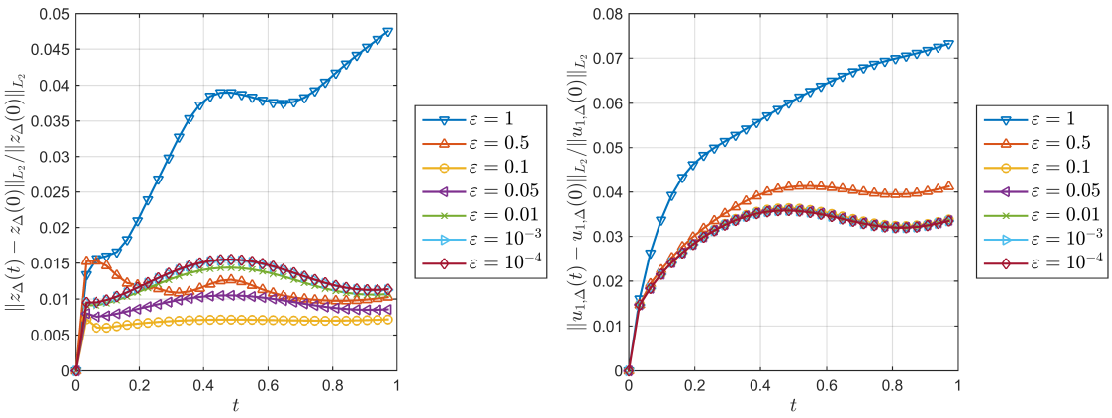


(a) Contour of the surface perturbation.

(b) Velocity vector field.

Figure 6.11: Initial condition of Example (vi) for $\varepsilon = 1$.

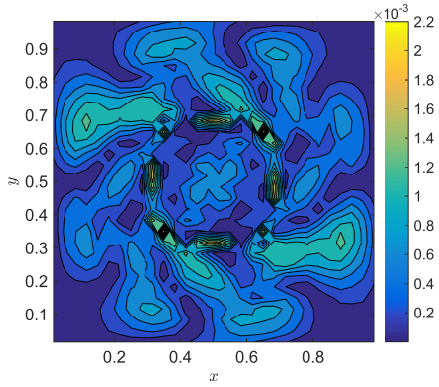
The accuracy of the scheme has been illustrated by Figure 6.12 in which the relative perturbation from the equilibrium has been plotted for $\varepsilon \in \{1, 0.5, 0.1, 0.05, 0.01, 10^{-3}, 10^{-4}\}$. It appears that, as [AKNV11], the error does not increase as $\varepsilon \to 0$. Moreover, Figures 6.13 and 6.14 present the absolute error of the solution for different ε and T_f, and confirm the accuracy of the RS-IMEX scheme even in the (geostrophic) limit of the solution and compared with the results of [AKNV11, AKO09]. The error does not decrease with ε, but with the mesh refinement. This can be explained by the fact that in the procedure of constructing the initial reference velocity field by $\overline{z}_0$, there is an $\mathcal{O}(\Delta x)$ error since this polar initial condition cannot be presented exactly on a Cartesian grid, which leads to $\|\boldsymbol{V}^0_{2,\Delta}\|_{\ell_\infty}, \|\boldsymbol{V}^0_{3,\Delta}\|_{\ell_\infty} = \mathcal{O}(\Delta x)$ rather than $\mathcal{O}(\varepsilon)$. This issue does not affect the solution for $\varepsilon = \mathcal{O}(1)$ as $\|\boldsymbol{V}^0_{2,\Delta}\|_{\ell_\infty}, \|\boldsymbol{V}^0_{3,\Delta}\|_{\ell_\infty} = \mathcal{O}(1)$, which covers the $\mathcal{O}(\Delta x)$ error. Moreover, similar to the discussion in Chapter 5, the reference solver adds some ε-independent error to the solution. Note that in comparison to [AKNV11, AKO09], we are considering the error for z rather than for the water height $h = \Theta z - b$. So, one expects that the error in water height vanishes as $\varepsilon \to 0$.



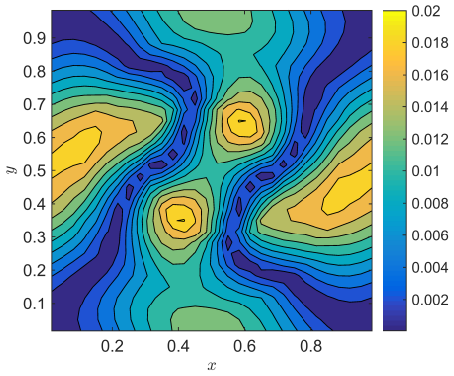
(a) Relative error for z.

(b) Relative error for u_1.

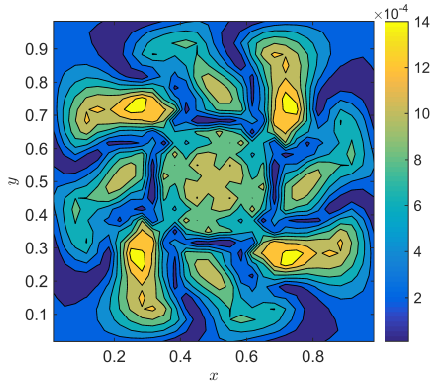
Figure 6.12: Evolution of the relative error for the RS-IMEX scheme in Example (vi), computed on the 30×30 grid, and for different ε.



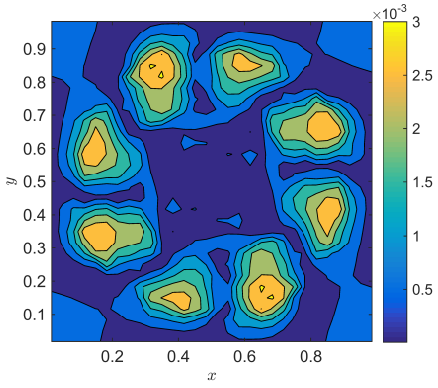
(a) $t = 1$ and $\varepsilon = 0.1$.



(b) $t = 10$ and $\varepsilon = 0.1$.



(c) $t = 1$ and $\varepsilon = 10^{-4}$.



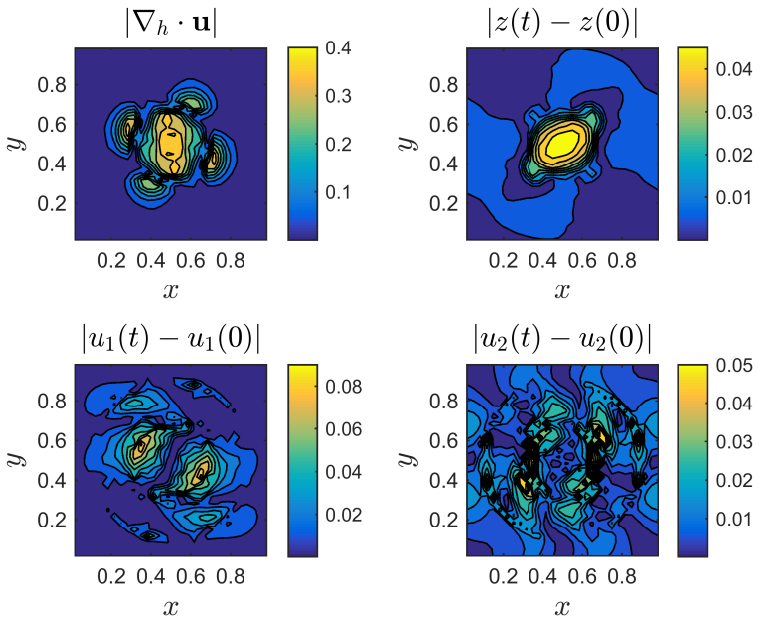
(d) $t = 10$ and $\varepsilon = 10^{-4}$.

Figure 6.13: Perturbation from the equilibrium, $|z_\Delta(t) - z_\Delta(0)|$, for the RS-IMEX scheme in Example (vi), computed on the 30×30 grid and for $\varepsilon = 0.1, 10^{-4}$ and $t = 1, 10$.

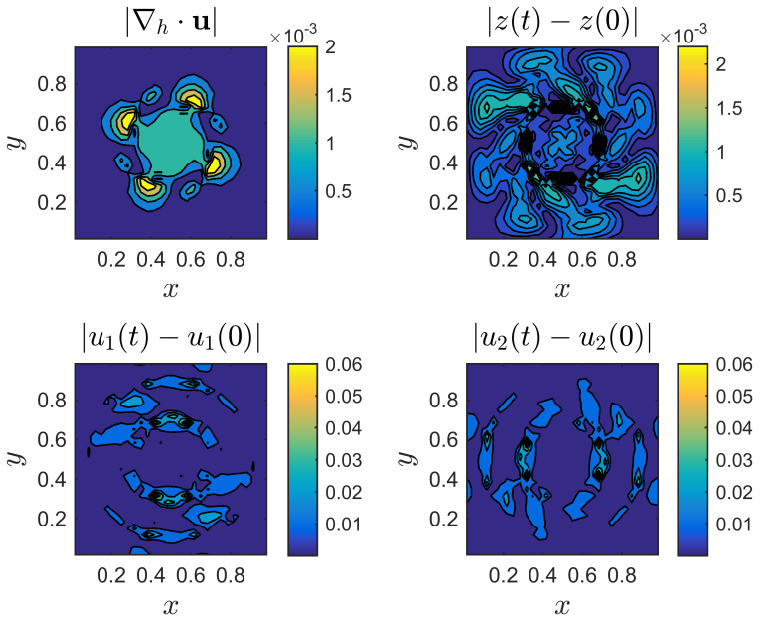
From Figure 6.14, one can see that the scheme is asymptotically stable and accurate. Regarding the asymptotic consistency, the divergence of the velocity field is vanishing with $\varepsilon \to 0$, as suggested by the asymptotic analysis in (6.22). Also, Figure 6.15 shows that the deviation from the geostrophic balance is $\mathcal{O}(\varepsilon)$, as proved in (6.23). Thus, the scheme is asymptotic preserving.

In Figure 6.16 a cut of the solution along the x-axis at $y = 0.5$ is presented to compare the results for different ε. For a small ε, the two lines are not distinguishable from each other anymore. Also, in Figure 6.17, we have illustrated the stability of the Arakawa method (for the case $\varepsilon = 1$). It has been proved by [Ara66] that, for the semi-discrete Arakawa method applied to the barotropic vorticity equation, the mean kinetic energy, the mean PV and the mean square PV are conserved in time. For the QGE the kinetic energy is no longer conserved, but the total energy, *cf.* [Dur13]. Figure 6.17 illustrates this for the fully-discrete scheme, also shows that the stream function is bounded.

Effects of the reference solution Finding the evolution of the reference solution in time requires additional computational costs, which should be justified. For this example, we show that using the quasi-geostrophic reference solution leads to better accuracy; thus, it is reasonable

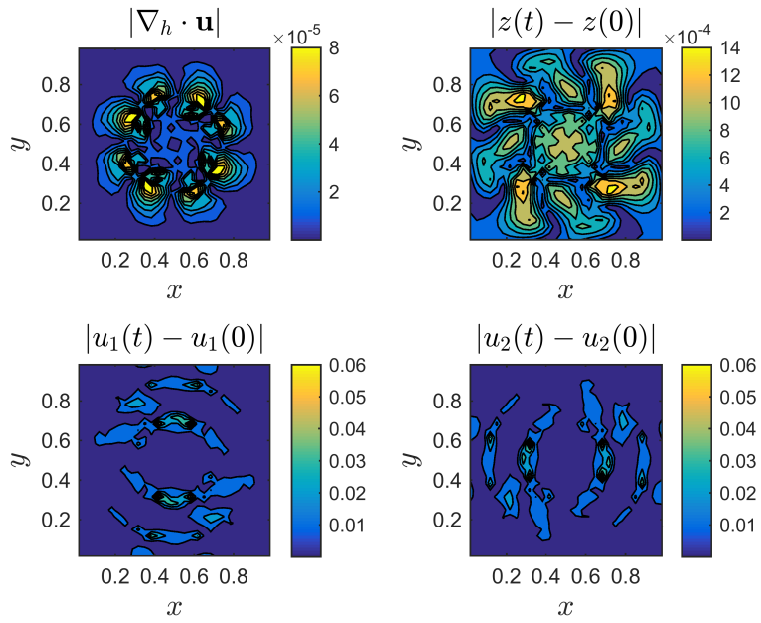


(a) $\varepsilon = 1$.

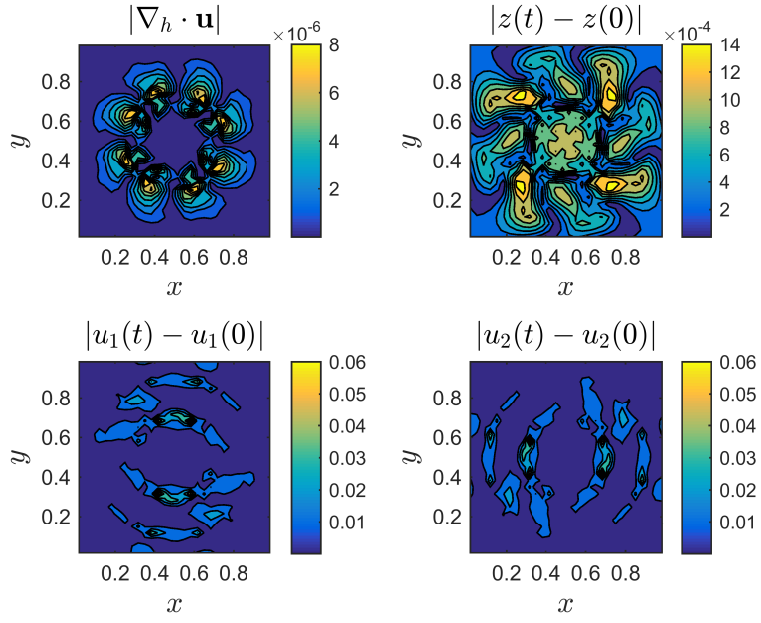


(b) $\varepsilon = 0.1$.

Figure 6.14: Absolute perturbation from the equilibrium for the RS-IMEX solution in Example (vi), computed on the 30×30 grid and for different ε.



(c) $\varepsilon = 0.01$.



(d) $\varepsilon = 0.001$.

Figure 6.14: Absolute perturbation from the equilibrium for the RS-IMEX solution in Example (vi), computed on the 30×30 grid and for different ε. (cont.)

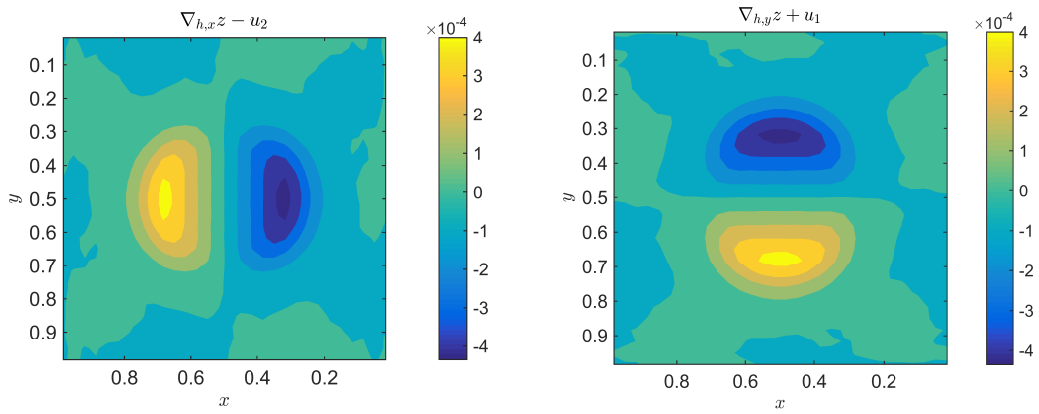


Figure 6.15: Components of the geostrophic equilibrium $\nabla_{h,\boldsymbol{x}} z = \boldsymbol{u}^{\perp}$ for the RS-IMEX solution in Example (vi), computed on the 30×30 grid, with $\varepsilon = 10^{-4}$ and for $T_f = 1$.

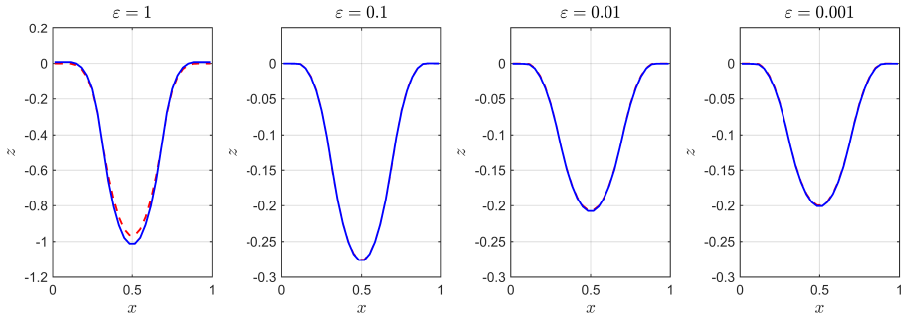


Figure 6.16: Surface perturbation along the cut $y = 0.5$ for the RS-IMEX scheme in Example (vi), computed on the 30×30 grid, for $T_f = 1$ and different ε: Dotted red line is the initial (exact) solution and the continuous blue line is the solution of the RS-IMEX scheme.

to invest in finding a suitable reference solution. Figure 6.18 illustrates the error of the RS-IMEX solution with the zero reference solution. One can clearly observe that compared to Figure 6.16 and Figure 6.14, the scheme is much more diffusive and less accurate.

Efficiency of the RS-IMEX scheme Since the time evolution of the reference solution should be computed in time, it is of interest to see how much time this computation requires. As shown below in Table 6.1, the cost of computing the reference solution is a bit higher than the non-rotating case (see Table 5.1), though, it is still not comparable to the total cost.

Table 6.1: CPU time (in seconds) for different steps of the RS-IMEX scheme in Example (vi), computed on the 30×30 grid, with CFL $= 0.45$, $T_f = 1$ and for $\varepsilon = 1$.

Total	Implicit step LSE solver	Elliptic solver for $\psi_{\Delta}^{n+1/2}$ and ψ_{Δ}^{n+1}
11.110	2.346 (21.1%)	0.17 (1.53%)

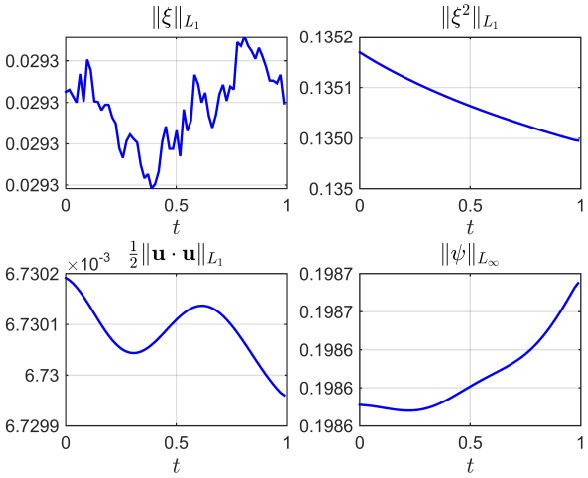
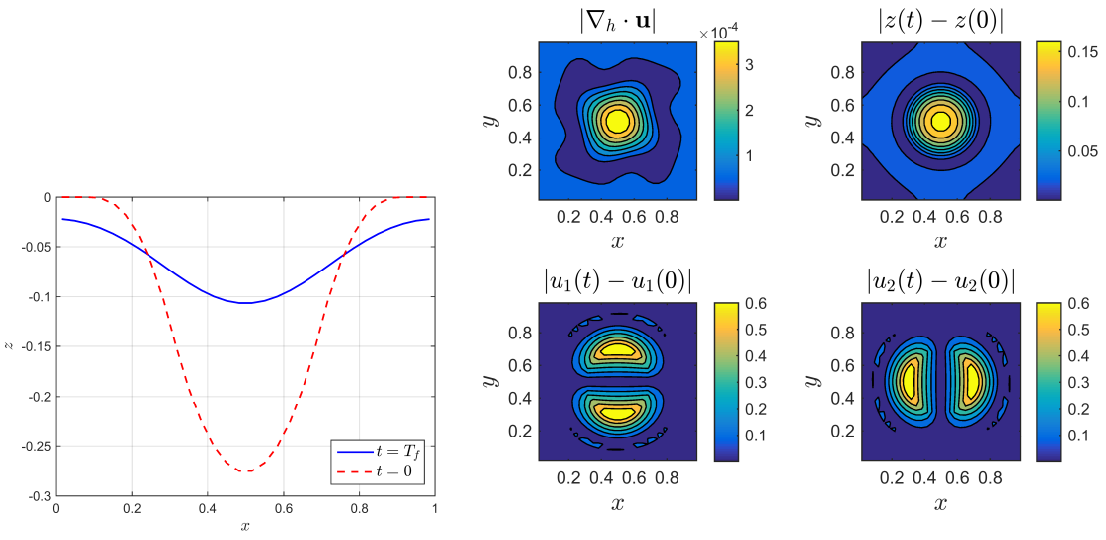


Figure 6.17: Non-linear stability of the fully-discrete Arakawa method in Example (vi), computed on the 30×30 grid and with the initial data as for the case $\varepsilon = 1$.



(a) Surface perturbation along the cut $y = 0.5$.

(b) Distance from the exact solution.

Figure 6.18: Error of the RS-IMEX scheme in Example (vi), with the zero reference solution and computed on the 30×30 grid for $T_f = 1$ and $\varepsilon = 0.1$. The results should be compared with Figure 6.16 and Figure 6.14.

Conclusion & perspectives

Throughout this manuscript, we have studied implicit-explicit (IMEX) asymptotic preserving (AP) finite volume schemes for the low-Froude shallow water equations, in terms of rigorous asymptotic consistency and stability. Regarding the source terms, we have taken the bottom topography and the Coriolis force into account so that the model is suitable for geophysical flows. At first, we studied the Lagrange-projection IMEX (LP-IMEX) scheme in one space dimension. Then, we introduced the so-called reference solution IMEX (RS-IMEX) scheme and discussed its asymptotic analysis starting from the rather simple one-dimensional case to the two-dimensional system with the Coriolis force. The focus for each case has been put on the rigorous proof of the asymptotic consistency and stability, and to verify these properties by numerical examples. More precisely, each chapter can be concluded as follows.

In Chapter 2, we have extended the stability results of the LP-IMEX scheme for the one-dimensional isentropic Euler equations, as in [CNPT10], regarding the uniformity in terms of the Mach number with well-prepared initial data. We have shown that the scheme is asymptotically consistent, rigorously. The key step for this result was to show the boundedness of the implicit solution operator w.r.t. the Mach number. Also, we have obtained a Mach-uniform time step restriction, which provides the discrete entropy stability and the density positivity, as well as the stability of the computed solution in the ℓ_∞-norm. Moreover, we have applied a similar analysis to the shallow water equations with a non-flat bottom topography as an important example of balance laws.

In Chapter 3, generalising [SN14], we introduced the low-frequency assumption to justify the use of truncated modified equations, reviewed the strict stability framework of Majda and Pego [MP85] and employed it for the stability of modified equations. We also showed that for symmetric splittings the viscosity matrix is positive, leading to stability. Furthermore, we discussed a general class of splittings and showed that positivity of the viscosity matrix, after being transformed by the matrix of eigenvectors of the (linearised) flux Jacobian, is sufficient for stability; this matches the results of [MP85] and motivated the RS-IMEX splitting introduced in Chapter 4. This criterium has been used to show the stability (of the modified equations) of several flux-splitting IMEX schemes for stiff systems of hyperbolic conservation laws: the Haack–Jin–Liu splitting [HJL12] and the Degond–Tang splitting [DT11] for the isentropic Euler equations, and the RS-IMEX splitting for the shallow water equations with a flat bottom topography. For the full Euler equation, we discovered a small region of instability for Klein's so-called auxiliary splitting [Kle95], for the colliding pulses example. This seems to confirm the computational results in [NBA$^+$14], and indicates a discrepancy between the stability analysis of the modified equation and numerical results of [NBA$^+$14], when one imposes the low-frequency assumption.

In Chapter 4, we have analysed the RS-IMEX scheme for the one-dimensional shallow water equations in the zero-Froude singular limit, and with two reference solutions, the lake at rest and the zero-Froude limit. The quality of solutions computed by the scheme has been guaranteed by numerical analysis as well as several numerical tests. We have proved that the scheme is uniformly consistent and well-balanced regarding the lake at rest equilibrium state. Indeed, the asymptotic consistency analysis is not only formal but also rigorous by virtue of a uniform bound for the implicit solution operator. Moreover, we have proved the asymptotic stability of the scheme, however, for a finite time, a fixed grid, and under a smallness assumption for the initial datum.

In Chapter 5, we have extended the asymptotic consistency and stability results from the one-dimensional case in Chapter 4 to the two-dimensional case. The formal asymptotic consistency analysis has been provided while the rigorous analysis is more subtle due to the non-stationary reference solution and could be verified based on some numerical evidence on the structure of the implicit solution operator. The proof has been presented only for a simplified case with the zero reference velocity field, $\overline{\boldsymbol{u}} = \mathbf{0}$, and for the flat bottom topography. The asymptotic analysis of the scheme has been corroborated by several numerical experiments. The study has raised an important question about the well-balancing analysis of the scheme, which is not a trivial issue and a well-balancing remedy has been proposed by a compatibility analysis.

Finally, in Chapter 6, we have discussed the RS-IMEX scheme for the two-dimensional shallow water equations with the additional Coriolis force in the quasi-geostrophic distinguished limit as a characterisation of Rossby and Froude numbers. We have analysed the asymptotic consistency of the scheme formally and, based on some numerical evidence, the rigorous proof is justified. The computed solutions have been shown to be well-qualified in several numerical experiments.

There are some immediate extensions to be considered. Regarding the LP-IMEX scheme, it is interesting to extend the analysis to the full Euler equations or multiple space dimensions, which are formidable tasks, particularly the latter as has been discussed to some extent in [CGK16, DJOR16]. Also, along the lines of [LS01], it is of interest to prove the convergence of the scheme to the unique entropy solution by the compensated compactness approach. For the RS-IMEX scheme, a more detailed well-balancing analysis is of quite an importance. Also, studying other types of boundary conditions and higher order versions of the scheme are truly desirable. A more delicate task is to improve the asymptotic stability results, *e.g.*, using an energy estimate, *cf.* [Gie15, BLMY17], or to refine the asymptotic consistency arguments in terms of asymptotic preserving error estimates, *cf.* [GHMN17, FLMN$^+$16, Fis15].

Bibliography

[ABB+04] Audusse, E., Bouchut, F., Bristeau, M.-O., Klein, R., and Perthame, B. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM Journal on Scientific Computing*, 25(6):2050–2065, 2004.

[ADDMHP15] Audusse, E., Dellacherie, S., Do Minh Hieu, P. O., and Penel, Y. Godunov type scheme for the linear wave equation with Coriolis source term. *HAL: hal-01254888*, 2015.

[ADG89] Abarbanel, S., Duth, P., and Gottlieb, D. Splitting methods for low Mach number Euler and Navier–Stokes equations. *Computers & Fluids*, 17(1):1–12, 1989.

[AKNV11] Audusse, E., Klein, R., Nguyen, D. D., and Vater, S. *Preservation of the discrete geostrophic equilibrium in shallow water flows*, pages 59–67. Springer Berlin Heidelberg, 2011.

[AKO09] Audusse, E., Klein, R., and Owinoh, A. Z. Conservative discretization of Coriolis force in a finite volume framework. *Journal of Computational Physics*, 228(8):2934–2950, 2009.

[AN12] Arun, K. R. and Noelle, S. An asymptotic preserving scheme for low Froude number shallow flows. IGPM report 352, RWTH Aachen University, 2012.

[Ara66] Arakawa, A. Computational design for long-term numerical integration of the equations of fluid motion: Two-dimensional incompressible flow. Part I. *Journal of Computational Physics*, 1(1):119–143, 1966.

[ARS97] Ascher, U. M., Ruuth, S. J., and Spiteri, R. J. Implicit-explicit Runge–Kutta methods for time-dependent partial differential equations. *Applied Numerical Mathematics*, 25(2-3):151–167, 1997.

[BALMN14] Bispen, G., Arun, K. R., Lukáčová-Medvid'ová, M., and Noelle, S. IMEX large time step finite volume methods for low Froude number shallow water flows. *Communications in Computational Physics*, 16:307–347, 2014.

[BBCM16] Berthon, C., Bessemoulin-Chatard, M., and Mathis, H. Numerical convergence rate for a diffusive limit of hyperbolic systems: p-system with damping. *SMAI-Journal of Computational Mathematics*, 36(2):99–119, 2016.

[BdSV71] Barré de Saint-Venant, A. J. C. Théorie du mouvement non permanent des eaux, avec application aux crues des rivières et à l'introduction des marées dans leurs lits. *Comptes Rendus des séances de l'Académie des Sciences*, 73:237–240, 1871.

[BEK+16] Barsukow, W., Edelmann, P. V. F., Klingenberg, C., Miczek, F., and Roepke, F. K. A numerical scheme for the compressible low-Mach number regime of ideal fluid dynamics. *arXiv preprint arXiv:1612.03910*, 2016.

[Ben02] Bendixson, I. Sur les racines d'une équation fondamentale. *Acta Mathematica*, 25(1):359–365, 1902.

[Ber09] Bernstein, D. S. *Matrix mathematics: Theory, facts, and formulas*. Princeton University Press, 2009.

[Bis15] Bispen, G. *IMEX finite volume methods for the shallow water equations*. PhD thesis, Johannes Gutenberg-Universität Mainz, 2015.

[BKL11] Bresch, D., Klein, R., and Lucas, C. Multiscale analyses for the shallow water equations. In *Computational Science and High Performance Computing IV*, pages 149–164. Springer, 2011.

[BKLL04] Botta, N., Klein, R., Langenberg, S., and Lützenkirchen, S. Well balanced finite volume methods for nearly hydrostatic flows. *Journal of Computational Physics*, 196(2):539–565, 2004.

[Bla16] Blachère, F. *Schémas numériques d'ordre élevé et préservant l'asymptotique pour l'hydrodynamique radiative*. PhD thesis, Université Nantes; Université Bretagne Loire, 2016.

[BLMY16] Bispen, G., Lukáčová-Medvid'ová, M., and Yelash, L. IMEX finite volume evolution Galerkin scheme for three-dimensional weakly compressible flows. In *Proceedings of the Conference Algoritmy*, pages 62–73, 2016.

[BLMY17] Bispen, G., Lukáčová-Medvid'ová, M., and Yelash, L. Asymptotic preserving IMEX finite volume schemes for low mach number Euler equations with gravitation. *Journal of Computational Physics*, 335:222–248, 2017.

[BLSZ04] Bouchut, F., Le Sommer, J., and Zeitlin, V. Frontal geostrophic adjustment and nonlinear wave phenomena in one-dimensional rotating shallow water. Part 2. High-resolution numerical simulations. *Journal of Fluid Mechanics*, 514:35–63, 2004.

[BM05a] Birken, P. and Meister, A. On low Mach number preconditioning of finite volume schemes. *PAMM*, 5(1):759–760, 2005.

[BM05b] Birken, P. and Meister, A. Stability of preconditioned finite volume schemes at low Mach numbers. *BIT Numerical Mathematics*, 45(3):463–480, 2005.

[BMCPV03] Bouchut, F., Mangeney-Castelnau, A., Perthame, B., and Vilotte, J.-P. A new model of Saint-Venant and Savage–Hutter type for gravity driven shallow water flows. *Comptes Rendus Mathematique*, 336(6):531–536, 2003.

[Bou04] Bouchut, F. *Nonlinear stability of finite volume methods for hyperbolic conservation laws: And well-balanced schemes for sources*. Springer Science & Business Media, 2004.

[BR09] Boscarino, S. and Russo, G. On a class of uniformly accurate IMEX Runge–Kutta schemes and applications to hyperbolic systems with relaxation. *SIAM Journal on Scientific Computing*, 31(3):1926–1945, 2009.

[Bro30] Browne, E. T. The characteristic roots of a matrix. *Bulletin of the American Mathematical Society*, 36(10):705–710, 1930.

[BT16] Blachère, F. and Turpault, R. An admissibility and asymptotic-preserving scheme for systems of conservation laws with source term on 2D unstructured meshes. *Journal of Computational Physics*, 315:98–123, 2016.

[Buc15] Buckingham, E. Model experiments and the forms of empirical equations. *Transactions of the American Society of Mechanical Engineers*, 37:263–296, 1915.

[BV94] Bermudez, A. and Vazquez, M. E. Upwind methods for hyperbolic conservation laws with source terms. *Computers & Fluids*, 23(8):1049–1071, 1994.

[BW98] Bijl, H. and Wesseling, P. A unified method for computing incompressible and compressible flows in boundary-fitted coordinates. *Journal of Computational Physics*, 141(2):153–173, 1998.

[BW04] Bouchut, F. and Westdickenberg, M. Gravity driven shallow water models for arbitrary topography. *Communications in Mathematical Sciences*, 2(3):359–389, 2004.

[CC08] Chalons, C. and Coulombel, J.-F. Relaxation approximation of the Euler equations. *Journal of Mathematical Analysis and Applications*, 348(2):872–893, 2008.

[CCG+10] Chalons, C., Coquel, F., Godlewski, E., Raviart, P.-A., and Seguin, N. Godunov-type schemes for hyperbolic systems with parameter-dependent source: The case of Euler system with friction. *Mathematical Models and Methods in Applied Sciences*, 20(11):2109–2166, 2010.

[CDGG06] Chemin, J.-Y., Desjardins, B., Gallagher, I., and Grenier, E. *Mathematical geophysics: An introduction to rotating fluids and the Navier–Stokes equations*. The Clarendon Press Oxford University Press, Oxford, 2006.

[CDK12] Cordier, F., Degond, P., and Kumbaro, A. An asymptotic-preserving all-speed scheme for the Euler and Navier–Stokes equations. *Journal of Computational Physics*, 231(17):5685–5704, 2012.

[CDKLM14] Chertock, A., Dudzinski, M., Kurganov, A., and Lukáčová-Medvid'ová, M. Well-balanced schemes for the shallow water equations with Coriolis forces. In *European Conference on Mathematics for Industry*, page 496, 2014.

[CDV16] Crouseilles, N., Dimarco, G., and Vignal, M.-H. Multiscale schemes for the BGK–Vlasov–Poisson system in the quasi-neutral and fluid limits: Stability analysis and first order schemes. *Multiscale Modeling & Simulation*, 14(1):65–95, 2016.

[CDV17] Couderc, F., Duran, A., and Vila, J.-P. An explicit asymptotic preserving low Froude scheme for the multilayer shallow water model with density stratification. *Journal of Computational Physics*, 343:235–270, 2017.

[CFvN50] Charney, J. G., Fjörtoft, R., and von Neumann, J. Numerical integration of the barotropic vorticity equation. *Tellus*, 2(4):237–254, 1950.

[CG84] Casulli, V. and Greenspan, D. Pressure method for the numerical solution of transient, compressible fluid flows. *International Journal for Numerical Methods in Fluids*, 4(11):1001–1012, 1984.

[CGK13] Chalons, C., Girardin, M., and Kokh, S. Large time step and asymptotic preserving numerical schemes for the gas dynamics equations with source terms. *SIAM Journal on Scientific Computing*, 35(6):A2874–A2902, 2013.

[CGK16] Chalons, C., Girardin, M., and Kokh, S. An all-regime Lagrange–projection like scheme for the gas dynamics equations on unstructured meshes. *Communications in Computational Physics*, 20(1):188–233, 2016.

[CGS07] Coquel, F., Godlewski, E., and Seguin, N. Regularization and relaxation tools for interface coupling. In *Proceedings of XX Congresso de Ecuaciones Differenciales Y Aplicaicones, CEDYA*, 2007.

[Cha48] Charney, J. G. On the scale of atmospheric motions. *Geofysiske Publikasjoner*, 17(2), 1948.

[Cha49] Charney, J. G. On a physical basis for numerical prediction of large-scale motions in the atmosphere. *Journal of Meteorology*, 6(6):372–385, 1949.

[Cho67] Chorin, A. J. A numerical method for solving incompressible viscous flow problems. *Journal of Computational Physics*, 2(1):12–26, 1967.

[Cho68] Chorin, A. J. Numerical solution of the Navier–Stokes equations. *Mathematics of Computation*, 22(104):745–762, 1968.

[Cho69] Chorin, A. J. On the convergence of discrete approximations to the Navier–Stokes equations. *Mathematics of Computation*, 23(106):341–353, 1969.

[CJ16] Corless, R. M. and Jankowski, J. E. Variations on a theme of Euler. *SIAM Review*, 58(4):775–792, 2016.

[CJR97] Caflisch, R. E., Jin, S., and Russo, G. Uniformly accurate schemes for hyperbolic systems with relaxation. *SIAM Journal on Numerical Analysis*, 34(1):246–281, 1997.

[CK16] Chiodaroli, E. and Kreml, O. An overview of some recent results on the Euler system of isentropic gas dynamics. *Bulletin of the Brazilian Mathematical Society*, 47(1):241–253, 2016.

[CKKS16] Chalons, C., Kestener, P., Kokh, S., and Stauffert, M. A large time-step and well-balanced Lagrange-projection type scheme for the shallow water equations. *HAL: hal-01297043*, 2016.

[CLL94] Chen, G.-Q., Levermore, C. D., and Liu, T.-P. Hyperbolic conservation laws with stiff relaxation terms and entropy. *Communications on Pure and Applied Mathematics*, 47(6):787–830, 1994.

[CLP08] Castro, M. J., López, J. A., and Parés, C. Finite volume simulation of the geostrophic adjustment in a rotating shallow water system. *SIAM Journal on Scientific Computing*, 31(1):444–477, 2008.

[CMS13] Cances, C., Mathis, H., and Seguin, N. Relative entropy for the finite volume approximation of strong solutions to systems of conservation laws. *HAL: hal-00798287*, 2013.

[CMV15] Chalons, C., Massot, M., and Vié, A. On the Eulerian large eddy simulation of disperse phase flows: An asymptotic preserving scheme for small Stokes number flows. *Multiscale Modeling & Simulation*, 13(1):291–315, 2015.

[CNPT10] Coquel, F., Nguyen, Q., Postel, M., and Tran, Q. Entropy-satisfying relaxation method with large time-steps for Euler IBVPs. *Mathematics of Computation*, 79(271):1493–1533, 2010.

[Cor35] Coriolis, G. G. Mémoire sur les équations du mouvement relatif des systèmes de corps. *Journal de lÉcole Polytechnique*, 15:142–154, 1835.

[CS70] Conley, C. C. and Smoller, J. A. Viscosity matrices for two-dimensional nonlinear hyperbolic systems. *Communications on Pure and Applied Mathematics*, 23(6):867–884, 1970.

[CSJT98] Cockburn, B., Shu, C.-W., Johnson, C., and Tadmor, E. *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*. Lecture Notes in Mathematics. Springer-Verlag, 1998.

[Daf10] Dafermos, C. M. *Hyperbolic conservation laws in continuum physics*, volume 325 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2010.

[Dan05] Danchin, R. Low Mach number limit for viscous compressible flows. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 39(3):459–475, 2005.

[DD16] Degond, P. and Deluzet, F. Asymptotic-preserving methods and multiscale models for plasma physics. *arXiv preprint arXiv:1603.08820*, 2016.

[DDN+10] Degond, P., Deluzet, F., Navoret, L., Sun, A.-B., and Vignal, M.-H. Asymptotic-preserving particle-in-cell method for the Vlasov–Poisson system near quasineutrality. *Journal of Computational Physics*, 229(16):5630–5652, 2010.

[Del10] Dellacherie, S. Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number. *Journal of Computational Physics*, 229(4):978–1016, 2010.

[DHPRF04] Donea, J., Huerta, A., Ponthot, J.-P., and Rodríguez-Ferran, A. Arbitrary Lagrangian-Eulerian methods. In *Encyclopedia of Computational Mechanics*. John Wiley & Sons, Ltd, 2004.

[DiP85] DiPerna, R. J. Measure-valued solutions to conservation laws. *Archive for Rational Mechanics and Analysis*, 88(3):223–270, 1985.

[DJOR16] Dellacherie, S., Jung, J., Omnes, P., and Raviart, P.-A. Construction of modified Godunov-type schemes accurate at any Mach number for the compressible euler system. *Mathematical Models and Methods in Applied Sciences*, 26(13):2525–2615, 2016.

[DLP93] Demirdžić, I., Lilek, Ž., and Perić, M. A collocated finite volume method for predicting flows at all speeds. *International Journal for Numerical Methods in Fluids*, 16(12):1029–1050, 1993.

[DLS10] De Lellis, C. and Székelyhidi, L. On admissibility criteria for weak solutions of the Euler equations. *Archive for Rational Mechanics and Analysis*, 195(1):225, 2010.

[DLV17] Dimarco, G., Loubère, R., and Vignal, M.-H. Study of a new asymptotic preserving scheme for the Euler system in the low Mach number limit. *SIAM Journal on Scientific Computing*, 39(5):A2099–A2128, 2017.

[DM87] DiPerna, R. J. and Majda, A. Oscillations and concentrations in weak solutions of the incompressible fluid equations. *Communications in mathematical physics*, 108(4):667–689, 1987.

[DM15] Duran, A. and Marche, F. Discontinuous-Galerkin discretization of a new class of Green–Naghdi equations. *Communications in Computational Physics*, 17(03):721–760, 2015.

[DM16] Duran, A. and Marche, F. A discontinuous Galerkin method for a new class of Green–Naghdi equations on simplicial unstructured meshes. *arXiv preprint arXiv:1604.05227*, 2016.

[DMLM95] Dal Maso, G., Lefloch, P. G., and Murat, F. Definition and weak stability of nonconservative products. *Journal de Mathématiques Pures et Appliquées*, 74(6):483–548, 1995.

[DMTB15] Duran, A., Marche, F., Turpault, R., and Berthon, C. Asymptotic preserving scheme for the shallow water equations with source terms on unstructured meshes. *Journal of Computational Physics*, 287:184–206, 2015.

[DOR10] Dellacherie, S., Omnes, P., and Rieper, F. The influence of cell geometry on the Godunov scheme applied to the linear wave equation. *Journal of Computational Physics*, 229(14):5315–5338, 2010.

[DP14] Dimarco, G. and Pareschi, L. Numerical methods for kinetic equations. *Acta Numerica*, 23:369–520, 2014.

[DR06] Drikakis, D. and Rider, W. *High-resolution methods for incompressible and low-speed flows*. Springer Science & Business Media, 2006.

[DT11] Degond, P. and Tang, M. All speed scheme for the low Mach number limit of the isentropic Euler equation. *Communications in Computational Physics*, 10(1):1–31, 2011.

[Dur13] Durran, D. R. *Numerical methods for wave equations in geophysical fluid dynamics*, volume 32. Springer Science & Business Media, 2013.

[EDMS17a] Even-Dar Mandel, L. and Schochet, S. Convergence of solutions to finite difference schemes for singular limits of nonlinear evolutionary PDEs. *ESAIM: Mathematical Modelling and Numerical Analysis*, 51(2):587–614, 2017.

[EDMS17b] Even-Dar Mandel, L. and Schochet, S. Uniform discrete Sobolev estimates of solutions to finite difference schemes for singular limits of nonlinear PDEs. *ESAIM: Mathematical Modelling and Numerical Analysis*, 51(2):727–757, 2017.

[EGH00] Eymard, R., Gallouët, T., and Herbin, R. Finite volume methods. *Handbook of Numerical Analysis*, 7:713–1018, 2000.

[Ell06] Elling, V. A possible counterexample to well posedness of entropy solutions and to Godunov scheme convergence. *Mathematics of Computation*, 75(256):1721–1733, 2006.

[EM96] Embid, P. F. and Majda, A. Averaging over fast gravity waves for geophysical flows with arbitary potential vorticity. *Communications in Partial Differential Equations*, 21(3-4):619–658, 1996.

[FGGVN12] Feireisl, E., Gallagher, I., Gerard-Varet, D., and Novotnỳ, A. Multi-scale analysis of compressible viscous and rotating fluids. *Communications in Mathematical Physics*, 314(3):641–670, 2012.

[FGN12] Feireisl, E., Gallagher, I., and Novotnỳ, A. A singular limit for compressible rotating fluids. *SIAM Journal on Mathematical Analysis*, 44(1):192–205, 2012.

[Fis15] Fischer, J. A posteriori modeling error estimates for the assumption of perfect incompressibility in the Navier–Stokes equation. *SIAM Journal on Numerical Analysis*, 53(5):2178–2205, 2015.

[FJ10] Filbet, F. and Jin, S. A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources. *Journal of Computational Physics*, 229(20):7625–7648, 2010.

[FLMN$^+$16] Feireisl, E., Lukáčová-Medvid'ová, M., Nečasová, v., Novotnỳ, A., and She, B. Asymptotic preserving error estimates for numerical solutions of compressible Navier–Stokes equations in the low Mach number regime. Preprint 49-2016, Czech Academy of Sciences, 2016.

[FN09] Feireisl, E. and Novotnỳ, A. *Singular limits in thermodynamics of viscous fluids*. Springer Science & Business Media, 2009.

[FN14a] Feireisl, E. and Novotnỳ, A. Multiple scales and singular limits for compressible rotating fluids with general initial data. *Communications in Partial Differential Equations*, 39(6):1104–1127, 2014.

[FN14b] Feireisl, E. and Novotnỳ, A. Scale interactions in compressible rotating fluids. *Annali di Matematica Pura ed Applicata*, 193(6):1703–1725, 2014.

[FN16] Fedele, B. and Negulescu, C. Numerical study of an anisotropic Vlasov equation arising in plasma physics. *arXiv preprint arXiv:1610.01592*, 2016.

[Fos13] Foster, E. L. *Finite Elements for the quasi-geostrophic equations of the ocean*. PhD thesis, Virginia Polytechnic Institute and State University, 2013.

[FR13] Filbet, F. and Rambaud, A. Analysis of an asymptotic preserving scheme for relaxation systems. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 47(2):609–633, 2013.

[FR16] Filbet, F. and Rodrigues, L. M. Asymptotically stable particle-in-cell methods for the Vlasov–Poisson system with a strong external magnetic field. *SIAM Journal on Numerical Analysis*, 54(2):1120–1146, 2016.

[Fra12] Franck, E. *Design and numerical analysis of asymptotic preserving schemes on unstructured meshes: Application to the linear transport and Friedrichs systems*. Theses, Université Pierre et Marie Curie - Paris VI, October 2012.

[Fro10] Frobenius, F. G. Über die mit einer matrix vertauschbaren matrizen. In *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin*, pages 3–15, 1910.

[Gal02] Gallice, G. Solveurs simples positifs et entropiques pour les systèmes hyperboliques avec terme source. *Comptes Rendus Mathématique*, 334(8):713–716, 2002.

[Gal03] Gallice, G. Positive and entropy stable Godunov-type schemes for gas dynamics and MHD equations in Lagrangian or Eulerian coordinates. *Numerische Mathematik*, 94(4):673–713, 2003.

[Gel59] Gel'fand, I. M. Some problems in the theory of quasi-linear equations. *Uspekhi Matematicheskikh Nauk*, 14(2):87–158, 1959.

[GGHL08] Gallouët, T., Gastaldo, L., Herbin, R., and Latché, J.-C. An unconditionally stable pressure correction scheme for the compressible barotropic Navier–Stokes equations. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 42(2):303–331, 2008.

[GHK$^+$11] Gastaldo, L., Herbin, R., Kheriji, W., Lapuerta, C., and Latché, J.-C. Staggered discretizations, pressure correction schemes and all speed barotropic flows. In *Finite Volumes for Complex Applications VI Problems & Perspectives*, pages 839–855. Springer, 2011.

[GHKL15] Grapsas, D., Herbin, R., Kheriji, W., and Latché, J.-C. An unconditionally stable staggered pressure correction scheme for the compressible Navier–Stokes equations. *HAL: hal-01115250*, 2015.

[GHMN16] Gallouët, T., Herbin, R., Maltese, D., and Novotnỳ, A. Error estimates for a numerical approximation to the compressible barotropic Navier–Stokes equations. *IMA Journal of Numerical Analysis*, 36(2):543–592, 2016.

[GHMN17] Gallouët, T., Herbin, R., Maltese, D., and Novotnỳ, A. Implicit MAC scheme for compressible Navier–Stokes equations: Low Mach asymptotic error estimates. *hal-01462822*, 2017.

[Gie15] Giesselmann, J. Low Mach asymptotic-preserving scheme for the Euler–Korteweg model. *IMA Journal of Numerical Analysis*, 35(2):802–833, 2015.

[GJL99] Golse, F., Jin, S., and Levermore, C. D. The convergence of numerical transfer schemes in diffusive regimes I: Discrete-ordinate method. *SIAM Journal on Numerical Analysis*, 36(5):1333–1369, 1999.

[GL96] Greenberg, J. M. and Leroux, A.-Y. A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. *SIAM Journal on Numerical Analysis*, 33(1):1–16, 1996.

[GM04] Guillard, H. and Murrone, A. On the behavior of upwind schemes in the low Mach number limit: II. Godunov type schemes. *Computers & Fluids*, 33(4):655–675, 2004.

[GMS06] Guermond, J.-L., Minev, P., and Shen, J. An overview of projection methods for incompressible flows. *Computer Methods in Applied Mechanics and Engineering*, 195(44):6011–6045, 2006.

[Gos11] Gosse, L. Transient radiative transfer in the grey case: Well-balanced and asymptotic-preserving schemes built on Case's elementary solutions. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 112(12):1995–2012, 2011.

[Gos13] Gosse, L. *Computing qualitatively correct approximations of balance laws*, volume 2 of *SIMAI Springer Series*. Springer, 2013.

[GR96] Godlewski, E. and Raviart, P.-A. *Numerical approximation of hyperbolic systems of conservation laws*, volume 118. Springer, 1996.

[GR97] Gustafson, K. E. and Rao, D. K. M. *Numerical range*. Springer, 1997.

[GR10] Giraldo, F. X. and Restelli, M. High-order semi-implicit time-integrators for a triangular discontinuous Galerkin oceanic shallow water model. *International Journal for Numerical Methods in Fluids*, 63(9):1077–1102, 2010.

[Gra06] Gray, R. M. *Toeplitz and circulant matrices: A review*. Now Publishers Inc., 2006.

[Gre97] Grenier, E. Oscillatory perturbations of the Navier–Stokes equations. *Journal de Mathématiques Pures et Appliquées*, 76(6):477–498, 1997.

[GSR07] Gallagher, I. and Saint-Raymond, L. On the influence of the earths rotation on geophysical flows. *Handbook of Mathematical Fluid Dynamics*, 4:201–329, 2007.

[GSS86] Griffiths, D. F. and Sanz-Serna, J. M. On the scope of the method of modified equations. *SIAM Journal on Scientific and Statistical Computing*, 7(3):994–1008, 1986.

[GT02] Gosse, L. and Toscani, G. An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations. *Comptes Rendus Mathematique*, 334(4):337–342, 2002.

[GT03] Gosse, L. and Toscani, G. Space localization and well-balanced schemes for discrete kinetic models in diffusive regimes. *SIAM Journal on Numerical Analysis*, 41(2):641–658, 2003.

[GT04] Gosse, L. and Toscani, G. Asymptotic-preserving & well-balanced schemes for radiative transfer and the Rosseland approximation. *Numerische Mathematik*, 98(2):223–250, 2004.

[GV99] Guillard, H. and Viozat, C. On the behaviour of upwind schemes in the low Mach number limit. *Computers & Fluids*, 28(1):63–86, 1999.

[GVL12] Golub, G. H. and Van Loan, C. F. *Matrix computations*, volume 3. JHU Press, 2012.

[HA68] Harlow, F. H. and Amsden, A. A. Numerical calculation of almost incompressible flow. *Journal of Computational Physics*, 3(1):80–93, 1968.

[HA71] Harlow, F. H. and Amsden, A. A. A numerical fluid dynamics calculation method for all flow speeds. *Journal of Computational Physics*, 8(2):197–213, 1971.

[Hir02] Hirsch, M. A. Sur les racines d'une équation fondamentale. *Acta Mathematica*, 25(1):367–370, 1902.

[Hir68] Hirt, C. W. Heuristic stability theory for finite-difference equations. *Journal of Computational Physics*, 2(4):339–355, 1968.

[HJ86] Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge University Press, New York, NY, USA, 1986.

[HJ91] Hom, R. A. and Johnson, C. R. Topics in matrix analysis. *Cambridge UP, New York*, 1991.

[HJL12] Haack, J., Jin, S., and Liu, J.-G. An all-speed asymptotic-preserving method for the isentropic Euler and Navier–Stokes equations. *Communications in Computational Physics*, 12(4):955–980, 2012.

[HJL16] Hu, J., Jin, S., and Li, Q. Asymptotic-preserving schemes for multiscale hyperbolic and kinetic equations. *Handbook of Numerical Analysis*, 2016.

[HKL12] Herbin, R., Kheriji, W., and Latché, J.-C. Staggered schemes for all speed flows. In *ESAIM: Proceedings*, volume 35, pages 122–150. EDP Sciences, 2012.

[HKL13] Herbin, R., Kheriji, W., and Latché, J.-C. Consistent semi-implicit staggered schemes for compressible flows. Part II: Euler equations. *HAL: hal-00805514*, 2013.

[HKL14] Herbin, R., Kheriji, W., and Latché, J.-C. On some implicit and semi-implicit staggered schemes for the shallow water and Euler equations. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 48(6):1807–1857, 2014.

[HKLR10] Holden, H., Karlsen, K. H., Lie, K.-A., and Risebro, N. H. *Splitting methods for partial differential equations with rough solutions*. European Math. Soc. Publishing House, 2010.

[HLN13a] Herbin, R., Latché, J.-C., and Nguyen, T. T. Consistent explicit staggered schemes for compressible flows Part I: the barotropic Euler equations. *HAL: hal-00821069*, 2013.

[HLN13b] Herbin, R., Latché, J.-C., and Nguyen, T. T. Consistent explicit staggered schemes for compressible flows Part II: the Euler equation. *HAL: hal-00821070*, 2013.

[HLN13c] Herbin, R., Latché, J.-C., and Nguyen, T. T. Consistent semi-implicit staggered schemes for compressible flows. Part I: the barotropic Euler equations. *HAL: hal-00821069*, 2013.

[HLN13d] Herbin, R., Latché, J.-C., and Nguyen, T. T. Explicit staggered schemes for the compressible Euler equations. In *ESAIM: Proceedings*, volume 40, pages 83–102. EDP Sciences, 2013.

[HLS17] Herbin, R., Latché, J.-C., and Saleh, K. Low Mach number limit of a pressure correction MAC scheme for compressible barotropic flows. In *Finite Volumes for Complex Applications 8*, 2017.

[HLVL97] Harten, A., Lax, P. D., and Van Leer, B. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. In *Upwind and High-Resolution Schemes*, pages 53–79. Springer, 1997.

[HM14] Hiltebrand, A. and Mishra, S. Efficient computation of all speed flows using an entropy stable shock-capturing space-time discontinuous Galerkin method. In *Seminar for Applied Mathematics, ETH Zürich*, volume 17, pages 1–21, 2014.

[HP94] Hardin, J. C. and Pope, D. S. An acoustic/viscous splitting technique for computational aeroacoustics. *Theoretical and Computational Fluid Dynamics*, 6(5):323–340, 1994.

[HW65] Harlow, F. H. and Welch, J. E. Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. *Physics of Fluids*, 8(12):2182, 1965.

[HW96] Hairer, E. and Wanner, G. *Solving ordinary differential equations. II*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 1996. Stiff and differential-algebraic problems.

[HZLMP11] Hundertmark-Zaušková, A., Lukáčová-Medvid'ová, M., and Prill, F. Large time step finite volume evolution Galerkin methods. *Journal of Scientific Computing*, 48(1):227–240, 2011.

[Il'69] Il'in, A. M. Differencing scheme for a differential equation with a small parameter affecting the highest derivative. *Mathematical Notes of the Academy of Sciences of the USSR*, 6(2):596–602, 1969.

[Jin95] Jin, S. Runge–Kutta methods for hyperbolic conservation laws with stiff relaxation terms. *Journal of Computational Physics*, 122(1):51–67, 1995.

[Jin99] Jin, S. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM Journal on Scientific Computing*, 21(2):441–454, 1999.

[Jin10] Jin, S. Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: A review. *Lecture Notes for Summer School on "Methods and Models of Kinetic Theory" (M&MKT), Porto Ercole (Grosseto, Italy)*, pages 177–216, 2010.

[Jin16] Jin, S. private communication, 2016.

[JL91] Jin, S. and Levermore, C. D. The discrete-ordinate method in diffusive regimes. *Transport Theory and Statistical Physics*, 20(5-6):413–439, 1991.

[JL93] Jin, S. and Levermore, C. D. Fully-discrete numerical transfer in diffusive regimes. *Transport Theory and Statistical Physics*, 22(6):739–791, 1993.

[JL96] Jin, S. and Levermore, C. D. Numerical schemes for hyperbolic conservation laws with stiff relaxation terms. *Journal of Computational Physics*, 126(2):449–467, 1996.

[JLQX14] Jang, J., Li, F., Qiu, J.-M., and Xiong, T. Analysis of asymptotic preserving DG-IMEX schemes for linear kinetic transport equations in a diffusive scaling. *SIAM Journal on Numerical Analysis*, 52(4):2048–2072, 2014.

[JPT98] Jin, S., Pareschi, L., and Toscani, G. Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations. *SIAM Journal on Numerical Analysis*, 35(6):2405–2439, 1998.

[JX95] Jin, S. and Xin, Z. The relaxation schemes for systems of conservation laws in arbitrary space dimensions. *Communications on Pure and Applied Mathematics*, 48(3):235–276, 1995.

[KAK11] Kacimi, A., Aliziane, T., and Khouider, B. The Arakawa Jacobian method and a fourth-order essentially nonoscillatory scheme for the beta-plane barotropic equations. *International Journal of Numerical Analysis and Modeling*, 2011.

[KBS+01] Klein, R., Botta, N., Schneider, T., Munz, C.-D., Roller, S., Meister, A., Hoffmann, L., and Sonar, T. Asymptotic adaptive methods for multi-scale problems in fluid mechanics. In *Practical Asymptotics*, pages 261–343. Springer, 2001.

[KFJ16] Küpper, K., Frank, M., and Jin, S. An asymptotic preserving two-dimensional staggered grid method for multiscale transport equations. *SIAM Journal on Numerical Analysis*, 54(1):440–461, 2016.

[Kla98] Klar, A. An asymptotic-induced scheme for nonstationary transport equations in the diffusive limit. *SIAM Journal on Numerical Analysis*, 35(3):1073–1094, 1998.

[Kle95] Klein, R. Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics I: One-dimensional flow. *Journal of Computational Physics*, 121(2):213–237, 1995.

[KLN91] Kreiss, H.-O., Lorenz, J., and Naughton, M. J. Convergence of the solutions of the compressible to the solutions of the incompressible Navier–Stokes equations. *Advances in Applied Mathematics*, 12(2):187–214, 1991.

[KM81] Klainerman, S. and Majda, A. Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids. *Communications on Pure and Applied Mathematics*, 34(4):481–524, 1981.

[KM82] Klainerman, S. and Majda, A. Compressible and incompressible fluids. *Communications on Pure and Applied Mathematics*, 35(5):629–651, 1982.

[KM95] Klein, R. and Munz, C.-D. The multiple pressure variables (MPV) for the numerical approximation of weakly compressible fluid flow. In *Proceedings of Numerical Modelling in Continuum Mechanics, Charles University Prague*, 1995.

[KM05] Khouider, B. and Majda, A. A non-oscillatory balanced scheme for an idealized tropical climate model. *Theoretical and Computational Fluid Dynamics*, 19(5):331–354, 2005.

[KM14] Käppeli, R. and Mishra, S. Well-balanced schemes for the Euler equations with gravitation. *Journal of Computational Physics*, 259:199–219, 2014.

[KP96] Kobayashi, M. H. and Pereira, J. C. F. Characteristic-based pressure correction at all speeds. *AIAA journal*, 34(2):272–280, 1996.

[KS17] Kaiser, K. and Schütz, J. A high-order method for weakly compressible flows. *Communications in Computational Physics*, 22(4):1150–1174, 2017.

[KSSN16] Kaiser, K., Schütz, J., Schöbel, R., and Noelle, S. A new stable splitting for the isentropic Euler equations. *Journal of Scientific Computing*, pages 1–18, 2016.

[KVPR10] Klein, R., Vater, S., Paeschke, E., and Ruprecht, D. Multiple scales methods in meteorology. In *Asymptotic Methods in Fluid Mechanics: Survey and Recent Advances*, pages 127–196. Springer, 2010.

[Lan13] Lannes, D. The water waves problem. *Mathematical Surveys and Monographs*, 188, 2013.

[LE88] Lomov, S. A. and Eliseev, A. G. Asymptotic integration of singularly perturbed problems. *Russian Mathematical Surveys*, 43(3):1–63, 1988.

[LeV02] LeVeque, R. J. *Finite volume methods for hyperbolic problems*, volume 31. Cambridge university press, 2002.

[LeV07] LeVeque, R. J. *Finite difference methods for ordinary and partial differential equations: Steady-state and time-dependent problems*, volume 98. SIAM, 2007.

[Lio96] Lions, P.-L. *Mathematical topics in fluid mechanics*, volume 1. The Clarendon Press Oxford University Press, New York, 1996.

[Liu87] Liu, T.-P. Hyperbolic conservation laws with relaxation. *Communications in Mathematical Physics*, 108(1):153–175, 1987.

[LL98] Lax, P. D. and Liu, X.-D. Solution of two-dimensional riemann problems of gas dynamics by positive schemes. *SIAM Journal on Scientific Computing*, 19(2):319–340, 1998.

[LM89] Larsen, E. W. and Morel, J. E. Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes II. *Journal of Computational Physics*, 83(1), 1989.

[LM08] Lemou, M. and Mieussens, L. A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit. *SIAM Journal on Scientific Computing*, 31(1):334–368, 2008.

[LM15] Lannes, D. and Marche, F. A new class of fully nonlinear and weakly dispersive Green–Naghdi models for efficient 2D simulations. *Journal of Computational Physics*, 282:238–268, 2015.

[LMM87] Larsen, E. W., Morel, J. E., and Miller, W. F. Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. *Journal of Computational Physics*, 69(2):283–324, 1987.

[LMNK07] Lukáčová-Medvid'ová, M., Noelle, S., and Kraft, M. Well-balanced finite volume evolution Galerkin methods for the shallow water equations. *Journal of Computational Physics*, 221(1):122–147, 2007.

[LOT96] Levermore, C. D., Oliver, M., and Titi, E. S. Global well-posedness for the lake equations. *Physica D: Nonlinear Phenomena*, 98(2):492–509, 1996.

[LS01] Lattanzio, C. and Serre, D. Convergence of a relaxation scheme for hyperbolic systems of conservation laws. *Numerische Mathematik*, 88(1):121–134, 2001.

[LW07] LeFloch, P. G. and Westdickenberg, M. Finite energy solutions to the isentropic Euler equations with geometric effects. *Journal de Mathématiques Pures et Appliquées*, 88(5):389–429, 2007.

[Maj03] Majda, A. *Introduction to PDEs and waves for the atmosphere and ocean*, volume 9. American Mathematical Society, 2003.

[Mar49] Marden, M. *Geometry of polynomials*. Number 3. American Mathematical Society, 1949.

[Mas07] Masmoudi, N. Examples of singular limits in hydrodynamics. *Handbook of Differential Equations: Evolutionary Equations*, 3:195–275, 2007.

[MD01] Moukalled, F. and Darwish, M. A high-resolution pressure-based algorithm for fluid flow at all speeds. *Journal of Computational Physics*, 168(1):101–130, 2001.

[MD16] Michel-Dansac, V. *Development of high-order well-balanced schemes for geophysical flows*. PhD thesis, Université de Nantes, 2016.

[MDBCF16] Michel-Dansac, V., Berthon, C., Clain, S., and Foucher, F. A well-balanced scheme for the shallow-water equations with topography. *Computers & Mathematics with Applications*, 72(3):568–593, 2016.

[MDR07] Munz, C.-D., Dumbser, M., and Roller, S. Linearized acoustic perturbation equations for low Mach number flow with variable density and temperature. *Journal of Computational Physics*, 224(1):352–364, 2007.

[Mei99] Meister, A. Asymptotic single and multiple scale expansions in the low Mach number limit. *SIAM Journal on Applied Mathematics*, 60(1):256–271, 1999.

[MM98] Morton, K. W. and Mayers, D. F. Numerical solution of partial differential equations. *Journal of Fluid Mechanics*, 363:349–349, 1998.

[Moc80] Mock, M. S. A topological degree for orbits connecting critical points of autonomous systems. *Journal of Differential Equations*, 38(2):176–191, 1980.

[MP85] Majda, A. and Pego, R. L. Stable viscosity matrices for systems of conservation laws. *Journal of Differential equations*, 56(2):229–262, 1985.

[MRKG03] Munz, C.-D., Roller, S., Klein, R., and Geratz, K. J. The extension of incompressible flow solvers to the weakly compressible regime. *Computers & Fluids*, 32(2):173–196, 2003.

[MS01] Métivier, G. and Schochet, S. The incompressible limit of the non-isentropic Euler equations. *Archive for Rational Mechanics and Analysis*, 158(1):61–90, 2001.

[MW06] Majda, A. and Wang, X. *Nonlinear dynamics and statistical theories for basic geophysical flows*. Cambridge University Press, 2006.

[MYO90] Munson, B. R., Young, D. F., and Okiishi, T. H. Fundamentals of fluid mechanics. *New York*, 3(4), 1990.

[NBA+14] Noelle, S., Bispen, G., Arun, K. R., Lukáčová-Medvid'ová, M., and Munz, C.-D. A weakly asymptotic preserving low Mach number scheme for the Euler equations of gas dynamics. *SIAM Journal on Scientific Computing*, 36(6):B989–B1024, 2014.

[NMRR96] Nečas, J., Málek, J., Rokyta, M., and Růžička, M. *Weak and measure-valued solutions to evolutionary PDEs*, volume 13. CRC Press, 1996.

[NPPN06] Noelle, S., Pankratz, N., Puppo, G., and Natvig, J. R. Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows. *Journal of Computational Physics*, 213(2):474–499, 2006.

[NT92] Nessyahu, H. and Tadmor, E. The convergence rate of approximate solutions for nonlinear scalar conservation laws. *SIAM Journal on Numerical Analysis*, 29(6):1505–1519, 1992.

[NXS07] Noelle, S., Xing, Y., and Shu, C.-W. High-order well-balanced finite volume WENO schemes for shallow water equation with moving water. *Journal of Computational Physics*, 226(1):29–58, 2007.

[OSB+16] Oßwald, K., Siegmund, A., Birken, P., Hannemann, V., and Meister, A. L^{2}Roe: a low dissipation version of Roe's approximate Riemann solver for low Mach numbers. *International Journal for Numerical Methods in Fluids*, 81(2):71–86, 2016.

[Ost66] Ostrowski, A. M. *Solution of equations and systems of equations*, volume 9. Academic Press New York, 1966.

[PDZ+14] Panda, N., Dawson, C., Zhang, Y., Kennedy, A. B., Westerink, J. J., and Donahue, A. S. Discontinuous Galerkin methods for solving Boussinesq–Green–Naghdi equations in resolving non-linear and dispersive surface water waves. *Journal of Computational Physics*, 273:572–588, 2014.

[Ped13] Pedlosky, J. *Geophysical fluid dynamics*. Springer Science & Business Media, 2013.

[Per98] Persson, A. How do we understand the Coriolis force? *Bulletin of the American Meteorological Society*, 79(7):1373, 1998.

[Phi59] Phillips, N. A. An example of non-linear computational instability. *The Atmosphere and the Sea in motion*, 501:501–504, 1959.

[PM05] Park, J. H. and Munz, C.-D. Multiple pressure variables methods for fluid flow at all Mach numbers. *International Journal of Numerical Methods in Fluids*, 49:905–931, 2005.

[PR05] Pareschi, L. and Russo, G. Implicit-explicit Runge–Kutta schemes and applications to hyperbolic systems with relaxation. *Journal of Scientific Computing*, 25(1-2):129–155, 2005.

[PTA12] Pletcher, R. H., Tannehill, J. C., and Anderson, D. *Computational fluid mechanics and heat transfer*. CRC Press, 2012.

[RB09a] Ricchiuto, M. and Bollermann, A. Stabilized residual distribution for shallow water simulations. *Journal of Computational Physics*, 228(4):1071–1115, 2009.

[RB09b] Rieper, F. and Bader, G. The influence of cell geometry on the accuracy of upwind schemes in the low Mach number regime. *Journal of Computational Physics*, 228(8):2918–2933, 2009.

[Rie10] Rieper, F. On the dissipation mechanism of upwind-schemes in the low Mach number regime: A comparison between Roe and HLL. *Journal of Computational Physics*, 229(2):221–232, 2010.

[Rie11] Rieper, F. A low-Mach number fix for Roe's approximate Riemann solver. *Journal of Computational Physics*, 230(13):5263–5287, 2011.

[RM67] Richtmyer, R. D. and Morton, K. W. *Difference methods for initial-value problems*. Interscience Publishers John Wiley & Sons, Inc., Academia Publishing House of the Czechoslovak Acad, 1967.

[Ros38] Rossby, C.-G. On the mutual adjustment of pressure and velocity distributions in certain simple current systems, II. *Journal of Marine Research*, 1(3):239–263, 1938.

[Ros63] Rosenbrock, H. H. Some general implicit processes for the numerical solution of differential equations. *The Computer Journal*, 5(4):329–330, 1963.

[Sch05] Schochet, S. The mathematical theory of low Mach number flows. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 39(3):441–458, 2005.

[Sil00] Silvester, J. R. Determinants of block matrices. *The Mathematical Gazette*, pages 460–467, 2000.

[SK16] Schütz, J. and Kaiser, K. A new stable splitting for singularly perturbed ODEs. *Applied Numerical Mathematics*, 107:18–33, 2016.

[SN14] Schütz, J. and Noelle, S. Flux splitting for stiff equations: A notion on stability. *Journal of Scientific Computing*, pages 1–19, 2014.

[Svä15] Svärd, M. Entropy solutions of the compressible Euler equations. *BIT Numerical Mathematics*, pages 1–18, 2015.

[Svä16] Svärd, M. A convergent numerical scheme for the compressible Navier–Stokes equations. *SIAM Journal on Numerical Analysis*, 54(3):1484–1506, 2016.

[Tad87] Tadmor, E. The numerical viscosity of entropy stable schemes for systems of conservation laws. I. *Mathematics of Computation*, 49(179):91–103, 1987.

[Tad03] Tadmor, E. Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. *Acta Numerica*, 12:451–512, 2003.

[Tan12] Tang, M. Second order all speed method for the isentropic Euler equations. *Kinetic and Related Models*, 5:155–184, 2012.

[Tao12] Tao, T. *Topics in random matrix theory*, volume 132. American Mathematical Society, 2012.

[Tem69] Temam, R. Sur l'approximation de la solution des équations de Navier–Stokes par la méthode des pas fractionnaires (II). *Archive for Rational Mechanics and Analysis*, 33(5):377–385, 1969.

[TFVL93] Turkel, E., Fiterman, A., and Van Leer, B. Preconditioning and the limit to the incompressible flow equations. Technical report, DTIC Document, 1993.

[TKK16] Touma, R., Koley, U., and Klingenberg, C. Well-balanced unstaggered central schemes for the Euler equations with gravitation. *SIAM Journal on Scientific Computing*, 38(5):B773–B807, 2016.

[Tra09] Trangenstein, J. A. *Numerical solution of hyperbolic partial differential equations*. Cambridge University Press, 2009.

[Tre96] Trefethen, L. N. *Finite difference and spectral methods for ordinary and partial differential equations*. Cornell University, 1996.

[Tur87] Turkel, E. Preconditioned methods for solving the incompressible and low speed compressible equations. *Journal of Computational Physics*, 72(2):277–298, 1987.

[TZ59] Taussky, O. and Zassenhaus, H. On the similarity transformation between a matrix and its transpose. *Pacific J. Math*, 9(3):893–896, 1959.

[Var75] Varah, J. M. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, 11(1):3–5, 1975.

[Vas94] Vasilieva, A. B. On the development of singular perturbation theory at Moscow State University and elsewhere. *SIAM Review*, 36(3):440–452, 1994.

[Vat13] Vater, S. *A multigrid-based multiscale numerical scheme for shallow water flows at low Froude number*. PhD thesis, Freie Universität Berlin, 2013.

[vdHVW03] van der Heul, D. R., Vuik, C., and Wesseling, P. A conservative pressure-correction method for flow at all speeds. *Computers & Fluids*, 32(8):1113–1132, 2003.

[VK09] Vater, S. and Klein, R. Stability of a Cartesian grid projection method for zero Froude number shallow water flows. *Numerische Mathematik*, 113(1):123–161, 2009.

[VL92] Van Loan, C. *Computational frameworks for the fast Fourier transform*, volume 10. SIAM, 1992.

[VR99] Villatoro, F. R. and Ramos, J. I. On the method of modified equations. I: Asymptotic analysis of the Euler forward difference method. *Applied mathematics and computation*, 103(2):111–139, 1999.

[WEL+07] Weinan, E., Engquist, B., Li, X., Ren, W., and Vanden-Eijnden, E. Heterogeneous multiscale methods: A review. *Communications in Computational Physics*, 2(3):367–450, 2007.

[Wey49] Weyl, H. Shock waves in arbitrary fluids. *Communications on Pure and Applied Mathematics*, 2(2-3):103–122, 1949.

[WH74] Warming, R. F. and Hyett, B. J. The modified equation approach to the stability and accuracy analysis of finite-difference methods. *Journal of Computational Physics*, 14(2):159–179, 1974.

[WSW02] Wenneker, I., Segal, A., and Wesseling, P. A Mach-uniform unstructured staggered grid method. *International Journal for Numerical Methods in Fluids*, 40(9):1209–1235, 2002.

[WW65] Willett, D. and Wong, J. S. W. On the discrete analogues of some generalizations of Gronwall's inequality. *Monatshefte für Mathematik*, 69(4):362–367, 1965.

[Zak16a] Zakerzadeh, H. Asymptotic analysis of the RS-IMEX scheme for the shallow water equations in one space dimension. *ESAIM: Mathematical Modelling and Numerical Analysis (to appear)*, 2016. HAL: hal-01491450.

[Zak16b] Zakerzadeh, H. Asymptotic consistency of the RS-IMEX scheme for the low-Froude shallow water equations: Analysis and numerics. In *Proceedings of XVI International Conference on Hyperbolic Problems, Aachen*, 2016.

[Zak17a] Zakerzadeh, H. On the Mach-uniformity of the Lagrange-projection scheme. *ESAIM: Mathematical Modelling and Numerical Analysis*, 51(4):1343–1366, 2017.

[Zak17b] Zakerzadeh, H. *The RS-IMEX Scheme for the rotating shallow water equations with the Coriolis force*, pages 199–207. Springer International Publishing, Cham, 2017.

[ZF16] Zakerzadeh, H. and Fjordholm, U. S. High-order accurate, fully discrete entropy stable schemes for scalar conservation laws. *IMA Journal of Numerical Analysis*, 36(2):633–654, 2016.

[ZKD+14] Zhang, Y., Kennedy, A. B., Donahue, A. S., Westerink, J. J., Panda, N., and Dawson, C. Rotational surf zone modeling for $O(\mu^4)$ Boussinesq–Green–Naghdi systems. *Ocean Modelling*, 79:43–53, 2014.

[ZM16] Zakerzadeh, M. and May, G. On the convergence of a shock capturing discontinuous Galerkin method for nonlinear hyperbolic systems of conservation laws. *SIAM Journal of Numerical Analysis*, 54(2):874–898, 2016.

[ZN17] Zakerzadeh, H. and Noelle, S. A note on the stability of implicit-explicit flux-splittings for stiff systems of hyperbolic conservation laws. *Communications in Mathematical Sciences (to appear)*, 2017. IGPM report 449, RWTH Aachen University.

List of Figures

List of Tables

Curriculum Vitae

Hamed Zakerzadeh

Institut de Mathématiques de Toulouse
Université Toulouse III - Paul Sabatier
Toulouse, France
118 route de Narbonne, 31400
Bat. IR3, Bureau 222
Email: szakerza@math.univ-toulouse.fr

Date of birth: 24.08.1988 (Tehran, Iran)
Citizenship: Iran

Agent Contractuel de Recherche (PostDoc), Institut de Mathématiques de Toulouse
Université Toulouse III - Paul Sabatier, Toulouse, France *since Oct. 2017*

Mitarbeiter (Researcher), Institut für Geometrie und Praktische Mathematik (IGPM)
RWTH Aachen, Aachen, Germany *Apr. 2017 – Sept. 2017*

PhD student, Institut für Geometrie und Praktische Mathematik (IGPM)
RWTH Aachen, Aachen, Germany *2014 – 2017*

Master of Science, German Research School for Simulation Sciences (GRS)
RWTH Aachen, Aachen, Germany *2011 – 2013*

Bachelor of Science, Mechanical Engineering Department
Sharif University of Technology, Tehran, Iran *2006 – 2010*