Asymptotic analysis of the RS-IMEX scheme for the shallow water equations in one space dimension

Hamed Zakerzadeh

Mathematical Modelling and Numerical Analysis Will be set by the publisher
Modélisation Mathématique et Analyse Numérique

ASYMPTOTIC ANALYSIS OF THE RS-IMEX SCHEME FOR THE SHALLOW WATER EQUATIONS IN ONE SPACE DIMENSION *

Hamed Zakerzadeh[1]

Abstract. We introduce and analyse the so-called *Reference Solution IMplicit-EXplicit scheme* as a flux-splitting method 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 confirm the analysis.

Résumé. ...

1991 Mathematics Subject Classification. Primary 35L65, 65M08; Secondary 35L81, 65M12.

January 11, 2018.

1. Introduction

Singular limits of conservation laws (or more generally PDEs), characterised by the singular 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 [48]. For instance, when the Mach number, denoted by Ma, approaches zero for the Euler equations, the sound speed (the characteristic speed) goes to infinity and the PDE changes to be hyperbolic-elliptic, in the so-called incompressible limit. To prove the convergence of the solution of the compressible Euler equations to the incompressible system is very demanding; we refer the reader to consult [14, 43, 44, 48, 58] to review the existing results.

Tackling such singular problems numerically is also complicated. For example, the weakly compressible Euler system (with $Ma \sim \varepsilon \ll 1$) is stiff due to very fast acoustic waves, which makes the Courant–Friedrichs–Lewy (CFL) condition to restrict the time step varies non-uniformly with ε, *i.e.*, $\Delta t \lesssim \varepsilon\, \Delta x$. This leads to very small time steps, thus a huge computational cost. Generally speaking, the usual numerical schemes also lose their accuracy in the limit for under-resolved grids; see [17, 18, 31, 30, 55, 56, 54].

Keywords and phrases: IMEX scheme, Asymptotic preserving, Flux splitting, Stability analysis

* *The author's research was supported by the scholarship of RWTH Aachen university through Graduiertenförderung nach Richtlinien zur Förderung des wissenschaftlichen Nachwuchses (RFwN).*

[1] Institut für Geometrie und Praktische Mathematik, RWTH Aachen University
Templergraben 55, 52056, Aachen, Germany e-mail: h.zakerzadeh@igpm.rwth-aachen.de

In the sequel, we mainly consider well-prepared initial data, which are consistent with the limit $\varepsilon \to 0$, in order to simplify the problem and to eliminate spurious initial layers (see Definition 3.1 and [48]). We also assume that the *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 in the discrete level. This is, in fact, the idea of Asymptotic Preserving (AP) schemes, which has been introduced by Jin in [39, 40] for relaxation systems; see also [41] for a general review and [47] for older works (without being named AP). Illustrating the definition, assume that $\mathcal{M}^\varepsilon$ stands for a continuous physical model with the (singular) perturbation 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, then the scheme is called to be AP. More precisely, we define an AP scheme as below.

Definition 1.1. *A scheme is called AP, provided that the following conditions are fulfilled for the scheme:*

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

(ii) Asymptotic Efficiency (AEf): It 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) Asymptotic Stability (AS): It is stable in some suitable sense, uniformly in ε.

Remark 1.2. *(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. [41, 27]. Also, sometimes, the uniformity of the CFL condition is classified as the asymptotic stability rather than asymptotic efficiency.*

(ii) As mentioned in [41], the asymptotic consistency suggests that the solution belongs to a manifold approaching the limit manifold as $\varepsilon \to 0$ (up to some discretisation error).

(iii) 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 [6]. Moreover and very recently, the authors in [21] have addressed this point more fundamentally for some toy models related to Vlasov–Maxwell equations.

(iv) Mostly, we consider the well-prepared initial data; however, we will also show in Appendix C, that the scheme projects the ill-prepared initial data on the limit manifold for $\varepsilon \ll 1$. So, one would expect to observe good results even with an ill-prepared initial datum. This is verified by a numerical example in Section 5.1.2.

The AP property has been studied extensively for the kinetic equations (see [41, 38] for a review) while its application for the conservation laws is more recent; just to name a few see [7, 50, 19, 13, 16, 32]. There are also several related works without using the initialism AP; see [45] as one of the first examples for the Euler equations and [46, 24]. Note that while the (formal) asymptotic consistency of schemes is often studied and proved in the literature, there are only a few results regarding the uniform (asymptotic) stability, particularly in the context of conservation laws, like [19, 66] for the isentropic Euler equations; see also [15] and [27] for the Euler–Poisson and Euler–Korteweg systems, respectively.

The bottom-line of these AP schemes is the implicit-explicit (IMEX) strategy, *e.g.*, to split the flux (or its Jacobian) into stiff and non-stiff parts, and to treat the latter explicitly in time and the former implicitly in time; see [3, 8, 12, 51] for more details about the use of IMEX methods for constructing AP schemes. Some kind of implicit treatment is definitely necessary to find schemes with an ε-uniform CFL condition. But, such a *CFL stability* is not sufficient for the asymptotic stability in a norm; see for example [1], where the scheme is unconditionally unstable in L_2-norm, though, both split parts are stable in terms of the CFL condition. On the other hand, IMEX schemes are L_2-stable as long as each step is so, as shown in [32]. One can also think of fully-implicit schemes like finite volume schemes [26], mixed finite element-finite volume schemes [22], and space-time dG schemes [34] (see also [69] for its modified variant without the streamline diffusion). Fully-implicit schemes have 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.

In [50], the authors applied a flux-splitting scheme to the full Euler equations, which uses a variant of Klein's auxiliary splitting [45]. The scheme required an ε-dependent time step for stability. Motivated by this, the authors in [60] 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; see also [68] 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 paper.

The key idea is the linearisation around an *(asymptotic) reference solution* such that the resulting modified equation is stable. In fact, using the asymptotic reference solution gives a small commutator, which provides a heuristic argument for the stability of the modified equation (*cf.* [68]); see Remark 2.2 below for a discussion on this. Note that in the work of our collaborators, the RS-IMEX scheme is shown to be quantitatively well-behaved in practice; see [59] and [42] for the application of the scheme to the Van der Pol equation and the 2d isentropic Euler system, respectively.

In the present article, 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 3.10). These make a solid background for the future works which extend the scheme to the multi-dimensional SWE with different source terms; see [65, 67] for instance. Note that broadly speaking, the splitting developed in [7] can be considered as a particular example of the RS-IMEX scheme, with the zero reference solution; see Remark 3.4 for more details. We would also like to mention that a somewhat similar idea to the RS-IMEX scheme has been used in [25] (as the so-called *penalisation method* [38]) for the kinetic equations in the low-Knudsen regime, where the authors split the collision operator using a linearisation around the Maxwellian distribution.

The remainder of this paper is organised as follows. In Section 2, we present a short introduction to the RS-IMEX scheme for a general hyperbolic system of balance laws, followed by the rigorous AP analysis (consistency and stability) of the RS-IMEX scheme for the 1d SWE with the lake at rest or LaR (constant water surface and zero velocity) and the zero-Froude limit reference solutions, in Section 3 and Section 4. Section 5 provides some numerical evidence to confirm the AP analysis and to test the quality of the solutions. The results of this manuscript supply some necessary elements for the more interesting case of the 2d SWE in [65, 67].

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 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):

$$\partial_t \boldsymbol{U}(t, \boldsymbol{x}; \varepsilon) + \operatorname{div}_{\boldsymbol{x}} \boldsymbol{F}(\boldsymbol{U}, t, \boldsymbol{x}; \varepsilon) = \boldsymbol{S}(\boldsymbol{U}, t, \boldsymbol{x}; \varepsilon), \tag{1}$$

where $\boldsymbol{U} : [0, +\infty) \times \Omega \to \mathbb{R}^q$ is the vector of unknowns, $\boldsymbol{F} : \mathbb{R}^q \times [0, +\infty) \times \Omega \to \mathbb{R}^{q \times d}$ 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 $\boldsymbol{F}' := \partial_{\boldsymbol{U}} \boldsymbol{F}$, *i.e.*, for all directions $\boldsymbol{n} = (n_1, \dots, n_d)^T$ in $\mathbb{R}^d$, the 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.

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

$$\overline{\boldsymbol{U}} : [0, +\infty) \times \mathbb{R}^d \to \mathbb{R}^q, \quad (t, \boldsymbol{x}) \mapsto \overline{\boldsymbol{U}}(t, \boldsymbol{x}). \tag{2}$$

Typically, $\overline{\boldsymbol{U}}$ is a steady state solution of the balance law, or the solution of the asymptotic limit equation, derived from (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 (1) into the reference solution $\overline{\boldsymbol{U}}$ and a perturbation $\boldsymbol{U}_{pert}$, that is $\boldsymbol{U}(t,\boldsymbol{x};\varepsilon) = \overline{\boldsymbol{U}}(t,\boldsymbol{x}) + \boldsymbol{U}_{pert}(t,\boldsymbol{x};\varepsilon)$. 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 $\overline{\boldsymbol{U}}$, to a reference, stiff and non-stiff parts respectively:

$$\boldsymbol{F}(\boldsymbol{U}) = \boldsymbol{F}(\overline{\boldsymbol{U}}) + \boldsymbol{F}'(\overline{\boldsymbol{U}})\boldsymbol{U}_{pert} + \big(\boldsymbol{F}(\boldsymbol{U}) - \boldsymbol{F}(\overline{\boldsymbol{U}}) - \boldsymbol{F}'(\overline{\boldsymbol{U}})\boldsymbol{U}_{pert}\big) =: \overline{\boldsymbol{F}} + \widetilde{\boldsymbol{F}} + \widehat{\boldsymbol{F}}, \tag{3a}$$

$$\boldsymbol{S}(\boldsymbol{U}) = \boldsymbol{S}(\overline{\boldsymbol{U}}) + \boldsymbol{S}'(\overline{\boldsymbol{U}})\boldsymbol{U}_{pert} + \big(\boldsymbol{S}(\boldsymbol{U}) - \boldsymbol{S}(\overline{\boldsymbol{U}}) - \boldsymbol{S}'(\overline{\boldsymbol{U}})\boldsymbol{U}_{pert}\big) =: \overline{\boldsymbol{S}} + \widetilde{\boldsymbol{S}} + \widehat{\boldsymbol{S}}. \tag{3b}$$

Note that the stiff part of the splitting $(\widetilde{\boldsymbol{F}}, \widetilde{\boldsymbol{S}})$ is linear by construction, which is very advantageous in terms of computational cost, compared to splittings with non-linear stiff parts like [32, 16]. Hence, there is no need for solving non-linear systems, *e.g.*, by the Newton iteration method. The idea of such a linearisation goes back to [57] (see also [33, Chap IV.7]) for ODEs (the so-called linearly-implicit methods) and has been used later in [7] for the SWE, motivated by [52], which shares the same concept but using the semi-Lagrangian method. 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 (in the sense of a Poincaré or asymptotic expansion). Later on, we will see that an appropriate choice of the scaling matrix, not only makes the analysis more illustrative (see Remark 3.9) but also may affect the numerical solution (see Remark 4.2). For this reason, we introduce the diagonal matrix $D := \mathrm{diag}(\varepsilon^{d_j})$ with some integer d_j for $j = 1, \dots, q$, and we can 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 (3a) and (3b) can be re-written:

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

We refer the reader to Section 3 for an explicit example of such a scaling for the shallow water system.

Remark 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 $\boldsymbol{V} = (v_1, \dots, v_q)^T$:

$$\partial_t \boldsymbol{V} + \mathrm{div}_{\boldsymbol{x}}\Big(\widetilde{\boldsymbol{G}}(\overline{\boldsymbol{U}}, \boldsymbol{V}) + \widehat{\boldsymbol{G}}(\overline{\boldsymbol{U}}, \boldsymbol{V})\Big) = \widetilde{\boldsymbol{Z}}(\overline{\boldsymbol{U}}, \boldsymbol{V}) + \widehat{\boldsymbol{Z}}(\overline{\boldsymbol{U}}, \boldsymbol{V}) - \overline{\boldsymbol{T}}(\overline{\boldsymbol{U}}), \tag{4}$$

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

$$\overline{\boldsymbol{T}}(\overline{\boldsymbol{U}}) := D^{-1}\partial_t\overline{\boldsymbol{U}} + \mathrm{div}_{\boldsymbol{x}}\overline{\boldsymbol{G}}(\overline{\boldsymbol{U}}) - \overline{\boldsymbol{Z}}(\overline{\boldsymbol{U}}). \tag{5}$$

Remark 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*

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

By choosing the reference solution as the asymptotic limit, $\boldsymbol{U} - \overline{\boldsymbol{U}}$ *is small for* $\varepsilon \ll 1$*, which suggests that the commutator is small and leads to the stability in the sense of modified equations [68].*[1] *Moreover, we have*

[1] The modified equation is the *"actual"* equation solved by a numerical approximation when applied to a differential equation [63]; see also [68, 60].

shown in [68] that if the eigenspaces have similar structures, the modified equation is likely to be stable. Since the eigenvectors of $\boldsymbol{G}'(\boldsymbol{U})$ and $\boldsymbol{G}'(\overline{\boldsymbol{U}})$ are close to each other in the structure (regardless of $\|\boldsymbol{U}-\overline{\boldsymbol{U}}\|$), the modified equation is stable, which suggests that it is the linearly-implicit strategy which makes the modified equation stable, not the asymptotic reference solution. We will see in Section 3 that even with a large commutator, the modified equation can be stable. Although it seems that the reference solution does 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 below).

Defining $\boldsymbol{R} := -\mathrm{div}_{\boldsymbol{x}}\boldsymbol{G} + \boldsymbol{Z}$ (with analogous definitions for $\overline{\boldsymbol{R}}$, $\widetilde{\boldsymbol{R}}$ and $\widehat{\boldsymbol{R}}$), one can reformulate (4) as

$$\partial_t \boldsymbol{V} = -\overline{\boldsymbol{T}} + \widetilde{\boldsymbol{R}} + \widehat{\boldsymbol{R}}, \tag{6}$$

which is a balance law for the scaled perturbation $\boldsymbol{V}$. Note that using the forms (4) and (6) are not indispensable for the numerical scheme, but it is more convenient 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 (1).

2.1. Numerical scheme

The Jacobian $\widetilde{\boldsymbol{F}}'$ in (3a) (and $\widetilde{\boldsymbol{G}}'$ due to Remark 2.1) has stiff eigenvalues. So, to solve (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;[2] 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.

Definition 2.3. *Given the reference solution $\overline{\boldsymbol{U}}$, the fully-discrete RS-IMEX scheme for (6) is given by*

$$D_t \boldsymbol{V}_\Delta^n = -\overline{\boldsymbol{T}}_\Delta^{n+1} + \widetilde{\boldsymbol{R}}_\Delta^{n+1} + \widehat{\boldsymbol{R}}_\Delta^n, \tag{7}$$

where $D_t\phi(\boldsymbol{x},t) := \frac{\phi(\boldsymbol{x},t+\Delta t)-\phi(\boldsymbol{x},t)}{\Delta t}$, and the subscript Δ stands for a choice of spatial discretisation.

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 $x_{i+1/2}$. The numerical viscosity α is originally chosen such that $\alpha_{i+1/2} \geq \max_{u\in[u_i,u_{i+1}]} f'(u)$. However, as the stiff system is treated implicitly and implicit schemes are rather diffusive, we choose α for the stiff subsystem rather arbitrary not to add too much diffusion; that is to say that $\widetilde{\alpha}, \widehat{\alpha} = \mathcal{O}(1)$. The extension of this numerical flux to systems and/or in multi-dimensions is obvious. Note that the source term should be discretised appropriately so that the scheme preserves the equilibrium (well-balancing, *cf.* [10]). We will see in Section 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.

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 (1), we follow the procedure described in Section 2 to derive the scheme for the 1d SWE with topography and the LaR reference solution. Then, in Section 4, we use the zero-Froude limit reference solution.

[2] 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.

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.

The non-dimensionalised SWE in one space dimension, using $h = z - b$ (with $b < 0$) and $m = hu$, can be written as in [7]:

$$\boldsymbol{U} = \begin{bmatrix} z \\ m \end{bmatrix}, \qquad \boldsymbol{F} = \begin{bmatrix} m \\ \dfrac{m^2}{z-b} + \dfrac{z^2 - 2zb}{2\varepsilon^2} \end{bmatrix}, \qquad \boldsymbol{S} = \begin{bmatrix} 0 \\ -\dfrac{z\partial_x b}{\varepsilon^2} \end{bmatrix}. \tag{8}$$

In this notation, z is the surface elevation from some chosen constant surface level H_{ref}, m is the momentum and b is the water depth measured from H_{ref} with a negative sign (see Figure 1). The singular parameter $\varepsilon \in (0, 1]$ is called the *Froude number*, defined as the ratio of the characteristic bulk velocity over the characteristic velocity of gravity waves, which is analogous to acoustic waves for the isentropic Euler system. Note that for this shallow water model to be valid, the bottom slope or $\partial_x b$ should be small enough; see [9] for details.

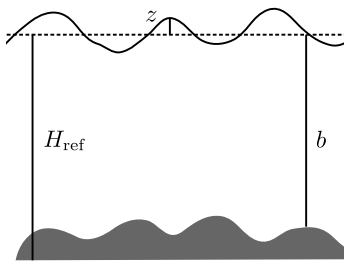


FIGURE 1. Variables used in the shallow water formulation (8). The water height is written as $h = z - b$.

We set the reference state as the LaR, $\overline{\boldsymbol{U}} := (\overline{z}, \overline{m})^T$ with $\overline{z}$ constant in space and $\overline{m} = 0$. Therefore, due to (3a)–(3b), the splitting reads

$$\overline{\boldsymbol{F}} = \begin{bmatrix} 0 \\ \dfrac{1}{2\varepsilon^2}\overline{z}(\overline{z} - 2b) \end{bmatrix}, \qquad \widetilde{\boldsymbol{F}} = \begin{bmatrix} m_{pert} \\ \dfrac{(\overline{z} - b)}{\varepsilon^2} z_{pert} \end{bmatrix}, \qquad \widehat{\boldsymbol{F}} = \begin{bmatrix} 0 \\ \dfrac{m_{pert}^2}{\overline{z} + z_{pert} - b} + \dfrac{z_{pert}^2}{2\varepsilon^2} \end{bmatrix},$$

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

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

$$\begin{aligned} \widetilde{\boldsymbol{F}}' &= \begin{bmatrix} 0 & 1 \\ \dfrac{\overline{z}-b}{\varepsilon^2} & 0 \end{bmatrix}, & \widetilde{\lambda} &= \pm\frac{\sqrt{\overline{z}-b}}{\varepsilon}, \\ \widehat{\boldsymbol{F}}' &= \begin{bmatrix} 0 & 0 \\ -u_{pert}^2 + \dfrac{z_{pert}}{\varepsilon^2} & 2u_{pert} \end{bmatrix}, & \widehat{\lambda} &= 0, 2\,u_{pert}, \end{aligned} \tag{9}$$

with $u_{pert} := m_{pert}/(\overline{z} + z_{pert} - b)$. Thus, the splitting is admissible in the sense of [60]. Note that the case $\overline{\boldsymbol{U}} = \boldsymbol{0}$ gives the same splitting as in [7, 6].

Finding the scaling matrix D, we employ the formal asymptotic analysis in Appendix A (see also [43, 44] for the rigorous justification for the flat bottom case), which suggests the formal zero-Froude limit in Definition 3.1 below, with the following asymptotic (Poincaré) expansion

$$\begin{aligned} z(t,x) &= z_{(0)} + \varepsilon z_{(1)} + \varepsilon^2 z_{(2)}, \\ m(t,x) &= m_{(0)} + \varepsilon m_{(1)} + \varepsilon^2 m_{(2)}. \end{aligned} \tag{10}$$

Definition 3.1. *The formal zero-Froude limit of the shallow water system (8) gives the so-called "lake equations" and writes*

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

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

Definition 3.2. *For the 1d SWE (8), we call the initial data $(z_{0,\varepsilon}, m_{0,\varepsilon})$ well-prepared if it holds that*

$$\begin{aligned} z(0,\cdot) &= z_{0,\varepsilon} = z_{(0)}^0 + \varepsilon^2 z_{(2),\varepsilon}^0, \\ m(0,\cdot) &= m_{0,\varepsilon} = m_{(0)}^0 + \varepsilon m_{(1),\varepsilon}^0, \end{aligned} \tag{11}$$

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 (11), 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.

3.1. RS-IMEX scheme

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

$$\widehat{\boldsymbol{G}} = \begin{bmatrix} 0 \\ \dfrac{v_2^2}{\overline{z} + \varepsilon^2 v_1 - b} + \dfrac{\varepsilon^2}{2} v_1^2 \end{bmatrix}, \qquad \widetilde{\boldsymbol{G}} = \begin{bmatrix} v_2/\varepsilon^2 \\ (\overline{z} - b) v_1 \end{bmatrix}, \tag{12a}$$

$$\widehat{\boldsymbol{Z}} = \boldsymbol{0}, \qquad \widetilde{\boldsymbol{Z}} = \begin{bmatrix} 0 \\ -v_1 \partial_x b \end{bmatrix}. \tag{12b}$$

Owing to (9) and Remark 2.1, this splitting is also admissible even with an ill-prepared initial datum (as defined in Appendix C).

Since the LaR 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

$$\partial_t \boldsymbol{V} = -\partial_x \begin{bmatrix} v_2/\varepsilon^2 \\ (\overline{z}-b)v_1 \end{bmatrix} - \partial_x \begin{bmatrix} 0 \\ \dfrac{v_2^2}{\overline{z}+\varepsilon^2 v_1 - b} + \varepsilon^2 \dfrac{v_1^2}{2} \end{bmatrix} + \begin{bmatrix} 0 \\ -v_1 \partial_x b \end{bmatrix}. \tag{13}$$

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

$$\boldsymbol{V}_i^{n+1/2} = \boldsymbol{V}_i^n - \frac{\Delta t}{\Delta x}\left(\widehat{\boldsymbol{G}}_{i+1/2}^n - \widehat{\boldsymbol{G}}_{i-1/2}^n\right) \qquad \text{(Explicit step)}, \tag{14a}$$

$$\boldsymbol{V}_i^{n+1} = \boldsymbol{V}_i^{n+1/2} - \frac{\Delta t}{\Delta x}\left(\widetilde{\boldsymbol{G}}_{i+1/2}^{n+1} - \widetilde{\boldsymbol{G}}_{i-1/2}^{n+1}\right) + \Delta t\, \widetilde{\boldsymbol{Z}}_i^{n+1} \qquad \text{(Implicit step)}, \tag{14b}$$

for each cell $i \in \{1, 2, \dots, N\}$ in the computational domain Ω_N with N cells, where $\widetilde{\boldsymbol{G}}_{i+1/2}$ and $\widehat{\boldsymbol{G}}_{i+1/2}$ denote the Rusanov flux (as defined in Section 2) with $\widehat{\boldsymbol{G}}$ and $\widetilde{\boldsymbol{G}}$ as defined in (12a), and $\widehat{\boldsymbol{Z}}_i^n$ is the central discretisation of the source terms (12b). Denoting ∇_h and Δ_h respectively as the central discretisation of the first and second derivatives, one can re-write (14a)–(14b) as

$$\boldsymbol{V}_i^{n+1/2} = \boldsymbol{V}_i^n - \Delta t \nabla_h \begin{bmatrix} 0 \\ \dfrac{v_{2,i}^{n,2}}{\overline{z}+\varepsilon^2 v_{1,i}^n - b_i} + \dfrac{\varepsilon^2}{2} v_{1.i}^{n,2} \end{bmatrix} + \frac{\widehat{\alpha}\Delta x}{2} \Delta t \Delta_h \boldsymbol{V}_i^n, \tag{15a}$$

$$\boldsymbol{V}_i^{n+1} = \boldsymbol{V}_i^{n+1/2} - \Delta t \nabla_h \begin{bmatrix} v_{2,i}^{n+1}/\varepsilon^2 \\ (\overline{z}-b_i)\, v_{1,i}^{n+1} \end{bmatrix} + \frac{\widetilde{\alpha}\Delta x}{2} \Delta t \Delta_h \boldsymbol{V}_i^{n+1} - \Delta t \begin{bmatrix} 0 \\ v_{1,i}^{n+1} \nabla_h b_i \end{bmatrix}. \tag{15b}$$

Due to Remark 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 $\mathcal{O}(1/\varepsilon^2)$:

$$\left[\widetilde{\boldsymbol{G}}', \widehat{\boldsymbol{G}}'\right] := \widetilde{\boldsymbol{G}}'\widehat{\boldsymbol{G}}' - \widehat{\boldsymbol{G}}'\widetilde{\boldsymbol{G}}' = \begin{bmatrix} v_1 - \dfrac{v_2^2}{(\overline{z}+\varepsilon^2 v_1 - b)^2} & \dfrac{2v_2/\varepsilon^2}{\overline{z}+\varepsilon^2 v_1 - b} \\ \dfrac{-2(\overline{z}-b)v_2}{\overline{z}+\varepsilon^2 v_1 - b} & -v_1 + \dfrac{v_2^2}{(\overline{z}+\varepsilon^2 v_1 - b)^2} \end{bmatrix}. \tag{16}$$

However, as shown in [68], the modified equation is asymptotically stable.

3.2. Numerical analysis of the scheme

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

Theorem 3.3. *For the shallow water equations with topography and well-prepared initial data in the sense of Definition 3.2, the RS-IMEX scheme (15a)–(15b), with (12a)–(12b), 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 convergent subservience 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 $\|\boldsymbol{V}_\Delta^n\|_{\ell_2} \leq C_{N,T_f}\|\boldsymbol{V}_\Delta^0\|_{\ell_2}$.
(v) preserves the lake at rest equilibrium state, i.e., it is well-balanced.

We present the proof of Theorem 3.3 in the next sections.

Remark 3.4. *As we have already mentioned, the scheme in [7, 6] is a particular example of the RS-IMEX scheme with the zero reference solution. So, one may expect that the analysis in [6] coincides with Theorem 3.3. The difference is that the analysis of [6] is, basically, for the flat bottom case and a detailed analysis has been done for various high-order reconstructions. By contrast, throughout this paper, we focus on a first-order scheme in one space dimension and prove asymptotic consistency for a non-flat topography. In Section 4, we show that a similar analysis can be used for more general reference solutions.*

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 define $\overline{h} := \overline{z} - b$ and $\beta := \frac{\Delta t}{2\Delta x}$.

(i) Constant b. Owing to (15b), 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

$$J_\varepsilon := \begin{bmatrix} P & \frac{\beta}{\varepsilon^2}Q \\ \beta\overline{h}Q & P \end{bmatrix} \tag{17}$$

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 [29], and one knows from [61, Thm. 1] (see also [5, 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 [37, Chap. 6], the numerical range [36, 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.* [36, 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

$$J_\varepsilon = \begin{bmatrix} P & \frac{\beta}{\varepsilon^2}Q \\ \beta R_b & P \end{bmatrix}, \tag{18}$$

where R_b is an *almost* circular 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 as well, *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 [29]). Using this fact, one can consider the transformed matrix Ξ_ε for showing solvability where

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

From [5, 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 [37, 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 (18) is invertible as well.

3.2.2. *ε-stability of the solution*

In this section, we aim to justify the validity of the formal Poincaré expansion, which will be used for the formal asymptotic consistency analysis, *i.e.*, we prove that the *implicit operator* is bounded in terms of ε. We call such a property *ε-stability* hereinafter. A similar idea has been used in [6] in the context of the Finite Volume Evolution Galerkin (FVEG) scheme [7], and in [66] for the Lagrange–projection scheme. 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 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

$$\Xi_\varepsilon := \begin{bmatrix} \Lambda_P & \frac{\beta}{\varepsilon^2}\Lambda_Q \\ \beta\overline{h}\Lambda_Q & \Lambda_P \end{bmatrix}. \tag{19}$$

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 a suitable 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.*, [5, Prop. 2.8.7].

Lemma 3.5 (Schur complement)**.** *Consider the portioned matrix* $M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$. *Then, the inverse of M exists and writes*

$$M^{-1} = \begin{bmatrix} \left(M_{11} - M_{12}M_{22}^{-1}M_{21}\right)^{-1} & -M_{11}^{-1}M_{12}\left(M_{22} - M_{21}M_{11}^{-1}M_{12}\right)^{-1} \\ -M_{22}^{-1}M_{21}\left(M_{11} - M_{12}M_{22}^{-1}M_{21}\right)^{-1} & \left(M_{22} - M_{21}M_{11}^{-1}M_{12}\right)^{-1} \end{bmatrix} \tag{20}$$

if all the inverses exist.

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

Lemma 3.6. *The inverse of matrix Ξ_ε in (19) has a bounded norm for $\varepsilon \to 0$.*

Proof. From Lemma 3.5, the inverse of Ξ_ε reads

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

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

Remark 3.7. *Lemma 3.6 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 $\boldsymbol{V}_\Delta$ 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 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 (19)) and the procedure of Lemma 3.6 does not seem to be fruitful. Using Lemma 3.5 for the inversion of partitioned matrices, one gets (with $\widetilde{\alpha} = 0$ for simplicity)

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

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 [66] 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 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 3.8. *For J_ε as in (18), 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 (18), one can write $J_\varepsilon^* J_\varepsilon$ as

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

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 [36, Chap. 1], one can write the numerical range of $J_\varepsilon^* J_\varepsilon$ as

$$W(J_\varepsilon^* J_\varepsilon) = \|\beta R_b\boldsymbol{w}_1 + \boldsymbol{w}_2\|_{\ell_2}^2 + \left\|\frac{\beta}{\varepsilon^2}Q\boldsymbol{w}_2 + \boldsymbol{w}_1\right\|_{\ell_2}^2. \tag{21}$$

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,

$$\boldsymbol{w}_1 = -\frac{\beta}{\varepsilon^2} Q\boldsymbol{w}_2 + o(1), \tag{22a}$$

$$\boldsymbol{w}_2 = -\beta R_b \boldsymbol{w}_1 + o(1). \tag{22b}$$

Multiplying (22a) 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 (22a) into (22b) yields

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

which can be re-written as $(\mathbb{I}_N - \frac{\beta^2}{\varepsilon^2} R_b Q)\boldsymbol{w}_2 = o(1)$. So, sending to the limit implies that $\boldsymbol{w}_2 \to \mathbf{0}$ as $\mathbb{I}_N - \frac{\beta^2}{\varepsilon^2} R_b Q$ is non-singular for a suitable choice of β. Alternatively, one can claim that in the limit $\boldsymbol{w}_2^{(0)} = \beta^2 R_b Q \boldsymbol{w}_2^{(2)}$, and it is shown in Appendix B 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 (22b) 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 B), $\boldsymbol{w}_1$ should tend to a *constant*. But from (22a), $\boldsymbol{w}_1$ has a difference structure, thus a vanishing mean. As discussed in Appendix B, 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 3.8 verifies that the scaled perturbation $\boldsymbol{V}_\Delta$ is $\mathcal{O}(1)$, which justifies the formal asymptotic consistency of the next section.

Remark 3.9. *So far, one important advantage of the RS-IMEX scheme based on (6) with a suitable scaling and reference solution has been to enrich Lemma 3.6 and Lemma 3.8 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 more difficult in general.*

Remark 3.10. *Note that 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 sub-sequence if necessary). To determine whether this limit is the correct zero-Froude limit will be the topic of the next section, Section 3.2.3.*

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 (10), 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 with the scaling matrix $\mathrm{diag}(\varepsilon^2, 1)$, the scaled variables write

$$v_1(t_n, x) = z_{(2)}(t_n, x), \qquad v_2(t_n, x) = m_{(0)}(t_n) + \varepsilon m_{(1)}(t_n) + \varepsilon^2 m_{(2)}(t_n, x). \tag{23}$$

Note that (23) 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 (23) into the momentum update of the explicit step (15a) 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 (23). 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 3.7 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 (15b), 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 5.2.

Lemma 3.11. *For the RS-IMEX scheme (15a)–(15b) 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:

$$J_\varepsilon \begin{bmatrix} [\![\boldsymbol{V}^{n+1}_{1,\Delta}]\!] \\ [\![\boldsymbol{V}^{n+1}_{2,\Delta}]\!] \end{bmatrix} = \begin{bmatrix} [\![\boldsymbol{V}^{n+1/2}_{1,\Delta}]\!] \\ [\![\boldsymbol{V}^{n+1/2}_{2,\Delta}]\!] \end{bmatrix}. \tag{24}$$

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

$$\|K_{11}\|, \|K_{12}\|, \|K_{22}\| = \mathcal{O}(1), \qquad \|K_{21}\| = \mathcal{O}(\varepsilon^2). \tag{25}$$

Then, owing to (23), $\|[\![\boldsymbol{V}^{n+1/2}_{1,\Delta}]\!]\| = \mathcal{O}(1)$ and $\|[\![\boldsymbol{V}^{n+1/2}_{2,\Delta}]\!]\| = \mathcal{O}(\varepsilon^2)$. Combining it with (24) yields

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

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 (25), and in particular, K_{21} and K_{22}. Let us re-write the inverse K_ε as (using Lemma 3.5)

$$K_\varepsilon = \begin{bmatrix} \left(P - \frac{\beta\overline{h}}{\varepsilon^2}QP^{-1}Q\right)^{-1} & -\frac{\beta^2}{\varepsilon^2}P^{-1}Q\left(P - \frac{\beta^2\overline{h}}{\varepsilon^2}QP^{-1}Q\right)^{-1} \\ -\beta\overline{h}P^{-1}Q\left(P - \frac{\beta^2\overline{h}}{\varepsilon^2}QP^{-1}Q\right)^{-1} & \left(P - \frac{\beta^2\overline{h}}{\varepsilon^2}QP^{-1}Q\right)^{-1} \end{bmatrix}. \tag{26}$$

The, it is clear from Lemma 3.6 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 3.11. □

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 $\widetilde{\alpha}$):

$$\text{(Explicit step)} \quad \frac{v^{n+1/2}_{2(0),i} - v^n_{2(0),i}}{\Delta t} + \nabla_h\left[\frac{v^{2,n}_{2(0),i}}{\overline{h}_i + \varepsilon^2 v^n_{1(0),i}} + \frac{\varepsilon^2}{2}v^{2,n}_{(0)1,i}\right] - \frac{\widehat{\alpha}\Delta x}{2}\Delta_h v^n_{2(0),i} = 0. \tag{27a}$$

$$\text{(Implicit step)} \quad \frac{v^{n+1}_{2(0),i} - v^{n+1/2}_{2(0),i}}{\Delta t} + \nabla_h\left(\overline{h}_i v^{n+1}_{1(0),i}\right) - \frac{\widetilde{\alpha}\Delta x}{2}\Delta_h v^n_{2(0),i} = -v^{n+1}_{1(0),i}\nabla_h b_i. \tag{27b}$$

It is clear that (27a) and (27b) provide consistent discretisations of $\partial_t m_{(0)}$ for both explicit and implicit steps, (15a) and (15b). Thus, in the light of Lemma 3.8 and Lemma 3.11 for the ε-stability and smallness of checkerboard modes, the scheme is AC.

3.2.4. *Asymptotic stability*

In this section, we discuss the rigorous stability analysis of the RS-IMEX scheme in 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

$$\boldsymbol{Y}^k = \prod_{i=0}^{s-1}\mathcal{E}_{s-i}\,\boldsymbol{Y}^{k-1}, \qquad k = 0, 1, \ldots, n-1, \quad n = T_f/\Delta t, \tag{28}$$

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 [32, Lemma 3.1] (see also [53, 62]), one can show that the scheme (28) is stable for a finite time and in ℓ_p-norm, provided that there exist constants c_i independent of Δt such that

$$\|\mathcal{E}_i\|_{\ell_p} \leq 1 + c_i\Delta t, \qquad i = 1, \ldots, s. \tag{29}$$

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 (29) holds. Since the explicit step is nonlinear, obtaining (29) 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 (21) 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 $\boldsymbol{Y}^k = \left[\boldsymbol{V}_{1,\Delta}^k, \boldsymbol{V}_{2,\Delta}^k\right]^T$. Assuming $\widehat{\alpha} = 0$ for simplicity and from (15a), one can write $\|\mathcal{E}_1 \boldsymbol{Y}^k\|_{\ell_2}$ as

$$\begin{aligned}
\|\mathcal{E}_1 \boldsymbol{Y}^k\|_{\ell_2} &\leq \|\boldsymbol{Y}^k\|_{\ell_2} + \frac{2\beta}{h_{\min}^k} \|\langle \boldsymbol{V}_{2,\Delta}^k, \boldsymbol{V}_{2,\Delta}^k \rangle\|_{\ell_2} + \varepsilon^2 \beta \|\langle \boldsymbol{V}_{1,\Delta}^k, \boldsymbol{V}_{1,\Delta}^k \rangle\|_{\ell_2} \\
&\leq \|\boldsymbol{Y}^k\|_{\ell_2} + \beta \left(\frac{2}{h_{\min}^k} + \varepsilon^2 \right) \|\boldsymbol{Y}^k\|_{\ell_4}^2, \\
&\leq \left[1 + \beta \left(\frac{2}{h_{\min}^k} + \varepsilon^2 \right) \|\boldsymbol{Y}^k\|_{\ell_2} \right] \|\boldsymbol{Y}^k\|_{\ell_2}
\end{aligned} \tag{30}$$

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

For simplicity, one can re-write (30) as

$$y_{k+1} \leq y_k + \beta_k y_k^2, \qquad y_k := \|\boldsymbol{Y}^k\|_{\ell_2}, \quad \beta_k := \beta \left(\frac{2}{h_{\min}^k} + \varepsilon^2 \right). \tag{31}$$

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 [64].

Theorem 3.12 (Thm. 4 [64]). *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

$$y_{k+1} \leq \frac{1}{\psi(k)} \left(\sigma^q + q \sum_{i=0}^{k} \mu_k \psi^q(i) \right)^{1/q}, \qquad k = 0, 1, \ldots. \tag{32}$$

Using Theorem 3.12, the following corollary can be obtained.

Corollary 3.13. *Given a small enough initial datum, the sequence $\{y_k\}_{k=0,1,\ldots}$ defined in (31) is bounded uniformly in ε.*

Proof. One can re-write (31) as

$$y_{k+1} \leq y_0 + \sum_{i=0}^{k} \beta_i y_i^2 = \left(y_0 + \beta_0 y_0^2\right) + \sum_{i=1}^{k} \beta_i y_i^2. \tag{33}$$

Comparing (33) to Theorem 3.12, 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.*,

$$(1+\beta_0 y_0)\, y_0 \sum_{i=1}^{k} \beta_i < 1. \tag{34}$$

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

Remark 3.14. *Note that a simpler version of such bounds can be obtained by induction, like [49, eq. (3.11)].*

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

$$y_{k+1} \leq (1 + c_2\Delta t)(1 + c_1 \Delta t y_k) y_k, \tag{35}$$

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 3.12 yields the following stability result for the scheme.

Theorem 3.15. *Given a small enough initial datum and for $\varepsilon \ll 1$, the RS-IMEX scheme (15a)–(15b) is ℓ_2-bounded uniformly in ε and for a finite time.*

Remark 3.16. *One can read the smallness condition (34) 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 [27]; this is in the course of investigation.*

Remark 3.17. *As we have seen so far, the scheme is AC and AS. Due to Definition 1.1, 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 4.2). Although, literally speaking, the scheme is not AP in the sense of Definition 1.1, we call it AP (at least in a weaker sense) since it is AC and AS under a non-restrictive CFL condition.*

3.2.5. *Well-balancing*

To have the LaR equilibrium state at step n 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 $\widetilde{\alpha} = 0$)

$$\boldsymbol{V}_{1,\Delta}^{n+1} + \frac{\beta}{\varepsilon^2}\boldsymbol{V}_{2,\Delta}^{n+1} = c\,\mathbf{1}_N, \tag{36a}$$

$$\beta R_b \boldsymbol{V}_{1,\Delta}^{n+1} + \boldsymbol{V}_{2,\Delta}^{n+1} = \mathbf{0}_N, \tag{36b}$$

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

Putting (36a) into (36b) 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 3.2.1). Also, (36a) concludes that $\boldsymbol{V}_{1,\Delta}^{n+1}$ is constant, and completes the well-balancing proof.

Remark 3.18. *It is important to note that, generally speaking, having the solution at equilibrium does not necessarily imply that the reference solution or its perturbation are at equilibrium. This 1d case with the LaR reference solution is exceptional since the reference solution is constant and stationary.*

4. Shallow water equations with the zero-Froude limit reference solution

Here, we consider the SWE as in (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 (8). It can be obtained from Definition 3.1 and equation (11) 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 stem from its discretisation in the asymptotic analysis (as discussed in [65]). With this reference solution, the splitting can be obtained as

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

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

The splitting is admissible in the sense of [60]; the eigenvalues of $\widetilde{\boldsymbol{F}}'$ are stiff and those of $\widehat{\boldsymbol{F}}'$ are non-stiff:

$$\widetilde{\lambda} = \frac{\overline{m}}{\overline{z} - b} \pm \frac{\sqrt{\overline{z} - b}}{\varepsilon}, \qquad \widehat{\lambda} = 0, 2u_{pert}. \tag{37}$$

The asymptotic analysis presented in Appendix A and Definition 3.2 suggests the scaling matrix should be $D := \mathrm{diag}(\varepsilon^2, \varepsilon)$; so, the scaled RS-IMEX splitting reads

$$\widetilde{\boldsymbol{G}} := \begin{bmatrix} v_2/\varepsilon \\ -\dfrac{\overline{m}^2 v_1 \varepsilon}{(\overline{z}-b)^2} + \dfrac{(\overline{z}-b)v_1}{\varepsilon} + \dfrac{2\overline{m}v_2}{\overline{z}-b} \end{bmatrix}, \quad \widehat{\boldsymbol{G}} := \begin{bmatrix} 0 \\ \dfrac{(\overline{m}+\varepsilon v_2)^2}{\varepsilon\,(\overline{z}+\varepsilon^2 v_1 - b)} + \dfrac{\varepsilon v_1^2}{2} - \dfrac{\overline{m}^2}{\varepsilon\,(\overline{z}-b)} + \dfrac{\overline{m}^2 v_1 \varepsilon}{(\overline{z}-b)^2} - \dfrac{2\overline{m}v_2}{\overline{z}-b} \end{bmatrix}. \tag{38}$$

Due to the well-prepared initial velocity (11), 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.*,

$$\lim_{\varepsilon \to 0} \left[\widetilde{\boldsymbol{G}}', \widehat{\boldsymbol{G}}'\right] = \begin{bmatrix} v_1 & \dfrac{2v_2}{\overline{z}-b} \\ -2v_2 & -v_1 \end{bmatrix}. \tag{39}$$

Similar to the case of the LaR reference solution, the modified equation is stable for this splitting. For this case, the RS-IMEX scheme is defined as in (14a)–(14b) when $\widehat{\boldsymbol{G}}$ and $\widetilde{\boldsymbol{G}}$ change according to (38).

4.1. **Numerical analysis of the scheme**

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

Theorem 4.1. *For the shallow water equations with a flat bottom and well-prepared initial data in the sense of Definition 3.2, the RS-IMEX scheme (14a)–(14b), with (38) 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) its solution is ε-stable, i.e., it is bounded for $\varepsilon \ll 1$. So, there is convergent sub-sequence 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 $\|\boldsymbol{V}_\Delta^n\|_{\ell_2} \leq C_{N,T_f} \|\boldsymbol{V}_\Delta^0\|_{\ell_2}$.

We present the proof of Theorem 4.1 in the next sections.

4.1.1. *Solvability of the scheme*

Like Section 3.2, it is not difficult to see that J_ε reads

$$J_\varepsilon := \begin{bmatrix} P & \dfrac{\beta}{\varepsilon} Q \\ \left(-\dfrac{\overline{m}^2 \varepsilon}{\overline{h}^2} + \dfrac{\overline{h}}{\varepsilon}\right) \beta Q & P + \dfrac{2\beta\overline{m}}{\overline{h}} Q \end{bmatrix}. \tag{40}$$

The blocks of J_ε commute and from [61, 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 [4, 35]. 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.

4.1.2. *ε-stability of the solution*

Similar to Section 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 3.5, 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 4.1.3), 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 4.2. *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" [28, Sect. 3.5.2] in essence, and may improve the condition number of J_ε; see Table 1.*

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

TABLE 1. Comparison of different scaling for matrix J_ε.

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

$$\lim_{\varepsilon\to 0}\left[\frac{(\overline{m} + \varepsilon v_2^n)^2}{\varepsilon\,(\overline{z} + \varepsilon^2 v_1^n - b)} - \frac{\overline{m}^2}{\varepsilon\,(\overline{z} - b)}\right] = \mathcal{O}(1). \tag{41}$$

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 4.3. *The asymptotic stability analysis for the implicit step is very similar to Section 3.2.4. We just wish to stress that for the explicit step, one should use (41) to find an ε-uniform bound. Hence, one can conclude that the scheme is again AP (in a weaker sense than Definition 1.1), 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. *Comparing the results of this section to Section 3.2, both schemes are AC and AS. As we pointed out in Remark 2.2, the modified equation analysis in [68] suggests that the reference solution does not affect stability of the scheme. However, asymptotically smaller wave speeds for the zero-Froude case (compare (37) with (9)) indicates that the choice of reference solution affects the numerical diffusion, so the accuracy. We will illustrate this point in Section 5.1.1 for a numerical example.*

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. 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

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

5.1. Shallow water equations with 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 [16], which has been also discussed in [6]. Then, we discuss another colliding pulses example from [2].

5.1.1. *(i) Colliding pulses*

As [16, Example 6.1], we consider the following well-prepared initial data in the periodic domain $[0, 1)$:

$$\begin{aligned} h(0,x) &= \mathbf{1}_{[0\le x\le 0.2]} + (1+\varepsilon^2)\mathbf{1}_{[0.2<x\le 0.3]} + \mathbf{1}_{[0.3<x\le 0.7]} + (1-\varepsilon^2)\mathbf{1}_{[0.7<x\le 0.8]} + \mathbf{1}_{[0.8<x\le 1]},\\ m(0,x) &= \left(1-\frac{\varepsilon^2}{2}\right)\mathbf{1}_{[0\le x\le 0.2]} + \mathbf{1}_{[0.2<x\le 0.3]} + \left(1+\frac{\varepsilon^2}{2}\right)\mathbf{1}_{[0.3<x\le 0.7]} + \mathbf{1}_{[0.7<x\le 0.8]} + \left(1-\frac{\varepsilon^2}{2}\right)\mathbf{1}_{[0.8<x\le 1]} \end{aligned}$$

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 [16, Example 6.1] the pressure function $p(\varrho) = \varrho^2$ has been used; so, we compare the results of the RS-IMEX scheme with [6, Sect. 8.1], where the pressure function is the same as the SWE.

Figures 2 and 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 [6, Fig. 8.2], it is clear that the quality of computed solutions are fine. Note that for this example, the schemes in [6, Fig. 8.2] uses the same splitting as the RS-IMEX; but, they employ an elliptic approach for the surface perturbation update; see [6] for more details. As Figure 2 and Figure 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 a bit more accurate solution in terms of capturing the extrema. This is due to $\mathcal{O}(\varepsilon^2)$ wave speeds of the non-stiff system (compare (37) with (9)) which leads to smaller

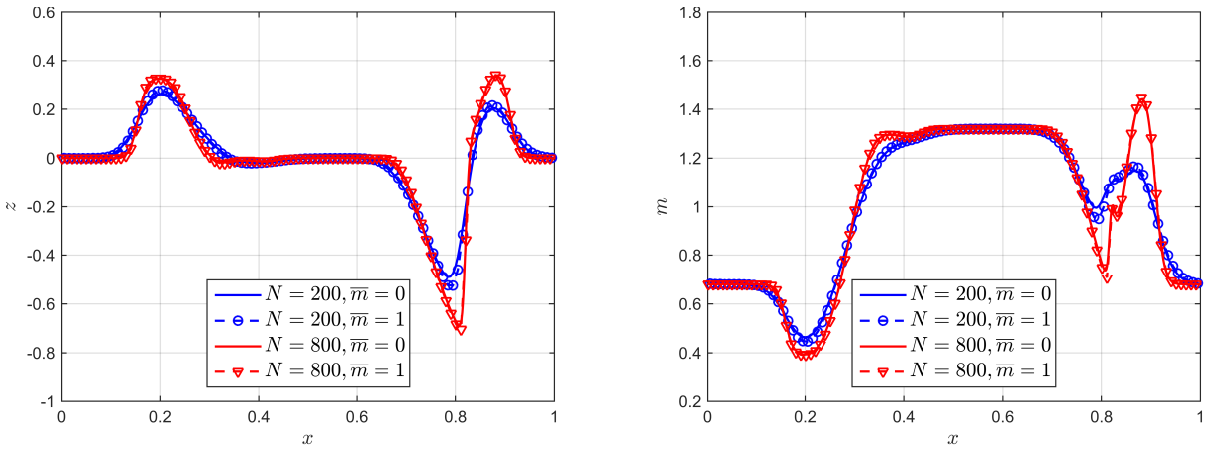
numerical diffusion; this can be clearly seen in Figure 3 where the solution is computed on a very fine mesh with $N = 6400$. Note that for $\varepsilon = 0.1$, both schemes cannot capture the details of the waves (micro-structures), which is also the case in [16, 6].

Figure 4 illustrates the experimental order of convergence (EOC) for different ε and $\overline{m} \in \{0, 1\}$, for a *"normalised"* version of the error $e(\phi_\Delta^{\text{num}})$, defined as

$$e(\phi_\Delta^{\text{num}}) := \|\phi_\Delta^{\text{num}} - \phi_\Delta^{\text{ref}}\|_{L_1(\Omega_{N_{\text{ref}}})} = \frac{1}{N_{\text{ref}}} \sum_{j \in \Omega_{N_{\text{ref}}}} \left|\phi_j^{\text{num}} - \phi_j^{\text{ref}}\right|, \tag{42}$$

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 $N = 3200$. Both for the surface perturbation and the momentum, as suggested to us by Rupert Klein, the error is normalised by $1/\varepsilon^2$ because the initial data consist of $\mathcal{O}(\varepsilon^2)$ fluctuations around zero for the surface perturbation and around a constant value for the momentum. The figures shows that the scheme, regardless of the reference momentum, has an almost uniform order of convergence for $\varepsilon = 0.8, 0.1, 0.05$, which coincides with the result of [16, Tab. 2].

Note that for this example, as well as all other examples in this paper, 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 4.2). Note also that for the zero-Froude limit reference state, due to $\mathcal{O}(\varepsilon)$ eigenvalues for the non-stiff system as in (37), $\Delta t_{\text{CFL}} = \mathcal{O}(1/\varepsilon)$; so, the time step imposed by the advective CFL condition 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 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. Comparing these figures with [6] verifies the accuracy of the scheme.

5.1.2. *(ii) Colliding pulses*

Consider the following ill-prepared initial data in the periodic domain $[-1, 1)$, as in [2] (motivated by an example of [45]):

$$\begin{aligned} h(0, x) &= 0.955 + \frac{\varepsilon}{2}\left(1 - \cos(2\pi x)\right), \\ u(0, x) &= -\text{sign}(x)\sqrt{2}\left(1 - \cos(2\pi x)\right). \end{aligned} \tag{ii$_a$}$$

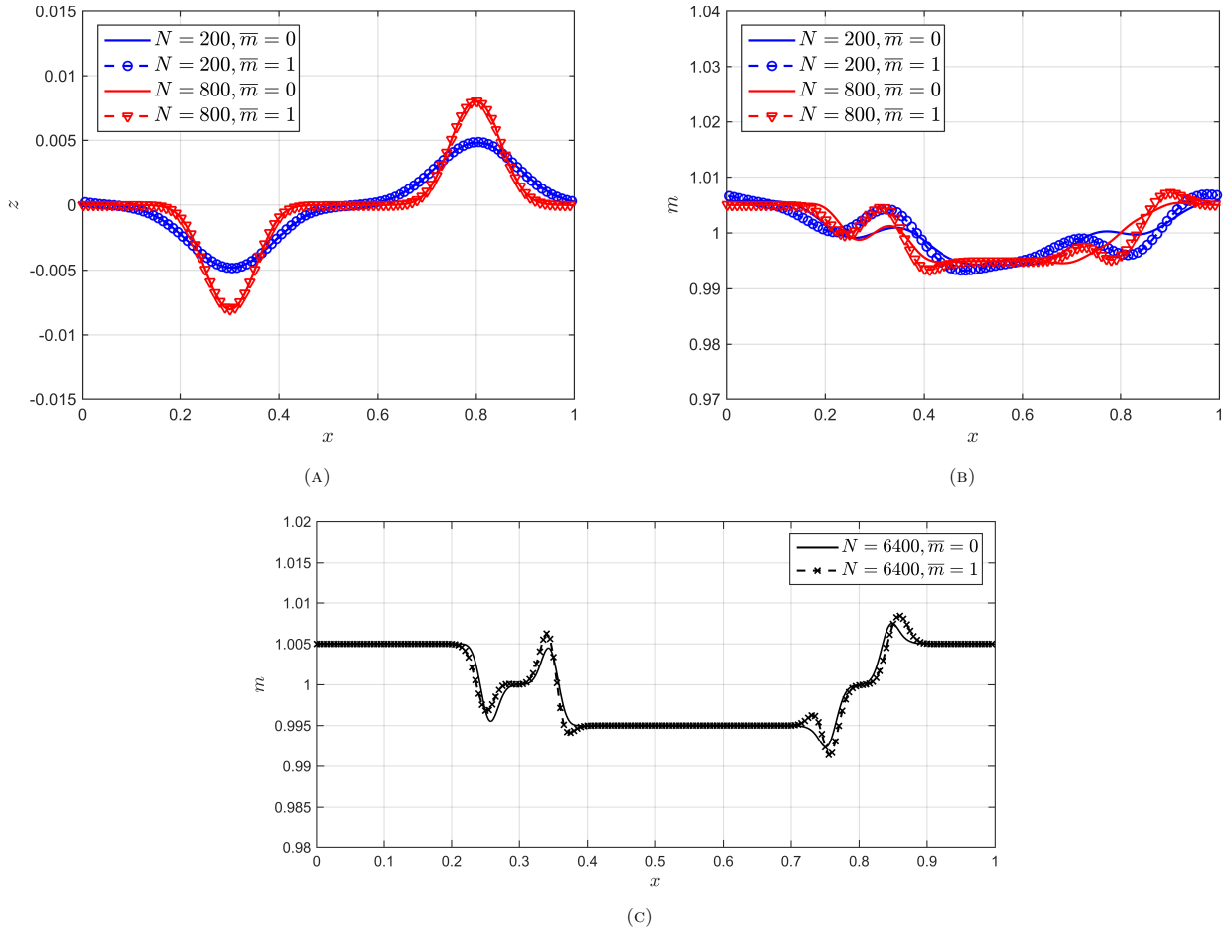


FIGURE 3. (A) and (B): The RS-IMEX solutions for Example (i), with $\varepsilon = 0.8$, CFL $= 0.45$, $T_f = 0.05$, and with two reference states: the LaR and the zero-Froude limit. (C) is like (B) but for a very fine mesh. The solutions are accurate in comparison to [6]. Also, the zero-Froude reference solution seems to provide better results.

Figure 5 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.*, $H_{\text{ref}} = 1$. The figure shows that, comparing to [2], the computed solution is accurate. Note that in [2], the height is computed by an elliptic approach. Moreover, Figure 6 confirms the ε-uniformity of the time step, and stability of the scheme in ℓ_2-norm, with the growth factor $\mathcal{G}_\phi$, defined as $\mathcal{G}_\phi^n := \|\phi_\Delta^n\|_{\ell_2}/\|\phi_\Delta^0\|_{\ell_2}$ for some quantity ϕ. As Figure 6 suggests, the scheme is stable uniformly in ε for variables like z, m and u. Also one can see that, as discussed in Appendix C, the scheme moves the solution toward the well-prepared (limit) manifold. Because the mean value of the momentum is zero, the analysis of Appendix C shows that the scheme makes the momentum $\mathcal{O}(\varepsilon^2)$, which is indicated by a very small $\mathcal{G}_m$ for a 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, we keep $\overline{z} = -0.045$ and change the reference momentum to $\overline{m} = \sqrt{2}$, which is not the zero-Froude limit. We call this setting case (ii$_b$). As Figure 7 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

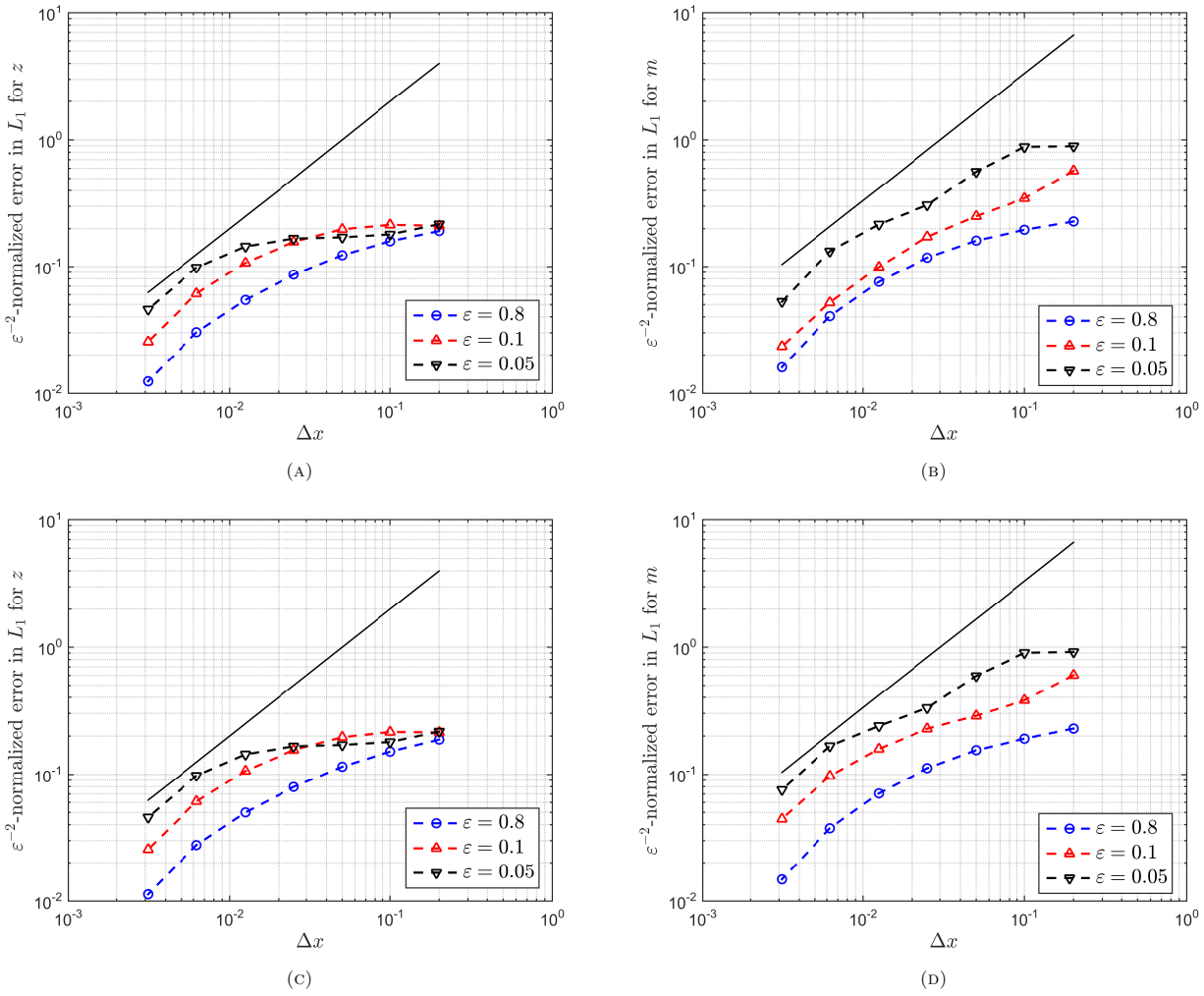


FIGURE 4. The EOC of the RS-IMEX scheme in Example (i), confirming the uniform order of convergence in ε, with CFL $= 0.45$, $T_f = 0.05$: (A) and (B) for the LaR ($\overline{m} = 0$) reference state, (C) and (D) for the zero-Froude limit ($\overline{m} = 1$) reference state. The black solid line is the line with slope one.

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 [20, p. 526]. Figure 7 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.

5.2. **Shallow water equations with 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 function $\eta^b(x) = 0.2\sin(3\pi x)$. We name this setting as Example (iii) hereinafter.

In Figure 8, the convergence rate of the scheme has been plotted, which shows the ε-uniform EOC for the scheme. Moreover, Table 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

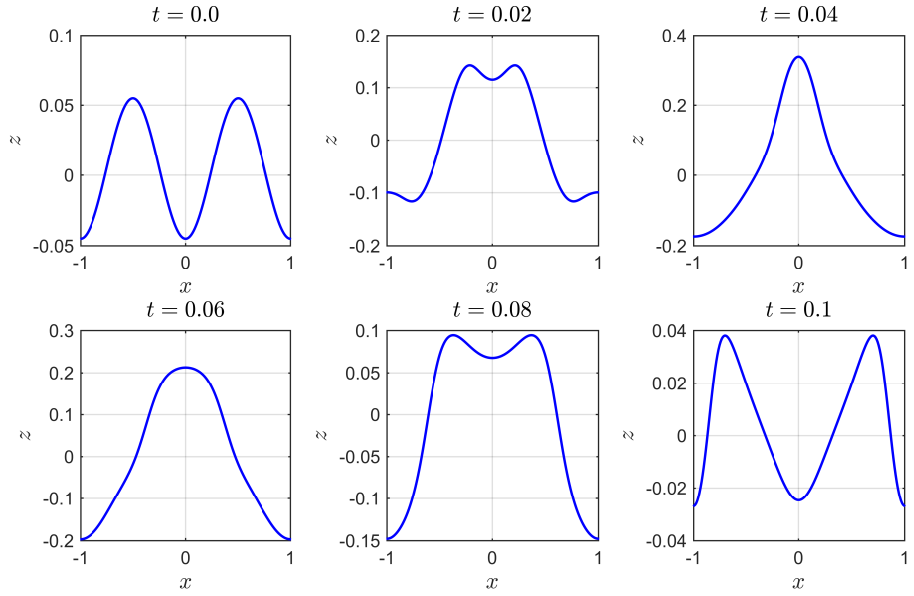
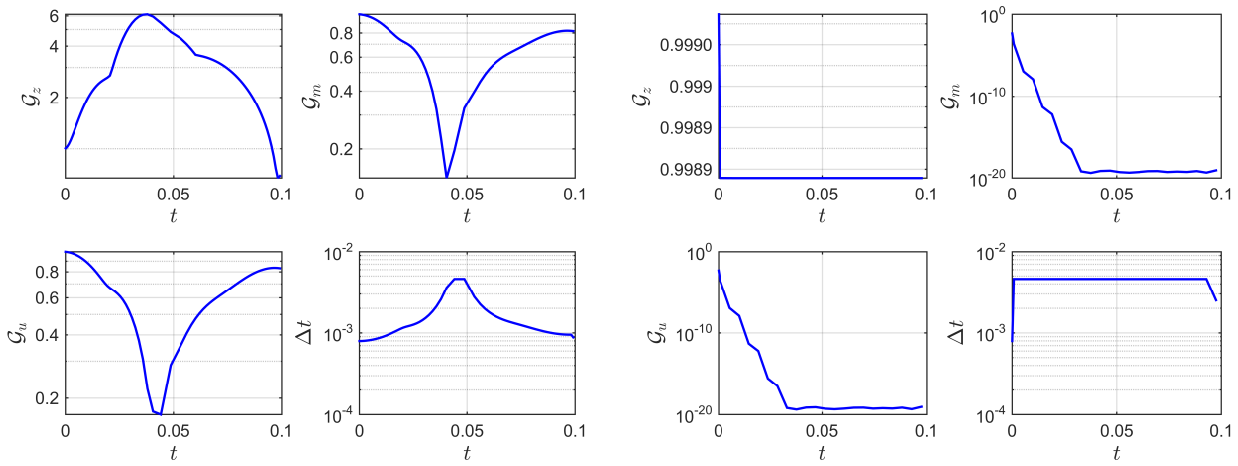


FIGURE 5. Evolution of the surface perturbation for the RS-IMEX solution in Example (ii$_a$), with $\varepsilon = 0.1$, CFL $= 0.45$, $N = 200$, and the LaR reference solution. The figure confirms the accuracy of the computed solution compared to [2].



(A) $\varepsilon = 10^{-1}$. (B) $\varepsilon = 10^{-4}$.

FIGURE 6. Growth factor and time step for different ε, to confirm the asymptotic stability of the scheme, for Example (ii$_a$) with the LaR reference solution.

some threshold in ε where the condition number of J_ε gets very large and affects the solution. 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.

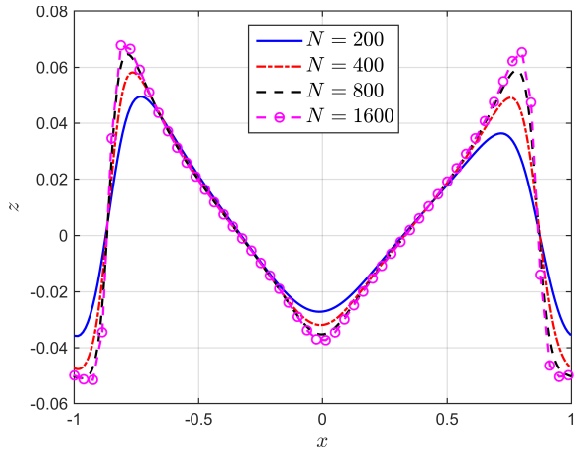


FIGURE 7. Vanishing effect of an unsuitable reference solution for Example (ii$_b$) as $\Delta x \to 0$, for $\varepsilon = 0.1$, $T_f = 0.1$ and $N = 200, 400, 800, 1600$.

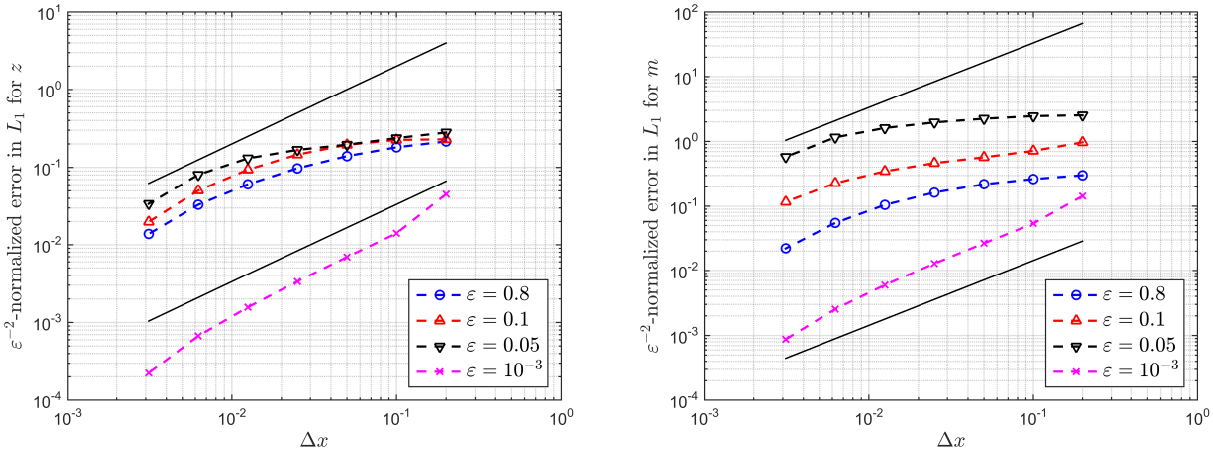


FIGURE 8. EOC of the RS-IMEX scheme for Example (iii), confirming the uniform order of convergence in ε, with $T_f = 0.05$, CFL $= 0.45$ and the LaR reference solution. The black solid line is the line with slope one.

TABLE 2. Smallness of the checker-board oscillations regarding ε and Δx for Example (iii).

ε	N	$\|\|[\![\boldsymbol{V}_{2,\Delta}^{n+1}]\!]\|\|_{\ell_\infty}$	$\mathrm{cond}_2(J_\varepsilon)$	ε	N	$\|\|[\![\boldsymbol{V}_{2,\Delta}^{n+1}]\!]\|\|_{\ell_\infty}$	$\mathrm{cond}_2(J_\varepsilon)$
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				

6. Concluding remarks

In this paper, we have analysed the RS-IMEX scheme for the shallow water equations w.r.t. the Froude number. The scheme has been presented in one space dimension and its quality is guaranteed by numerical analysis as well as several numerical tests. In practice, we have shown that the scheme is uniformly stable and consistent, when the analysis confirms the asymptotic preserving property for the scheme, as well as preservation of the lake at rest equilibrium state. Indeed, the asymptotic consistency and stability analyses are not only formal but also rigorous. These results are so far for two reference solutions, the lake at rest and the zero-Froude limit, and limited to one space dimension and first-order schemes on periodic domains. Of course, it is of a great interest to conduct a similar study for the two-dimensional case, for which a great challenge would be the more complicated structure of the matrix J_ε, which makes the rigorous asymptotic consistency analysis much more involved; we refer the reader to [6, 65] for more details.

Acknowledgement

The author would like to gratefully thank Negin Bagherpour and Mohammad Zakerzadeh for very useful discussions regarding some proofs in Section 3.2 and Appendix B.

A. Formal asymptotic analysis of 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 [11] for a more general formal analysis). Consider the usual formulation of the non-dimensionalised SWE with $\eta^b = H_{\text{ref}} + b$ and $h = z - b$ (compare it with system (8)):

$$\begin{aligned} &\partial_t h + \partial_x m = 0, \\ &\partial_t m + \partial_x \left(\frac{m^2}{h} + \frac{h^2}{2\varepsilon^2} \right) = -\frac{h \partial_x \eta^b}{\varepsilon^2}. \end{aligned} \tag{43}$$

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

$$h(t,x) = h_{(0)}(t,x) + \varepsilon h_{(1)}(t,x) + \varepsilon^2 h_{(2)}(t,x), \qquad m(t,x) = m_{(0)}(t,x) + \varepsilon m_{(1)}(t,x) + \varepsilon^2 m_{(2)}(t,x), \tag{44}$$

in (43), 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 3.1 as the formal asymptotic limit of the SWE.

B. On the proof of Lemma 3.8

In this section, we complete the proof of Lemma 3.8, 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 B.1 below.

For the non-flat bottom case, matrix R_b is defined as in (18) and $\beta^2 R_b Q \boldsymbol{w}_2^{(2)} = \boldsymbol{w}_2^{(0)}$ gives the following linear system of equations:

$$\beta^2 \begin{bmatrix} \overline{h}_N - \overline{h}_2 & \overline{h}_2 & & & & -\overline{h}_N \\ -\overline{h}_1 & \overline{h}_1 - \overline{h}_3 & \overline{h}_3 & & & \\ & -\overline{h}_2 & \overline{h}_2 - \overline{h}_4 & \overline{h}_4 & & \\ & & \ddots & \ddots & \ddots & \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_{2,2}^{(2)} - \boldsymbol{w}_{2,N}^{(2)} \\ \boldsymbol{w}_{2,3}^{(2)} - \boldsymbol{w}_{2,1}^{(2)} \\ \boldsymbol{w}_{2,4}^{(2)} - \boldsymbol{w}_{2,2}^{(2)} \\ \vdots \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_{2,odd}^{(0)} \\ \boldsymbol{w}_{2,even}^{(0)} \\ \boldsymbol{w}_{2,odd}^{(0)} \\ \vdots \end{bmatrix}. \tag{45}$$

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 B.1. *Consider the linear system of equations $M\boldsymbol{y} = \boldsymbol{c}$ of size N, in accordance with equation (45), and with the positive sequence $\{m_k\}_k > 0$, as:*

$$\begin{bmatrix} m_N - m_2 & m_2 & & & & -m_N \\ -m_1 & m_1 - m_3 & m_3 & & & \\ & -m_2 & m_2 - m_4 & m_4 & & \\ & & \ddots & \ddots & \ddots & \\ m_1 & & & -m_{N-1} & m_{N-1} - m_1 & \end{bmatrix} \begin{bmatrix} y_2 - y_N \\ y_3 - y_1 \\ y_4 - y_2 \\ \vdots \\ y_1 - y_{N-1} \end{bmatrix} = \begin{bmatrix} c_o \\ c_e \\ c_o \\ \vdots \end{bmatrix}. \tag{46}$$

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 (45) 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

$$\widetilde{M} := \begin{bmatrix} m_2 & & & & m_N \\ m_1 & m_3 & & & \\ & \ddots & \ddots & & \\ & & m_{N-1} & m_1 & \end{bmatrix}. \tag{47}$$

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 a 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 (46) 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 showed, $\widetilde{M}$ is non-singular and the system $\widetilde{M}\boldsymbol{\theta} = \boldsymbol{c}$ has the solution $\boldsymbol{\theta}^*$ such that for $k \in \Omega_N$

$$\begin{cases} m_2\theta_1^* + m_N\theta_N^* = c \\ m_1\theta_1^* + m_3\theta_2^* = c \\ \vdots \\ m_{N-1}\theta_{N-1}^* + m_1\theta_N^* = c \end{cases} \quad \Longrightarrow \quad \theta_{k+1}^* + \frac{m_k}{m_{k+2}}\theta_k^* = \frac{c}{m_{k+2}}. \tag{48}$$

Then, making a spatial sum on Ω_N for (48) 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 re-writing (48) as

$$\begin{cases} \vartheta_N\theta_N^* + \theta_1^* = \tilde{c}_N \\ \vartheta_1\theta_1^* + \theta_2^* = \tilde{c}_1 \\ \vdots \\ \vartheta_{N-1}\theta_{N-1}^* + \theta_N^* = \tilde{c}_{N-1} \end{cases} \Longrightarrow \begin{cases} \theta_2^* = \tilde{c}_1 - \vartheta_1\theta_1^* \\ \theta_3^* = -\tilde{c}_2 - \vartheta_2\theta_2^* = \tilde{c}_2 - \vartheta_2\tilde{c}_1 + \vartheta_1\vartheta_2\theta_1^* \\ \theta_4^* = \tilde{c}_3 - \vartheta_3\theta_3^* = \tilde{c}_3 - \vartheta_3\theta_2^* + \vartheta_2\vartheta_3\tilde{c}_1 - \vartheta_1\vartheta_2\vartheta_3\theta_1^* \\ \vdots \\ \theta_N^* = (-1)^{N-1}\Big[\prod_{j=1}^{N-1}\vartheta_j\theta_1^* - \sum_{j=1}^{N-1}(-1)^j\tilde{c}_j\prod_{\ell=j+1}^{N-1}\vartheta_\ell\Big] \end{cases}$$

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

$$\theta_1^* = \frac{\tilde{c}_N}{2} - \frac{\vartheta_N}{2}\sum_{j=1}^{N-1}(-1)^j\tilde{c}_j\prod_{\ell=j+1}^{N-1}\vartheta_\ell. \tag{49}$$

In equation (49), 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 $\mathcal{O}(\Delta x)$. 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

$$\sum_{k\in\Omega_N}\frac{m_k}{m_{k+2}}\theta_k^*=\sum_{(k=2j)\in\Omega_N}\frac{c_e}{m_{k+2}}+\sum_{(k=2j+1)\in\Omega_N}\frac{c_o}{m_{k+2}},\tag{50}$$

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,

$$\begin{cases}\vartheta_N\theta_N^*+\theta_1^*=-\tilde{c}_N\\ \vartheta_1\theta_1^*+\theta_2^*=\tilde{c}_1\\ \vdots\\ \vartheta_{N-1}\theta_{N-1}^*+\theta_N^*=\tilde{c}_{N-1}\end{cases}\quad\Longrightarrow\quad\begin{cases}\theta_2^*=\tilde{c}_1-\vartheta_1\theta_1^*\\ \theta_3^*=-\tilde{c}_2-\vartheta_2\theta_2^*=-\tilde{c}_2-\vartheta_2\tilde{c}_1+\vartheta_1\vartheta_2\theta_1^*\\ \theta_4^*=\tilde{c}_3-\vartheta_3\theta_3^*=\tilde{c}_3+\vartheta_3\theta_2^*+\vartheta_2\vartheta_3\tilde{c}_1-\vartheta_1\vartheta_2\vartheta_3\theta_1^*\\ \vdots\\ \theta_N^*=(-1)^{N-1}\Big[\prod_{j=1}^{N-1}\vartheta_j\theta_1^*-\sum_{j=1}^{N-1}\tilde{c}_j\prod_{\ell=j+1}^{N-1}\vartheta_\ell\Big]\end{cases}$$

One gets two different relations for θ_N^*, which should be the same:

$$\theta_N^*=-\vartheta_N^{-1}\theta_1^*-\vartheta_N^{-1}\tilde{c}_N,\qquad\qquad\theta_N^*=-\prod_{j=1}^{N-1}\vartheta_j\theta_1^*+\sum_{j=1}^{N-1}\tilde{c}_j\prod_{\ell=j+1}^{N-1}\vartheta_\ell.$$

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 B.1 implies the system (45) or (46) 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 (22a) 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.

C. Asymptotic consistency of the RS-IMEX scheme with the ill-prepared initial data

Regarding AP schemes for hyperbolic balance laws, the focus is often limited to the well-prepared initial data (Definition 3.2). Here, we briefly show that the rigorous asymptotic consistency analysis can also be done for the ill-prepared initial data (*cf.* [23, Sect. 4.6]), *i.e.*,

$$\begin{aligned}z_{0,\varepsilon}&=z_{(0)}^0+\varepsilon z_{(1),\varepsilon}^0,\\ m_{0,\varepsilon}&=m_{(0)}^0+\varepsilon m_{(1),\varepsilon}^0,\end{aligned}\tag{51}$$

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 (9) that the splitting is still admissible in the sense of [60]. 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 (51); 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 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 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 [12] 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 (51).

References

1. Saul Abarbanel, Pravir Duth, and David Gottlieb, *Splitting methods for low Mach number Euler and Navier–Stokes equations*, Computers & Fluids **17** (1989), no. 1, 1–12.
2. Koottungal Revi Arun and Sebastian Noelle, *An asymptotic preserving scheme for low Froude number shallow flows*, IGPM report 352, RWTH Aachen University, 2012.
3. Uri M. Ascher, Steven J. Ruuth, and Raymond J. Spiteri, *Implicit-explicit Runge–Kutta methods for time-dependent partial differential equations*, Applied Numerical Mathematics **25** (1997), no. 2-3, 151–167.
4. Ivar Bendixson, *Sur les racines d'une équation fondamentale*, Acta Mathematica **25** (1902), no. 1, 359–365.
5. Dennis S. Bernstein, *Matrix mathematics: Theory, facts, and formulas*, Princeton University Press, 2009.
6. Georgij Bispen, *IMEX finite volume methods for the shallow water equations*, Ph.D. thesis, Johannes Gutenberg-Universität Mainz, 2015.
7. Georgij Bispen, Koottungal Revi Arun, Mária Lukáčová-Medvid'ová, and Sebastian Noelle, *IMEX large time step finite volume methods for low Froude number shallow water flows*, Communications in Computational Physics **16** (2014), 307–347.
8. Sebastiano Boscarino and Giovanni Russo, *On a class of uniformly accurate IMEX Runge–Kutta schemes and applications to hyperbolic systems with relaxation*, SIAM Journal on Scientific Computing **31** (2009), no. 3, 1926–1945.
9. François Bouchut and Michael Westdickenberg, *Gravity driven shallow water models for arbitrary topography*, Communications in Mathematical Sciences **2** (2004), no. 3, 359–389.
10. François Bouchut, *Nonlinear stability of finite volume methods for hyperbolic conservation laws: And well-balanced schemes for sources*, Springer Science & Business Media, 2004.
11. Didier Bresch, Rupert Klein, and Carine Lucas, *Multiscale analyses for the shallow water equations*, Computational Science and High Performance Computing IV, Springer, 2011, pp. 149–164.
12. Russel E. Caflisch, Shi Jin, and Giovanni Russo, *Uniformly accurate schemes for hyperbolic systems with relaxation*, SIAM Journal on Numerical Analysis **34** (1997), no. 1, 246–281.
13. Floraine Cordier, Pierre Degond, and Anela Kumbaro, *An asymptotic-preserving all-speed scheme for the Euler and Navier–Stokes equations*, Journal of Computational Physics **231** (2012), no. 17, 5685–5704.
14. Raphaël Danchin, *Low Mach number limit for viscous compressible flows*, ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique **39** (2005), no. 3, 459–475.
15. Pierre Degond, Jian-Guo Liu, and Marie-Hélène Vignal, *Analysis of an asymptotic preserving scheme for the Euler–Poisson system in the quasineutral limit*, SIAM Journal on Numerical Analysis **46** (2008), no. 3, 1298–1322.
16. Pierre Degond and Min Tang, *All speed scheme for the low Mach number limit of the isentropic Euler equation*, Communications in Computational Physics **10** (2011), no. 1, 1–31.
17. Stéphane Dellacherie, *Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number*, Journal of Computational Physics **229** (2010), no. 4, 978–1016.
18. Stéphane Dellacherie, Pascal Omnes, and Felix Rieper, *The influence of cell geometry on the Godunov scheme applied to the linear wave equation*, Journal of Computational Physics **229** (2010), no. 14, 5315–5338.
19. Giacomo Dimarco, Raphaël Loubère, and Marie-Hélène Vignal, *Study of a new asymptotic preserving scheme for the Euler system in the low Mach number limit*, SIAM Journal on Scientific Computing **39** (2017), no. 5, A2099–A2128.
20. Dimitris Drikakis and William Rider, *High-resolution methods for incompressible and low-speed flows*, Springer Science & Business Media, 2006.
21. Baptiste Fedele and Claudia Negulescu, *Numerical study of an anisotropic Vlasov equation arising in plasma physics*, arXiv preprint arXiv:1610.01592 (2016).

22. Eduard Feireisl, Mária Lukáčová-Medvid'ová, Šárka Nečasová, Antonín Novotnỳ, and Bangwei She, *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.
23. Eduard Feireisl and Antonín Novotnỳ, *Singular limits in thermodynamics of viscous fluids*, Springer Science & Business Media, 2009.
24. Miloslav Feistauer, Vít Dolejší, and Václav Kučera, *On the discontinuous Galerkin method for the simulation of compressible flow with wide range of Mach numbers*, Computing and Visualization in Science **10** (2007), no. 1, 17–27.
25. Francis Filbet and Shi Jin, *A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources*, Journal of Computational Physics **229** (2010), no. 20, 7625–7648.
26. Thierry Gallouët, Raphaèle Herbin, David Maltese, and Antonín Novotnỳ, *Implicit MAC scheme for compressible Navier–Stokes equations: Low Mach asymptotic error estimates*, hal-01462822 (2017).
27. Jan Giesselmann, *Low Mach asymptotic-preserving scheme for the Euler–Korteweg model*, IMA Journal of Numerical Analysis **35** (2015), no. 2, 802–833.
28. Gene H. Golub and Charles F. Van Loan, *Matrix computations*, vol. 3, JHU Press, 2012.
29. Robert M. Gray, *Toeplitz and circulant matrices: A review*, Now Publishers Inc., 2006.
30. Hervé Guillard and Angelo Murrone, *On the behavior of upwind schemes in the low Mach number limit: II. Godunov type schemes*, Computers & Fluids **33** (2004), no. 4, 655–675.
31. Hervé Guillard and Cécile Viozat, *On the behaviour of upwind schemes in the low Mach number limit*, Computers & Fluids **28** (1999), no. 1, 63–86.
32. Jeffrey Haack, Shi Jin, and Jian-Guo Liu, *An all-speed asymptotic-preserving method for the isentropic Euler and Navier–Stokes equations*, Communications in Computational Physics **12** (2012), no. 4, 955–980.
33. Ernst Hairer and Gerhard Wanner, *Solving ordinary differential equations. II*, second ed., Springer Series in Computational Mathematics, vol. 14, Springer-Verlag, Berlin, 1996, Stiff and differential-algebraic problems. MR 1439506
34. Andreas Hiltebrand and Siddhartha Mishra, *Efficient computation of all speed flows using an entropy stable shock-capturing space-time discontinuous Galerkin method*, Seminar for Applied Mathematics, ETH Zürich, vol. 17, 2014, pp. 1–21.
35. M. A. Hirsch, *Sur les racines d'une équation fondamentale*, Acta Mathematica **25** (1902), no. 1, 367–370.
36. Roger A. Hom and Charles R. Johnson, *Topics in matrix analysis*, Cambridge UP, New York (1991).
37. Roger A. Horn and Charles R. Johnson, *Matrix analysis*, Cambridge University Press, New York, NY, USA, 1986.
38. Jingwei Hu, Shi Jin, and Qin Li, *Asymptotic-preserving schemes for multiscale hyperbolic and kinetic equations*, Handbook of Numerical Analysis (2016).
39. Shi Jin, *Runge–Kutta methods for hyperbolic conservation laws with stiff relaxation terms*, Journal of Computational Physics **122** (1995), no. 1, 51–67.
40. ———, *Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations*, SIAM Journal on Scientific Computing **21** (1999), no. 2, 441–454.
41. ———, *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) (2010), 177–216.
42. Klaus Kaiser, Jochen Schütz, Ruth Schöbel, and Sebastian Noelle, *A new stable splitting for the isentropic Euler equations*, Journal of Scientific Computing (2016), 1–18.
43. Sergiu Klainerman and Andrew Majda, *Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids*, Communications on Pure and Applied Mathematics **34** (1981), no. 4, 481–524.
44. ———, *Compressible and incompressible fluids*, Communications on Pure and Applied Mathematics **35** (1982), no. 5, 629–651.
45. Rupert Klein, *Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics I: One-dimensional flow*, Journal of Computational Physics **121** (1995), no. 2, 213–237.
46. Rupert Klein, Nicola Botta, Thomas Schneider, Claus-Dieter Munz, Sabine Roller, Andreas Meister, L. Hoffmann, and Thomas Sonar, *Asymptotic adaptive methods for multi-scale problems in fluid mechanics*, Practical Asymptotics, Springer, 2001, pp. 261–343.
47. Edward W. Larsen, J. E. Morel, and Warren F. Miller, *Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes*, Journal of Computational Physics **69** (1987), no. 2, 283–324.
48. Nader Masmoudi, *Examples of singular limits in hydrodynamics*, Handbook of Differential Equations: Evolutionary Equations **3** (2007), 195–275.
49. Haim Nessyahu and Eitan Tadmor, *The convergence rate of approximate solutions for nonlinear scalar conservation laws*, SIAM Journal on Numerical Analysis **29** (1992), no. 6, 1505–1519.
50. Sebastian Noelle, Georgij Bispen, Koottungal Revi Arun, Mária Lukáčová-Medvid'ová, and Claus-Dieter Munz, *A weakly asymptotic preserving low Mach number scheme for the Euler equations of gas dynamics*, SIAM Journal on Scientific Computing **36** (2014), no. 6, B989–B1024.
51. Lorenzo Pareschi and Giovanni Russo, *Implicit-explicit Runge–Kutta schemes and applications to hyperbolic systems with relaxation*, Journal of Scientific Computing **25** (2005), no. 1-2, 129–155.
52. Marco Restelli, *Semi-Lagrangian and semi-implicit discontinuous Galerkin methods for atmospheric modeling applications*, Ph.D. thesis, 2007.

53. Robert D. Richtmyer and Keith W. Morton, *Difference methods for initial-value problems*, Interscience Publishers John Wiley & Sons, Inc., Academia Publishing House of the Czechoslovak Acad, 1967.
54. Felix Rieper, *On the dissipation mechanism of upwind-schemes in the low Mach number regime: A comparison between Roe and HLL*, Journal of Computational Physics **229** (2010), no. 2, 221–232.
55. ———, *A low-Mach number fix for Roe's approximate Riemann solver*, Journal of Computational Physics **230** (2011), no. 13, 5263–5287.
56. Felix Rieper and Georg Bader, *The influence of cell geometry on the accuracy of upwind schemes in the low Mach number regime*, Journal of Computational Physics **228** (2009), no. 8, 2918–2933.
57. Howard H. Rosenbrock, *Some general implicit processes for the numerical solution of differential equations*, The Computer Journal **5** (1963), no. 4, 329–330.
58. Steven Schochet, *The mathematical theory of low Mach number flows*, ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique **39** (2005), no. 3, 441–458.
59. Jochen Schütz and Klaus Kaiser, *A new stable splitting for singularly perturbed ODEs*, Applied Numerical Mathematics **107** (2016), 18–33.
60. Jochen Schütz and Sebastian Noelle, *Flux splitting for stiff equations: A notion on stability*, Journal of Scientific Computing (2014), 1–19.
61. John R. Silvester, *Determinants of block matrices*, The Mathematical Gazette (2000), 460–467.
62. Lloyd N. Trefethen, *Finite difference and spectral methods for ordinary and partial differential equations*, Cornell University, 1996.
63. R. F. Warming and B. J. Hyett, *The modified equation approach to the stability and accuracy analysis of finite-difference methods*, Journal of Computational Physics **14** (1974), no. 2, 159–179.
64. D. Willett and J. S. W. Wong, *On the discrete analogues of some generalizations of Gronwall's inequality*, Monatshefte für Mathematik **69** (1965), no. 4, 362–367.
65. Hamed Zakerzadeh, *Asymptotic consistency of the RS-IMEX scheme for the low-Froude shallow water equations: Analysis and numerics*, Proceedings of XVI International Conference on Hyperbolic Problems, Aachen, 2016.
66. ———, *On the Mach-uniformity of the Lagrange-projection scheme*, ESAIM: Mathematical Modelling and Numerical Analysis **51** (2017), no. 4, 1343–1366.
67. ———, *The RS-IMEX scheme for the rotating shallow water equations with the Coriolis force*, pp. 199–207, Springer International Publishing, Cham, 2017.
68. Hamed Zakerzadeh and Sebastian Noelle, *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.
69. Mohammad Zakerzadeh and Georg May, *On the convergence of a shock capturing discontinuous Galerkin method for nonlinear hyperbolic systems of conservation laws*, SIAM Journal of Numerical Analysis **54** (2016), no. 2, 874–898.