

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

HAMED ZAKERZADEH<sup>\*,#</sup>

**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  $\varepsilon$  tending to zero. We prove the asymptotic consistency, asymptotic stability, solvability and well-balancing of the scheme for the case of the one-dimensional shallow water equations when the singular parameter is the Froude number. We will also study several test cases to illustrate the quality of the computed solutions and to confirm the analysis.

**Mathematics Subject Classification.** 35L65, 65M08, 35L81, 65M12.

Received March 16, 2017. Accepted January 26, 2019.

### 1. INTRODUCTION

Singular limits of conservation laws (or more generally PDEs), characterised by the singular parameter  $\varepsilon$  approaching zero (usually scaled such that  $\varepsilon \in (0, 1]$ ), 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 [50]. 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 [15, 44, 45, 50, 60] for 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, so that the Courant–Friedrichs–Lewy (CFL) condition restricts the timestep varies non-uniformly with  $\varepsilon$ , *i.e.*,  $\Delta t \lesssim \varepsilon \Delta x$ . This leads to very small timesteps, thus a huge computational cost. Generally speaking, the usual numerical schemes also lose their accuracy in the limit for under-resolved grids; see [18, 19, 31, 32, 56–58].

In the sequel, we mainly consider well-prepared initial data, which are consistent with the limit  $\varepsilon \rightarrow 0$ , in order to simplify the problem and to eliminate spurious initial layers (see Def. 3.1 and [50]). We also assume

---

*Keywords and phrases.* IMEX scheme, asymptotic preserving, flux splitting, stability analysis.

Institut für Geometrie und Praktische Mathematik, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany.

<sup>#</sup>Current address: Université Toulouse III – Paul Sabatier, UMR CNRS 5219, Institut de Mathématiques de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex, France.

<sup>\*</sup>Corresponding author: [seyed\\_hamed.zakerzadeh@math.univ-toulouse.fr](mailto:seyed_hamed.zakerzadeh@math.univ-toulouse.fr); [h.zakerzadeh@igpm.rwth-aachen.de](mailto:h.zakerzadeh@igpm.rwth-aachen.de)

that the *solution* of the PDE with the singular parameter  $\varepsilon$  converges to the *solution* of the limit PDE as  $\varepsilon \rightarrow 0$ , 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 have been introduced by Jin [41] for relaxation systems; see also [42] for a general review and [40, 46, 49] for older works (without being named AP). Clarifying the definition, assume that  $\mathcal{M}^\varepsilon$  stands for a continuous physical model with the (singular) perturbation parameter  $\varepsilon$ , and  $\mathcal{M}_\Delta^\varepsilon$  is a discrete-level model which provides a consistent discretisation of  $\mathcal{M}^\varepsilon$ . If  $\mathcal{M}_\Delta^\varepsilon$  is a *suitable* and *efficient* scheme for  $\mathcal{M}^\varepsilon$  uniformly in  $\varepsilon$ , 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:

- (i) Asymptotic Consistency (AC):  $\mathcal{M}_\Delta^\varepsilon$  gives a consistent discretisation of  $\mathcal{M}^\varepsilon$  for all  $\varepsilon > 0$ , in particular, for the limit problem  $\mathcal{M}^0$ .
- (ii) Asymptotic Efficiency (AEf):  $\mathcal{M}_\Delta^\varepsilon$  is efficient uniformly in  $\varepsilon$ , *e.g.*, the CFL condition is  $\varepsilon$ -uniform and the implicit step can be solved efficiently for all  $\varepsilon > 0$ .
- (iii) Asymptotic Stability (AS):  $\mathcal{M}_\Delta^\varepsilon$  is stable in some suitable sense, uniformly in  $\varepsilon$ .

**Remark 1.2.** (i) It is also prevalent in the literature to define the asymptotic stability as the stability of the limit scheme  $\mathcal{M}_\Delta^0$ , *cf.* [28, 42]. Also, sometimes, the uniformity of the CFL condition is classified as the asymptotic stability rather than asymptotic efficiency.

- (ii) As mentioned in [42], the asymptotic consistency suggests that the solution belongs to a manifold approaching the limit manifold as  $\varepsilon \rightarrow 0$  (up to some discretisation error).
- (iii) AEf implies the  $\varepsilon$ -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 [22] 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 [39, 42] for a review) while its application for the conservation laws is more recent; just to name a few see [7, 14, 16, 20, 33, 52]. There are also several related works without using the terminology AP; see [47] as one of the first examples for the Euler equations and [25, 48]. 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 [20, 68] for the isentropic Euler equation and [8] for the Euler system with gravity; see also [17, 28] 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, 9, 13, 53] 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  $\varepsilon$ -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, despite the fact that 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 [33]. One can also think of fully-implicit schemes like finite volume schemes [27], mixed finite element-finite volume schemes [24], and space-time dG schemes [35] (see also [72] 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 [52], the authors applied a flux-splitting scheme to the full Euler equations, which uses a variant of Klein’s auxiliary splitting [47]. The scheme required an  $\varepsilon$ -dependent timestep for stability. Motivated by this, the authors

in [62] 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 [71] 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.* [71]); see Remark 2.2 below for a discussion on this. Note that, apart from the RS-IMEX scheme for the shallow water equations [67, 69, 70], the RS-IMEX scheme is shown to be quantitatively well-behaved in practice; see [43, 61] 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 Rem. 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 [67, 69, 70] 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 [26] (as the so-called *penalisation method* [39]) 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 Sections 3 and 4. We will also question the asymptotic efficiency of the scheme by showing that the condition number of the linear system, which should be solved in the implicit step, is not  $\varepsilon$ -uniform and increases as  $\varepsilon \rightarrow 0$ ; see Remark 4.2. However, this issue could be avoided using pre-conditioning, like [7]. 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 [67, 69, 70].

## 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  $\varepsilon$  (*e.g.*, the Froude or Mach number):

$$\partial_t \mathbf{U}^\varepsilon + \operatorname{div}_{\mathbf{x}} \mathbf{F}^\varepsilon = \mathbf{S}^\varepsilon, \quad (2.1)$$

where  $\mathbf{U} \in \mathbb{R}^q$  is the vector of unknowns,  $\mathbf{F} = \mathbf{F}(\mathbf{U}) \in \mathbb{R}^{q \times d}$  is the flux (in  $d$  space dimensions), *i.e.*,  $\mathbf{F} = [\mathbf{f}_1, \dots, \mathbf{f}_d]$  with  $\mathbf{f}_k \in \mathbb{R}^q$ , and  $\mathbf{S} \in \mathbb{R}^q$  is the source term, *e.g.*, due to the gravitational force, Coriolis force, or bottom friction. Note that we often suppress the dependence of  $\mathbf{U}$ ,  $\mathbf{F}$  and  $\mathbf{S}$  on  $\varepsilon$ .  $\Omega$  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  $\mathbf{F}$  has a real diagonalisable Jacobian  $\mathbf{F}' := \partial_{\mathbf{U}} \mathbf{F}$ , *i.e.*, for all directions  $\mathbf{n} = (n_1, \dots, n_d)^T$  in  $\mathbb{R}^d$ , the matrix  $\mathbf{A}_{\mathbf{n}} := \sum_{i=1}^d \partial_{\mathbf{U}} \mathbf{f}_i n_i$  has  $q$  real eigenvalues  $\lambda_1 \geq \dots \geq \lambda_q$  with linearly-independent eigenvectors.

Let us assume that system (2.1) is stiff due to the singular parameter, namely as  $\varepsilon$  approaches zero, the source term blows up and the characteristics of the flux Jacobian diverge. We aim to decompose this system into stiff (fast) and non-stiff (slow) parts, for which there is no unique way due to non-linearity. Here, we find such a fast-slow decomposition by considering the given  $\varepsilon$ -independent function  $\bar{\mathbf{U}}$  as the *reference solution*<sup>1</sup>.

<sup>1</sup>Typically,  $\bar{\mathbf{U}}$  is a steady state solution of the balance law, or the solution of the asymptotic limit equation, derived from (2.1) as  $\varepsilon \rightarrow 0$ , *e.g.*, it can be the lake at rest state for the SWE or the incompressible limit for the Euler equations.

We split the solution  $\mathbf{U}$  of the balance law (2.1) into the reference solution  $\bar{\mathbf{U}}$  and a perturbation  $\mathbf{U}_{\text{pert}}$ , *i.e.*,  $\mathbf{U} = \bar{\mathbf{U}} + \mathbf{U}_{\text{pert}}$ . Then, we split the flux and source terms using the Taylor expansion (linearisation) around  $\bar{\mathbf{U}}$ , to the reference, stiff and non-stiff parts respectively:

$$\mathbf{F}(\mathbf{U}) = \mathbf{F}(\bar{\mathbf{U}}) + \mathbf{F}'(\bar{\mathbf{U}})\mathbf{U}_{\text{pert}} + (\mathbf{F}(\mathbf{U}) - \mathbf{F}(\bar{\mathbf{U}}) - \mathbf{F}'(\bar{\mathbf{U}})\mathbf{U}_{\text{pert}}) =: \bar{\mathbf{F}} + \tilde{\mathbf{F}} + \hat{\mathbf{F}}, \quad (2.2a)$$

$$\mathbf{S}(\mathbf{U}) = \mathbf{S}(\bar{\mathbf{U}}) + \mathbf{S}'(\bar{\mathbf{U}})\mathbf{U}_{\text{pert}} + (\mathbf{S}(\mathbf{U}) - \mathbf{S}(\bar{\mathbf{U}}) - \mathbf{S}'(\bar{\mathbf{U}})\mathbf{U}_{\text{pert}}) =: \bar{\mathbf{S}} + \tilde{\mathbf{S}} + \hat{\mathbf{S}}. \quad (2.2b)$$

Since  $\bar{\mathbf{U}}$  is given, the real unknown is  $\mathbf{U}_{\text{pert}}$  in regard of which the stiff part of the splitting  $(\tilde{\mathbf{F}}, \tilde{\mathbf{S}})$  is linear. This would be truly advantageous for the numerical treatment of this split form as, unlike the splittings in [16, 33], there is no need for solving non-linear systems by the Newton iteration method. The idea of such a linearisation goes back to [59] (see also [34], Chap. IV.7) for ODEs (the so-called linearly-implicit methods) and has been used later in [7] for the SWE, motivated by Restelli [54], 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 generalised linearisation state.

It is 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)<sup>2</sup>. So, we introduce the diagonal matrix  $D := \text{diag}(\varepsilon^{d_1}, \dots, \varepsilon^{d_q})$  with some integer  $d_j$ , and define the *scaled* (preconditioned) perturbation  $\mathbf{V}$  as  $\mathbf{V} := D^{-1}\mathbf{U}_{\text{pert}}$ . We also denote the corresponding *scaled* flux and source terms by  $\mathbf{G} := D^{-1}\mathbf{F}$  and  $\mathbf{Z} := D^{-1}\mathbf{S}$  such that the splittings (2.2a) and (2.2b) can be re-written as

$$\mathbf{G} = \bar{\mathbf{G}} + \tilde{\mathbf{G}} + \hat{\mathbf{G}}, \quad \mathbf{Z} = \bar{\mathbf{Z}} + \tilde{\mathbf{Z}} + \hat{\mathbf{Z}}.$$

We refer the reader to Sections 3 and 4 for two explicit examples of such a scaling (and fast-slow decomposition) for the shallow water system.

**Remark 2.1.** Note that the eigenvalues of  $\tilde{\mathbf{F}}' := \partial_{\mathbf{U}_{\text{pert}}}\tilde{\mathbf{F}}$  and  $\hat{\mathbf{F}}' := \partial_{\mathbf{U}_{\text{pert}}}\hat{\mathbf{F}}$  are exactly the same as the eigenvalues of  $\tilde{\mathbf{G}}' := \partial_{\mathbf{V}}\tilde{\mathbf{G}}$  and  $\hat{\mathbf{G}}' := \partial_{\mathbf{V}}\hat{\mathbf{G}}$ , respectively. This should be clear because these matrices can be transformed into each other by a similarity transformation with  $D$ . So, the scaling does not change the eigenvalues of the flux Jacobians.

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

$$\partial_t \mathbf{V} + \text{div}_{\mathbf{x}} (\tilde{\mathbf{G}}(\bar{\mathbf{U}}, \mathbf{V}) + \hat{\mathbf{G}}(\bar{\mathbf{U}}, \mathbf{V})) = \tilde{\mathbf{Z}}(\bar{\mathbf{U}}, \mathbf{V}) + \hat{\mathbf{Z}}(\bar{\mathbf{U}}, \mathbf{V}) - \bar{\mathbf{T}}(\bar{\mathbf{U}}), \quad (2.3)$$

with stiff and non-stiff scaled fluxes  $(\tilde{\mathbf{G}}, \hat{\mathbf{G}})$  and scaled source terms  $(\tilde{\mathbf{Z}}, \hat{\mathbf{Z}})$ , and the contributions of the reference solution collected in  $\bar{\mathbf{T}}(\bar{\mathbf{U}})$ , which is the (*a priori*-known) *residual of the reference solution* and reads

$$\bar{\mathbf{T}}(\bar{\mathbf{U}}) := D^{-1}\partial_t \bar{\mathbf{U}} + \text{div}_{\mathbf{x}} \bar{\mathbf{G}}(\bar{\mathbf{U}}) - \bar{\mathbf{Z}}(\bar{\mathbf{U}}). \quad (2.4)$$

Note that  $\bar{\mathbf{T}} \equiv 0$  if and only if the reference solution  $\bar{\mathbf{U}}$  satisfies the original system (2.1).

For brevity of notations, we collect the slow and fast fluxes together and define  $\mathbf{R} := -\text{div}_{\mathbf{x}} \mathbf{G} + \mathbf{Z}$  (with analogous definitions for  $\bar{\mathbf{R}}$ ,  $\tilde{\mathbf{R}}$  and  $\hat{\mathbf{R}}$ ) so that to rewrite (2.3) as

$$\partial_t \mathbf{V} = -\bar{\mathbf{T}} + \tilde{\mathbf{R}} + \hat{\mathbf{R}}. \quad (2.5)$$

**Remark 2.2.** One can confirm that the non-stiff Jacobian is  $\hat{\mathbf{G}}' = \mathbf{G}'(\mathbf{U}) - \mathbf{G}'(\bar{\mathbf{U}})$  while the stiff one reads  $\tilde{\mathbf{G}}' = \mathbf{G}'(\bar{\mathbf{U}})$ . So, the commutator (of the fast and slow flux Jacobian matrices) can be obtained as

$$[\hat{\mathbf{G}}', \tilde{\mathbf{G}}'] := \hat{\mathbf{G}}'\tilde{\mathbf{G}}' - \tilde{\mathbf{G}}'\hat{\mathbf{G}}' = [\mathbf{G}'(\mathbf{U}), \mathbf{G}'(\bar{\mathbf{U}})].$$

<sup>2</sup>Later on, we will see that an appropriate choice of the scaling matrix, not only makes the analysis more illustrative (see Rem. 3.9) but also may affect the numerical solution (see Rem. 4.2).

By choosing the reference solution as the asymptotic limit,  $\mathbf{U} - \bar{\mathbf{U}}$  is small for  $\varepsilon \ll 1$ , which suggests that the commutator is small and leads to the stability of modified equations [71]<sup>3</sup>. Moreover, we have shown in [71] that if the eigenspaces have similar structures, the modified equation is likely to be stable. This is, in fact, the case for the RS-IMEX splitting as the eigenvectors of  $\mathbf{G}'(\mathbf{U})$  and  $\mathbf{G}'(\bar{\mathbf{U}})$  are close to each other in the structure (regardless of  $\|\mathbf{U} - \bar{\mathbf{U}}\|$ ), 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 Rem. 4.4 below).

## 2.1. Numerical scheme

The Jacobian  $\tilde{\mathbf{F}}'$  in (2.2a) (and  $\tilde{\mathbf{G}}'$ , due to Rem. 2.1) has stiff eigenvalues. So, to solve (2.5) numerically, we treat the stiff part  $\tilde{\mathbf{R}}$  implicitly in time to avoid restrictive timesteps, by using the implicit Euler time integration. The term  $\tilde{\mathbf{R}}$  is *expected* to be non-stiff<sup>4</sup>; so, the explicit Euler scheme is employed. Note that in the sequel, we limit ourselves to first-order schemes. To compute the residual  $\bar{\mathbf{T}}$ , we need an independent solver for the reference solution, *e.g.*, an incompressible Euler solver when the reference solution is the solution of the incompressible Euler system. In general, computing  $\bar{\mathbf{U}}$  is cheap and an appropriate method should be chosen for each case separately. Now, we can define the RS-IMEX scheme as follows.

**Definition 2.3.** Given the reference solution  $\bar{\mathbf{U}}$ , the fully-discrete RS-IMEX scheme for (2.5) is given by

$$D_t \mathbf{V}_\Delta^n = -\bar{\mathbf{T}}_\Delta^{n+1} + \tilde{\mathbf{R}}_\Delta^{n+1} + \hat{\mathbf{R}}_\Delta^n, \quad (2.6)$$

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

The numerical flux is a Rusanov-type numerical flux, which in 1d and for the scalar case can be written as

$$f_{i+1/2} := \frac{f(u_i) + f(u_{i+1})}{2} - \frac{\alpha_{i+1/2}}{2} (u_{i+1} - u_i),$$

for the flux  $f(u)$  at the interface  $x_{i+1/2}$ . The extension of this numerical flux to systems or in multi-dimensions is obvious. The numerical viscosity  $\alpha$  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 quite diffusive, we choose  $\alpha$  for the stiff sub-system rather arbitrary not to add too much diffusion; that is to say that  $\tilde{\alpha}, \hat{\alpha} = \mathcal{O}(1)$ . 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  $\mathbf{V}_\Delta^{n+1}$ , while the reference state  $\bar{\mathbf{U}}_\Delta^{n+1}$  may evolve over time and should be computed independently. At the end of each step, the solution can be computed as  $\bar{\mathbf{U}}_\Delta^{n+1} + D\mathbf{V}_\Delta^{n+1}$ . The RS-IMEX procedure has been summarised in Algorithm 1.

**Remark 2.4.** The RS-IMEX scheme introduced here is a bit different from [43, 61] mainly in two aspects: firstly, in those series of papers, the reference solution is computed implicitly as the limit of the singularly-perturbed system while here we employ off-the-shelf methods for this purpose. Moreover, here we use the reformulation (2.5) with the scaled perturbation, which makes the analysis easier and more illustrative as we will see later on.

<sup>3</sup>The modified equation is the “actual” equation solved by a numerical approximation when applied to a differential equation [65]; see also [62, 71].

<sup>4</sup>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  $\bar{\mathbf{U}}_\Delta^n$  and  $\mathbf{V}_\Delta^n$ .
  - 2: **Reference step:** Find the updated reference state  $\bar{\mathbf{U}}_\Delta^{n+1}$ .
  - 3: **Explicit step:** Solve  $D_t \mathbf{V}_\Delta^n = \hat{\mathbf{R}}_\Delta^n$  to find  $\mathbf{V}_\Delta^{n+1/2}$ .
  - 4: **Implicit step:** Solve  $D_t \mathbf{V}_\Delta^{n+1/2} = -\bar{\mathbf{T}}_\Delta^{n+1} + \tilde{\mathbf{R}}_\Delta^{n+1}$  to find the updated perturbation  $\mathbf{V}_\Delta^{n+1}$ .
  - 5: Find the updated solution as  $\mathbf{U}_\Delta^{n+1} = \bar{\mathbf{U}}_\Delta^{n+1} + D \mathbf{V}_\Delta^{n+1}$ .
  - 6: Continue with step 2.
- 

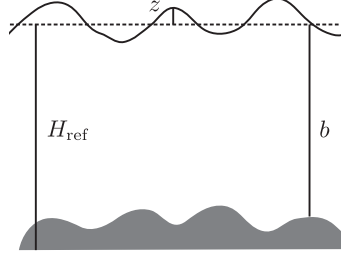


FIGURE 1. Variables used in the shallow water formulation (3.1). The water height is written as  $h = z - b$ .

### 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 (2.1), 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.

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

$$\mathbf{U} = \begin{bmatrix} z \\ m \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} m \\ \frac{m^2}{z-b} + \frac{z^2 - 2zb}{2\varepsilon^2} \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 0 \\ -\frac{z\partial_x b}{\varepsilon^2} \end{bmatrix}. \quad (3.1)$$

In this notation,  $z$  is the surface elevation from some chosen constant surface level  $H_{\text{ref}}$ ,  $m$  is the momentum and  $b$  is the water depth measured from  $H_{\text{ref}}$  with a negative sign (see Fig. 1). The singular parameter  $\varepsilon$  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  $\partial_x b$  is assumed to be small enough; see [11] for details.

We set the reference state as the LaR,  $\bar{\mathbf{U}} := (\bar{z}, \bar{m})^T$  with  $\bar{z}$  constant in space and  $\bar{m} = 0$ . Therefore, due to (2.2a) and (2.2b), the splitting reads

$$\begin{aligned} \bar{\mathbf{F}} &= \begin{bmatrix} 0 \\ \frac{1}{2\varepsilon^2} \bar{z}(\bar{z} - 2b) \end{bmatrix}, & \tilde{\mathbf{F}} &= \begin{bmatrix} m_{\text{pert}} \\ \frac{(\bar{z} - b)}{\varepsilon^2} z_{\text{pert}} \end{bmatrix}, & \hat{\mathbf{F}} &= \begin{bmatrix} 0 \\ \frac{m_{\text{pert}}^2}{\bar{z} + z_{\text{pert}} - b} + \frac{z_{\text{pert}}^2}{2\varepsilon^2} \end{bmatrix}, \\ \bar{\mathbf{S}} &= \begin{bmatrix} 0 \\ -\frac{\bar{z}b_x}{\varepsilon^2} \end{bmatrix}, & \tilde{\mathbf{S}} &= \begin{bmatrix} 0 \\ -\frac{z_{\text{pert}}\partial_x b}{\varepsilon^2} \end{bmatrix}, & \hat{\mathbf{S}} &= \mathbf{0}. \end{aligned}$$

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

$$\begin{aligned}\tilde{\mathbf{F}}' &= \begin{bmatrix} 0 & 1 \\ \frac{\bar{z}-b}{\varepsilon^2} & 0 \end{bmatrix}, & \tilde{\lambda} &= \pm \frac{\sqrt{\bar{z}-b}}{\varepsilon}, \\ \hat{\mathbf{F}}' &= \begin{bmatrix} 0 & 0 \\ -u_{\text{pert}}^2 + \frac{z_{\text{pert}}}{\varepsilon^2} & 2u_{\text{pert}} \end{bmatrix}, & \hat{\lambda} &= 0, 2u_{\text{pert}},\end{aligned}\tag{3.2}$$

with  $u_{\text{pert}} := m_{\text{pert}}/(\bar{z} + z_{\text{pert}} - b)$ . Thus, the splitting is admissible in the sense of [62], *i.e.*, both split systems are hyperbolic and the eigenvalues of the slow Jacobian are  $\mathcal{O}(1)$ . Note that the case  $\bar{\mathbf{U}} = \mathbf{0}$  gives the same splitting as in [6, 7].

To determine the scaling matrix  $D$ , we employ the formal asymptotic analysis in Appendix A (see also [44, 45] for the rigorous justification for the flat bottom case), which yields the formal zero-Froude limit in Definition 3.1 below, with the following asymptotic (Poincaré) expansion

$$\begin{aligned}z(t, x) &= z_{(0)} + \varepsilon z_{(1)} + \varepsilon^2 z_{(2)}, \\ m(t, x) &= m_{(0)} + \varepsilon m_{(1)} + \varepsilon^2 m_{(2)}.\end{aligned}\tag{3.3}$$

**Definition 3.1.** The formal zero-Froude limit of the shallow water system (3.1) gives the so-called “lake equations”, with  $p_{(2)}$  as the “incompressible pressure” or the second-order contribution of the pressure function, and writes

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

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

**Definition 3.2.** For the 1d SWE (3.1), we call the initial data  $(z_{0,\varepsilon}, m_{0,\varepsilon})$  well-prepared if it holds that

$$\begin{aligned}z(0, \cdot) &= z_{0,\varepsilon} = z_{(0)}^0 + \varepsilon^2 z_{(2),\varepsilon}^0, \\ m(0, \cdot) &= m_{0,\varepsilon} = m_{(0)}^0 + \varepsilon m_{(1),\varepsilon}^0,\end{aligned}\tag{3.4}$$

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 (3.4), we pick  $\bar{z} = z_{(0)}^0$ , which implies  $D := \text{diag}(\varepsilon^2, 1)$ . For simplicity, we stick to this particular choice of  $\bar{z}$  throughout this section. Nonetheless, it is rather straightforward to confirm that the asymptotic analysis we are going to present holds for every constant  $\bar{z}$ , while the choice may affect the numerical diffusion of the scheme, thus the solution.

### 3.1. RS-IMEX scheme

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

$$\tilde{\mathbf{G}} = \begin{bmatrix} v_2/\varepsilon^2 \\ (\bar{z}-b)v_1 \end{bmatrix}, \quad \hat{\mathbf{G}} = \begin{bmatrix} 0 \\ \frac{v_2^2}{\bar{z} + \varepsilon^2 v_1 - b} + \frac{\varepsilon^2}{2} v_1^2 \end{bmatrix},\tag{3.5a}$$

$$\tilde{\mathbf{Z}} = \begin{bmatrix} 0 \\ -v_1 \partial_x b \end{bmatrix}, \quad \hat{\mathbf{Z}} = \mathbf{0}.\tag{3.5b}$$

Owing to (3.2) 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,  $\bar{z}$  is constant in time and the reference solution needs not to be updated. Thus,  $\bar{\mathbf{T}} \equiv 0$ , and one can reformulate the 1d SWE as

$$\partial_t \mathbf{V} = -\partial_x \left[ \frac{v_2/\varepsilon^2}{(\bar{z}-b)v_1} \right] - \partial_x \left[ \frac{0}{\bar{z} + \varepsilon^2 v_1 - b} + \varepsilon^2 \frac{v_1^2}{2} \right] + \begin{bmatrix} 0 \\ -v_1 \partial_x b \end{bmatrix}. \quad (3.6)$$

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

$$\mathbf{V}_i^{n+1/2} = \mathbf{V}_i^n - \frac{\Delta t}{\Delta x} \left( \hat{\mathbf{G}}_{i+1/2}^n - \hat{\mathbf{G}}_{i-1/2}^n \right) \quad (\text{Explicit step}), \quad (3.7a)$$

$$\mathbf{V}_i^{n+1} = \mathbf{V}_i^{n+1/2} - \frac{\Delta t}{\Delta x} \left( \tilde{\mathbf{G}}_{i+1/2}^{n+1} - \tilde{\mathbf{G}}_{i-1/2}^{n+1} \right) + \Delta t \tilde{\mathbf{Z}}_i^{n+1} \quad (\text{Implicit step}), \quad (3.7b)$$

for each cell  $i \in \{1, 2, \dots, N\}$  in the computational domain  $\Omega_N$  with  $N$  cells, where  $\tilde{\mathbf{G}}_{i+1/2}^n$  and  $\hat{\mathbf{G}}_{i+1/2}^n$  denote the Rusanov flux (as defined in Sect. 2) with  $\hat{\mathbf{G}}$  and  $\tilde{\mathbf{G}}$  as defined in (3.5a), and  $\tilde{\mathbf{Z}}_i^n$  is the central discretisation of the fast source term in (3.5b). Denoting  $\nabla_h$  and  $\Delta_h$  respectively as the central discretisation of the first and second derivatives, one can re-write (3.7a) and (3.7b) as

$$\mathbf{V}_i^{n+1/2} = \mathbf{V}_i^n - \Delta t \nabla_h \left[ \frac{0}{\bar{z} + \varepsilon^2 v_{1,i}^n - b_i} + \frac{\varepsilon^2}{2} v_{1,i}^{n,2} \right] + \frac{\hat{\alpha} \Delta x}{2} \Delta t \Delta_h \mathbf{V}_i^n, \quad (3.8a)$$

$$\mathbf{V}_i^{n+1} = \mathbf{V}_i^{n+1/2} - \Delta t \nabla_h \left[ \frac{v_{2,i}^{n+1}/\varepsilon^2}{(\bar{z} - b_i) v_{1,i}^{n+1}} \right] + \frac{\tilde{\alpha} \Delta x}{2} \Delta t \Delta_h \mathbf{V}_i^{n+1} - \Delta t \left[ \frac{0}{v_{1,i}^{n+1} \nabla_h b_i} \right]. \quad (3.8b)$$

Due to Remark 2.1, the eigenvalues of  $\mathbf{F}'$  and  $\mathbf{G}'$  (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 which was claimed to be needed for the stability of the modified equation, cf. [71]. Indeed, the commutator is formally  $\mathcal{O}(1/\varepsilon^2)$ :

$$[\tilde{\mathbf{G}}', \hat{\mathbf{G}}'] := \tilde{\mathbf{G}}' \hat{\mathbf{G}}' - \hat{\mathbf{G}}' \tilde{\mathbf{G}}' = \begin{bmatrix} v_1 - \frac{v_2^2}{(\bar{z} + \varepsilon^2 v_1 - b)^2} & \frac{2v_2/\varepsilon^2}{\bar{z} + \varepsilon^2 v_1 - b} \\ \frac{-2(\bar{z} - b)v_2}{\bar{z} + \varepsilon^2 v_1 - b} & -v_1 + \frac{v_2^2}{(\bar{z} + \varepsilon^2 v_1 - b)^2} \end{bmatrix}. \quad (3.9)$$

However, as shown in [71], 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 (3.8a) and (3.8b), with (3.5a) and (3.5b), a constant  $\tilde{\alpha}$ , and under an  $\varepsilon$ -uniform timestep restriction*

- (i) *is solvable, i.e., it has a unique solution for all  $\varepsilon > 0$ .*
- (ii) *has an  $\varepsilon$ -stable solution, i.e., it is bounded for  $\varepsilon \ll 1$ . So, there is convergent sub-sequence of the discrete solutions as  $\varepsilon \rightarrow 0$ .*
- (iii) *is consistent with the asymptotic limit in the fully-discrete settings, i.e., it is asymptotically consistent.*

- (iv) *is asymptotically  $\ell_2$ -stable for the fixed grid  $\Delta x$ , 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  $\|\mathbf{V}_\Delta^n\|_{\ell_2} \leq C_{N,T_f} \|\mathbf{V}_\Delta^0\|_{\ell_2}$ .*
- (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 [6, 7] 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 results of [6] are, basically, for the flat bottom case with a detailed analysis 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  $\bar{h} := \bar{z} - b$  and  $\beta := \frac{\Delta t}{2\Delta x}$ .

- (i) Constant  $b$ . Owing to (3.8b), we write the implicit step as  $J_\varepsilon \mathbf{V}_\Delta^{n+1} = \mathbf{V}_\Delta^{n+1/2}$ , i.e., the implicit solution operator is  $J_\varepsilon^{-1}$ . The matrix  $J_\varepsilon \in \mathbb{R}^{2N \times 2N}$  writes

$$J_\varepsilon := \begin{bmatrix} P & \frac{\beta}{\varepsilon^2} Q \\ \beta \bar{h} Q & P \end{bmatrix}, \quad (3.10)$$

where  $P$  and  $Q$  are circulant matrices defined as

$$P := \mathbf{Circ}(1 + 2\tilde{\alpha}\beta, -\tilde{\alpha}\beta, 0, \dots, 0, -\tilde{\alpha}\beta), \quad Q := \mathbf{Circ}(0, 1, 0, \dots, 0, -1).$$

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 [30], and one knows from Theorem 1 of [63] (see also [5], Sect. 2.14) that since all blocks of  $J_\varepsilon$  commute with each other, the determinant of  $J_\varepsilon$  can be computed as

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

Due to Gerschgorin's circle theorem ([37], Chap. 6), the numerical range ([38], Chap. 1) of  $-\frac{\bar{h}\beta^2}{\varepsilon^2} Q^2$  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. [38], Chap. 1), the eigenvalues of the sum cannot be zero. Thus,  $J_\varepsilon$  is non-singular, and there exists a unique solution for the scheme.

- (ii) Non-constant  $b$ . For this case, one of the blocks of  $J_\varepsilon$  is not circulant and the matrix  $J_\varepsilon$  is written as

$$J_\varepsilon = \begin{bmatrix} P & \frac{\beta}{\varepsilon^2} Q \\ \beta R_b & P \end{bmatrix}, \quad (3.11)$$

where  $R_b$  is an *almost* circular matrix such that its  $i$ -th row is  $(R_b)_i = (b_{i+1} - b_{i-1}, \bar{h}_{i+1}, 0, \dots, 0, -\bar{h}_{i-1})$ , 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  $\Lambda_M$  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 [30]). Using this fact, one can consider the transformed matrix  $\Xi_\varepsilon$ , instead of  $J_\varepsilon$  itself, for showing solvability where

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

From Fact 2.14.13 of [5] and since the blocks  $\Xi_{11}$  and  $\Xi_{12}$  are commutable, the determinant of  $\Xi_\varepsilon$  can be written as

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

Matrix  $P$  is SDD and invertible ([37], Thm. 6.1.10); thus,  $\Lambda_P$  does not have a zero on the diagonal. So, as the matrix  $\Lambda_Q F_N^* R_b F_N$  does depend neither on  $\varepsilon$  nor on  $\beta$ , a suitable choice for  $\beta$  makes  $\Xi_\varepsilon$  invertible and concludes that  $J_\varepsilon$  in (3.11) is invertible as well.

### 3.2.2. $\varepsilon$ -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  $\varepsilon$ . We call such a property  $\varepsilon$ -stability hereinafter. A similar idea has been used in [6], in the context of the Finite Volume Evolution Galerkin (FVEG) scheme [7], and in [68] for the Lagrange–projection scheme. Note that the  $\varepsilon$ -stability of the implicit operator does not provide  $\varepsilon$ -stability of the solution *per se*. For that, one also needs the  $\varepsilon$ -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  $\Xi_\varepsilon$  can be obtained as

$$\Xi_\varepsilon := \begin{bmatrix} \Lambda_P & \frac{\beta}{\varepsilon^2} \Lambda_Q \\ \beta \bar{h} \Lambda_Q & \Lambda_P \end{bmatrix}. \quad (3.12)$$

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

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

for a suitable natural matrix norm. This bound depends on  $\varepsilon$  only through  $\|\Xi_\varepsilon^{-1}\|$ ; so, we have to show that  $\Xi_\varepsilon^{-1}$  is uniformly bounded in  $\varepsilon$ . 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.*, Proposition 2.8.7 of [5].

**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*

$$M^{-1} = \begin{bmatrix} (M_{11} - M_{12} M_{22}^{-1} M_{21})^{-1} & -M_{11}^{-1} M_{12} (M_{22} - M_{21} M_{11}^{-1} M_{12})^{-1} \\ -M_{22}^{-1} M_{21} (M_{11} - M_{12} M_{22}^{-1} M_{21})^{-1} & (M_{22} - M_{21} M_{11}^{-1} M_{12})^{-1} \end{bmatrix} \quad (3.13)$$

*if all the inverses exist.*

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

**Lemma 3.6.** *The inverse of matrix  $\Xi_\varepsilon$  in (3.12) has a bounded norm for  $\varepsilon \rightarrow 0$ , for a suitable and non-restrictive choice of  $\Delta t$ .*

*Proof.* From Lemma 3.5, the inverse of  $\Xi_\varepsilon$  reads

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

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

**Remark 3.7.** Lemma 3.6 concludes that the implicit solution operator  $J_\varepsilon^{-1}$  is bounded in terms of  $\varepsilon$ . The immediate result of this  $\varepsilon$ -stability is that the scaled perturbation  $\mathbf{V}_\Delta$  should be  $\mathcal{O}(1)$  as long as the explicit step is  $\varepsilon$ -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  $J_\varepsilon^{-1}$  (unlike the matrix in (3.12)) 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  $\tilde{\alpha} = 0$  for simplicity)

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

As  $R_b$  is close to  $Q$ , it is plausible to guess that the block  $(\mathbb{I}_N - \frac{\beta^2}{\varepsilon^2} Q R_b)^{-1}$  is a constant matrix with some  $\mathcal{O}(\varepsilon^2)$  fluctuations (see [68] 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  $\sigma$  denoting the singular values, and show that the smallest singular value of  $J_\varepsilon$  does not approach zero in the limit. From Section 3.2.1,  $J_\varepsilon$  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  $J_\varepsilon^{-1}$ .

**Lemma 3.8.** *For  $J_\varepsilon$  as in (3.11), there exists a constant  $C$  independent of  $\varepsilon$ , such that  $\lim_{\varepsilon \rightarrow 0} \|J_\varepsilon^{-1}\| \leq C$ .*

*Proof.* Here, we consider  $\tilde{\alpha} = 0$  to simplify the analysis; however, the analysis for  $\tilde{\alpha} \neq 0$  can be done similarly. Using (3.11), 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  $\mathbf{w} := (\mathbf{w}_1, \mathbf{w}_2)^T \in \mathbb{C}^{2N}$  living on the unit sphere, i.e.,  $\|\mathbf{w}\|_{\ell_2} = 1$ , where both  $\mathbf{w}_1$  and  $\mathbf{w}_2$  are vectors of size  $N$  with complex entries. Then, by the definition of numerical range ([38], Chap. 1), one can write the numerical range of  $J_\varepsilon^* J_\varepsilon$  as

$$W(J_\varepsilon^* J_\varepsilon) = \|\beta R_b \mathbf{w}_1 + \mathbf{w}_2\|_{\ell_2}^2 + \left\| \frac{\beta}{\varepsilon^2} Q \mathbf{w}_2 + \mathbf{w}_1 \right\|_{\ell_2}^2. \quad (3.14)$$

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

$$\mathbf{w}_1 = -\frac{\beta}{\varepsilon^2}Q\mathbf{w}_2 + o(1), \quad (3.15a)$$

$$\mathbf{w}_2 = -\beta R_b \mathbf{w}_1 + o(1). \quad (3.15b)$$

Multiplying (3.15a) by  $\varepsilon^2$  yields  $\beta Q\mathbf{w}_2 = o(\varepsilon^2) - \varepsilon^2 \mathbf{w}_1$ . For  $\varepsilon \rightarrow 0$ , both terms in the right-hand side have a limit (note that  $\mathbf{w}_2$  is a bounded function); so, the limit of  $\mathbf{w}_2$  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  $\mathbf{w}_2 = \mathbf{w}_2^{(0)} + \varepsilon^2 \mathbf{w}_2^{(2)}$  where  $\mathbf{w}_2^{(0)}$  consists, in general, of two constants for odd and even indices.

Now, putting  $\mathbf{w}_1$  from (3.15a) into (3.15b) yields

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

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

The equation (3.15b) implies that  $R_b \mathbf{w}_1 \rightarrow \mathbf{0}$ . Since the kernel of  $R_b$  consists of vectors with a checker-board like (or CB) structure (see Appendix B),  $\mathbf{w}_1$  should tend to a *constant*. But from (3.15a),  $\mathbf{w}_1$  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 no checker-board like structure and  $\mathbf{w}_1 \rightarrow \mathbf{0}$ . Hence,  $(\mathbf{w}_1, \mathbf{w}_2)$  is tending to zero, which contradicts the assumption that  $\mathbf{w}$  lives on the unit sphere. This concludes the lemma.  $\square$

Assuming the  $\varepsilon$ -stability of the explicit step, Lemma 3.8 verifies that the scaled perturbation  $\mathbf{V}_\Delta$  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 (2.5) with a suitable scaling and reference solution has been to enrich Lemmas 3.6 and 3.8 to conclude the  $\varepsilon$ -stability of the numerical solution since we directly work with perturbations. Otherwise, one needs to study the structure of  $J_\varepsilon^{-1}$ , *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  $\varepsilon$ -stability of the solution implies that there exists a sequence  $\{\mathbf{V}_{\Delta, \varepsilon_k}^{n+1}\}_{k \in \mathbb{N}}$  ( $\varepsilon_k \rightarrow 0$  as  $k \rightarrow \infty$ ) 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  $\varepsilon$ -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  $\varepsilon$ , thanks to the  $\varepsilon$ -stability.

Now, consider the discrete version of the asymptotic expansion (3.3), for all  $i \in \Omega_N$  and the temporal step  $n$ :

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

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

$$v_1(t_n, x) = z_{(2)}(t_n, x), \quad v_2(t_n, x) = m_{(0)}(t_n) + \varepsilon m_{(1)}(t_n) + \varepsilon^2 m_{(2)}(t_n, x). \quad (3.16)$$

Note that (3.16) and  $\varepsilon$ -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 (3.16) into the momentum update of the explicit step (3.8a) yields (with  $k = 0, 1$ )

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

where  $v_{2(k)c}^n$  is a constant from (3.16). So, the explicit step for the momentum does not introduce an  $\mathcal{O}(1/\varepsilon)$  term into the scheme, i.e.,  $\|\mathbf{V}_{\Delta}^{n+1/2}\| = \mathcal{O}(1)$ . Remark 3.7 implies that the boundedness of  $\mathbf{V}_{\Delta}^{n+1/2}$  leads to the  $\varepsilon$ -stability of the implicit solution. Thus, from the implicit  $v_1$  update (3.8b), one can (rigorously) conclude that for all  $i \in \Omega_N$

$$v_{2(0)i+1}^{n+1} = v_{2(0)i-1}^{n+1}, \quad v_{2(1)i+1}^{n+1} = v_{2(1)i-1}^{n+1}.$$

So, the updated momentum is *almost* constant, i.e., the discrete divergence operator vanishes in the limit  $\nabla_h v_{2,\Delta}^{n+1} = \mathcal{O}(\varepsilon^2)$ . 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 (3.8a) and (3.8b) with a constant  $\tilde{\alpha}$ , applied to the 1d SWE with flat bottom, the deviations of the computed momentum is  $\mathcal{O}(\varepsilon^2)$ , as  $\varepsilon \rightarrow 0$ . 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  $\llbracket v_{k,i} \rrbracket := v_{k,i} - v_{k,i-1}$  with  $k = 1, 2$ , the following holds:

$$J_{\varepsilon} \begin{bmatrix} \llbracket \mathbf{V}_{1,\Delta}^{n+1} \rrbracket \\ \llbracket \mathbf{V}_{2,\Delta}^{n+1} \rrbracket \end{bmatrix} = \begin{bmatrix} \llbracket \mathbf{V}_{1,\Delta}^{n+1/2} \rrbracket \\ \llbracket \mathbf{V}_{2,\Delta}^{n+1/2} \rrbracket \end{bmatrix}. \quad (3.17)$$

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

$$\|K_{11}\|, \|K_{12}\|, \|K_{22}\| = \mathcal{O}(1), \quad \|K_{21}\| = \mathcal{O}(\varepsilon^2). \quad (3.18)$$

Then, owing to (3.16),  $\|\llbracket \mathbf{V}_{1,\Delta}^{n+1/2} \rrbracket\| = \mathcal{O}(1)$  and  $\|\llbracket \mathbf{V}_{2,\Delta}^{n+1/2} \rrbracket\| = \mathcal{O}(\varepsilon^2)$ . Combining it with (3.17) yields

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

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 (3.18), and in particular,  $K_{21}$  and  $K_{22}$ . Let us re-write the inverse  $K_{\varepsilon}$  as (using Lem. 3.5)

$$K_{\varepsilon} = \begin{bmatrix} \left( P - \frac{\beta \bar{h}}{\varepsilon^2} Q P^{-1} Q \right)^{-1} & -\frac{\beta^2}{\varepsilon^2} P^{-1} Q \left( P - \frac{\beta^2 \bar{h}}{\varepsilon^2} Q P^{-1} Q \right)^{-1} \\ -\beta \bar{h} P^{-1} Q \left( P - \frac{\beta^2 \bar{h}}{\varepsilon^2} Q P^{-1} Q \right)^{-1} & \left( P - \frac{\beta^2 \bar{h}}{\varepsilon^2} Q P^{-1} Q \right)^{-1} \end{bmatrix}. \quad (3.19)$$

Then, it is clear from Lemma 3.6 and the structure of  $K_\varepsilon$ , that  $K_{12} = \frac{\beta}{\varepsilon^2 h} K_{21}$  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.  $\square$

To conclude the asymptotic consistency, it is also required to confirm 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  $\hat{\alpha}$  and  $\tilde{\alpha}$ ):

$$\frac{v_{2(0),i}^{n+1/2} - v_{2(0),i}^n}{\Delta t} + \nabla_h \left[ \frac{v_{2(0),i}^{2,n}}{\bar{h}_i + \varepsilon^2 v_{1(0),i}^n} + \frac{\varepsilon^2}{2} v_{(0)1,i}^{2,n} \right] - \frac{\hat{\alpha} \Delta x}{2} \Delta_h v_{2(0),i}^n = 0 \quad (\text{Explicit step}), \quad (3.20a)$$

$$\frac{v_{2(0),i}^{n+1} - v_{2(0),i}^{n+1/2}}{\Delta t} + \nabla_h \left( \bar{h}_i v_{1(0),i}^{n+1} \right) - \frac{\tilde{\alpha} \Delta x}{2} \Delta_h v_{2(0),i}^n = -v_{1(0),i}^{n+1} \nabla_h b_i \quad (\text{Implicit step}). \quad (3.20b)$$

It is clear that (3.20a) and (3.20b) provide consistent discretisations of  $\partial_t m_{(0)}$  for both explicit and implicit steps, (3.8a) and (3.8b). Thus, in the light of Lemmas 3.8 and 3.11 for the  $\varepsilon$ -stability and smallness of checker-board modes, the asymptotic consistency of the scheme is concluded.

### 3.2.4. Asymptotic stability

In this section, we discuss the rigorous stability analysis of the RS-IMEX scheme in  $\ell_2$ -norm, for a fixed grid and in finite time ( $T_f < \infty$ ). Consider the scheme as the following iteration with a constant  $\Delta t$

$$\mathbf{Y}^k = \prod_{i=0}^{s-1} \mathcal{E}_{s-i} \mathbf{Y}^{k-1}, \quad k = 0, 1, \dots, n-1, \quad n = T_f / \Delta t, \quad (3.21)$$

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

Motivated by Lemma 3.1 of [33] (see also [55, 64]), one can show that the scheme (3.21) is stable for a finite time and in  $\ell_p$ -norm, provided that there exist constants  $c_i$  independent of  $\Delta t$  such that

$$\|\mathcal{E}_i\|_{\ell_p} \leq 1 + c_i \Delta t, \quad i = 1, \dots, s. \quad (3.22)$$

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

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 (3.22) holds. Since the explicit step is non-linear, obtaining (3.22) 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  $J_\varepsilon^{-1}$ . 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  $\ell_2$ . 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 (3.14) 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  $\varepsilon$ -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  $\varepsilon$ -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  $\mathbf{Y}^k = [\mathbf{V}_{1,\Delta}^k, \mathbf{V}_{2,\Delta}^k]^T$ .

Assuming  $\hat{\alpha} = 0$  for simplicity and from (3.8a), one can write  $\|\mathcal{E}_1 \mathbf{Y}^k\|_{\ell_2}$  as

$$\begin{aligned} \|\mathcal{E}_1 \mathbf{Y}^k\|_{\ell_2} &\leq \|\mathbf{Y}^k\|_{\ell_2} + \frac{2\beta}{h_{\min}^k} \|\langle \mathbf{V}_{2,\Delta}^k, \mathbf{V}_{2,\Delta}^k \rangle\|_{\ell_2} + \varepsilon^2 \beta \|\langle \mathbf{V}_{1,\Delta}^k, \mathbf{V}_{1,\Delta}^k \rangle\|_{\ell_2} \\ &\leq \|\mathbf{Y}^k\|_{\ell_2} + \beta \left( \frac{2}{h_{\min}^k} + \varepsilon^2 \right) \|\mathbf{Y}^k\|_{\ell_2}^2, \\ &\leq \left[ 1 + \beta \left( \frac{2}{h_{\min}^k} + \varepsilon^2 \right) \|\mathbf{Y}^k\|_{\ell_2} \right] \|\mathbf{Y}^k\|_{\ell_2} \end{aligned} \quad (3.23)$$

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

For simplicity, one can re-write (3.23) as

$$y_{k+1} \leq y_k + \beta_k y_k^2, \quad y_k := \|\mathbf{Y}^k\|_{\ell_2}, \quad \beta_k := \beta \left( \frac{2}{h_{\min}^k} + \varepsilon^2 \right). \quad (3.24)$$

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 [66].

**Theorem 3.12** ([66], Thm. 4). *Consider the sequences  $\{\mu_k\}_{k>0}$ ,  $\{\nu_k\}_{k>0} \geq 0$  for  $k = 0, 1, \dots$  while  $\mu_0 = \nu_0 = 0$ . If for the non-negative sequence  $\{y_k\}_{k=0,1,\dots}$  it holds that*

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

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

$$y_{k+1} \leq \frac{1}{\psi(k)} \left( \sigma^q + q \sum_{i=0}^k \mu_i \psi^q(i) \right)^{1/q}, \quad k = 0, 1, \dots \quad (3.25)$$

Using Theorem 3.12, the following corollary can be obtained.

**Corollary 3.13.** *Given a small enough initial datum, the sequence  $\{y_k\}_{k=0,1,\dots}$  defined in (3.24) is bounded uniformly in  $\varepsilon$ .*

*Proof.* One can re-write (3.24) as

$$y_{k+1} \leq y_0 + \sum_{i=0}^k \beta_i y_i^2 = (y_0 + \beta_0 y_0^2) + \sum_{i=1}^k \beta_i y_i^2. \quad (3.26)$$

Comparing (3.26) 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.*,

$$(1 + \beta_0 y_0) y_0 \sum_{i=1}^k \beta_i < 1. \quad (3.27)$$

So, the norm of solution of the explicit step  $y_k$  is bounded under an “smallness assumption”.  $\square$

**Remark 3.14.** Note that a simpler version of such bounds can be obtained by induction, like equation (3.11) of [51].

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

$$y_{k+1} \leq (1 + c_2 \Delta t)(1 + c_1 \Delta t y_k) y_k, \quad (3.28)$$

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 (3.8a) and (3.8b) is  $\ell_2$ -bounded uniformly in  $\varepsilon$  and for a finite time.*

**Remark 3.16.** One can read the smallness condition (3.27) as a timestep restriction. This condition is restrictive, not in  $\varepsilon$ , 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 [28]; 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  $\varepsilon$ -uniform (with material velocity), but the condition number of  $J_\varepsilon$  increases as  $\varepsilon \rightarrow 0$  (see Rem. 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, and since one can simply combine the method with a pre-conditioner (as in [7]) to regain the asymptotic efficiency.

### 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  $\widehat{\mathbf{G}}$  is also constant; so,  $\mathbf{V}_\Delta^{n+1/2} = \mathbf{V}_\Delta^n$ . 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.*,  $\mathbf{V}_\Delta^{n+1} = \mathbf{V}_\Delta^n$ ; thus, the scheme is well-balanced. For a more rigorous proof, we write the implicit step as (assuming  $\tilde{\alpha} = 0$ )

$$\mathbf{V}_{1,\Delta}^{n+1} + \frac{\beta}{\varepsilon^2} \mathbf{V}_{2,\Delta}^{n+1} = c \mathbf{1}_N, \quad (3.29a)$$

$$\beta R_b \mathbf{V}_{1,\Delta}^{n+1} + \mathbf{V}_{2,\Delta}^{n+1} = \mathbf{0}_N, \quad (3.29b)$$

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

Putting (3.29a) into (3.29b) gives

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

which implies that  $\mathbf{V}_{2,\Delta}^{n+1} = \mathbf{0}$  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 Sect. 3.2.1). Also, (3.29a) concludes that  $\mathbf{V}_{1,\Delta}^{n+1}$  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 is 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 (3.1) in a periodic domain and with a flat bottom topography while the reference solution  $\bar{\mathbf{U}} = (\bar{z}, \bar{m})^T$  is chosen as the zero-Froude limit solution of (3.1). It can be obtained from Definition 3.1 and equation (3.4) that  $\bar{z} = z_{(0)}^0$  and  $\bar{m} = m_{(0)}^0$ , 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  $\bar{\mathbf{T}}$  to vanish and avoids the difficulties stem from its discretisation in the asymptotic analysis (as discussed in [70]). With this reference solution, the splitting can be obtained as

$$\begin{aligned} \bar{\mathbf{F}} &:= \left[ \frac{\bar{m}^2}{\bar{z} - b} + \frac{\bar{m}}{2\varepsilon^2} \right], \quad \tilde{\mathbf{F}} := \left[ -\frac{\bar{m}^2 z_{\text{pert}}}{(\bar{z} - b)^2} + \frac{m_{\text{pert}}}{\varepsilon^2} z_{\text{pert}} + \frac{2\bar{m}m_{\text{pert}}}{\bar{z} - b} \right], \\ \hat{\mathbf{F}} &:= \left[ \frac{(\bar{m} + m_{\text{pert}})^2}{\bar{z} + z_{\text{pert}} - b} + \frac{z_{\text{pert}}^2}{2\varepsilon^2} - \frac{\bar{m}^2}{\bar{z} - b} + \frac{\bar{m}^2 z_{\text{pert}}}{(\bar{z} - b)^2} - \frac{2\bar{m}m_{\text{pert}}}{\bar{z} - b} \right]. \end{aligned}$$

The splitting is admissible in the sense of [62]; the eigenvalues of  $\tilde{\mathbf{F}}'$  are stiff and those of  $\hat{\mathbf{F}}'$  are non-stiff:

$$\tilde{\lambda} = \frac{\bar{m}}{\bar{z} - b} \pm \frac{\sqrt{\bar{z} - b}}{\varepsilon}, \quad \hat{\lambda} = 0, 2u_{\text{pert}}. \quad (4.1)$$

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

$$\tilde{\mathbf{G}} := \left[ -\frac{\bar{m}^2 v_1 \varepsilon}{(\bar{z} - b)^2} + \frac{v_2/\varepsilon}{(\bar{z} - b)} + \frac{2\bar{m}v_2}{\bar{z} - b} \right], \quad \hat{\mathbf{G}} := \left[ \frac{(\bar{m} + \varepsilon v_2)^2}{\varepsilon(\bar{z} + \varepsilon^2 v_1 - b)} + \frac{\varepsilon v_1^2}{2} - \frac{\bar{m}^2}{\varepsilon(\bar{z} - b)} + \frac{\bar{m}^2 v_1 \varepsilon}{(\bar{z} - b)^2} - \frac{2\bar{m}v_2}{\bar{z} - b} \right]. \quad (4.2)$$

Due to the well-prepared initial velocity (3.4), the zero-Froude limit reference state makes the wave speeds of the slow system really small, *i.e.*,  $\mathcal{O}(\varepsilon)$ , as  $u_{\text{pert}} = \mathcal{O}(\varepsilon)$ . Also,  $\bar{\mathbf{U}}$  is asymptotically close to the solution. Thus, the commutator would be  $\mathcal{O}(1)$ , *i.e.*,

$$\lim_{\varepsilon \rightarrow 0} [\tilde{\mathbf{G}}', \hat{\mathbf{G}}'] = \begin{bmatrix} v_1 & \frac{2v_2}{\bar{z} - b} \\ -2v_2 & -v_1 \end{bmatrix}. \quad (4.3)$$

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 (3.7a) and (3.7b) when  $\hat{\mathbf{G}}$  and  $\tilde{\mathbf{G}}$  change according to (4.2).

#### 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 (3.7a) and (3.7b), with (4.2) a constant  $\tilde{\alpha}$ , and under an  $\varepsilon$ -uniform timestep restriction*

- (i) *is solvable, i.e., it has a unique solution for all  $\varepsilon > 0$ .*
- (ii) *its solution is  $\varepsilon$ -stable, i.e., it is bounded for  $\varepsilon \ll 1$ . So, there is convergent sub-sequence of the discrete solutions as  $\varepsilon \rightarrow 0$ .*
- (iii) *is consistent with the asymptotic limit in the fully-discrete settings, i.e., it is asymptotically consistent.*
- (iv) *is asymptotically  $\ell_2$ -stable for the fixed grid  $\Delta x$ , 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  $\|\mathbf{V}_\Delta^n\|_{\ell_2} \leq C_{N,T_f} \|\mathbf{V}_\Delta^0\|_{\ell_2}$ .*

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  $J_\varepsilon$  reads

$$J_\varepsilon := \begin{bmatrix} P & \frac{\beta}{\varepsilon} Q \\ \left(-\frac{\bar{m}^2 \varepsilon}{\bar{h}^2} + \frac{\bar{h}}{\varepsilon}\right) \beta Q & P + \frac{2\beta \bar{m}}{\bar{h}} Q \end{bmatrix}. \quad (4.4)$$

The blocks of  $J_\varepsilon$  commute and from Theorem 1 of [63] the determinant of  $J_\varepsilon$  can be computed as

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

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 theorem [4, 36]. Thus, it remains to show that  $\mathfrak{A}$  has only non-zero eigenvalues. Note that the eigenvalues of  $P^2 + \frac{\beta^2 \bar{m}^2}{\bar{h}^2} Q^2$  can be set positive, by a suitable and  $\varepsilon$ -uniform choice of  $\beta$ . Using the sub-additivity of the numerical range (spectrum for symmetric matrices), adding  $-\frac{\beta^2 \bar{h}}{\varepsilon^2} Q^2$  with non-negative eigenvalues makes  $J_\varepsilon$  non-singular.

##### 4.1.2. $\varepsilon$ -stability of the solution

Similar to Section 3.2, we can find  $\Xi_\varepsilon$  as

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

TABLE 1. Comparison of different scaling for matrix  $J_\varepsilon$ .

|                 | Scaling by $\text{diag}(\varepsilon^2, 1)$                            | Scaling by $\text{diag}(\varepsilon^2, \varepsilon)$                                         |
|-----------------|-----------------------------------------------------------------------|----------------------------------------------------------------------------------------------|
| $J_\varepsilon$ | $\begin{bmatrix} 1 \mathcal{O}(1/\varepsilon^2) \\ 1 1 \end{bmatrix}$ | $\begin{bmatrix} 1 \mathcal{O}(1/\varepsilon) \\ \mathcal{O}(1/\varepsilon) 1 \end{bmatrix}$ |

One can then confirm that  $\Xi_\varepsilon^{-1}$  has a bounded norm in terms of  $\varepsilon$  as, thanks to Lemma 3.5, the blocks of  $\Xi_\varepsilon^{-1}$  can be obtained as

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

which are all bounded. Assuming the  $\varepsilon$ -stability of the explicit step (see Sect. 4.1.3), the solution of the implicit step (thus the whole scheme) can be shown to be  $\varepsilon$ -stable. The  $\varepsilon$ -stability of the solution implies that the scaled perturbation  $\mathbf{V}_\Delta$  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  $J_\varepsilon$  depends on the scaling matrix. For example, one can confirm that using  $\text{diag}(\varepsilon^2, 1)$  and  $\text{diag}(\varepsilon^2, \varepsilon)$  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” ([29], Sect. 3.5.2) in essence, and may improve the condition number of  $J_\varepsilon$ ; see Table 1.

#### 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  $\varepsilon$ -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

$$\lim_{\varepsilon \rightarrow 0} \left[ \frac{(\bar{m} + \varepsilon v_2^n)^2}{\varepsilon (\bar{z} + \varepsilon^2 v_1^n - b)} - \frac{\bar{m}^2}{\varepsilon (\bar{z} - b)} \right] = \mathcal{O}(1). \quad (4.5)$$

So, it is asymptotically consistent (and  $\varepsilon$ -stable). This implies that for the implicit step, as shown in the previous section,  $\mathbf{V}_\Delta^{n+1} = \mathcal{O}(1)$ . 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 (4.5) to find an  $\varepsilon$ -uniform bound. Hence, one can conclude that the scheme is again AP (in a weaker sense than Def. 1.1), *i.e.*, it is AC and AS under a non-restrictive CFL condition while the condition number of  $J_\varepsilon$  increases as  $\varepsilon \rightarrow 0$ .

**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 [71] suggests that the reference solution does not affect stability of the scheme. However, asymptotically smaller wave speeds for the zero-Froude case (compare (4.1) with (3.2)) 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  $\hat{\alpha}$  like in the Lax–Friedrichs scheme as the maximum value of all wave speeds over the whole domain, and  $\tilde{\alpha}$  is likewise but computed for  $\varepsilon = 1$  avoiding excessive diffusion. The timestep has been computed as  $\Delta t := \min(\Delta t_{\text{CFL}}, \Delta t_{\text{Aux}})$  where the CFL timestep  $\Delta t_{\text{CFL}}$  and the auxiliary timestep  $\Delta t_{\text{Aux}}$  are defined as

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

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

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

where  $\mathbf{1}_\omega$  is the characteristic function in the domain  $\omega$ , and  $H_{\text{ref}} = 1$ ; so,  $\bar{z} = 0$ . We also set the final time  $T_f = 0.05$  and  $\text{CFL} = 0.45$ . In Example 6.1 of [16] the pressure function  $p(\varrho) = \varrho^2$  has been used; so, we compare the results of the RS-IMEX scheme with Section 8.1 of [6], where the pressure function is the same as the SWE.

Figures 2 and 3 show the results of the RS-IMEX scheme with  $\bar{m} = 0$  (LaR) and  $\bar{m} = 1$  (zero-Froude limit) for  $\varepsilon = 0.8$  and  $\varepsilon = 0.1$ . Compared to Figure 8.2 of [6], it is clear that the quality of computed solutions is fine. Note that for this example, the schemes in Figure 8.2 of [6] use the same splitting as the RS-IMEX; but, they employ an elliptic approach for the surface perturbation update; see [6] for more details. As Figures 2 and 3 suggest, the computed surface perturbation  $z$  does not change considerably with the choice of the reference momentum, particularly for  $\varepsilon = 0.1$ . For the momentum, the  $\bar{m} = 1$  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 (4.1) with (3.2)) which leads to the 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 [6, 16].

Figure 4 illustrates the experimental order of convergence (EOC) for different  $\varepsilon$  and  $\bar{m} \in \{0, 1\}$ , for the “mean” error  $e(\phi_\Delta^{\text{num}})$  in the discrete  $L_1$ -norm, defined as

$$e(\phi_\Delta^{\text{num}}) := \|\phi_\Delta^{\text{num}} - \phi_\Delta^{\text{ref}}\|_{L_1(\Omega_{N_{\text{ref}}})} = \frac{1}{N_{\text{ref}}} \sum_{j \in \Omega_{N_{\text{ref}}}} |\phi_j^{\text{num}} - \phi_j^{\text{ref}}|, \quad (5.1)$$

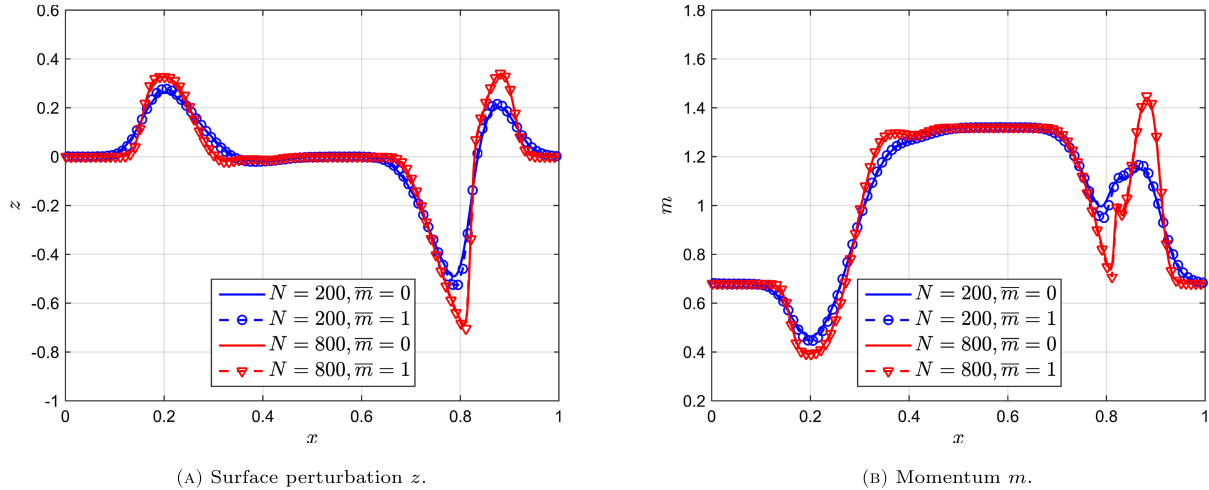


FIGURE 2. The RS-IMEX solutions for Example (i), with  $\varepsilon = 0.8$ , CFL = 0.45,  $T_f = 0.05$ , and with two reference states: the LaR and the zero-Froude limit. Comparing these figures with [6] verifies the accuracy of the scheme.

where  $\phi$  is the variable of interest (momentum, height, etc.), and  $\phi_{\Delta}^{\text{num}}$  and  $\phi_{\Delta}^{\text{ref}}$  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 show that the scheme, regardless of the reference momentum and in agreement with the result of Table 2 from [16], has an almost uniform order of convergence for the selected values of  $\varepsilon$  for a small enough  $\Delta x$ , which is in this case  $\Delta x \approx 0.01$ . Note that for this example and with a small  $\varepsilon$ , namely  $\varepsilon = 0.001$ , the computed surface perturbation, so the error, are as small as the machine accuracy. So, to confirm the uniform EOC for smaller values of  $\varepsilon$ , we refer to another example in Section 5.2.

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  $J_{\varepsilon}$  to be  $\mathcal{O}(1/\varepsilon)$  (see Rem. 4.2). Note also that for the zero-Froude limit reference state, due to  $\mathcal{O}(\varepsilon)$  eigenvalues for the non-stiff system as in (4.1),  $\Delta t_{\text{CFL}} = \mathcal{O}(1/\varepsilon)$ ; so, the timestep imposed by the advective CFL condition gets larger as  $\varepsilon$  decreases. For this example, since there are only  $\mathcal{O}(\varepsilon^2)$  deviations of the initial momentum from  $\bar{m}$ , one expects  $\Delta t_{\text{CFL}} = \mathcal{O}(1/\varepsilon^2)$ .

### 5.1.2. (ii) Colliding pulses

Consider the following ill-prepared initial data in the periodic domain  $[-1, 1)$ , as in Section 5.2 from [2] (motivated by an example of [47]):

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

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  $H_{\text{ref}} = 1$ , i.e.,  $\bar{z} = -0.045$ . The figure shows that, comparing to Figure 1 from [2], the quality of the computed solution is fine. Note that in [2], the height is computed by an elliptic approach. Moreover, Figure 6 confirms the  $\varepsilon$ -uniformity of the timestep, and stability of the scheme in  $\ell_2$ -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  $\phi$ . As Figure 6 suggests, the scheme is stable uniformly in  $\varepsilon$  for variables like  $z$ ,  $m$  and  $u$ . Also one can see that, as

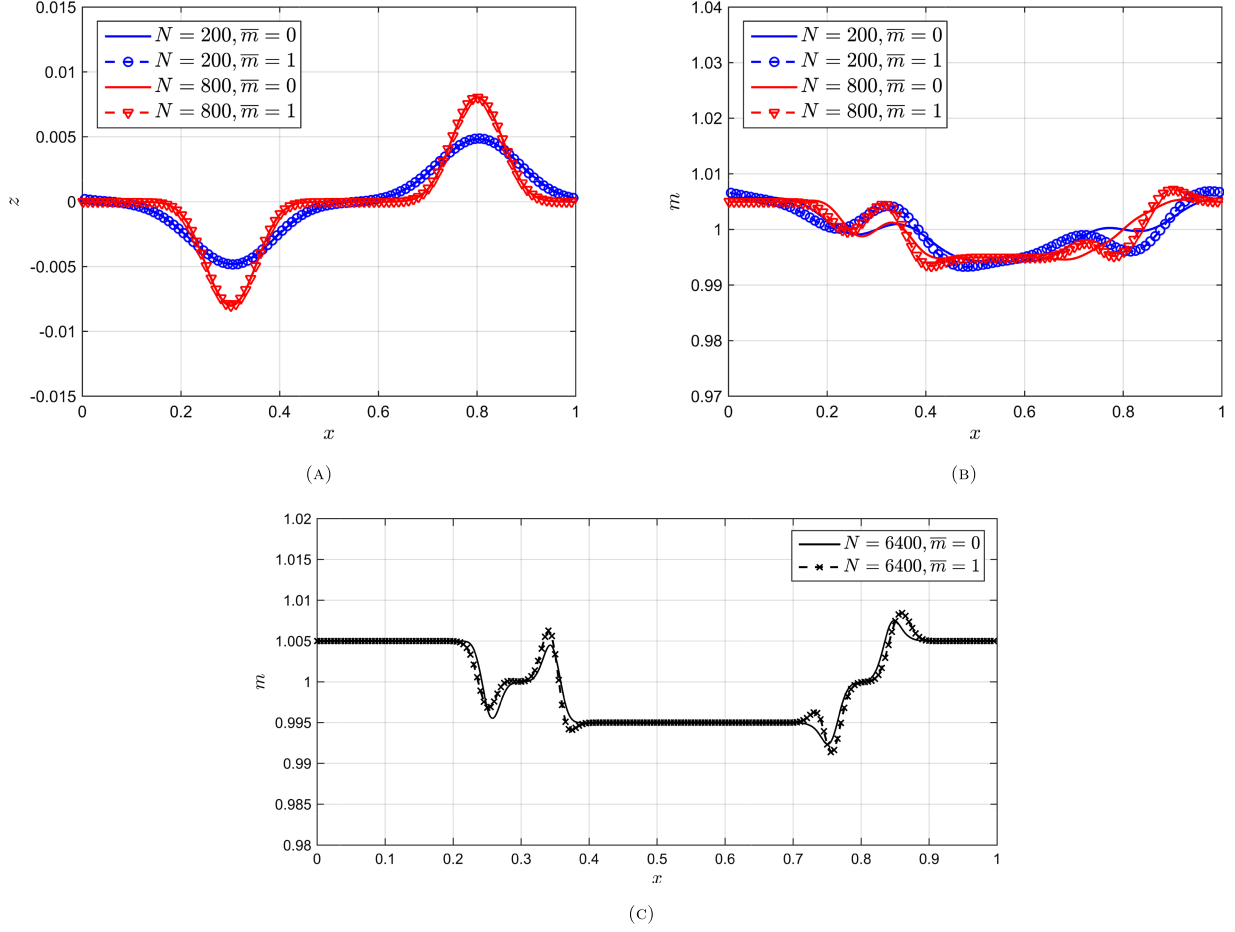


FIGURE 3. *Panels A and B:* The RS-IMEX solutions for Example (i), with  $\varepsilon = 0.1$ , CFL = 0.45,  $T_f = 0.05$ , and with two reference states: the LaR and the zero-Froude limit. *Panel C* is like *panel B*, but for a very fine mesh. The solutions are accurate in comparison to [6]. Also, the zero-Froude reference solution seems to provide better results.

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  $\varepsilon$ . Note that after the second step, it is  $\hat{\alpha} = \mathcal{O}(1)$  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  $\bar{z} = -0.045$  and change the reference momentum to  $\bar{m} = \sqrt{2}$ , which is not the zero-Froude limit. We call this setting case (ii<sub>b</sub>). 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 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 [21], p. 526). Figure 7 confirms this conjecture, as it shows that the solution gets more symmetric as one refines the grid, *i.e.*, as the operator splitting error gets smaller.

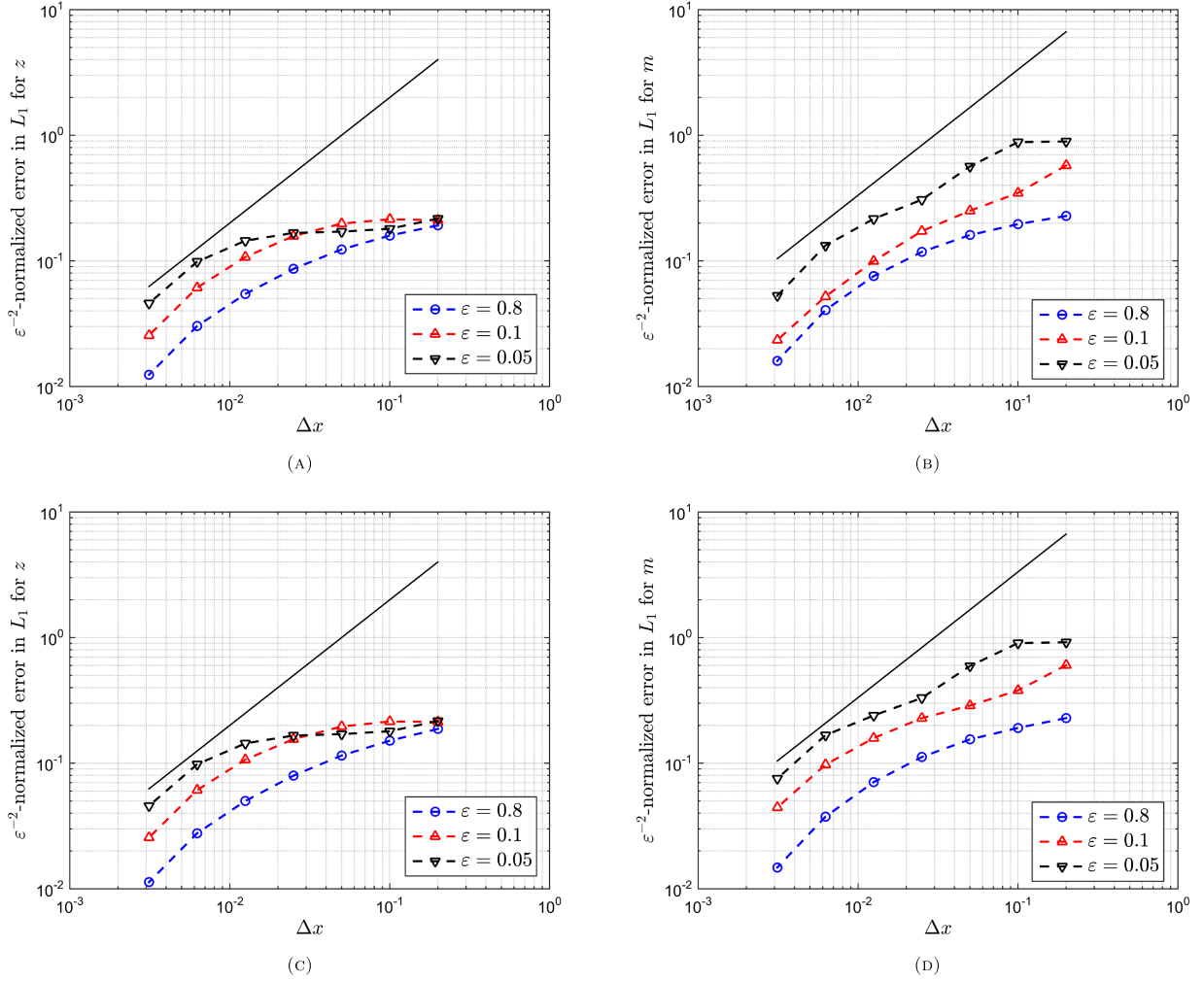


FIGURE 4. The EOC of the RS-IMEX scheme in Example (i), confirming the theoretical order of convergence, with  $\text{CFL} = 0.45$ ,  $T_f = 0.05$ : *Panels A and B* for the LaR ( $\bar{m} = 0$ ) reference state, *Panels C and D* for the zero-Froude limit ( $\bar{m} = 1$ ) reference state. The black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

## 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 smallness of checker-board oscillations, 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  $\varepsilon$ -uniform EOC for the scheme. Moreover, Figure 9 shows the smallness of the checker-board oscillations for the momentum, without any need for a special fix. It can be seen that  $\|[\mathbf{V}_{2,\Delta}]\|_{\ell_\infty}$ , which indicates the amplitude of possible checker-board oscillations, decreases like  $\mathcal{O}(\varepsilon)$  as  $\varepsilon \rightarrow 0$ . It can also be seen that the condition number of  $J_\varepsilon$  is very large, *i.e.*,  $\text{cond}_2(J_\varepsilon) = \mathcal{O}(1/\varepsilon)$ .

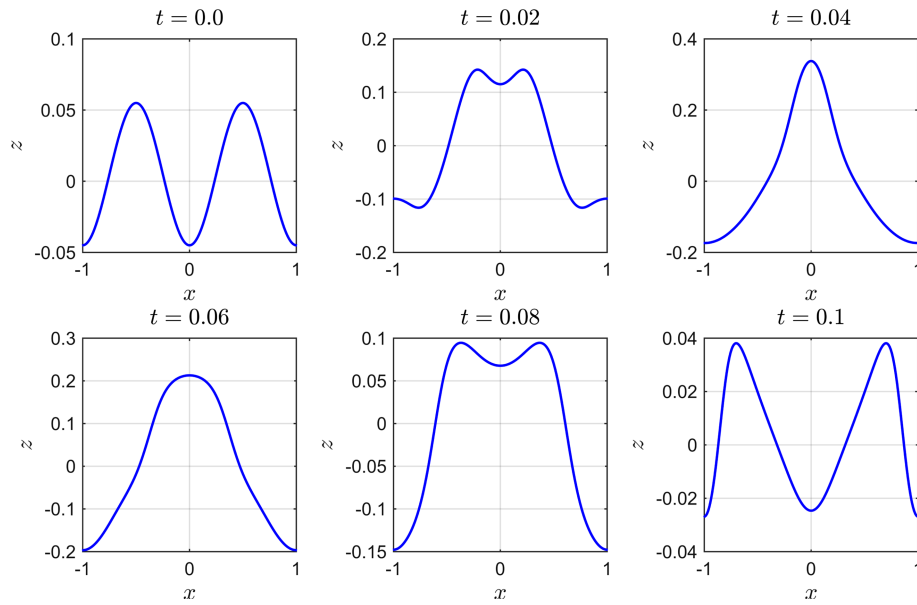


FIGURE 5. Evolution of the surface perturbation for the RS-IMEX solution in Example (ii<sub>a</sub>), with  $\varepsilon = 0.1$ , CFL = 0.45,  $N = 200$ , and the LaR reference solution. The figure confirms the accuracy of the computed solution compared to [2].

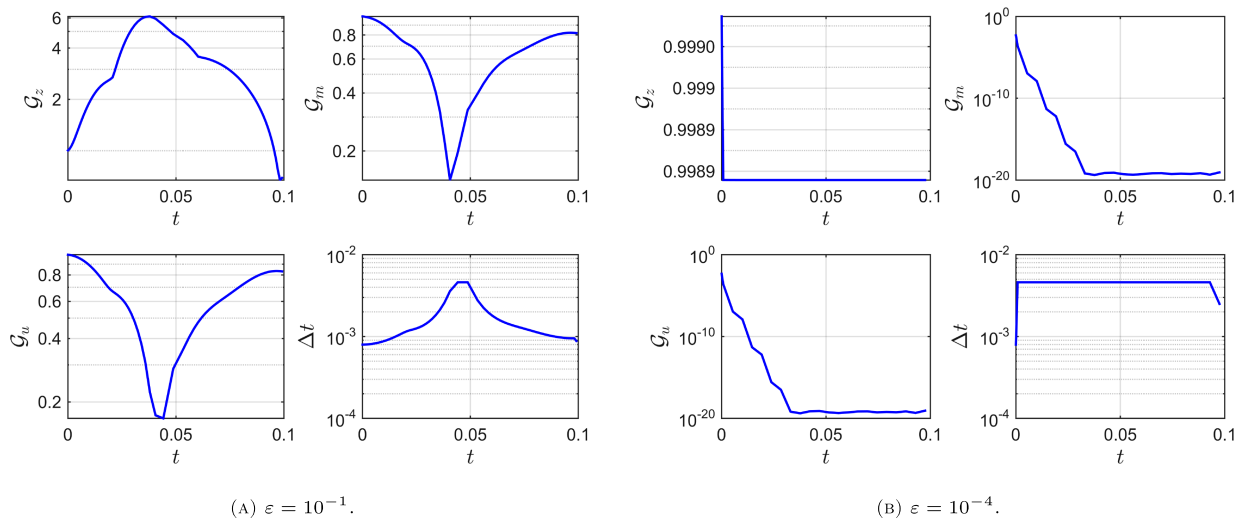


FIGURE 6. Growth factor and timestep for different  $\varepsilon$ , to confirm the asymptotic stability of the scheme, for Example (ii<sub>a</sub>) with the LaR reference solution.

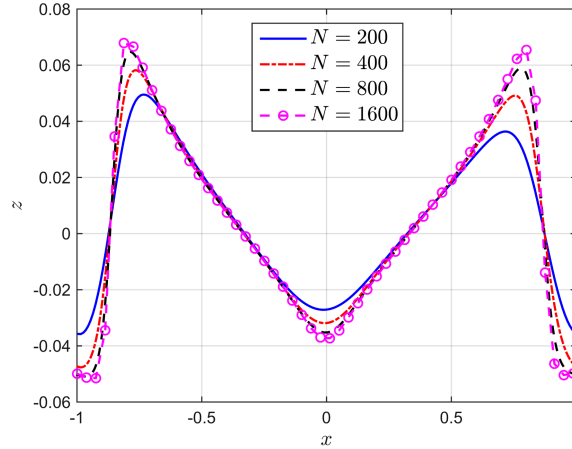


FIGURE 7. Vanishing effect of an unsuitable reference solution for Example (ii<sub>b</sub>) as  $\Delta x \rightarrow 0$ , for  $\varepsilon = 0.1$ ,  $T_f = 0.1$  and  $N = 200, 400, 800, 1600$ .

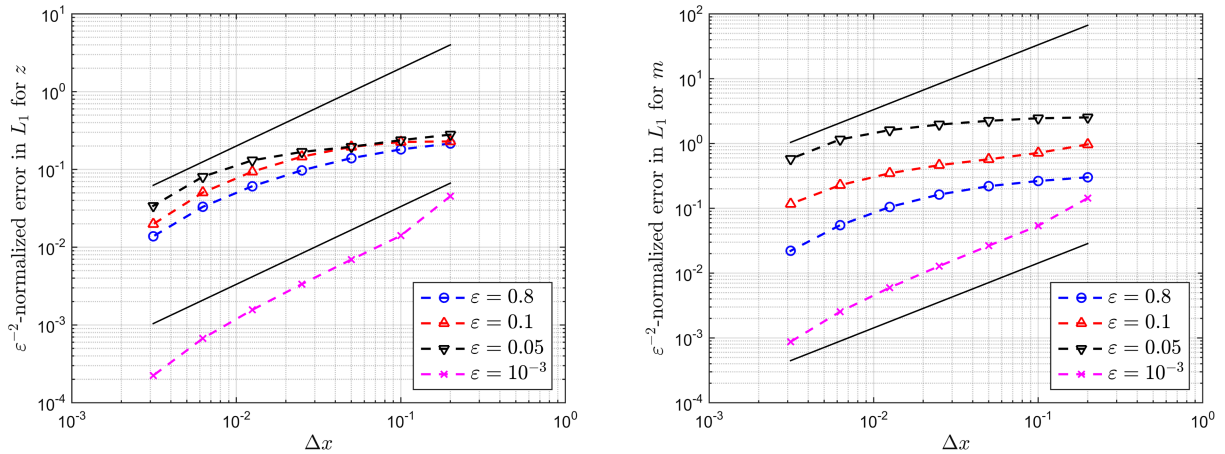


FIGURE 8. EOC of the RS-IMEX scheme for Example (iii), confirming the uniform order of convergence in  $\varepsilon$ , with  $T_f = 0.05$ ,  $\text{CFL} = 0.45$  and the LaR reference solution. The black solid line is the line with slope one.

## 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 the asymptotic consistency and stability, as well as the quality of the computed solution have been illustrated by numerical analysis and several numerical tests. Although, strictly speaking, the implicit step is not uniformly-efficient due to the large condition number of the linear system associated to that, we claim that the scheme is AP as, practically speaking, the scheme is efficient, thanks to the simple structure of the linear system to be solved. Moreover, the analysis confirms the 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 presented, 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.

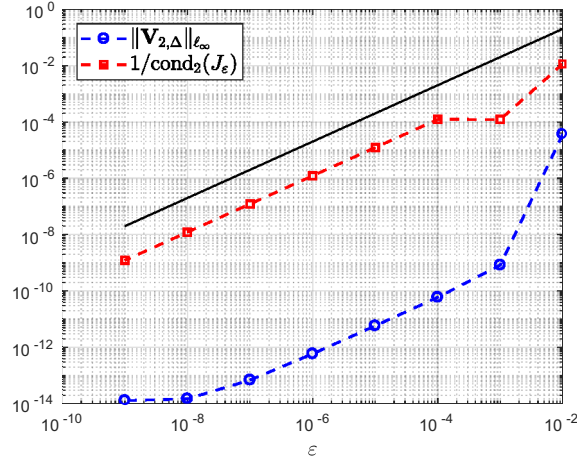


FIGURE 9. Behaviour of the condition number of  $J_\varepsilon$  and the amplitude of checker-board like oscillations regarding  $\varepsilon$ , computed for Example (iii) with  $T_f = 0.05$ , CFL = 0.45 and the LaR reference solution. The black solid line is the line with slope one.

Of course, it is of a great interest to conduct a similar study for the two-dimensional case, for which a significant challenge would be the more complicated structure of the implicit operator (or equivalently matrix  $J_\varepsilon$ ), which makes the rigorous asymptotic consistency analysis much more involved; we refer the reader to [6, 67, 70] for more details.

## APPENDIX 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  $\Omega$  (see also [12] 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 (3.1)):

$$\begin{aligned} \partial_t h + \partial_x m &= 0, \\ \partial_t m + \partial_x \left( \frac{m^2}{h} + \frac{h^2}{2\varepsilon^2} \right) &= -\frac{h \partial_x \eta^b}{\varepsilon^2}. \end{aligned} \quad (\text{A.1})$$

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

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

in (A.1), and balance equal powers of  $\varepsilon$ .  $\mathcal{O}(\varepsilon^{-2})$  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  $\eta_{(0)}^s := h_{(0)} + \eta^b = \eta_{(0)}^s(t)$ . 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

$$\frac{d}{dt} \int_{\Omega} h_{(0)} dx = - \int_{\Omega} \partial_x m_{(0)} dx = 0,$$

owing to the divergence theorem and the assumption of periodic boundary conditions. Thus,  $\partial_t h_{(0)} = 0$  and  $\eta_{(0)}^s = \text{const.}$ , which give  $h_{(0)} = h_{(0)}(x) = \eta_{(0)}^s - \eta^b(x)$  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

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

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.

## APPENDIX 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  $\beta^2 R_b Q \mathbf{w}_2^{(2)} = \mathbf{w}_2^{(0)}$  implies that  $\mathbf{w}_2^{(0)}$  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 (3.11) and  $\beta^2 R_b Q \mathbf{w}_2^{(2)} = \mathbf{w}_2^{(0)}$  gives the following linear system of equations:

$$\beta^2 \begin{bmatrix} \bar{h}_N - \bar{h}_2 & \bar{h}_2 & & & -\bar{h}_N \\ -\bar{h}_1 & \bar{h}_1 - \bar{h}_3 & \bar{h}_3 & & \\ & -\bar{h}_2 & \bar{h}_2 - \bar{h}_4 & \bar{h}_4 & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} \mathbf{w}_{2,2}^{(2)} - \mathbf{w}_{2,N}^{(2)} \\ \mathbf{w}_{2,3}^{(2)} - \mathbf{w}_{2,1}^{(2)} \\ \mathbf{w}_{2,4}^{(2)} - \mathbf{w}_{2,2}^{(2)} \\ \vdots \end{bmatrix} = \begin{bmatrix} \mathbf{w}_{2,\text{odd}}^{(0)} \\ \mathbf{w}_{2,\text{even}}^{(0)} \\ \mathbf{w}_{2,\text{odd}}^{(0)} \\ \vdots \end{bmatrix}. \quad (\text{B.1})$$

We want to characterise the null space of the coefficient matrix and show that the system has a solution only if  $\mathbf{w}_{2,\text{odd}}^{(0)}, \mathbf{w}_{2,\text{even}}^{(0)} = \mathbf{0}$ . One can pose the following lemma.

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

$$\begin{bmatrix} m_N - m_2 & m_2 & & & -m_N \\ -m_1 & m_1 - m_3 & m_3 & & \\ & -m_2 & m_2 - m_4 & m_4 & \\ & & \ddots & \ddots & \ddots \\ m_1 & & & -m_{N-1} & m_{N-1} - m_1 \end{bmatrix} \begin{bmatrix} y_2 - y_N \\ y_3 - y_1 \\ y_4 - y_2 \\ \vdots \\ y_1 - y_{N-1} \end{bmatrix} = \begin{bmatrix} c_o \\ c_e \\ c_o \\ \vdots \end{bmatrix}. \quad (\text{B.2})$$

Then,

- (i) every  $\mathbf{y} \in \mathcal{N}_M$  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 (B.1) 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  $\mathbf{y}$  such that  $M\mathbf{y} = \mathbf{0}$ . We rewrite the system as  $\widetilde{M}\boldsymbol{\theta} = \mathbf{0}$  by denoting  $\theta_j := y_{j+1} - y_j$ , where

$$\widetilde{M} := \begin{bmatrix} m_2 & & & m_N \\ m_1 & m_3 & & \\ & \ddots & \ddots & \\ & & m_{N-1} & m_1 \end{bmatrix}. \quad (\text{B.3})$$

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  $\mathbf{y}$  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  $\theta^*$  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.*,  $\{m_k\}_k \geq \min_{k \in \Omega_N} \bar{h}_k > 0$ . So, the quotient  $m_{k-1}/m_{k+1} \approx 1$  is also smooth for a small enough  $\Delta x$ . This implies that, like for  $\theta^*$ , the components  $y_k$  should have a CB-like structure.

Note that for such a  $\theta^*$  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 (B.2) is equivalent to the auxiliary system  $\widetilde{M}\theta = c$  as soon as  $\theta$  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  $\widetilde{M}\theta = c$  has the solution  $\theta^*$  such that for  $k \in \Omega_N$

$$\begin{cases} m_2\theta_1^* + m_N\theta_N^* = c \\ m_1\theta_1^* + m_3\theta_2^* = c \\ \vdots \\ m_{N-1}\theta_{N-1}^* + m_1\theta_N^* = c \end{cases} \implies \theta_{k+1}^* + \frac{m_k}{m_{k+2}}\theta_k^* = \frac{c}{m_{k+2}}. \quad (\text{B.4})$$

Then, making a spatial sum on  $\Omega_N$  for (B.4) 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 (B.4) as

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

So, one gets two different relations for  $\theta_N^*$ , which, indeed, should be the same:

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

It is not difficult to confirm that  $\prod_{j=1}^{N-1} \vartheta_j = \vartheta_N^{-1}$ ; so

$$\theta_1^* = \frac{\tilde{c}_N}{2} - \frac{\vartheta_N}{2} \sum_{j=1}^{N-1} (-1)^j \tilde{c}_j \prod_{\ell=j+1}^{N-1} \vartheta_\ell. \quad (\text{B.5})$$

In equation (B.5), since  $\{\tilde{c}_k\}_k$  and  $\{\vartheta_k\}_k$  vary smoothly, there are some cancellations for the second term, which suggests that it is of  $\mathcal{O}(\Delta x)$ . A similar procedure for every  $k \in \Omega_N$  shows that the leading order of  $\theta_k^*$  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  $\theta^*$ , one gets

$$\sum_{k \in \Omega_N} \frac{m_k}{m_{k+2}} \theta_k^* = \sum_{(k=2j) \in \Omega_N} \frac{c_e}{m_{k+2}} + \sum_{(k=2j+1) \in \Omega_N} \frac{c_o}{m_{k+2}}, \quad (\text{B.6})$$

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  $c_e = c_o = 0$  should hold which matches the statement of the lemma. Consider the same definition of  $\vartheta_k$  and  $\tilde{c}_k$  as above, with  $c_e = -c_o = c$ . So,

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

One gets two different relations for  $\theta_N^*$ , which should be the same:

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

Because  $\prod_{j=1}^{N-1} \vartheta_j = \vartheta_N^{-1}$  and the second term has a different sign,  $\tilde{c}_k = 0$  for all  $k$ , *i.e.*,  $\mathbf{c} = \mathbf{0}$ .  $\square$

Lemma B.1 implies that the system (B.1) or (B.2) is consistent only if  $\mathbf{w}_{2,\text{odd}}^{(0)}, \mathbf{w}_{2,\text{even}}^{(0)} = \mathbf{0}$ . Also, it confirms that  $R_b \mathbf{w}_1 \rightarrow 0$  if and only if  $\mathbf{w}_1$  tends to a vector with the CB-like structure. The relation (3.15a) shows that the mean of  $\mathbf{w}_1^{(0)}$  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  $\mathbf{w}_1^{(0)}$  should have different signs. This concludes that  $\mathbf{w}_1^{(0)}$  is the zero vector.

## APPENDIX 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 (Def. 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.*,

$$\begin{aligned} z_{0,\varepsilon} &= z_{(0)}^0 + \varepsilon z_{(1),\varepsilon}^0, \\ m_{0,\varepsilon} &= m_{(0)}^0 + \varepsilon m_{(1),\varepsilon}^0, \end{aligned} \quad (\text{C.1})$$

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 (3.2) that the splitting is still admissible in the sense of [62]. Also, without scaling the perturbation, we pick  $\mathbf{V} = \mathbf{U}_{\text{pert}}$ .

At first we show the  $\varepsilon$ -stability of the updated solution to justify the use of asymptotic expansion. From the definition of  $\tilde{\mathbf{F}}$  and  $\hat{\mathbf{F}}$ , one can simply check that the intermediate step solution is  $\varepsilon$ -stable as the pressure term

$v_1^2/2\varepsilon^2$  is  $\mathcal{O}(1)$ , owing to (C.1); this implies that  $\|\mathbf{V}_\Delta^{n+1/2}\| = \mathcal{O}(1)$ . Since  $J_\varepsilon^{-1}$  is  $\varepsilon$ -stable (with similar arguments as in Sect. 3.2.2),  $\|\mathbf{V}_\Delta^{n+1}\| = \mathcal{O}(1)$  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  $\nabla_h(\bar{h}v_{1,i}^{n+1}) = \mathcal{O}(\varepsilon^2)$ . This, combined with  $\|\mathbf{V}_{1,\Delta}^{n+1/2}\| = \mathcal{O}(\varepsilon)$  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  $\|\mathbf{V}_{1,\Delta}^{n+3/2}\| = \mathcal{O}(\varepsilon^2)$ , 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 [13] 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 (C.1).

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

## REFERENCES

- [1] S. Abarbanel, P. Duth and D. Gottlieb, Splitting methods for low Mach number Euler and Navier–Stokes equations. *Comput. Fluids* **17** (1989) 1–12.
- [2] K.R. Arun and S. Noelle, An Asymptotic Preserving Scheme for Low Froude Number Shallow Flows. IGPM Report 352. RWTH Aachen University (2012).
- [3] U.M. Ascher, S.J. Ruuth and R.J. Spiteri, Implicit-explicit Runge–Kutta methods for time-dependent partial differential equations. *Appl. Numer. Math.* **25** (1997) 151–167.
- [4] I. Bendixson, Sur les racines d'une équation fondamentale. *Acta Math.* **25** (1902) 359–365.
- [5] D.S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, Princeton, NJ (2009).
- [6] G. Bispen, *IMEX finite volume methods for the shallow water equations*. Ph.D. thesis, Johannes Gutenberg-Universität Mainz (2015).
- [7] G. Bispen, K.R. Arun, M. Lukáčová-Medvid'ová and S. Noelle, IMEX large time step finite volume methods for low Froude number shallow water flows. *Commun. Comput. Phys.* **16** (2014) 307–347.
- [8] G. Bispen, M. Lukáčová-Medvid'ová and L. Yelash, Asymptotic preserving IMEX finite volume schemes for low mach number Euler equations with gravitation. *J. Comput. Phys.* **335** (2017) 222–248.
- [9] S. Boscarino and G. Russo, On a class of uniformly accurate IMEX Runge–Kutta schemes and applications to hyperbolic systems with relaxation. *SIAM J. Sci. Comput.* **31** (2009) 1926–1945.
- [10] F. Bouchut, *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws: And Well-Balanced Schemes for Sources*. Springer Science & Business Media, Berlin (2004).
- [11] F. Bouchut and M. Westdickenberg, Gravity driven shallow water models for arbitrary topography. *Commun. Math. Sci.* **2** (2004) 359–389.
- [12] D. Bresch, R. Klein and C. Lucas, Multiscale analyses for the shallow water equations. In: *Computational Science and High Performance Computing IV*. Springer, Berlin (2011) 149–164.
- [13] R.E. Caflisch, S. Jin and G. Russo, Uniformly accurate schemes for hyperbolic systems with relaxation. *SIAM J. Numer. Anal.* **34** (1997) 246–281.
- [14] F. Cordier, P. Degond and A. Kumbaro, An asymptotic-preserving all-speed scheme for the Euler and Navier–Stokes equations. *J. Comput. Phys.* **231** (2012) 5685–5704.
- [15] R. Danchin, Low Mach number limit for viscous compressible flows. *ESAIM: M2AN* **39** (2005) 459–475.
- [16] P. Degond and M. Tang, All speed scheme for the low Mach number limit of the isentropic Euler equation. *Commun. Comput. Phys.* **10** (2011) 1–31.
- [17] P. Degond, J.-G. Liu and M.-H. Vignal, Analysis of an asymptotic preserving scheme for the Euler–Poisson system in the quasineutral limit. *SIAM J. Numer. Anal.* **46** (2008) 1298–1322.
- [18] S. Dellacherie, Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number. *J. Comput. Phys.* **229** (2010) 978–1016.
- [19] S. Dellacherie, P. Omnes and F. Rieper, The influence of cell geometry on the Godunov scheme applied to the linear wave equation. *J. Comput. Phys.* **229** (2010) 5315–5338.
- [20] G. Dimarco, R. Loubère and M.-H. Vignal, Study of a new asymptotic preserving scheme for the Euler system in the low Mach number limit. *SIAM J. Sci. Comput.* **39** (2017) A2099–A2128.

- [21] D. Drikakis and W. Rider, High-resolution Methods for Incompressible and Low-speed Flows. Springer Science & Business Media, Berlin (2006).
- [22] B. Fedele and C. Negulescu, Numerical study of an anisotropic Vlasov equation arising in plasma physics. Preprint arXiv:[1610.01592](https://arxiv.org/abs/1610.01592) (2016).
- [23] E. Feireisl and A. Novotný, Singular Limits in Thermodynamics of Viscous Fluids. Springer Science & Business Media, Berlin (2009).
- [24] E. Feireisl, M. Lukáčová-Medvid'ová, v. Nečasová, A. Novotný, and B. She, Asymptotic preserving error estimates for numerical solutions of compressible Navier–Stokes equations in the low Mach number regime. *Multiscale Model. Simul.* **16** (2018) 150–183.
- [25] M. Feistauer, V. Dolejší and V. Kučera, On the discontinuous Galerkin method for the simulation of compressible flow with wide range of Mach numbers. *Comput. Visual. Sci.* **10** (2007) 17–27.
- [26] F. Filbet and S. Jin, A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources. *J. Comput. Phys.* **229** (2010) 7625–7648.
- [27] T. Gallouët, R. Herbin, D. Maltese and A. Novotný, mplicit MAC scheme for compressible Navier–Stokes equations: low Mach asymptotic error estimates. Preprint [hal-01462822](https://arxiv.org/abs/1701.01462) (2017).
- [28] J. Giesselmann, Low Mach asymptotic-preserving scheme for the Euler–Korteweg model. *IMA J. Numer. Anal.* **35** (2015) 802–833.
- [29] G.H. Golub and C.F. Van Loan, Matrix Computations. JHU Press, Baltimore, MD **3** (2012).
- [30] R.M. Gray, Toeplitz and Circulant Matrices: A Review. Now Publishers Inc., Boston, MA (2006).
- [31] H. Guillard and A. Murrone, On the behavior of upwind schemes in the low Mach number limit: II. Godunov type schemes. *Comput. Fluids* **33** (2004) 655–675.
- [32] H. Guillard and C. Viozat, On the behaviour of upwind schemes in the low Mach number limit. *Comput. Fluids* **28** (1999) 63–86.
- [33] J. Haack, S. Jin, and J.-G. Liu, An all-speed asymptotic-preserving method for the isentropic Euler and Navier–Stokes equations. *Commun. Comput. Phys.* **12** (2012) 955–980.
- [34] E. Hairer and G. Wanner, Solving ordinary differential equations. II, Stiff and differential-algebraic problems. 2nd edition. In: vol. 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin (1996).
- [35] A. Hildebrand and S. Mishra, Efficient computation of all speed flows using an entropy stable shock-capturing space-time discontinuous Galerkin method. In: *Non-Linear Partial Differential Equations, Mathematical Physics, and Stochastic Analysis*. European Mathematical Society (2018) 287–318.
- [36] M.A. Hirsch, Sur les racines d'une équation fondamentale. *Acta Math.* **25** (1902) 367–370.
- [37] R.A. Horn and C.R. Johnson, Matrix Analysis. Cambridge University Press, New York, NY, USA (1986).
- [38] R.A. Horn and C.R. Johnson, Topics in Matrix Analysis. Cambridge UP, New York, NY (1991).
- [39] J. Hu, S. Jin and Q. Li, Asymptotic-preserving schemes for multiscale hyperbolic and kinetic equations. In: *Handbook of Numerical Analysis* (2016).
- [40] S. Jin, Runge–Kutta methods for hyperbolic conservation laws with stiff relaxation terms. *J. Comput. Phys.* **122** (1995) 51–67.
- [41] S. Jin, Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM J. Sci. Comput.* **21** (1999) 441–454.
- [42] S. Jin, Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review. In: *Lecture Notes for Summer School on “Methods and Models of Kinetic Theory” (M&MKT), Porto Ercole (Grosseto, Italy)* (2010) 177–216.
- [43] K. Kaiser, J. Schütz, R. Schöbel and S. Noelle, A new stable splitting for the isentropic Euler equations. *J. Sci. Comput.* **70** (2017) 1390–1407.
- [44] S. Klainerman and A. Majda, Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids. *Commun. Pure Appl. Math.* **34** (1981) 481–524.
- [45] S. Klainerman and A. Majda, Compressible and incompressible fluids. *Commun. Pure Appl. Math.* **35** (1982) 629–651.
- [46] A. Klar, A numerical method for nonstationary transport equations in diffusive regimes. *Transp. Theory Stat. Phys.* **27** (1998) 653–666.
- [47] R. Klein, Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics I: one-dimensional flow. *J. Comput. Phys.* **121** (1995) 213–237.
- [48] R. Klein, N. Botta, T. Schneider, C.-D. Munz, S. Roller, A. Meister, L. Hoffmann and T. Sonar, Asymptotic adaptive methods for multi-scale problems in fluid mechanics. In: *Practical Asymptotics*. Springer, Berlin (2001) 261–343.
- [49] E.W. Larsen, J.E. Morel and W.F. Miller, Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. *J. Comput. Phys.* **69** (1987) 283–324.
- [50] N. Masmoudi, Examples of singular limits in hydrodynamics. In: Vol. 3 of *Handbook of Differential Equations: Evolutionary Equations* (2007) 195–275.
- [51] H. Nessyahu and E. Tadmor, The convergence rate of approximate solutions for nonlinear scalar conservation laws. *SIAM J. Numer. Anal.* **29** (1992) 1505–1519.
- [52] S. Noelle, G. Bispen, K.R. Arun, M. Lukáčová-Medvid'ová and C.-D. Munz, A weakly asymptotic preserving low Mach number scheme for the Euler equations of gas dynamics. *SIAM J. Sci. Comput.* **36** (2014) B989–B1024.
- [53] L. Pareschi and G. Russo, Implicit-explicit Runge–Kutta schemes and applications to hyperbolic systems with relaxation. *J. Sci. Comput.* **25** (2005) 129–155.
- [54] M. Restelli, *Semi-Lagrangian and semi-implicit discontinuous Galerkin methods for atmospheric modeling applications*. Ph.D. thesis, Politecnico di Milano (2007).

- [55] R.D. Richtmyer and K.W. Morton, Difference Methods for Initial-value Problems. Interscience Publishers John Wiley & Sons, Inc., Academia Publishing House of the Czechoslovak Acad (1967).
- [56] F. Rieper, On the dissipation mechanism of upwind-schemes in the low Mach number regime: a comparison between Roe and HLL. *J. Comput. Phys.* **229** (2010) 221–232.
- [57] F. Rieper, A low-Mach number fix for Roe’s approximate Riemann solver. *J. Comput. Phys.* **230** (2011) 5263–5287.
- [58] F. Rieper and G. Bader, The influence of cell geometry on the accuracy of upwind schemes in the low Mach number regime. *J. Comput. Phys.* **228** (2009) 2918–2933.
- [59] H.H. Rosenbrock, Some general implicit processes for the numerical solution of differential equations. *Comput. J.* **5** (1963) 329–330.
- [60] S. Schochet, The mathematical theory of low Mach number flows. *ESAIM: M2AN* **39** (2005) 441–458.
- [61] J. Schütz and K. Kaiser, A new stable splitting for singularly perturbed ODEs. *Appl. Numer. Math.* **107** (2016) 18–33.
- [62] J. Schütz and S. Noelle, Flux splitting for stiff equations: a notion on stability. *J. Sci. Comput.* **64** (2015) 522–540.
- [63] J.R. Silvester, Determinants of block matrices. *Math. Gazette* **84** (2000) 460–467.
- [64] L.N. Trefethen, Finite Difference and Spectral Methods for Ordinary and Partial Differential Equations. Cornell University (1996).
- [65] R.F. Warming and B.J. Hyett, The modified equation approach to the stability and accuracy analysis of finite-difference methods. *J. Comput. Phys.* **14** (1974) 159–179.
- [66] D. Willett and J.S.W. Wong, On the discrete analogues of some generalizations of Gronwall’s inequality. *Monatsh. Math.* **69** (1965) 362–367.
- [67] H. Zakerzadeh, *Asymptotic preserving finite volume schemes for the singularly-perturbed shallow water equations with source terms*. PhD thesis, RWTH Aachen University (2017).
- [68] H. Zakerzadeh, On the Mach-uniformity of the Lagrange-projection scheme. *ESAIM: M2AN* **51** (2017) 1343–1366.
- [69] H. Zakerzadeh, The RS-IMEX Scheme for the Rotating Shallow Water Equations with the Coriolis Force. Springer International Publishing, Cham (2017) 199–207.
- [70] H. Zakerzadeh, Asymptotic Consistency of the RS-IMEX Scheme for the low-Froude Shallow Water Equations: Analysis and Numerics. Springer International Publishing, Cham (2018) 665–675.
- [71] H. Zakerzadeh and S. Noelle, A note on the stability of implicit-explicit flux-splittings for stiff systems of hyperbolic conservation laws. *Commun. Math. Sci.* **16** (2018) 1–15.
- [72] M. Zakerzadeh and G. May, On the convergence of a shock capturing discontinuous Galerkin method for nonlinear hyperbolic systems of conservation laws. *SIAM J. Numer. Anal.* **54** (2016) 874–898.