PREPRINT 455

igpm | RWTH AACHEN UNIVERSITY

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

Hamed Zakerzadeh†

Institut für Geometrie und Praktische Mathematik
Templergraben 55, 52062 Aachen, Germany

JULY 2016

* The author's research was supported by the scholarship of RWTH Aachen university through Graduiertenförderung nach Richtlinien zur Förderung des wissenschaftlichen Nachwuchses (RFwN).
† Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55, 52056 Aachen, Germany

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

HAMED ZAKERZADEH†

Abstract. In this article, we analyze a recently-presented scheme for singularly-perturbed systems of balance laws, the so-called Reference Solution Implicit Explicit scheme. RS-IMEX scheme's bottom-line is to use the Taylor expansion of the flux function and the source term around a reference solution (typically the asymptotic limit or an equilibrium solution) to decompose the flux and the source into stiff and non-stiff parts so that the resulting IMEX scheme is Asymptotic Preserving (AP) w.r.t. the singular parameter ϵ as $\epsilon \to 0$. After a brief introduction to the scheme, we prove the asymptotic consistency, asymptotic ℓ_2-stability, solvability and well-balancing of the scheme for the case of the one-dimensional shallow water equations and with two reference solutions (the lake at rest and the zero-Froude limit). Thus, the scheme is AP and can be used for flows with various Froude numbers. Finally we will test the scheme numerically for several test cases to show the quality of the solutions and confirm the analysis.

Key words. IMEX scheme, Asymptotic preserving, Flux splitting, Stability analysis

AMS subject classifications. 35L65, 65M08, 35L81, 65M12

1. Introduction.

Singular limits of conservation laws (or more generally PDEs), 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 [46], e.g., when the Mach number approaches zero for the Euler equations. This is a singular limit, since the sound speed (the characteristic speed) goes to the infinity and the PDE changes to be hyperbolic-elliptic, in the so-called incompressible limit. So, the convergence of the solution of the compressible Euler equations to the incompressible one is not straightforward to be shown (see [34,35] for the first justification using the theory of singular limits of symmetric hyperbolic systems [44]). We also refer the reader to consult with [10, 45, 56] to review the results for the compressible-incompressible limit, and with [46] for a nice review of different examples of singular limits in hydrodynamics.

Tackling such singular problems numerically is also complicated. For example in the case of the compressible-incompressible limit problem, the Courant–Friedrichs–Lewy (CFL) condition restricts the time step non-uniformly such that it should tend to zero, i.e., $\Delta t \lesssim \epsilon\, \Delta x$. This leads to very small time steps thus huge computational cost. Generally speaking, the usual numerical schemes also lose their accuracy in the limit for under-resolved mesh sizes; see [13, 14, 22, 23, 51–53].

In the sequel and for the sake of simplicity, we only consider well-prepared initial data to eliminate spurious initial layers (see Definition 3.2 and [21, 38, 47]). We also assume that the *solution* of the PDE with the singular parameter ϵ converges to the *solution* of the limit PDE as $\epsilon \to 0$, and aim to show that the counterpart of such a convergence exists at the discrete level. This is in fact the idea of Asymptotic Preserving (AP) schemes, which has been introduced by Jin in [29, 30] for relaxation systems; see also [31] for a general review and [39] for older works (without being named AP). Figure 1 illustrates this definition; $\mathcal{M}^\epsilon$ stands for a continuous physical model with the (singular) perturbation parameter $\epsilon \in (0, 1]$, and $\mathcal{M}^\epsilon_\Delta$ is a discrete-level model which provides a consistent discretization of $\mathcal{M}^\epsilon$. As in [31], if $\mathcal{M}^0_\Delta$ is a

*This work was supported by the scholarship of RWTH Aachen university through *Graduiertenförderung nach Richtlinien zur Förderung des wissenschaftlichen Nachwuchses (RFwN).*

†Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55, 52056 Aachen, Germany. (h.zakerzadeh@igpm.rwth-aachen.de).

suitable and *efficient* scheme for $\mathcal{M}^0$, then the scheme is called to be AP.

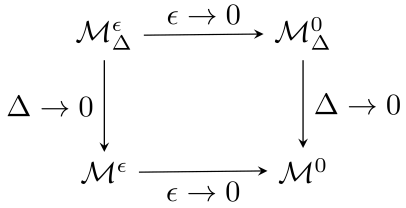


Fig. 1: Illustration of Asymptotic Preserving schemes.

Different interpretations of a *suitable* and *efficient* scheme give rise to various definitions of an AP scheme. So, we define an AP scheme for the framework of this article more precisely.

DEFINITION 1.1. *[AP schemes] A scheme is called to be AP, provided that the following conditions are fulfilled for the scheme.*

(i) It gives a consistent discretization of $\mathcal{M}^\epsilon$ for all ϵ, in particular for the limit problem $\mathcal{M}^0$.

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

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

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

REMARK 1.2. *As mentioned in [31], the asymptotic consistency suggests that the solution belongs to a manifold which is driven to the limit manifold as $\epsilon \to 0$ (up to some discretization error). For instance the velocity space of the solutions of the weakly compressible Euler equations should be in an $\mathcal{O}(\epsilon)$-neighborhood of the div-free (solenoidal) manifold, though there is a discretization error for the limit velocity itself since $\Delta > 0$.*

The AP property has been studied extensively for conservation laws as well as kinetic equations, and several AP schemes have been developed and analyzed; just to name a few see [6,9,12,15,18,24,28,32,40,43,48]. There are also several related works without using the initialism AP; see [36] as one of the first examples for the Euler equations and the review [37]. Note that while the uniform (asymptotic) consistency of the scheme is often studied and proved in the literature, particularly in the context of conservation laws there are only few results regarding the uniform (asymptotic) stability like [15, 62] for the isentropic Euler equations; see also [11] and [16] for the Euler–Poisson and Euler–Korteweg systems, respectively.

The bottom-line of these AP schemes is a mixed implicit-explicit (IMEX) strategy, e.g., to split the flux (or its Jacobian) into two parts and treat one part explicitly in time and the other one implicitly in time. This approach is definitely necessary to find schemes with ϵ-uniform CFL conditions; but, not sufficient at all for asymptotic stability; see for example [1] where it is shown that for an Explicit-Explicit splitting with the Lax–Wendroff scheme, even if both split parts are stable in terms of CFL condition, the resulting scheme is unconditionally unstable in L_2-norm. On the other hand, IMEX schemes are L_2-stable as long as each step is so, as shown in [24]. Thus, there is a huge gap between these two cases.

In [48], the authors applied a flux-splitting scheme to the full Euler equations, which uses a variant of Klein's auxiliary splitting [36]. The scheme required an ϵ-

dependent time step for stability. This motivated the authors of [58] to begin a stability analysis of the modified equation of linearized systems in Fourier variables, which identified the importance of the commutator of the stiff and non-stiff flux Jacobian matrices (see also [63] for a generalization of the analysis). That investigation leads Noelle and his collaborators to the main idea of the RS-IMEX scheme [49] whose rigorous asymptotic analysis is the core topic of this paper.

The key idea is the linearization around an asymptotic reference solution, which results in a particularly small slow Jacobian $\widehat{A}$, thus makes the commutator of the stiff and non-stiff Jacobian matrices small such that the modified equation can be shown to be stable. This heuristic argument makes the foundation of the RS-IMEX scheme [49]; see also [63] for a detailed discussion about the stability in the sense of modified equations. Note that in the work of our collaborators the RS-IMEX scheme is shown to be quantitatively well-behaved in practice; see [57] and [33] for the application of the scheme to the Van der Pol equation and the two-dimensional isentropic Euler system, respectively.

In the present article, we restrict our attention to the rigorous AP analysis for the case of one-dimensional shallow water system, i.e., asymptotic consistency, asymptotic stability and convergence to the limit for fixed grids (see Remark 3.12). These make a solid background for the future works which extend the scheme to the multi-dimensional shallow water system with different source terms; see [61] for instance. Note that broadly speaking, the splitting developed in [6] can also be considered as a particular example of the RS-IMEX scheme, with the lake at rest reference solution; see Section 3 in general and Remark 3.5 in particular.

The remainder of this paper is organized as follows. In Section 2 we present a short introduction to the RS-IMEX scheme, which follows in Section 3 and Section 4 with the rigorous AP analysis (consistency and stability) of the RS-IMEX scheme for the one-dimensional shallow water equations, with the lake at rest and the zero-Froude limit solution as the reference solution. We see that although the reference solution is rather simple, the rigorous analysis is not too straightforward. Section 5 provides some numerical evidences to confirm the AP analysis and test the quality of the solutions. The results of this manuscript supply some necessary elements for the treatment of the more interesting case of the two-dimensional shallow water equations in [61].

Acknowledgment. The author acknowledges the discussions and collaborations of Arun K.R., Georgij Bispen, Klaus Kaiser, Rupert Klein, Mária Lukáčová-Medviďová, Claus-Dieter Munz, Sebastian Noelle and Jochen Schütz, leading to the RS-IMEX approach. Also he would like to gratefully thank Negin Bagherpour and Mohammad Zakerzadeh for very useful discussions regarding Section 3.2.

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 [49] (to be applied to the shallow water equations in Section 3). Let us consider the general hyperbolic system of balance laws

$$\partial_t \mathbf{U}(\mathbf{x},t;\epsilon) + \nabla_{\mathbf{x}|t} \cdot \mathbf{F}(\mathbf{U},\mathbf{x},t;\epsilon) = \mathbf{S}(\mathbf{U},\mathbf{x},t;\epsilon), \tag{2.1}$$

where $\nabla_{\alpha|\beta}$ denotes the partial derivative with respect to α when β is fixed, and

$$\begin{aligned} &\mathbf{U} : \Omega \times \mathbb{R}_+ \times (0,1] \to \mathbb{R}^n, \\ &\mathbf{F} : \mathbb{R}^n \times \Omega \times \mathbb{R}_+ \times (0,1] \to \mathbb{R}^{n\times d}, \\ &\mathbf{S} : \mathbb{R}^n \times \Omega \times \mathbb{R}_+ \times (0,1] \to \mathbb{R}^n, \end{aligned} \tag{2.2}$$

where $\mathbf{U}$ is the vector of conserved quantities, $\mathbf{F}$ is the flux matrix (in d space dimensions), $\epsilon \in (0,1]$ is the singular parameter (e.g., the Froude or Mach number), and $\mathbf{S}$ 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 ϵ as well as β in $\nabla_{\alpha|\beta}$. Ω is a subspace of $\mathbb{R}^d$, usually be chosen to be periodic (a torus), i.e., $\Omega := \mathbb{T}^d$. To have a hyperbolic system, we also assume that $\mathbf{F}$ has a real diagonalizable Jacobian

$$\mathbf{F}'(\mathbf{U}, \mathbf{x}, t) := \partial_{\mathbf{U}|\mathbf{x},t}\mathbf{F}(\mathbf{U}, \mathbf{x}, t).$$

Let us consider the given function $\overline{\mathbf{U}}$ as the *reference solution*:

$$\overline{\mathbf{U}} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}^n, \quad (\mathbf{x}, t) \mapsto \overline{\mathbf{U}}(\mathbf{x}, t). \tag{2.3}$$

Typically, it is a steady state solution of the balance law, or the solution of the asymptotic limit equation, derived from (2.1) as $\epsilon \to 0$, e.g., it can be the *lake at rest* (LaR) state for the shallow water equations or the incompressible limit for the Euler equations.

Given the reference solution, we split the solution $\mathbf{U}$ of the balance law (2.1) into the reference solution $\overline{\mathbf{U}}$ and a perturbation $\mathbf{U}_{pert}$,

$$\mathbf{U}(\mathbf{x}, t; \epsilon) = \overline{\mathbf{U}}(\mathbf{x}, t; \epsilon) + \mathbf{U}_{pert}(\mathbf{x}, t; \epsilon). \tag{2.4}$$

Our goal is to design an algorithm for the perturbation $\mathbf{U}_{pert}$ which is asymptotically stable and consistent. The algorithm uses the IMEX approach and the CFL number for the explicit part which shall be independent of the small parameter ϵ.

For flux-splitting, we use the Taylor expansion of $\mathbf{F}$ around $\overline{\mathbf{U}}(\mathbf{x}, t)$:

$$\begin{aligned} \mathbf{F}(\mathbf{U}, \mathbf{x}, t) &= \mathbf{F}(\overline{\mathbf{U}}, \mathbf{x}, t) + \mathbf{F}'(\overline{\mathbf{U}}, \mathbf{x}, t)\, \mathbf{U}_{pert} + \widehat{\mathbf{F}}(\overline{\mathbf{U}}, \mathbf{U}_{pert}, \mathbf{x}, t) \\ &= \overline{\mathbf{F}} + \widetilde{\mathbf{F}} + \widehat{\mathbf{F}}, \end{aligned} \tag{2.5}$$

where we have used the shortcuts

$$\begin{aligned} \overline{\mathbf{F}} &:= \mathbf{F}(\overline{\mathbf{U}}(\mathbf{x}, t), \mathbf{x}, t), \\ \widetilde{\mathbf{F}} &:= \mathbf{F}'(\overline{\mathbf{U}}(\mathbf{x}, t), \mathbf{x}, t)\, \mathbf{U}_{pert}, \\ \widehat{\mathbf{F}} &:= \mathbf{F}(\overline{\mathbf{U}}(\mathbf{x}, t) + \mathbf{U}_{pert}, \mathbf{x}, t) - \overline{\mathbf{F}} - \widetilde{\mathbf{F}}. \end{aligned}$$

Analogously, one can do the expansion for the source term to get the following splitting with similar definitions.

$$\mathbf{S}(\mathbf{U}, \mathbf{x}, t) = \overline{\mathbf{S}} + \widetilde{\mathbf{S}} + \widehat{\mathbf{S}}. \tag{2.6}$$

It may be useful to scale the components of the perturbed solution, since by the suitable scaling one can work with $\mathcal{O}(1)$ terms in the analysis of the scheme. Later on in Section 3, we see that a physically appropriate choice of the scaling matrix, not

only makes the analysis more illustrative (see Remark 3.11) but also may affect the numerical solution (see Remark 4.4). For this reason, we introduce the diagonal matrix $D := \mathrm{diag}(\epsilon^{d_j})$ with $j = 1, \dots, n$ and define the scaled (preconditioned) perturbed solution vector $\mathbf{V}(\mathbf{x}, t)$ as $\mathbf{V} := D^{-1}\,\mathbf{U}_{pert}$ and denote the corresponding scaled flux and source terms by

$$\mathbf{G}(\overline{\mathbf{U}}, \mathbf{V}, \mathbf{x}, t) := D^{-1}\mathbf{F}\big(\overline{\mathbf{U}}(\mathbf{x}, t) + D\mathbf{V}(\mathbf{x}, t)\big), \tag{2.7}$$

$$\mathbf{Z}(\overline{\mathbf{U}}, \mathbf{V}, \mathbf{x}, t) := D^{-1}\mathbf{S}\big(\overline{\mathbf{U}}(\mathbf{x}, t) + D\mathbf{V}(\mathbf{x}, t)\big). \tag{2.8}$$

So, with $\overline{\mathbf{G}}, \widetilde{\mathbf{G}}, \widehat{\mathbf{G}}, \overline{\mathbf{Z}}, \widetilde{\mathbf{Z}}$ and $\widehat{\mathbf{Z}}$ defined analogously as for $\mathbf{F}$ and $\mathbf{S}$, the splittings (2.5) and (2.6) can be re-written as

$$\mathbf{G}(\overline{\mathbf{U}}, \mathbf{V}, \mathbf{x}, t) = \overline{\mathbf{G}} + \widetilde{\mathbf{G}} + \widehat{\mathbf{G}},$$
$$\mathbf{Z}(\overline{\mathbf{U}}, \mathbf{V}, \mathbf{x}, t) = \overline{\mathbf{Z}} + \widetilde{\mathbf{Z}} + \widehat{\mathbf{Z}}.$$

REMARK 2.1. *Note that the eigenvalues of $\widetilde{\mathbf{F}}'$ and $\widehat{\mathbf{F}}'$ are exactly the same as the eigenvalues of $\widetilde{\mathbf{G}}'$ and $\widehat{\mathbf{G}}'$, respectively.*

Defining $\mathbf{R} := -\nabla \cdot \mathbf{G} + \mathbf{Z}$ (with analogous definitions for $\overline{\mathbf{R}}$, $\widetilde{\mathbf{R}}$ and $\widehat{\mathbf{R}}$), and also $\overline{\mathbf{T}}(\mathbf{x}, t)$ as the (a priori-known) scaled model truncation error of the reference solution

$$\overline{\mathbf{T}} := D^{-1}\partial_t \overline{\mathbf{U}} - \overline{\mathbf{R}}, \tag{2.9}$$

one can reformulate the original balance law (2.1) as

$$\partial_t \mathbf{V} = -\overline{\mathbf{T}} + \widetilde{\mathbf{R}} + \widehat{\mathbf{R}}, \tag{2.10}$$

which is a balance law for the scaled perturbed solution $\mathbf{V}$ around the reference solution $\overline{\mathbf{U}}$. Note that this reformulation is not necessary for the numerical scheme, but it is suitable notably for the asymptotic consistency analysis. Note also that $\overline{\mathbf{T}} \equiv 0$ if and only if the reference solution $\overline{\mathbf{U}}$ is an exact solution of (2.1). This may be the case, e.g., when the reference solution is a stationary solution of the system.

2.1. RS-IMEX scheme. In (2.5), $\widetilde{\mathbf{F}}$, and $\widetilde{\mathbf{G}}$ due to Remark 2.1, has stiff eigenvalues. So, to solve (2.10) numerically, it is reasonable to treat the stiff part $\widetilde{\mathbf{R}}$ implicitly in time to avoid restrictive time steps in the limit, e.g., by using the implicit Euler time integration. The term $\widehat{\mathbf{R}}$ is expected to be non-stiff; so, an explicit scheme (like explicit Euler scheme) is a suitable choice. The scaled model truncation error $\overline{\mathbf{T}}$, is computed independently. For instance for the Euler system and incompressible reference solution, an appropriate incompressible solver is employed to compute $\overline{\mathbf{T}}$. Thus, we can define the RS-IMEX scheme as follows.

DEFINITION 2.2. *Given the reference solution $\overline{\mathbf{U}}(\mathbf{x}, t)$, the RS-IMEX scheme for (2.10) is given by*

$$D_t \mathbf{V}_\Delta^n = -\overline{\mathbf{T}}_\Delta^{n+1} + \widetilde{\mathbf{R}}_\Delta^{n+1} + \widehat{\mathbf{R}}_\Delta^n, \tag{2.11}$$

where D_t and the subscript Δ stand respectively for a choice of discretization method in time and space.

From now on, we limit ourselves to the Rusanov-type numerical flux in space and the implicit/explicit Euler time integration. So, given a time step Δt and a vector $\phi(\mathbf{x}, t)$, the Euler time discretization of $\partial_t \phi$, denoted by $D_t \phi(\mathbf{x}, t)$, is defined as

$$D_t \phi(\mathbf{x}, t) := \frac{\phi(\mathbf{x}, t + \Delta t) - \phi(\mathbf{x}, t)}{\Delta t}. \tag{2.12}$$

For the spatial discretization in one dimension, the Rusanov-type flux function for the scalar flux $f(u)$ at the interface $i+1/2$ is defined as

$$f_{i+1/2} := \frac{f(u_i)+f(u_{i+1})}{2} - \frac{\alpha_{i+1/2}}{2}\left(u_{i+1}-u_i\right), \tag{2.13}$$

when α is originally chosen such that $\alpha_{i+1/2} \geq \max_{u\in[u_i,u_{i+1}]}\left(f'(u)\right)$. However in this paper, we choose α for the stiff sub-system rather arbitrary not to make the scheme too dissipative. The extension of this numerical flux to systems and/or in multi-dimensions is obvious. Note that the residual includes also a source term, which should be discretized appropriately so that the scheme preserves the equilibrium (C-property or well-balancing, cf. [8]). Since this property depends heavily on the structure of the system being studied, we postpone it to Section 3 where we discuss the RS-IMEX scheme for the shallow water equations with topography.

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

Algorithm 1 RS-IMEX scheme

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

In the next section, we apply the RS-IMEX algorithm to the one-dimensional shallow water system. The ultimate goal is to check if the stability and consistency of the scheme are uniform in ϵ, under a CFL condition independent of the maximal eigenvalue of the stiff flux Jacobian $\widetilde{\mathbf{G}}'$.

3. Shallow water equations with the LaR as reference solution. In this section and as an example of the RS-IMEX scheme for the general balance law (2.1), we follow the procedure described in Section 2 to derive the scheme for one-dimensional shallow water equation with topography. Then we discuss solvability and well-balancing for the lake at rest equilibrium state, as well as asymptotic consistency and convergence to the limit. We also show that the scheme is asymptotically stable in ℓ_2-norm. Note that we present these analyses with the lake at rest reference solution. Then in Section 4 we consider the reference state to be the zero-Froude limit solution. From now on, we assume the spatial domain Ω to be periodic unless stated otherwise.

The non-dimensionalized shallow water equations in one space dimension, using $z = h + b$ (with $b < 0$) and $m = hu$, can be written as has been proposed in [6]:

$$\mathbf{U} = \begin{bmatrix} z \\ m \end{bmatrix}, \quad \mathbf{F}(\mathbf{U}) = \begin{bmatrix} m \\ \dfrac{m^2}{z-b} + \dfrac{z^2-2zb}{2\epsilon^2} \end{bmatrix}, \quad \mathbf{S}(\mathbf{U}) = \begin{bmatrix} 0 \\ -\dfrac{zb_x}{\epsilon^2} \end{bmatrix}. \tag{3.1}$$

In this notation, z is the surface elevation from some chosen constant surface level H_{ref}, m is the momentum and b is the water depth measured from H_{ref} with a negative sign (see Figure 2). The singular parameter $\epsilon \in (0,1]$ is called the Froude number for simplicity (though it is different from the Froude number by a constant factor, cf. [35]). We also consider the given $z^0(x) := z(x,0)$ and $m^0(x) := m(x,0)$ as initial conditions.

This formulation of the shallow water equations (with the opposite sign for the bottom topography) has been introduced at first in [54,55] to give a balanced system and circumvent the need to devise any specific source term discretization for well-balancing. Because (3.1) cannot be readily used for the cases involving wet-dry fronts, a modified (but similar) formulation has been introduced later on in [17,41,42].

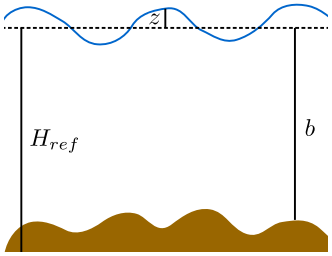


Fig. 2: Variables used in the shallow water formulation (3.1).

In this section, we set the reference state as the lake at rest, $\overline{\mathbf{U}} := (z_\circ, m_\circ)^T$ with $z_\circ$ constant in space and $m_\circ = 0$. Therefore, due to (2.5), the flux splitting is as:

$$\overline{\mathbf{F}} = \begin{bmatrix} 0 \\ \dfrac{1}{2\epsilon^2} z_\circ (z_\circ - 2b) \end{bmatrix}, \qquad \mathbf{F}'(\overline{\mathbf{U}}) = \begin{bmatrix} 0 & 1 \\ \dfrac{z_\circ - b}{\epsilon^2} & 0 \end{bmatrix},$$

$$\widetilde{\mathbf{F}} = \begin{bmatrix} m_{pert} \\ \dfrac{(z_\circ - b)}{\epsilon^2} z_{pert} \end{bmatrix}, \qquad \widehat{\mathbf{F}} = \begin{bmatrix} 0 \\ \dfrac{m_{pert}^2}{z_\circ + z_{pert} - b} + \dfrac{z_{pert}^2}{2\epsilon^2} \end{bmatrix}.$$

Owing to (2.6), the source term is split analogously as

$$\overline{\mathbf{S}} = \begin{bmatrix} 0 \\ -\dfrac{z_\circ b_x}{\epsilon^2} \end{bmatrix}, \qquad \mathbf{S}'(\overline{\mathbf{U}}) = \begin{bmatrix} 0 & 0 \\ -\dfrac{b_x}{\epsilon^2} & 0 \end{bmatrix},$$

$$\widetilde{\mathbf{S}} = \begin{bmatrix} 0 \\ -\dfrac{z_{pert} b_x}{\epsilon^2} \end{bmatrix}, \qquad \widehat{\mathbf{S}} = \mathbf{0}.$$

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

$$\widetilde{\mathbf{F}}' = \begin{bmatrix} 0 & 1 \\ \dfrac{z_\circ - b}{\epsilon^2} & 0 \end{bmatrix}, \qquad \widetilde{\mu} := \pm \frac{\sqrt{z_\circ - b}}{\epsilon},$$

$$\widehat{\mathbf{F}}' = \begin{bmatrix} 0 & 0 \\ -u_{pert}^2 + \dfrac{z_{pert}}{\epsilon^2} & 2u_{pert} \end{bmatrix}, \quad \widehat{\mu} = 0, 2\, u_{pert},$$

where $u_{pert} := \dfrac{m_{pert}}{z_\circ + z_{pert} - b}$. Thus, the splitting is admissible in the sense of [58].

REMARK 3.1. *Note that the RS-IMEX splitting with* $\overline{\mathbf{U}} = \mathbf{0}$ *gives the same splitting as in [5, 6].*

Before continuing, we should find the scaling matrix D by the asymptotic analysis of the system, which has been done in Appendix A. This analysis justifies the following definition for the formal asymptotic limit of the shallow water equations. We present it for one-dimensional case, however it can be generalized effortlessly to multi-dimensions (see [5, Sect. 2.3]).

DEFINITION 3.2. *The formal zero-Froude limit of the shallow water equations is defined as*

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

where $p(h) = \frac{h^2}{2}$ *is the (hydrostatic) pressure function and with the following asymptotic (Poincaré) expansion*

$$
\begin{aligned}
z(x,t) &= z_{(0)} + \epsilon z_{(1)} + \epsilon^2 z_{(2)},\\
m(x,t) &= m_{(0)} + \epsilon m_{(1)} + \epsilon^2 m_{(2)}.
\end{aligned}
\tag{3.2}
$$

Thus, the well-prepared initial data can be defined as

$$
\begin{aligned}
z^0_{\mathrm{WP}}(x) &:= z^0_{(0)} + \epsilon^2 z^0_{(2)}(x),\\
m^0_{\mathrm{WP}}(x) &:= m^0_{(0)} + \epsilon\, m^0_{(1)}(x).
\end{aligned}
\tag{3.3}
$$

where $z^0_{(0)}$ *and* $m^0_{(0)}$ *are constant values.*

The motivation for scaling the equations was to work with $\mathcal{O}(1)$ quantities. So, due to (3.3), we pick $z_\circ = z^0_{(0)}$, which implies $D := \mathrm{diag}(\epsilon^2, 1)$. From now on, such a scaling matrix is denoted by D_2. For the sake of simplicity, we stick to this particular choice of $z_\circ$ throughout this section. Nonetheless, it is rather straightforward to confirm that the asymptotic analysis we are going to present holds for every constant $z_\circ$, while the choice may affect the numerical solution for $\epsilon = \mathcal{O}(1)$.

REMARK 3.3. *Note that the analysis in Appendix A has been done for periodic domains. However, with some other boundary conditions such as open boundary conditions with a fast decaying momentum,* $h_{(0)}$ *and* $h_{(1)}$ *would have similar asymptotic expansion as the periodic case and* D_2 *is again a relevant scaling.*

3.1. RS-IMEX scheme. For the scaling matrix D_2, the scaled split fluxes and source terms read

$$
\widehat{\mathbf{G}} = \begin{bmatrix} 0 \\ \dfrac{v_2^2}{z_\circ + \epsilon^2 v_1 - b} + \epsilon^2 \dfrac{v_1^2}{2} \end{bmatrix}, \qquad \widetilde{\mathbf{G}} = \begin{bmatrix} \dfrac{v_2}{\epsilon^2} \\ (z_\circ - b) v_1 \end{bmatrix},
\tag{3.4}
$$

$$
\widehat{\mathbf{Z}} = \mathbf{0}, \qquad \widetilde{\mathbf{Z}} = \begin{bmatrix} 0 \\ -b_x v_1 \end{bmatrix},
\tag{3.5}
$$

where

$$\mathbf{V} := \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} := D_2^{-1}\mathbf{U}_{pert} = \begin{bmatrix} z_{pert}/\epsilon^2 \\ m_{pert} \end{bmatrix} \tag{3.6}$$

Since the lake at rest is a stationary solution of the system, $z_\circ$ is constant in time and the reference solution needs not to be updated. Thus $\overline{\mathbf{T}} \equiv 0$, and one can reformulate the one-dimensional shallow water equations as

$$\partial_t \mathbf{V} = -\partial_x \begin{bmatrix} v_2/\epsilon^2 \\ (z_\circ - b)v_1 \end{bmatrix} - \partial_x \begin{bmatrix} 0 \\ \dfrac{v_2^2}{z_\circ + \epsilon^2 v_1 - b} + \epsilon^2 \dfrac{v_1^2}{2} \end{bmatrix} + \begin{bmatrix} 0 \\ -b_x v_1 \end{bmatrix}. \tag{3.7}$$

This is the system that one solves by the RS-IMEX scheme, with unknowns v_1 and v_2. For the numerical scheme, one should treat the stiff parts of (3.7) implicitly and the rest explicitly. So, the RS-IMEX scheme for the shallow water equations with the lake at rest reference solution can be written as

$$\mathbf{V}^{n\dagger}_{\Delta,i} = \mathbf{V}^n_{\Delta,i} - \frac{\Delta t}{\Delta x}\left(\widehat{\mathbf{G}}^n_{i+1/2} - \widehat{\mathbf{G}}^n_{i-1/2}\right) + \Delta t\, \widehat{\mathbf{Z}}^n_i \qquad \text{Explicit step,} \tag{3.8}$$

$$\mathbf{V}^{n+1}_{\Delta,i} = \mathbf{V}^{n\dagger}_{\Delta,i} - \frac{\Delta t}{\Delta x}\left(\widetilde{\mathbf{G}}^{n+1}_{i+1/2} - \widetilde{\mathbf{G}}^{n+1}_{i-1/2}\right) + \Delta t\, \widetilde{\mathbf{Z}}^{n+1}_i \qquad \text{Implicit step,} \tag{3.9}$$

for each cell $i \in \{1, 2, \ldots, N\}$ in the computational domain Ω_N of size N, where $\widetilde{\mathbf{G}}_{i+1/2}$ and $\widehat{\mathbf{G}}_{i+1/2}$ denote the Rusanov flux (as defined in (2.13)) with $\widehat{\mathbf{G}}$ and $\widetilde{\mathbf{G}}$ as defined in (3.4), and $\widehat{\mathbf{Z}}^n_i$ and $\widehat{\mathbf{Z}}^n_i$ are central discretizations of the source terms in (3.5). One can re-write (3.8)-(3.9) as

$$\begin{aligned} \mathbf{V}^{n\dagger}_{\Delta,i} = \mathbf{V}^n_{\Delta,i} &- \frac{\Delta t}{2\Delta x}\left[\frac{v^{n,2}_{2,i+1}}{z_\circ + \epsilon^2 v^n_{1,i+1} - b_{i+1}} - \frac{v^{n,2}_{2,i-1}}{z_\circ + \epsilon^2 v^n_{1,i-1} - b_{i-1}} + \frac{\epsilon^2}{2}\left(v^{n,2}_{1.i-1} - v^{n,2}_{1.i-1}\right)\right] \\ &+ \frac{\Delta t}{2\Delta x}\left(\widehat{\alpha}_{i+1/2}\mathbf{V}^n_{\Delta,i+1} - (\widehat{\alpha}_{i+1/2} + \widehat{\alpha}_{i-1/2})\mathbf{V}^n_{\Delta,i} + \widehat{\alpha}_{i-1/2}\mathbf{V}^n_{\Delta,i-1}\right), \end{aligned} \tag{3.10}$$

$$\begin{aligned} \mathbf{V}^{n+1}_{\Delta,i} = \mathbf{V}^{n\dagger}_{\Delta,i} &- \frac{\Delta t}{2\Delta x}\begin{bmatrix} \left(v^{n+1}_{2,i+1} - v^{n+1}_{2,i-1}\right)/\epsilon^2 \\ (z_\circ - b_{i+1})\, v^{n+1}_{1,i+1} - (z_\circ - b_{i-1})\, v^{n+1}_{1,i-1} \end{bmatrix} \\ &+ \frac{\Delta t}{2\Delta x}\left(\widetilde{\alpha}_{i+1/2}\mathbf{V}^{n+1}_{\Delta,i+1} - (\widetilde{\alpha}_{i+1/2} + \widetilde{\alpha}_{i-1/2})\mathbf{V}^{n+1}_{\Delta,i} + \widetilde{\alpha}_{i-1/2}\mathbf{V}^{n+1}_{\Delta,i-1}\right) \\ &- \frac{\Delta t}{2\Delta x}\begin{bmatrix} 0 \\ v^{n+1}_{1,i}\,(b_{i+1} - b_{i-1}) \end{bmatrix}. \end{aligned} \tag{3.11}$$

Note that for simplicity, we have suppressed the subscript Δ for the components of $\mathbf{V}_\Delta$.

Due to Remark 2.1, the eigenvalues of $\mathbf{F}$ and $\mathbf{G}$ (and their splittings) are the same. From above, one can clearly see that the eigenvalues of the non-stiff systems are $\mathcal{O}(1)$, so it may not give a small commutator needed for the heuristic stability based on the modified equation. Indeed, the commutator can be obtained as

$$\left[\widetilde{\mathbf{G}}', \widehat{\mathbf{G}}'\right] := \widetilde{\mathbf{G}}'\widehat{\mathbf{G}}' - \widehat{\mathbf{G}}'\widetilde{\mathbf{G}}' = \begin{bmatrix} v_1 - \dfrac{v_2^2}{(z_\circ + \epsilon^2 v_1 - b)^2} & \dfrac{2v_2/\epsilon^2}{z_\circ + \epsilon^2 v_1 - b} \\ \dfrac{-2(z_\circ - b)v_2}{z_\circ + \epsilon^2 v_1 - b} & -v_1 + \dfrac{v_2^2}{(z_\circ + \epsilon^2 v_1 - b)^2} \end{bmatrix}, \tag{3.12}$$

which is formally $\mathcal{O}(1/\epsilon^2)$. However, as shown in [63], the modified equation is asymptotically stable. In the next section and in Theorem 3.4, we prove the asymptotic stability of the scheme rigorously.

3.2. Stability analysis. We collect the stability properties of the RS-IMEX scheme in the following theorem.

THEOREM 3.4. *For the shallow water equations with topography and well-prepared initial data in the sense of Definition 3.2, the RS-IMEX scheme (3.10)-(3.11)*

(i) is solvable, i.e., it has a unique solution for all $\epsilon > 0$, which does not blow-up for $\epsilon \to 0$, under the assumption of constant $\widetilde{\alpha}$.

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

(iii) is asymptotically ℓ_2-stable, i.e., there exists a constant $C_{N,T}$ depending on the final time $T = n\Delta t$ and the number of grid points N such that $\|\mathbf{V}_\Delta^n\|_{\ell_2} \leq C_{N,T}\|\mathbf{V}_\Delta^0\|_{\ell_2}$.

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

We present the proof of Theorem 3.4 in the next sections.

REMARK 3.5. *As we already mentioned. the scheme in [5, 6] can be considered as a particular example of the RS-IMEX scheme with zero reference solution. So, one may expect that the AP analysis in [5] coincides with Theorem 3.4. The first difference which should be taken into account is that the analysis of [5] is for the two-dimensional shallow water system with the elliptic approach (in which the surface perturbation is computed by an elliptic equation [5, eq. (7.5a)]). Moreover, the rigorous asymptotic consistency proof in [5] is basically for the flat bottom case and a detailed analysis has been done for various high order reconstructions. By contrast, throughout this paper, we focus on the first-order schemes in one space dimension and prove asymptotic consistency for an arbitrarily topography. In addition, we also analyze asymptotic stability. In Section 4 and later on in [61], we show respectively that a similar analysis can be used for more general reference solutions as well as two-dimensional problems.*

3.2.1. Solvability of the scheme. Here, we aim to show that the scheme has a unique solution for all $\epsilon > 0$. Also we show that the solution does not blow-up for small ϵ; this implies incidentally the validity of the formal asymptotic consistency analysis since the solution is bounded for small ϵ. We take two cases into consideration: At first for simplification, we assume $\widetilde{\alpha}$ and the topography b to be constant (which makes the system similar to the isentropic Euler system). Afterwards, we generalize the arguments for the shallow water equations with varying bottom but again for constant $\widetilde{\alpha}$. To simplify the notation, we define $h_\circ := z_\circ - b$ and $\beta := \frac{\Delta t}{2\Delta x}$.

(i) constant b and constant $\widetilde{\alpha}$. It is not difficult to observe in (3.11) that $J_\epsilon \mathbf{V}_\Delta^{n+1} = \mathbf{V}_\Delta^{n\dagger}$. So, the implicit solution operator is J_ϵ^{-1}, where matrix $J_\epsilon \in \mathbb{R}^{2N\times 2N}$ is defined as

$$J_\epsilon := \begin{bmatrix} P & \frac{\beta}{\epsilon^2}Q \\ \beta h_\circ Q & P \end{bmatrix} \tag{3.13}$$

and P and Q are circulant matrices defined as

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

Note that P and Q are symmetric and skew-symmetric respectively, and P is strictly diagonally dominant (SDD). Also note that the matrix Q is the companion matrix for the central discretization.

In the following, we first show that J_ϵ is non-singular, thus there exists a unique solution for the implicit step (and so for the scheme). Then, we prove that the solution operator and the solution itself, are bounded in terms of ϵ. We call such a property ϵ-*stability* hereinafter. Note that the ϵ-stability of the solution operator does not provide ϵ-stability of the solution per se. For that, one also needs the ϵ-stability of the explicit step at the intermediate time $n\dagger$. For now, we simply assume the ϵ-stability of the explicit step and we show it in Section 3.2.2.

Existence of J_ϵ^{-1}. Since P and Q are circulant, they commute [20], and one knows from [59, Thm. 1] (see [4, Sect. 2.14] for more general cases) that since all blocks of J_ϵ commute with each other, the determinant of J_ϵ can be computed as

$$\det(J_\epsilon) = \det\left(P^2 - \frac{h_\circ \beta^2}{\epsilon^2} Q^2\right).$$

Due to Gerschgorin's circle theorem [26, Chap. 6], the numerical range of $-\frac{h_\circ\beta^2}{\epsilon^2}Q^2$ is located in the right half-plane while of P^2 is strictly positive, and both these parts are symmetric with real eigenvalues. So, using the sub-additivity of numerical range (Rayleigh quotient), the eigenvalues of the sum are bounded away from zero. Thus J_ϵ is not singular, and there exists a unique solution for the scheme.

Boundedness of J_ϵ^{-1} *(ϵ-stability).* Circulant matrices are commutable, so equivalently they are simultaneously diagonalizable. Thus, one can write J_ϵ as

$$J_\epsilon = \operatorname{diag}(F_N, F_N)\, \Xi_\epsilon\, \operatorname{diag}(F_N^*, F_N^*), \tag{3.14}$$

where $*$ denotes the conjugate transpose, and F_N is a (unique) unitary matrix, which consists of eigenvectors of circulant matrices of size N. It is important to mention that F_N does not depend on the entries of J_ϵ; it only depends on the size N (see [20]). Ξ_ϵ denotes a matrix containing the diagonal part (eigenvalues) of blocks of J_ϵ:

$$\Xi_\epsilon := \begin{bmatrix} \Gamma & \frac{\beta}{\epsilon^2}\Lambda \\ \beta h_\circ \Lambda & \Gamma \end{bmatrix}. \tag{3.15}$$

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

$$\begin{aligned} \|J_\epsilon^{-1}\| &\le \|\operatorname{diag}(F_N, F_N)\| \|\operatorname{diag}(F_N^*, F_N^*)\| \|\Xi_\epsilon^{-1}\| \\ &\le \operatorname{cond}(\operatorname{diag}(F_N, F_N))\, \|\Xi_\epsilon^{-1}\|, \end{aligned}$$

for a suitable natural matrix norm. This bound depends on ϵ only through $\|\Xi_\epsilon^{-1}\|$, so in the following lemma we show that Ξ_ϵ^{-1} is uniformly bounded in ϵ.

LEMMA 3.6. *The inverse of matrix* Ξ_ϵ*, has a bounded norm in terms of* ϵ.

Before we prove this lemma, 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 in the linear algebra; for example the reader can find it in [4, Prop. 2.8.7].

LEMMA 3.7 (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 it writes*

(3.16)
$$M^{-1} = \begin{bmatrix} \left(M_{11} - M_{12}M_{22}^{-1}M_{21}\right)^{-1} & -M_{11}^{-1}M_{12}\left(M_{22} - M_{21}M_{11}^{-1}M_{12}\right)^{-1} \\ -M_{22}^{-1}M_{21}\left(M_{11} - M_{12}M_{22}^{-1}M_{21}\right)^{-1} & \left(M_{22} - M_{21}M_{11}^{-1}M_{12}\right)^{-1} \end{bmatrix}$$

if all the inverses exist.

Proof. From Lemma 3.7, the inverse of Ξ_ϵ reads

$$\Xi^{-1} = \begin{bmatrix} \left(\Gamma - \frac{\beta^2 h_\circ}{\epsilon^2}\Lambda^2\Gamma^{-1}\right)^{-1} & -\frac{\beta}{\epsilon^2}\Gamma^{-1}\Lambda\left(\Gamma - \frac{\beta^2 h_\circ}{\epsilon^2}\Lambda^2\Gamma^{-1}\right)^{-1} \\ -b\beta\Gamma^{-1}\Lambda\left(\Gamma - \frac{\beta^2 h_\circ}{\epsilon^2}\Lambda^2\Gamma^{-1}\right)^{-1} & \left(\Gamma - \frac{\beta^2 h_\circ}{\epsilon^2}\Lambda^2\Gamma^{-1}\right)^{-1} \end{bmatrix}.$$

So, one can easily check that the entries of Ξ_ϵ^{-1} are bounded, thus is $\|\Xi_\epsilon^{-1}\|$. ∎

Due to this lemma, one can clearly conclude that the implicit solution operator J_ϵ^{-1} is bounded in terms of ϵ.

REMARK 3.8. *The immediate result of this* ϵ*-stability is that the scaled perturbation* $\mathbf{V}_\Delta$ *should be* $\mathcal{O}(1)$ *as long as the explicit step is* ϵ*-stable. This result justifies the asymptotic consistency analysis we are going to present in Section 3.2.2.*

(ii) constant $\widetilde{\alpha}$ *with varying* b. For this case, one of the blocks of J_ϵ is not circulant any longer; the matrix J_ϵ is written as

$$J_\epsilon = \begin{bmatrix} P & \frac{\beta}{\epsilon^2}Q \\ \beta R_b & P \end{bmatrix}, \tag{3.17}$$

where R_b is an *almost* circular matrix, i.e., its i-th row (up to a circulation) is

$$(R_b)_i = (b_{i+1} - b_{i-1}, h_{\circ,i+1}, 0, \dots, 0, -h_{\circ,i-1}).$$

Note that R_b is circulant if and only if its arguments are constant for all rows (or equivalently if the bottom is flat). This non-circulant structure makes the analysis of solvability and ϵ-stability more delicate.

Existence of J_ϵ^{-1}. In the following lemma we show that J_ϵ can be inverted, so it is non-singular and the scheme is again solvable.

LEMMA 3.9. *For* J_ϵ *as in* (3.17) *there exists an inverse.*

Proof. From Lemma 3.7, the inverse exists if all necessary inverses exist in (3.16). Matrix P is SDD, thus invertible [26, Thm. 6.1.10]. For $P - \frac{\beta^2}{\epsilon^2}QP^{-1}R_b$ and $P - \frac{\beta^2}{\epsilon^2}R_bP^{-1}Q$ the arguments for the invertibility are similar; we show the invertibility of the former in the following.

Assume $\widetilde{\alpha} = 0$, so $P = \mathbb{I}_N$. For an eigenvalue of $\mathbb{I}_N - \frac{\beta^2}{\epsilon^2}QR_b$ to be zero, there should exist an eigenvalue of QR_b to be $\frac{\beta^2}{\epsilon^2}$ (for every choice of β), which is not possible: Suppose that for a particular β and ϵ, denoted by β_0 and ϵ_0, one of the eigenvalues of QR_b is $\frac{\beta_0^2}{\epsilon_0^2}$. So, by changing β this equality does not hold anymore. Note that since there are only finite eigenvalues, such an appropriate choice of β always exists. For $\widetilde{\alpha} \neq 0$, the same argument works by factoring out P (since it is full-rank), and studying the eigenvalues of $\mathbb{I}_N - \frac{\beta^2}{\epsilon^2}P^{-1}QP^{-1}R_b$ in a similar way. ∎

Boundedness of J_ϵ^{-1} (ϵ-stability). Regarding the boundedness of J_ϵ^{-1}, employing the diagonal form of circulant matrices cannot simplify all the blocks of J_ϵ^{-1} (as in (3.15)) and the procedure of Lemma 3.6 does not seem to be fruitful. On the other hand, one knows that $\|J_\epsilon^{-1}\|_{\ell_2} = \sigma_{\min}^{-1}(J_\epsilon)$ [27, Fact 4.5], thus showing that $\sigma_{\min}(J_\epsilon)$ does not approach zero in the limit is enough to conclude the boundedness of J_ϵ^{-1}. From Lemma 3.9, J_ϵ is not singular for all $\epsilon > 0$. So the singular values are equal to the square root of the eigenvalues of $J_\epsilon^T J_\epsilon$. In the following lemma, Lemma 3.10, we prove the existence of a (positive) non-vanishing lower-bound for the eigenvalues of $J_\epsilon^T J_\epsilon$, which concludes the boundedness of J_ϵ^{-1}.

LEMMA 3.10. *For J_ϵ as in (3.17), there exists a constant C independent of ϵ, such that $\lim_{\epsilon\to 0}\|J_\epsilon^{-1}\| \le C$.*

Proof. Here, we consider $\widetilde{\alpha} = 0$ to simplify the analysis. However the analysis for $\widetilde{\alpha} \neq 0$ can be done similarly. Again, making use of the fact that all the circulant matrices have the same eigenspace, we can write J_ϵ as in (3.14) with

$$\Xi_\epsilon := \begin{bmatrix} \mathbb{I}_N & \frac{\beta}{\epsilon^2}\Lambda \\ \beta\mathcal{R}_b & \mathbb{I}_N \end{bmatrix}, \qquad \mathcal{R}_b := F_N^* R_b F_N.$$

Note that because $\mathrm{diag}(F_N, F_N)$ is unitary, and a unitary similarity transformation does not change the singular values, we find a lower-bound for the singular values of Ξ_ϵ rather than J_ϵ. $\Xi_\epsilon^*\Xi_\epsilon$ can be written as

$$\Xi_\epsilon^*\Xi_\epsilon = \begin{bmatrix} \mathbb{I}_N + \beta^2\mathcal{R}_b^*\mathcal{R}_b & \beta\left(\frac{\Lambda}{\epsilon^2} + \mathcal{R}_b^*\right) \\ \beta\left(\frac{\Lambda}{\epsilon^2} + \mathcal{R}_b^*\right)^* & \mathbb{I}_N + \frac{\beta^2}{\epsilon^4}\Lambda^*\Lambda \end{bmatrix}.$$

So we should analyze the numerical range of $\Xi_\epsilon^*\Xi_\epsilon$ to show that it is bounded away from zero with an $\mathcal{O}(1)$ bound. $\Xi_\epsilon^*\Xi_\epsilon$ can be re-written as

$$\Xi_\epsilon^*\Xi_\epsilon = \mathbb{I}_{2N} + \beta^2\begin{bmatrix} \mathcal{R}_b^*\mathcal{R}_b & O_N \\ O_N & \frac{1}{\epsilon^4}\Lambda^*\Lambda \end{bmatrix} + \frac{\beta}{\epsilon^2}\begin{bmatrix} O_N & \Lambda \\ \Lambda^* & O_N \end{bmatrix} + \beta\begin{bmatrix} O_N & \mathcal{R}_b^* \\ \mathcal{R}_b & O_N \end{bmatrix},$$

where O_N stands for the zero matrix of size N.

Now, consider the vector $\mathbf{z} := (\mathbf{u}, \mathbf{v})^T \in \mathbb{C}^{2N}$ with $\|\mathbf{z}\|_{\ell_2} = 1$, where both of $\mathbf{u}$ and $\mathbf{v}$ are vectors of the same size N with complex entries. Then by the definition of numerical range, one gets

$$W(\Xi_\epsilon^*\Xi_\epsilon) = \|\beta\mathcal{R}_b\mathbf{u} + \mathbf{v}\|_{\ell_2}^2 + \left\|\frac{\beta}{\epsilon^2}\Lambda\mathbf{v} + \mathbf{u}\right\|_{\ell_2}^2. \tag{3.18}$$

From this, it is clear that if $\mathbf{v} \notin \mathcal{N}_{\epsilon^2}(\Lambda) := \{\mathbf{u} \,|\, \|\Lambda\mathbf{u}\| = \mathcal{O}(\epsilon^2)\}$, then $\left\|\frac{\beta}{\epsilon^2}\Lambda\mathbf{v} + \mathbf{u}\right\|_{\ell_2}$ goes far from zero when $\epsilon \to 0$. Otherwise for $\mathbf{v} \in \mathcal{N}_{\epsilon^2}(\Lambda)$, since $(\mathbf{u}, \mathbf{v})$ lives on the unit sphere, adding $\frac{\beta}{\epsilon^2}\Lambda\mathbf{v}$ and $\beta\mathcal{R}_b\mathbf{u}$ (which are $\mathcal{O}(1)$) would perturb $(\mathbf{u}, \mathbf{v})$ from the unit circle by $\mathcal{O}(\beta)$; so, the numerical range, thereby the eigenvalues of $\Xi_\epsilon^*\Xi_\epsilon$, are bounded away from zero. This concludes the lemma. ∎

Assuming the ϵ-stability of the explicit step, Lemma 3.10 verifies that the scaled perturbation $\mathbf{V}_\Delta$ is $\mathcal{O}(1)$, which justifies the formal asymptotic consistency of the next section.

REMARK 3.11. *So far, one important advantage of the RS-IMEX scheme with a suitable scaling and reference solution has been to simplify the procedure of Lemma*

3.6 and Lemma 3.10 to conclude the ϵ-stability of the implicit solution operator (and of the numerical solution later on).

REMARK 3.12. *Note that the ϵ-stability of the solution implies that there exists a sequence $\{\mathbf{V}_{\Delta,\epsilon_k}^{n+1}\}_{k\in\mathbb{N}}$ ($\epsilon_k \to 0$ as $k \to \infty$) converging strongly to a limit (after extracting a subsequence if necessary). To determine whether this limit is the correct zero-Froude limit will be the topic of the next section, Section 3.2.2.*

3.2.2. Asymptotic consistency. AP analysis has three parts, as we have described in Definition 1.1: asymptotic consistency, asymptotic efficiency and asymptotic stability. We discuss the asymptotic consistency in this section. Because we have already discussed solvability and ϵ-stability of the implicit solution operator, the asymptotic consistency analysis we are going to present is not only formal (like [9, 12, 24]), but also rigorous, since —owing to the ϵ-stability— the coefficients of the asymptotic expansion have been shown to be bounded in terms of ϵ. Similar ideas has been used in [5] in the context of the Finite Volume Evolution Galerkin (FVEG) scheme [6], and in [62] for the Lagrange–projection scheme.

For the RS-IMEX scheme applied to the shallow water system, the asymptotic consistency requires that the zeroth- and first-order expansions of momentum to be constant in space. Due to the appropriate choice for the reference solution, the surface elevation z satisfies the continuous asymptotic expansions simply by construction.

We now consider the discrete version of the asymptotic expansion, which is similar to the continuous version (3.2):

$$\begin{aligned} z(x_i,t_n) &= z_{(0)} + \epsilon z_{(1)} + \epsilon^2 z_{(2)}(x_i,t_n),\\ m(x_i,t_n) &= m_{(0)}(t_n) + \epsilon m_{(1)}(t) + \epsilon^2 m_{(2)}(x_i,t_n). \end{aligned}$$

Since we assumed the reference state to be the lake at rest with the scaling matrix D_2, it turns out that the scaled variables at time t_n write

$$\begin{aligned} v_1(x,t_n) &= z_{(2)}(x,t_n),\\ v_2(x,t_n) &= m_{(0)}(t_n) + \epsilon m_{(1)}(t_n) + \epsilon^2 m_{(2)}(x,t_n). \end{aligned}$$

The goal is to determine whether or not the zeroth- and first-order expansions of the momentum are constant in space. Substituting (3.19) into the momentum update of the explicit step (3.10) yields

$$\begin{aligned} v_{2(0)i}^{n\dagger} &= m_{(0)i}^n - \frac{\Delta t}{2\Delta x}\frac{m_{(0)i}^{n,2}}{h_{\circ,i+1}h_{\circ,i-1}}(b_{i+1}-b_{i-1}) = m_{(0)c}^n - \frac{\Delta t}{2\Delta x}\frac{m_{(0)c}^{n,2}}{h_{\circ,i+1}h_{\circ,i-1}}(b_{i+1}-b_{i-1}),\\ v_{2(1)i}^{n\dagger} &= m_{(1)i}^n - \frac{\Delta t}{\Delta x}\frac{m_{(0)i}^n m_{(1)i}^n}{h_{\circ,i+1}h_{\circ,i-1}}(b_{i+1}-b_{i-1}) \;\; = m_{(1)c}^n - \frac{\Delta t}{\Delta x}\frac{m_{(0)c}^n m_{(1)c}^n}{h_{\circ,i+1}h_{\circ,i-1}}(b_{i+1}-b_{i-1}), \end{aligned}$$

where $m_{(0)c}^n$ and $m_{(1)c}^n$ are some constants. So, the explicit step for the momentum does not introduce an $\mathcal{O}(1/\epsilon)$ term into the scheme, i.e., $\mathbf{V}_\Delta^{n\dagger} = \mathcal{O}(1)$. Note that the surface perturbation does not change through the explicit step; see (3.4). Thus, the explicit step is asymptotically consistent up to $\mathcal{O}(\Delta x)$ provided that the bottom function is assumed to have a bounded derivative. Such a *small slope assumption* is usually imposed for the validity of the shallow water model (see [7]).

Incidentally, Remark 3.8 implies that the boundedness of $\mathbf{V}_\Delta^{n\dagger}$ leads to the ϵ-stability of the implicit solution in Section 3.2.1. Thus, from the implicit v_1 update

(3.11), one can (rigorously) conclude that

$$v^{n+1}_{2(0)i+1} = v^{n+1}_{2(0)i-1}, \qquad v^{n+1}_{2(1)i+1} = v^{n+1}_{2(1)i-1}. \tag{3.19}$$

So, the updated momentum is almost constant, i.e., the discrete divergence operator vanishes in the limit $\mathrm{div}_\Delta \mathbf{V}^{n+1}_{2,\Delta} = \mathcal{O}(\epsilon^2)$. Although this is often interpreted as the asymptotic consistency in the literature, it does not imply necessarily that the limit has been actually obtained. For example, one can confirm that although the discretization 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, if happens, is as small as $\mathcal{O}(\epsilon^2)$. Thus, it does not ruin the numerical solution in the limit.

LEMMA 3.13. *For the RS-IMEX scheme (3.10)-(3.11) with constant $\widetilde{\alpha}$, applied to the shallow water equations with flat bottom, the deviations of the computed momentum is $\mathcal{O}(\epsilon^2)$, as $\epsilon \to 0$. In other words, the possible checker-board oscillations for the computed momentum are at most $\mathcal{O}(\epsilon^2)$.*

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

$$J_\epsilon \begin{bmatrix} [\![\mathbf{V}^{n+1}_{1,\Delta}]\!] \\ [\![\mathbf{V}^{n+1}_{2,\Delta}]\!] \end{bmatrix} = \begin{bmatrix} [\![\mathbf{V}^{n\dagger}_{1,\Delta}]\!] \\ [\![\mathbf{V}^{n\dagger}_{2,\Delta}]\!] \end{bmatrix}. \tag{3.20}$$

For the case of flat bottom, we will show that the blocks of $K_\epsilon := J_\epsilon^{-1}$ behave as

$$\|K_{11}\|, \|K_{12}\|, \|K_{22}\| = \mathcal{O}(1), \qquad \|K_{21}\| = \mathcal{O}(\epsilon^2). \tag{3.21}$$

Then, it follows

$$\begin{aligned} \left\| [\![\mathbf{V}^{n+1}_{2,\Delta}]\!] \right\| &= \left\| K_{21} [\![\mathbf{V}^{n\dagger}_{1,\Delta}]\!] + K_{22} [\![\mathbf{V}^{n\dagger}_{2,\Delta}]\!] \right\| \\ &\leq C \left(\left\| [\![\mathbf{V}^{n\dagger}_{2,\Delta}]\!] \right\| + \epsilon^2 \left\| [\![\mathbf{V}^{n\dagger}_{1,\Delta}]\!] \right\| \right), \end{aligned}$$

and since $\|[\![\mathbf{V}^{n\dagger}_{1,\Delta}]\!]\| = \mathcal{O}(1)$ and $\|[\![\mathbf{V}^{n\dagger}_{2,\Delta}]\!]\| = \mathcal{O}(\epsilon^2)$ (as shown above for the case of flat bottom), it turns out that

$$\left\| [\![\mathbf{V}^{n+1}_{2,\Delta}]\!] \right\| = \mathcal{O}(\epsilon^2),$$

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

It only remains to study the behavior of the blocks in (3.21), and in particular K_{21} and K_{22}. Let us re-write the inverse K_ϵ as

$$K_\epsilon = \begin{bmatrix} \left(P - \frac{\beta h_\circ}{\epsilon^2} Q P^{-1} Q \right)^{-1} & -\frac{\beta^2}{\epsilon^2} P^{-1} Q \left(P - \frac{\beta^2 h_\circ}{\epsilon^2} Q P^{-1} Q \right)^{-1} \\ -\beta h_\circ P^{-1} Q \left(P - \frac{\beta^2 h_\circ}{\epsilon^2} Q P^{-1} Q \right)^{-1} & \left(P - \frac{\beta^2 h_\circ}{\epsilon^2} Q P^{-1} Q \right)^{-1} \end{bmatrix}.$$

It is clear from Lemma 3.6 and the structure of K_ϵ that since $K_{12} = \frac{\beta}{\epsilon^2 h_\circ} K_{21}$ and $\|K_{21}\| = \mathcal{O}(1)$, then

$$\|K_{21}\| = \mathcal{O}(\epsilon^2), \qquad \|K_{22}\| = \mathcal{O}(1),$$

which concludes the proof of Lemma 3.13.

When the bottom is non-flat, (3.20) is not valid anymore since the momentum equation has contributions of non-constant coefficients terms. However, one can confirm that (assuming $\widetilde{\alpha}=0$ for simplicity)

$$H_\epsilon \begin{bmatrix} [\![\mathbf{V}_{1,\Delta}^{n+1}]\!] \\ [\![\mathbf{V}_{2,\Delta}^{n+1}]\!] \end{bmatrix} = \begin{bmatrix} [\![\mathbf{V}_{1,\Delta}^{n\dagger}]\!] \\ [\![\mathbf{V}_{2,\Delta}^{n\dagger}]\!] \end{bmatrix}, \qquad H_\epsilon =: \begin{bmatrix} \mathbb{I}_N & \frac{\beta}{\epsilon^2}Q \\ \beta R_b^\Delta & \mathbb{I}_N \end{bmatrix}. \tag{3.22}$$

Defining R_b^Δ, which is the point of departure of H_ϵ from J_ϵ, we write the implicit momentum update for the jumps in (3.22)

$$[\![v_{2,i}^{n+1}]\!] + \beta\Big(h_{\circ,i+1}[\![v_{1,i+1}^{n+1}]\!] - \big(b_i - b_{,i-1}\big)[\![v_{1,i}^{n+1}]\!] - h_{\circ,i-2}[\![v_{1,i-1}^{n+1}]\!]\Big) = [\![v_{2,i}^{n\dagger}]\!], \tag{3.23}$$

which should be compared with what J_ϵ provides

$$[\![v_{2,i}^{n+1}]\!] + \beta\Big(h_{\circ,i+1}[\![v_{1,i+1}^{n+1}]\!] + \big(b_{i+1} - b_{i-1}\big)[\![v_{1,i}^{n+1}]\!] - h_{\circ,i-1}[\![v_{1,i-1}^{n+1}]\!]\Big) = [\![v_{2,i}^{n\dagger}]\!]. \tag{3.24}$$

Hence showing the smallness of checker-board oscillations may need more than a direct use of Lemma 3.10. Let us recall that since the proof of Lemma 3.10 does not depend on the structure of J_{21}, one can apply it to H_ϵ, so $\lim_{\epsilon\to 0}\|H_\epsilon^{-1}\|<\infty$, thus the blocks of H_ϵ^{-1}

$$H_\epsilon^{-1} = \begin{bmatrix} \left(\mathbb{I}_N - \frac{\beta^2}{\epsilon^2}QR_b^\Delta\right)^{-1} & -\frac{\beta}{\epsilon^2}Q\left(\mathbb{I}_N - \frac{\beta^2}{\epsilon^2}R_b^\Delta Q\right)^{-1} \\ -\beta R_b^\Delta\left(\mathbb{I}_N - \frac{\beta^2}{\epsilon^2}QR_b^\Delta\right)^{-1} & \left(\mathbb{I}_N - \frac{\beta^2}{\epsilon^2}R_b^\Delta Q\right)^{-1} \end{bmatrix}$$

are bounded in ϵ, i.e., they are all at most $\mathcal{O}(1)$. So, it can only be concluded that $\|[\![\mathbf{V}_{2,\Delta}^{n+1}]\!]\| = \mathcal{O}(1)$. However R_b^Δ is close to $h_\circ Q$ (with $\mathcal{O}(\Delta x)$ difference); so, it is plausible to claim that since $\|(H^{-1})_{12}\| = \mathcal{O}(1)$, one gets $\|(H^{-1})_{21}\| = \mathcal{O}(\epsilon^2)$. Because $\|[\![\mathbf{V}_{2,\Delta}^{n\dagger}]\!]\| = \mathcal{O}(\Delta x)$ there is an $\mathcal{O}(\Delta x)$ deviation from the result of the flat bottom, which gives $\|[\![\mathbf{V}_{2,\Delta}^{n+1}]\!]\| = \mathcal{O}(\epsilon^2) + \mathcal{O}(\Delta x)$.

Hence, for both cases, one can conclude that the momentum is close to a constant value in the limit.

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

Explicit step:

$$\begin{aligned} \frac{v_{2(0),i}^{n\dagger} - v_{2(0),i}^{n}}{\Delta t} + \frac{1}{2\Delta x}\Bigg[& \frac{v_{2(0),i+1}^{2,n}}{z_\circ + \epsilon^2 v_{1(0),i+1}^{n} - b_{i+1}} + \frac{\epsilon^2}{2} v_{(0)1,i+1}^{2,n} \\ & - \frac{v_{2(0),i-1}^{2,n}}{z_\circ \epsilon^2 v_{1(0),i-1}^{n} - b_{i-1}} - \frac{\epsilon^2}{2} v_{(0)1,i-1}^{2,n} \Bigg] \\ & - \frac{\widehat{\alpha}}{2\Delta x}\Big(v_{2(0),i+1}^{n} - 2v_{2(0),i}^{n} + v_{2(0),i-1}^{n}\Big) = 0. \end{aligned} \tag{3.25}$$

Implicit step:

$$\begin{aligned}\frac{v^{n+1}_{2(0),i}-v^{n\dagger}_{2(0),i}}{\Delta t}&+\frac{1}{2\Delta x}\left((z_\circ-b_{i+1})v^{n+1}_{1(0),i+1}-(z_\circ-b_{i+1})v^{n+1}_{1(0),i+1}\right)\\&-\frac{\widetilde{\alpha}}{2\Delta x}\left(v^{n}_{1(0),i+1}-2v^{n}_{1(0),i}+v^{n}_{1(0),i-1}\right)\\&=-\frac{1}{2\Delta x}v^{n+1}_{1(0),i}\left(b_{i+1}-b_{i-1}\right).\end{aligned}\tag{3.26}$$

It is clear that (3.25) and (3.26) provide consistent discretizations of $\partial_t m_{(0)}$ for both explicit and implicit steps. Thus, the scheme is AC.

3.2.3. Asymptotic stability. In this section, we discuss the rigorous stability analysis of the RS-IMEX scheme in ℓ_2-norm. Motivated by [24, Lemma 3.1] (see [50, 60] for further details), one can define the stability (in finite time) as follows.

DEFINITION 3.14. *Assume that $\mathcal{E}_i$ for $i=1,\dots,s$ are some discrete evolution operators, like explicit and implicit operators for the RS-IMEX, and suppose that the numerical solution at the step k (for $k=0,1,\dots,n-1$ and $n=T/\Delta t$) is obtained as*

$$\mathbf{Y}^k=\prod_{i=0}^{s-1}\mathcal{E}_{s-i}\,\mathbf{Y}^{k-1}.$$

Then, the numerical method is said to be stable in ℓ_p-norm (in finite time), i.e., $\|\mathbf{Y}^n\|_{\ell_p}\le CT\|\mathbf{Y}^0\|_{\ell_p}$ for all $n\in\mathbb{N}$ and with the constant C independent of Δt, provided that there exist constants c_i independent of Δt such that

$$\|\mathcal{E}_i\|_{\ell_p}\le 1+c_i\Delta t,\qquad i=1,\dots,s.\tag{3.27}$$

In what follows we aim to show that the condition (3.27) holds for the RS-IMEX scheme. Note that for the RS-IMEX scheme, $s=2$ and $\mathcal{E}_1$ and $\mathcal{E}_2$ denote the explicit and implicit operators, respectively.

Stability of the explicit step $\mathcal{E}_1$. To prove the stability of the explicit step requires to choose a suitable norm for the nonlinear operator $\mathcal{E}_1$. Let us define the operator norm for the operator $\mathcal{T}$ as $\|\mathcal{T}\|_{op,\ell_r}:=\max_{\|\mathbf{Y}\|_{\ell_r}=1}\|\mathcal{T}\mathbf{Y}\|_{\ell_r}$. For the explicit step, $\mathcal{T}=\mathcal{E}_1$ and r is chosen to be 2. Thus, $\|\mathcal{E}_1\mathbf{Y}\|_{\ell_2}\le\|\mathcal{E}_1\|_{op,\ell_2}\|\mathbf{Y}\|_{\ell_2}$. It remains to show that the condition (3.27) holds for $\|\mathcal{E}_1\|_{op,\ell_2}$. Assuming $\widehat{\alpha}=0$ for simplicity and from (3.10), one can write $\|\mathcal{E}_1\|_{op,\ell_2}$ as (note that $\mathbf{Y}=\begin{bmatrix}\mathbf{V}_{1,\Delta}\\ \mathbf{V}_{2,\Delta}\end{bmatrix}$)

$$\begin{aligned}\|\mathcal{E}_1\|_{op,\ell_2}&\le 1+\frac{2\beta}{h_{\min}}\|\langle\mathbf{V}_{2,\Delta},\mathbf{V}_{2,\Delta}\rangle\|_{\ell_2}+\epsilon^2\beta\|\langle\mathbf{V}_{1,\Delta},\mathbf{V}_{1,\Delta}\rangle\|_{\ell_2}\\&\le 1+\beta\left(\frac{2}{h_{\min}}+\epsilon^2\right)\|\mathbf{Y}\|^2_{\ell_4},\\&\le 1+\beta\left(\frac{2}{h_{\min}}+\epsilon^2\right),\end{aligned}$$

since for sequence spaces, $\|\mathbf{Y}\|_{\ell_q}\le\|\mathbf{Y}\|_{\ell_p}$ for $1\le p\le q$ and $\|\mathbf{Y}\|_{\ell_2}=1$ by definition. Here $h_{\min}$ is the lower-bound for the water height, i.e.

$$h_{\min}:=\min_{i\in\Omega_N}\left|z_\circ+\epsilon^2 v_{1,i}-b_i\right|\qquad\text{for}\quad\left\|\begin{bmatrix}\mathbf{V}_{1,\Delta}\\ \mathbf{V}_{2,\Delta}\end{bmatrix}\right\|_{\ell_2}=1.$$

This implies that for small enough ϵ, $h_{\min}$ is bounded away from zero; so, the explicit step is asymptotically stable (in finite time).

Stability of the implicit step $\mathcal{E}_2$. As we have mentioned earlier, the implicit operator is J_ϵ^{-1}. So, one should find some bound of the form $1 + c_1\Delta t$ for J_ϵ^{-1}. Let us assume the norm to be ℓ_2. So, one can write

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

where $\underline{\omega}(\Xi_\epsilon^*\Xi_\epsilon) := \min|W(\Xi_\epsilon^*\Xi_\epsilon)|$. On the other hand, as we have discussed in the proof of Lemma 3.10, $\underline{\omega}(\Xi_\epsilon^*\Xi_\epsilon)$ can be written as $1-\beta C_1$ with some positive ϵ-uniform constant C_1, which gives

$$\|\mathcal{E}_2\|_{\ell_2} \le \frac{1}{1-\beta C_2} \approx \sum_{k=0}^{\infty} (\beta C_2)^k \le 1 + \beta C,$$

due to the Taylor expansion around $\beta = 0$ and with another positive ϵ-uniform constant C . Thus, the implicit operator is asymptotically stable (in finite time).

Hence, combining these two results for explicit and implicit steps, one can conclude asymptotically stability. For non-small ϵ, one should add the positivity assumption to conclude that the positive solutions of RS-IMEX scheme are ℓ_2-stable.

Remark 3.15. *As we have seen so far, the scheme is AC and AS. Due to Definition 1.1, for the scheme to be AP, asymptotic efficiency is also necessary: The CFL condition is ϵ-uniform (with material velocity), but the condition number of J_ϵ increases as $\epsilon \to 0$ (see Remark 4.4). Although, literally speaking, the scheme is not AP in the sense of Definition 1.1, we call it AP (at least in a weaker sense) since it is AC and AS under a non-restrictive CFL condition.*

3.2.4. C-property (well-balancing). Considering the lake at rest, the C-property for the explicit step is boiled down to consistency of the numerical flux (due to lack of non-stiff source term) which is fulfilled by the construction.

For the implicit step, if one discretizes the source term central, i.e., $b_x(x_i) \approx \frac{b_{i+1}-b_{i-1}}{2\Delta x}$, the compatibility of the equilibrium solution is clear since there is exactly such a term in the difference of Rusanov fluxes as well. Because the implicit step has a unique solution, this compatibility confirms that the RS-IMEX scheme preserves the lake at rest equilibrium state, i.e., it is well-balanced.

4. Shallow water equations with the zero-Froude limit reference state. Here, we consider the same shallow water system as in (3.1) on a periodic domain, but with the zero-Froude limit solution of (3.1) as the reference solution, i.e., $\overline{\mathbf{U}} = (z_\circ, m_\circ)^T$. This can be formally obtained from the Definition 3.2 and (3.3), i.e., $z_\circ = z^0_{(0)}$ and $m_\circ = m^0_{(0)}$. Additionally, we assume the bottom to be flat to make the zero-Froude limit stationary (owing to periodic boundary conditions). This makes $\overline{\mathbf{T}}$ to vanish and avoids the difficulties stem from the truncation error of the reference solution $\overline{\mathbf{T}}$ in the asymptotic analysis (as will be discussed and analyzed in detail

in [61]). With this reference solution the splitting can be obtained as

$$\overline{\mathbf{F}} := \begin{bmatrix} m_\circ \\ \dfrac{m_\circ^2}{z_\circ - b} + \dfrac{z_\circ^2 - 2z_\circ b}{2\epsilon^2} \end{bmatrix}, \quad \widetilde{\mathbf{F}} := \begin{bmatrix} m_{pert} \\ -\dfrac{m_\circ^2}{(z_\circ - b)^2} z_{pert} + \dfrac{z_\circ - b}{\epsilon^2} z_{pert} + \dfrac{2m_\circ}{z_\circ - b} m_{pert} \end{bmatrix},$$

$$\widehat{\mathbf{F}} := \begin{bmatrix} 0 \\ \dfrac{(m_\circ + m_{pert})^2}{z_\circ + z_{pert} - b} + \dfrac{z_{pert}^2}{2\epsilon^2} - \dfrac{m_\circ^2}{z_\circ - b} + \dfrac{m_\circ^2}{(z_\circ - b)^2} z_{pert} - \dfrac{2m_\circ}{z_\circ - b} m_{pert} \end{bmatrix}.$$

Based on the asymptotic analysis presented in Appendix A, we choose the scaling matrix as $D_3 := \mathrm{diag}(\epsilon^2, \epsilon)$, and the scaled RS-IMEX splitting reads

$$\begin{aligned} \widetilde{\mathbf{G}} &:= \begin{bmatrix} v_2/\epsilon \\ -\dfrac{m_\circ^2 \epsilon v_1}{(z_\circ - b)^2} + \dfrac{(z_\circ - b)v_1}{\epsilon} + \dfrac{2m_\circ v_2}{z_\circ - b} \end{bmatrix}, \\ \widehat{\mathbf{G}} &:= \begin{bmatrix} 0 \\ \dfrac{(m_\circ + v_2\epsilon)^2}{\epsilon\,(z_\circ + \epsilon^2 v_1 - b)} + \dfrac{\epsilon v_1^2}{2} - \dfrac{m_\circ^2}{\epsilon\,(z_\circ - b)} + \dfrac{m_\circ^2 v_1 \epsilon}{(z_\circ - b)^2} - \dfrac{2m_\circ v_2}{z_\circ - b} \end{bmatrix}. \end{aligned} \tag{4.1}$$

One can see that the splitting is admissible in the sense of [58]. That is to say, the eigenvalues of $\widetilde{\mathbf{G}}$ are stiff and those of $\widehat{\mathbf{G}}$ are non-stiff:

$$\begin{aligned} \widetilde{\mu} &= \frac{m_\circ}{z_\circ - b} \pm \frac{\sqrt{z_\circ - b}}{\epsilon}, \\ \widehat{\mu} &= 0, \frac{2\epsilon\,(v_2(z_\circ - b) - \epsilon m_\circ v_1)}{(z_\circ - b)(z_\circ - b + \epsilon^2 v_1)}. \end{aligned} \tag{4.2}$$

So the zero-Froude limit reference state makes the wave speeds of the slow system of $\mathcal{O}(\epsilon)$, thus the commutator would be $\mathcal{O}(1)$. In fact, it can be obtained formally that (for $\epsilon \ll 1$)

$$\left[\widetilde{\mathbf{G}}', \widehat{\mathbf{G}}'\right] = \begin{bmatrix} v_1 + \mathcal{O}(\epsilon^2) & \dfrac{2v_2}{z_\circ - b} \\ -2v_2 + \mathcal{O}(\epsilon) & -v_1 + \mathcal{O}(\epsilon^2) \end{bmatrix}, \tag{4.3}$$

which is $\mathcal{O}(1)$. Similar to the case of the lake at rest reference solution, the modified equation is stable for this splitting.

For this case, the RS-IMEX scheme is defined as in (3.8)-(3.9) when $\widehat{\mathbf{G}}$ and $\widetilde{\mathbf{G}}$ change according to (4.1).

4.1. Stability analysis. We collect the stability 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 (3.8)-(3.9) and (4.1)*

(i) is solvable, i.e., it has a unique solution for all $\epsilon > 0$, which does not blow-up for $\epsilon \to 0$, under the assumption of constant $\widetilde{\alpha}$.

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

(iii) is asymptotically ℓ_2-stable, i.e., there exists a constant $C_{N,T}$ depending on the final time $T = n\Delta t$ and the number of grid points N such that $\|\mathbf{V}_\Delta^n\|_{\ell_2} \leq C_{N,T}\|\mathbf{V}_\Delta^0\|_{\ell_2}$.

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_ϵ reads

$$J_\epsilon := \begin{bmatrix} P & \frac{\beta}{\epsilon}Q \\ \left(-\frac{m_\circ^2\epsilon}{h_\circ^2}+\frac{h_\circ}{\epsilon}\right)\beta Q & P+\frac{2\beta m_\circ}{h_\circ}Q \end{bmatrix}. \tag{4.4}$$

In the following, we first show that J_ϵ is non-singular, thus there exists a unique solution for the implicit step (and so for the scheme). Then, we prove that the solution operator and the solution are ϵ-stable.

Existence of J_ϵ^{-1}. Like Section 3.2, Since P and Q are circulant, the blocks of J_ϵ commute and from [59, Thm. 1] the determinant of J_ϵ can be computed as

$$\det(J_\epsilon) = \det\left(\underbrace{P^2-\frac{\beta^2}{\epsilon}\left(-\frac{m_\circ^2\epsilon}{h_\circ^2}+\frac{h_\circ}{\epsilon}\right)Q^2}_{=:\mathfrak{A}} + \underbrace{\frac{2\beta m_\circ}{h_\circ}PQ}_{=:\mathfrak{B}}\right).$$

One can confirm that PQ is skew-symmetric; so, $\mathfrak{B}$ does not change the real eigenvalues of $\mathfrak{A}$ which is symmetric. This is a result from Bendixon [3,25]. Thus, it remains to show that $\mathfrak{A}$ has only non-zero eigenvalues, which can be done as in Section 3.2.1, by a suitable and ϵ-uniform choice of β. Hence J_ϵ is non-singular.

Boundedness of J_ϵ^{-1}. Again similar to Section 3.2, we can find Ξ_ϵ as

$$\Xi_\epsilon := \begin{bmatrix} \Gamma & \frac{\beta}{\epsilon}\Lambda \\ \left(-\frac{m_\circ^2\epsilon}{h_\circ}+\frac{h_\circ}{\epsilon}\right)\beta\Lambda & \Gamma+\frac{2\beta m_\circ}{h_\circ}\Lambda \end{bmatrix}.$$

In the following lemma we show that Ξ_ϵ^{-1} is ϵ-stable.

LEMMA 4.2. *The inverse of matrix Ξ_ϵ, has a bounded norm in terms of ϵ.*

Proof. From Lemma 3.7, the inverse of Ξ_ϵ reads

$$\begin{aligned}
\Xi_{11}^{-1} &= \left(\Gamma-\frac{\beta^2}{\epsilon}\left(-\frac{m_\circ^2\epsilon}{h_\circ}+\frac{h_\circ}{\epsilon}\right)\Lambda^2\left(\Gamma+2\frac{\beta m_\circ}{h_\circ}\Lambda\right)^{-1}\right)^{-1},\\
\Xi_{12}^{-1} &= -\frac{\beta}{\epsilon}\Gamma^{-1}\Lambda\Xi_{22}^{-1},\\
\Xi_{21}^{-1} &= -\left(\Gamma+\frac{2\beta m_\circ}{h_\circ}\Lambda\right)^{-1}\left(-\frac{m_\circ^2\epsilon}{h_\circ^2}+\frac{h_\circ}{\epsilon}\right)\beta\Lambda\Xi_{11}^{-1},\\
\Xi_{22}^{-1} &= \left(\Gamma+\frac{2\beta m_\circ}{z_\circ-b}\Lambda-\frac{\beta^2}{\epsilon}\left(-\frac{m_\circ^2\epsilon}{h_\circ^2}+\frac{h_\circ}{\epsilon}\right)\Gamma^{-1}\Lambda^2\right)^{-1}.
\end{aligned}$$

From this, and similar to Lemma 3.6 one can show that the blocks of Ξ_ϵ^{-1} are bounded, so is $\|\Xi_\epsilon^{-1}\|$. ∎

Due to this lemma and assuming the ϵ-stability of the explicit step (see Section 4.1.2), one can clearly conclude that the solution of the implicit step (thus the whole scheme) cannot blow-up as $\epsilon \to 0$, i.e., it is ϵ-stable.

REMARK 4.3. *The ϵ-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 4.4. *It is worth mentioning that the condition number of J_ϵ depends on the scaling matrix. For example, one can confirm that using D_2 and D_3 makes the condition number to be $\mathcal{O}(1/\epsilon^2)$ and $\mathcal{O}(1/\epsilon)$ respectively. The proof is straightforward from the definition of the condition number and the bound of J_ϵ (see Lemma 3.6 for the lake at rest reference solution and similar result for the zero-Froude limit reference solution in this section). We also wish to mention that in this sense, the scaling by the diagonal matrix D is the equilibration of matrices in essence, whose basic idea is to scale the matrix rows and columns by (possibly different) diagonal matrices to improve the condition number, cf. [19, Sect. 3.5.2]. This is exactly the advantage of D_3 over D_2 as shown in Table 1.*

	Scaling by D_2		Scaling by D_3	
J_ϵ	1	$\mathcal{O}(1/\epsilon^2)$	1	$\mathcal{O}(1/\epsilon)$
	1	1	$\mathcal{O}(1/\epsilon)$	1

Table 1: Comparison of different scaling for matrix J_ϵ.

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

For the explicit step and similarly to the case with the lake at rest reference solution, no $\mathcal{O}(1/\epsilon)$ contribution is associated with the explicit update since

$$\lim_{\epsilon\to 0}\left[\frac{(m_\circ+v_2^n\epsilon)^2}{\epsilon\,(z_\circ+\epsilon^2 v_1^n-b)}-\frac{m_\circ^2}{\epsilon\,(z_\circ-b)}\right]=\mathcal{O}(1). \tag{4.5}$$

So, it is asymptotically consistent (and ϵ-stable). This implies that for the implicit step, as shown in the previous section, $\mathbf{V}_\Delta^{n+1}=\mathcal{O}(1)$. So the perturbation vanishes in the limit and only $\overline{\mathbf{U}}$ remains, which concludes the asymptotic consistency of the scheme.

REMARK 4.5. *The asymptotic stability analysis for the implicit step is very similar to Section 3.2.3; so, we skip it here. We just wish to stress that for the explicit step, one should use (4.5) to find an ϵ-uniform bound for $\|\mathcal{E}_1\|_{op,\ell_2}$. Hence, one can conclude that the scheme is AP in a weaker sense than Definition 1.1, i.e., it is AC and AS under a non-restrictive CFL condition but the condition number of J_ϵ increases as $\epsilon\to 0$.*

5. Numerical results. In this section, we show that the RS-IMEX scheme, which has been discussed throughout this paper, has well-qualified solutions compared to existing schemes. Also we confirm numerically the AP property (asymptotic consistency and asymptotic stability) of the scheme.

At first, we consider the flat bottom case in two examples and provide numerical evidences regarding the convergence order and asymptotic consistency and stability. We also discuss the quality of the numerical solution by comparing it to other existing results like [2, 5]. Then, we continue with a non-flat bottom example.

Note that unless stated otherwise, the time step has been computed as $\Delta t:=\min\left(\Delta t_{\text{CFL}},\Delta t_{\text{Aux}}\right)$ where the CFL time step Δt_{CFL} and the auxiliary time step

Δt_{Aux} are defined as below:

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

5.1. Shallow water equations with flat bottom. In this section, we discuss numerical results for the case of shallow water equations with a flat bottom. Firstly, we consider a colliding pulses example of [12], which has been also discussed in [5]. Then, we discuss another colliding pulses example from [2]. In each case, we show that the quality of the solution is comparable and the convergence order is fine. Also we confirm the asymptotic consistency and stability, as well as the smallness of checkerboard oscillations.

5.1.1. Colliding pulses (I). Example 6.1 in [12] uses the pressure function $p(\varrho) = \varrho^2$ and the following well-prepared initial data:

$$\begin{aligned}
&x \in [0, 0.2] \cap (0.8, 1]: && h_0(x) = 1, && m_0(x) = 1 - \frac{\epsilon^2}{2},\\
&x \in (0.2, 0.3]: && h_0(x) = 1 + \epsilon^2, && m_0(x) = 1,\\
&x \in (0.3, 0.7]: && h_0(x) = 1, && m_0(x) = 1 + \frac{\epsilon^2}{2},\\
&x \in (0.7, 0.8]: && h_0(x) = 1 - \epsilon^2, && m_0(x) = 1,
\end{aligned}$$

with the final time $T = 0.05$, CFL $= 0.45$ and the periodic domain $[0, 1)$. Since, the pressure function for the shallow water equations is a bit different (by a factor of $\frac{1}{2}$), we compare the results of the RS-IMEX scheme with [5, Sect. 8.1].

Figures 3 and 4 show the results of the RS-IMEX scheme with $m_\circ = 0$ (LaR) and $m_\circ = 1$ (zero-Froude limit) for $\epsilon = 0.8$ and $\epsilon = 0.1$. Comparing to [5, Fig. 8.2], it is clear that the computed solutions are well-qualified. Note that for this example, the schemes in [5, Fig. 8.2] uses the same splitting as the RS-IMEX and they enjoy an elliptic approach for the surface perturbation update. For more details, the reader should consult with [5, 6]. As Figure 3 and Figure 4 suggest, the computed surface perturbation z does not depend on the reference solution particularly for $\epsilon = 0.1$. For the momentum, the $m_\circ = 1$ case gives a bit more accurate solution in terms of capturing the extrema; this can be clearly seen in Figure 4 where the *exact* solution is computed on a very fine mesh with $N = 6400$. Note that for $\epsilon = 0.1$ both schemes cannot capture the details of the waves (micro-structures), which is also the case in [5, 12].

Figure 5 illustrates the experimental order of convergence for different ϵ and $m_\circ = 0, 1$, for an error defined as

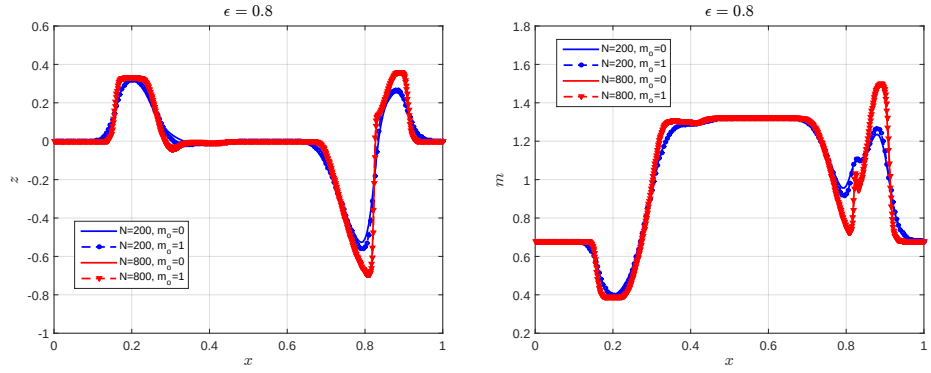
$$e(U_\Delta) := \|U_\Delta - U_{\text{exact}}\|_{L_1} = \frac{1}{N} \sum_{j\in\Omega_N} |U_{\Delta,j} - U_{\text{exact},j}|\,, \tag{5.1}$$

where U is the variable of interest (momentum, height, etc.) and U_Δ and U_{exact} are the computed solution and the exact solution respectively. For this example, the exact solution is computed on a finer mesh with $N = 3200$. Figure 5 shows that the scheme, regardless of the reference solution, has an almost uniform order of convergence for $\epsilon = 0.8, 0.1, 0.05$, which coincides with the result of [12, Tab. 2].

Verifying asymptotic consistency and stability, Figure 6 shows the solution for a small ϵ, namely $\epsilon = 10^{-8}$. It confirms that the solution is close to the incompressible

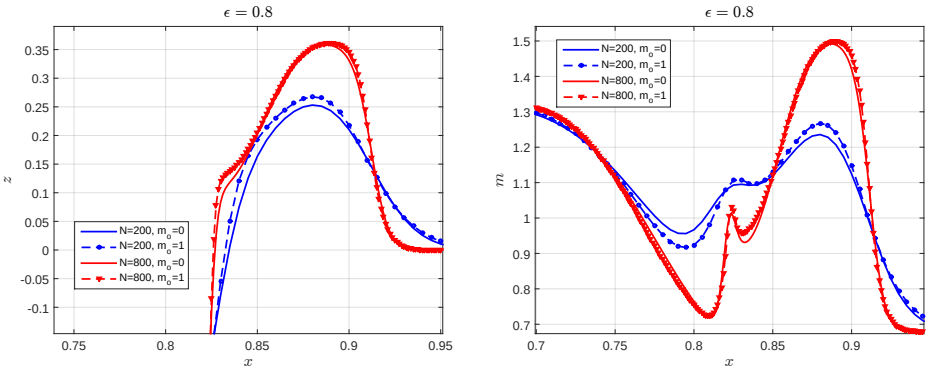
manifold. That is to say, the surface elevation is almost constant, and the momentum is *div*-free. It also confirms the smallness of the checker-board oscillations. Note that for the lake at rest reference solution the scheme uses D_2, which makes the condition number of J_ϵ to grow as $\mathcal{O}(1/\epsilon^2)$ (see Remark 4.4). This clearly affects the solution in the limit as one can see by comparing Figure 6(b) and Figure 6(d). If one changes the scaling matrix (for LaR) to D_3, the limit solution gets much closer to the limit manifold, as Figure 6(e) confirms.

Note that for the zero-Froude limit reference state, due to $\mathcal{O}(\epsilon)$ eigenvalues for the non-stiff system as in (4.2) , $\Delta t_{\text{CFL}} = \mathcal{O}(1/\epsilon)$; so, it gets larger as ϵ decreases. For this example, since there are only $\mathcal{O}(\epsilon^2)$ deviations of the initial momentum from $m_\circ$ one expects $\Delta t_{\text{CFL}} = \mathcal{O}(1/\epsilon^2)$. This is confirmed by Table 2.



(a) Comparison of surface perturbation.

(b) Comparison of momentum.



(c) Zoomed view of (a).

(d) Zoomed view of (b).

Fig. 3: The comparison of the RS-IMEX solutions for Example (I) with $\epsilon = 0.8$, CFL $= 0.45$, $T = 0.05$, and the LaR and the zero-Froude limit reference states.

5.1.2. Colliding pulses (II). This example, which has been discussed in [2] uses the initial data

$$(\text{II}_a) \qquad \begin{aligned} h_0(x) &= 0.955 + \frac{\epsilon}{2}\left(1 - \cos(2\pi x)\right), \\ u_0(x) &= -\text{sgn}(x)\sqrt{2}\left(1 - \cos(2\pi x)\right). \end{aligned}$$

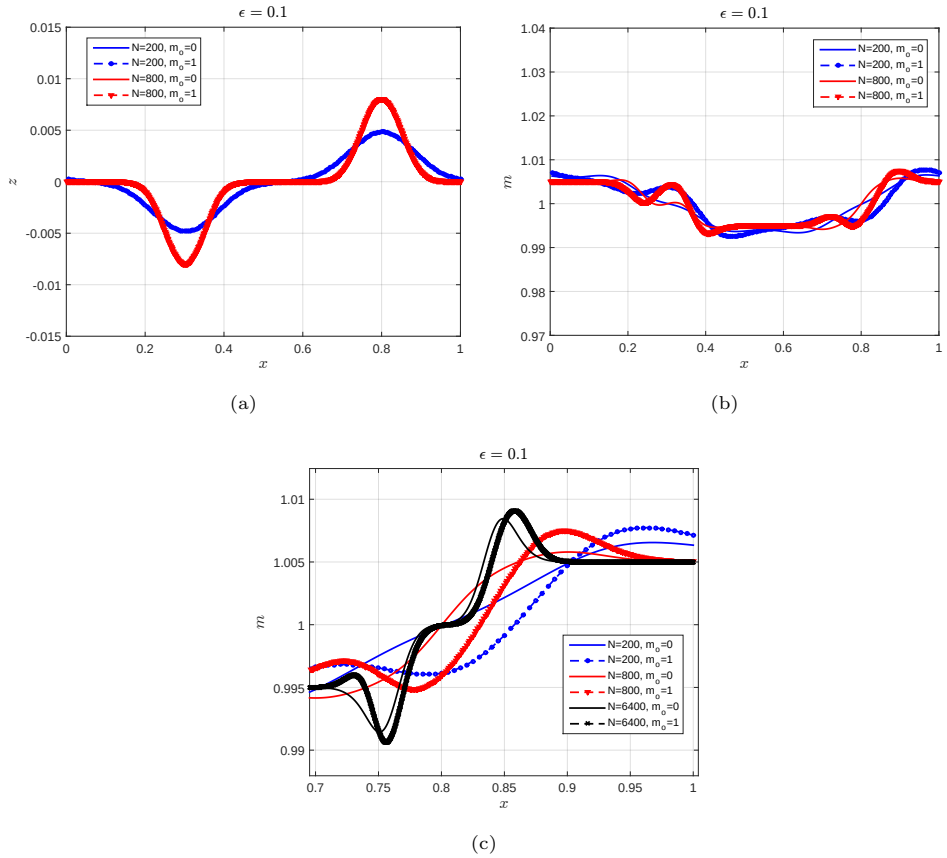


Fig. 4: (a) and (b): The comparison of the RS-IMEX solutions for Example (I) with $\epsilon = 0.1$, CFL $= 0.45$, $T = 0.05$, and the LaR and the zero-Froude limit reference states. (c) is a close-up of (b) with mesh refinement.

	ϵ		
	10^{-1}	10^{-3}	10^{-6}
LaR	1.06e-03	1.12e-03	1.12e-03
Constant	1.11e-01	1.12e+03	1.12e+09

Table 2: Comparison of Δt_{CFL} w.r.t. ϵ for different reference states in Example (I).

We set CFL $= 0.45$ and consider the problem in the periodic domain $[-1, 1)$.

Figure 7 shows the evolution of water height for the final time $T = 0.1$ and $\epsilon = 0.1$ with $N = 200$ and the lake at rest reference solution. We have also chosen $z_\circ = -0.045$. The figure shows that, comparing to [2], the computed solution is fine. Note that in [2], the height is computed by an elliptic approach; see [2] for more details.

Confirming the order of convergence, we keep the initial surface perturbation as in (II_a) and modify the initial velocity in case (II_a) to be solenoidal, i.e., div $u^0_{(0)} = 0$.

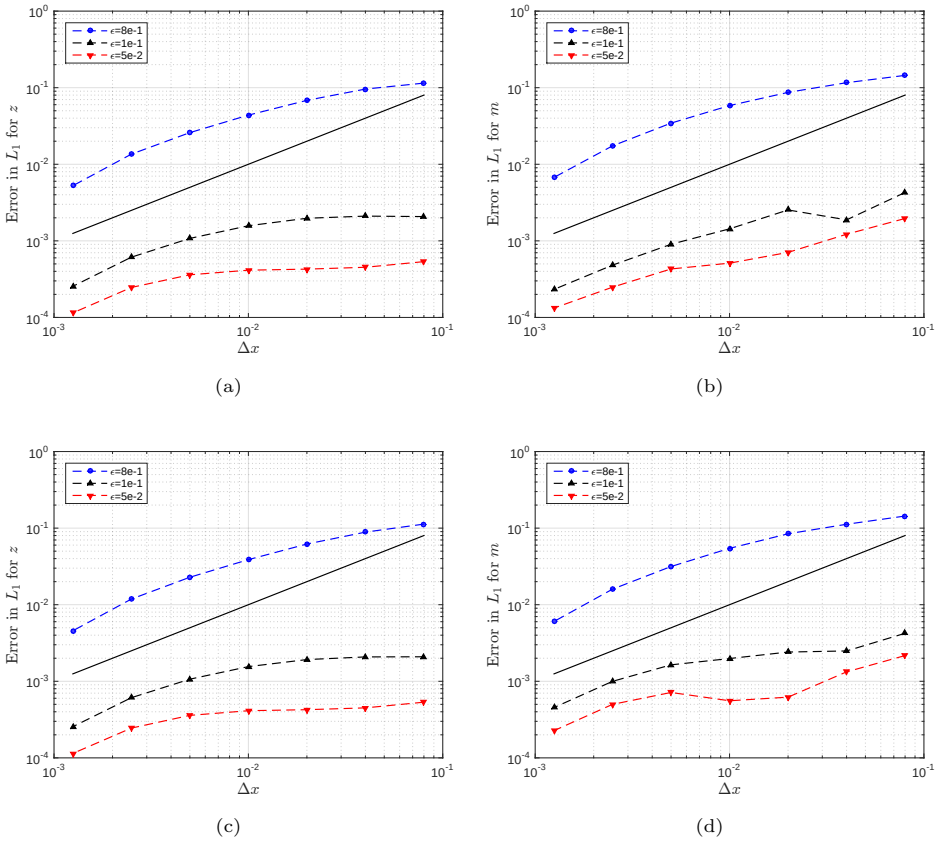


Fig. 5: The comparison of the order of convergence for the RS-IMEX solutions of Example (I) with CFL = 0.45, $T = 0.05$: (a) and (b) for the LaR ($m_\circ = 0$) reference state, (c) and (d) for the zero-Froude limit ($m_\circ = 1$) reference state.

For example:

$$u_0(x) = -\sqrt{2} + \sqrt{2}\,\epsilon\,\mathrm{sgn}(x)\cos(2\pi x). \tag{II$_b$}$$

As Figure 8 suggests, the experimental order of convergence is one, uniformly in ϵ, i.e., the scheme is uniformly consistent. Note that the error has been measured compared to the *exact* solution computed on a very fine mesh with $N = 3200$. Also note that the figure suggests that the error decreases as $\epsilon \to 0$, which is natural due to the well-prepared initial data. One may prefer to divide the error by ϵ^2 to see the *effective error reduction*; but, here we only care about the order of consistency.

Moreover, Figure 9 confirms the stability of the scheme in ℓ_2-norm, with the growth factor $\mathcal{G}_w$, which for the quantity w is defined as

$$\mathcal{G}_w^n := \frac{\|w_\Delta^n\|_{\ell_2}}{\|w_\Delta^{n-1}\|_{\ell_2}}. \tag{5.2}$$

As Figure 9 suggests, the scheme is stable uniformly in ϵ for variables like z, m and u. Also note that since there is an $\mathcal{O}(\epsilon)$ contribution in $h_0(x)$, v_1 is not $\mathcal{O}(1)$ as shown

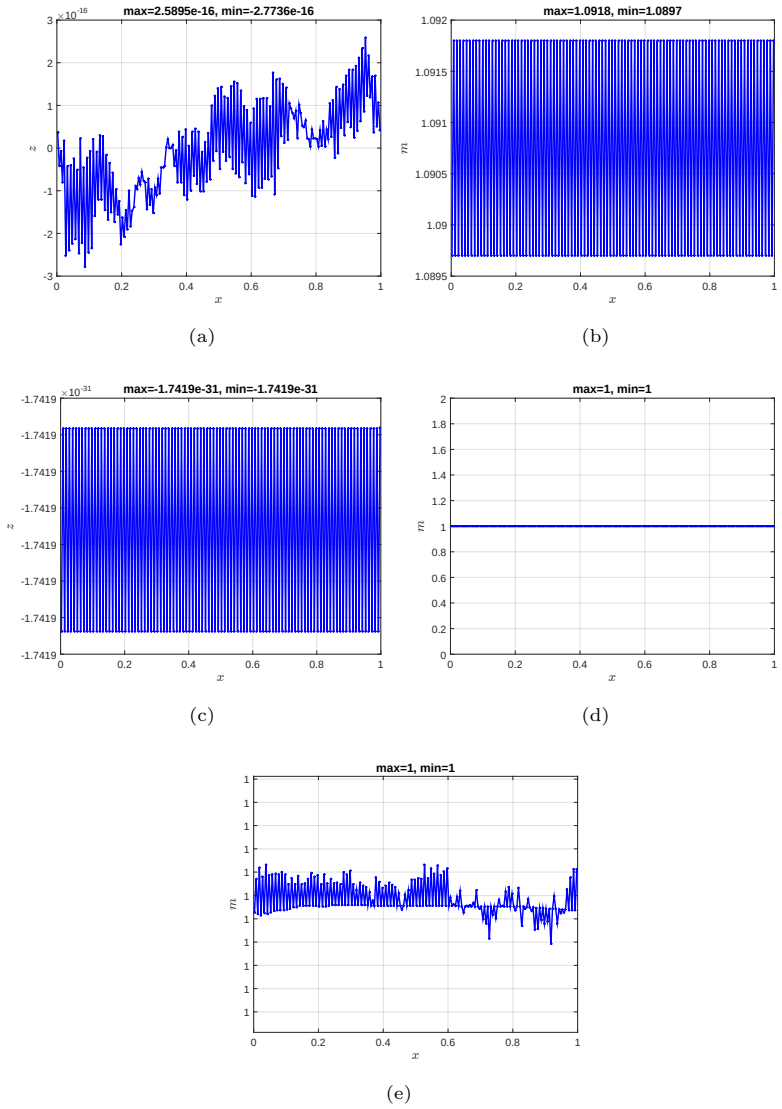
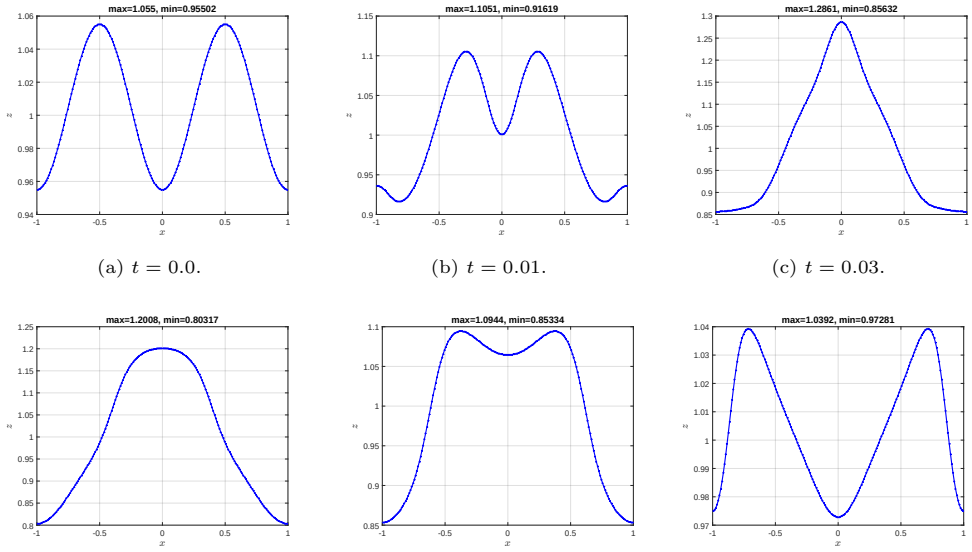


Fig. 6: Incompressible limit of the RS-IMEX solution for Example (I), with $N = 200$ and $\epsilon = 10^{-8}$. (a) and (b) are for the LaR reference solution with scaling matrix D_2, (c) and (d) are for the zero-Froude limit reference solution with scaling D_3. (e) is for the LaR reference solution with scaling matrix D_3.

in Lemma 3.6 and grows as $\mathcal{O}(1/\epsilon)$. The figures also show that the time step Δt does not depend on ϵ.

To compare the lake at rest and the zero-Froude limit reference solutions, for the case (II_a), we keep $z_\circ = -0.045$ and change the reference momentum to $m_\circ = \sqrt{2}$ (case II_c) (which is not the zero-Froude limit anymore!). As Figure 10 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 when the solution of each step has been perturbed significantly from the exact solution (here due to an unsuitable choice of the reference momentum). Figure 10 confirm this conjecture, as it shows that the solution tends to get symmetric with mesh refinement, i.e., as the operator splitting error gets smaller.



(a) $t = 0.0$. (b) $t = 0.01$. (c) $t = 0.03$.

(d) $t = 0.06$. (e) $t = 0.08$. (f) $t = 0.1$.

Fig. 7: RS-IMEX solution for Example (II_a) with $\epsilon = 0.1$, CFL $= 0.45$, $T = 0.1$ and the LaR reference solution.

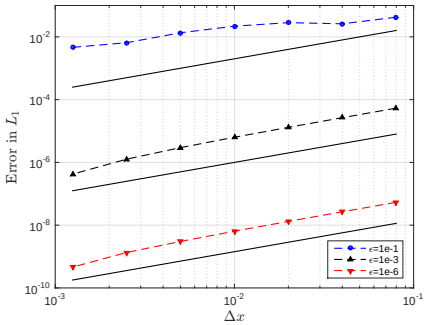
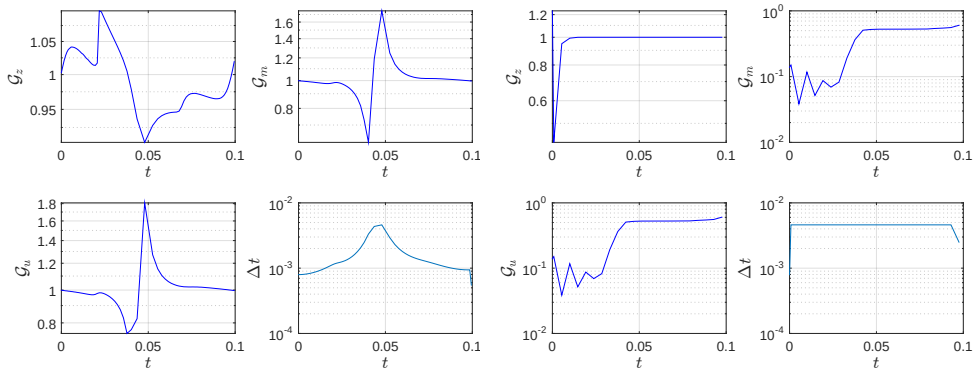


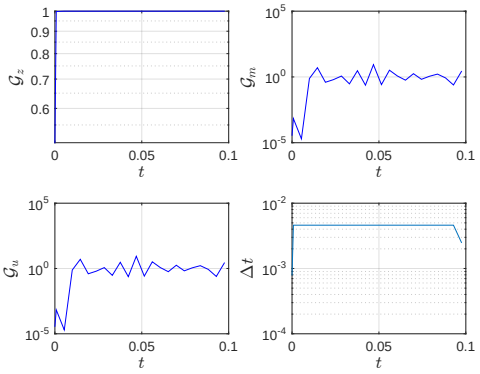
Fig. 8: Order of convergence of the RS-IMEX scheme for Example (II_b), for $\epsilon = 10^{-1}, 10^{-3}, 10^{-6}$. The error is measured for the momentum.

5.2. Shallow water equations with non-flat bottom (III). In this section, we study the result of the RS-IMEX scheme for the non-flat bottom case, and confirm the experimental order of convergence for a specific example. Also we verify the asymptotic consistency of the scheme, numerically. We set the initial condition as



(a) $\epsilon = 10^{-1}$.

(b) $\epsilon = 10^{-3}$.



(c) $\epsilon = 10^{-6}$.

Fig. 9: Growth factor and time step regarding ϵ, for Example (II_b) with the LaR reference solution.

follows

$$z^0(x) = \epsilon^2 \sin(2\pi x),$$
$$m^0(x) = -\sqrt{2} + \epsilon\sqrt{2}\,\mathrm{sgn}(x)\cos(2\pi x),$$

with the bottom function $b(x) = -(3 + \sin(\pi x))$, $\Delta t = 0.2\,\Delta x$, $T = 0.01$, $z_\circ = 0$, in a periodic domain $[-1, 1)$.

In Figure 11, the convergence rate of the scheme has been plotted, which shows super-convergence. Also Figure 12 shows the (almost) incompressible limit of the numerical solution with $\epsilon = 10^{-8}$; surface elevation is almost constant and the momentum is *div*-free. The figure clearly confirms the asymptotic consistency for the surface elevation, i.e., it is zero up to the machine accuracy. Also for the momentum, it can be obtained that the oscillations is of the order 10^{-14}. This justifies the asymptotic consistency of the scheme. Moreover, Table 3, shows the smallness of the checker-board oscillations for v_2. It can be seen that as ϵ approaches the limit, $\|[\![\mathbf{V}_{2,\Delta}]\!]\|_{\ell_\infty}$ (which indicates the amplitude of possible checker-board oscillations) decays with the rate of $\mathcal{O}(\epsilon)$, which is better than the analysis in Section 3.2.2, up to

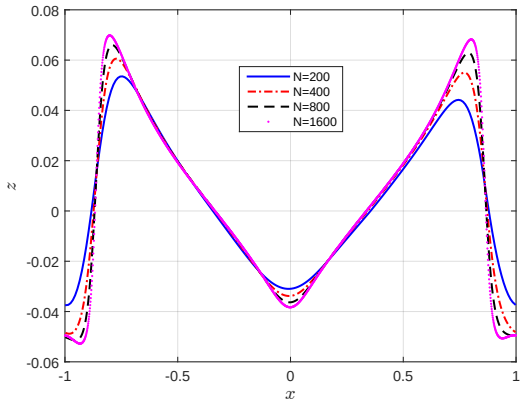
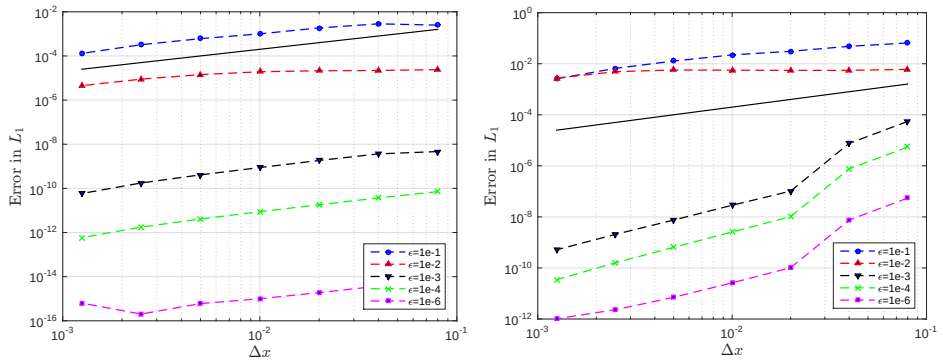


Fig. 10: Vanishing effect of an unsuitable reference solution for Example (II_c) as $\Delta x \to 0$, for $\epsilon = 10^{-1}$, $T = 0.1$ and $N = 200, 400, 800, 1600$.

some threshold ϵ where the condition number of J_ϵ gets very large and affects the solution. It can also seen that $\mathrm{cond}_2(J_\epsilon) = \mathcal{O}(1/\epsilon)$. Regarding the mesh refinement, the condition number is almost constant: The refinement can improve the oscillations to some extent (for rather coarse meshes); however after some point, the amplitude of the oscillations does not change with Δx.



(a) Error in z. (b) Error in m.

Fig. 11: Order of convergence of the RS-IMEX scheme for Example (III), with $T = 0.01$, $\Delta t = 0.2\,\Delta x$ and the LaR reference solution.

6. Concluding remarks. In this paper, we have analyzed the RS-IMEX scheme for the shallow water equations w.r.t. the Froude number. The scheme has been presented in one space dimension and its quality is guaranteed by numerical analysis as well as several numerical tests. In practice, we have shown that the scheme is uniformly stable and consistent, when the analysis confirms the asymptotic preserving property for the scheme, as well as C-property regarding the lake at rest equilibrium state. Indeed, the asymptotic consistency and stability analyses are not only formal

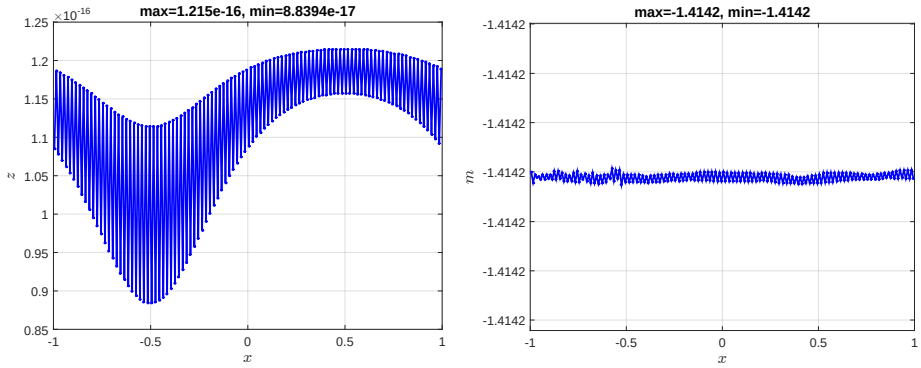


Fig. 12: Incompressible limit of the RS-IMEX scheme for Example (III), computed with $\epsilon = 10^{-8}$, $N = 200$, $T = 0.01$ and $\Delta t = 0.2\,\Delta x$.

ϵ	N	$\|\|[\![\mathbf{V}_{2,\Delta}^{n+1}]\!]\|\|_{\ell_\infty}$	$\mathrm{cond}_2(J_\epsilon)$	ϵ	N	$\|\|[\![\mathbf{V}_{2,\Delta}^{n+1}]\!]\|\|_{\ell_\infty}$	$\mathrm{cond}_2(J_\epsilon)$
10^{-2}	200	3.68e-05	8.72e+01	10^{-6}	50	1.52e-08	7.98e+05
10^{-3}	200	8.30e-10	8.18e+03	10^{-6}	100	1.75e-11	8.11e+05
10^{-4}	200	5.97e-11	8.17e+03	10^{-6}	200	5.89e-13	8.17e+05
10^{-5}	200	5.83e-12	8.17e+04	10^{-6}	400	1.95e-14	8.21e+05
10^{-6}	200	5.89e-13	8.17e+05	10^{-6}	800	7.73e-14	8.22e+05
10^{-7}	200	6.91e-14	8.17e+06	10^{-6}	1600	2.54e-13	8.23e+05
10^{-8}	200	1.49e-14	8.17e+07				
10^{-9}	200	1.29e-14	8.17e+08				

Table 3: Smallness of the checker-board oscillations regarding the refinement in ϵ and Δx for Example (III).

but also rigorous.

These results are so far for two reference solutions, the lake at rest and the zero-Froude limit, and limited to one space dimension and first-order schemes on periodic domains. As we have seen, even with these assumption the AP analysis is delicate. Extending the analysis to multi-dimensions, with more complicated source terms and boundary conditions are left for future works and is in progress. For example in [61] it would be shown that a similar analysis (as in Section 3.2) can be utilized for the two-dimensional RS-IMEX scheme applied to the shallow water equations.

A. Asymptotic analysis of shallow water equations. This section is to provide the formal asymptotic analysis for the low-Froude shallow water equations in one space dimension. On a periodic domain Ω, consider the usual formulation of the non-dimensionalized shallow water equations with η^b as the bottom function:

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

Then, we expand h and m in terms of the Froude number ϵ as

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

Substituting (A.2) in (A.1), $\mathcal{O}(\epsilon^{-2})$ 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 $\eta^{\mathrm{s}}_{(0)} := h_{(0)} + \eta^b = \eta^s_{(0)}(t)$. Using this, one can find for the higher order terms that $h_{(0)}\,\partial_x h_{(1)} = 0$ which leads to constant $h_{(1)}$ in space, i.e., $h_{(1)} = h_{(1)}(t)$.

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

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \left(h_{(0)} + \eta^b\right)\mathrm{d}x = -\int_{\partial\Omega} m_{(0)}.\mathbf{n}\mathrm{d}s = 0,$$

owing to the divergence theorem and the assumption of periodic boundary conditions. Thus $\partial_t h_{(0)} = 0$ and $\eta^s_{(0)} = \text{const.}$, which gives

$$h_{(0)} = h_{(0)}(x) = \eta^s_{(0)} - \eta^b(x),$$

and hence $m_{(0)} = m_{(0)}(t)$. 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

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

Thus the leading order momentum does not evolve in time when the bottom is flat, i.e., $\partial_t m_{(0)} = 0$.

Summing up all these results gives Definition 3.2 for the formal asymptotic limit of the shallow water equations.

REFERENCES

[1] Saul Abarbanel, Pravir Duth, and David Gottlieb, *Splitting methods for low Mach number Euler and Navier–Stokes equations*, Computers & fluids **17** (1989), no. 1, 1–12.

[2] Koottungal Revi Arun and Sebastian Noelle, *An asymptotic preserving scheme for low Froude number shallow flows*, IGPM report 352, RWTH Aachen University, 2012.

[3] Ivar Bendixson, *Sur les racines d'une équation fondamentale*, Acta Mathematica **25** (1902), no. 1, 359–365 (French).

[4] Dennis S. Bernstein, *Matrix mathematics: Theory, facts, and formulas*, Princeton University Press, 2009.

[5] Georgij Bispen, *IMEX finite volume methods for the shallow water equations*, Ph.D. thesis, Johannes Gutenberg-Universität, 2015.

[6] Georgij Bispen, Koottungal Revi Arun, Mária Lukáčová-Medviďová, and Sebastian Noelle, *IMEX large time step finite volume methods for low Froude number shallow water flows*, Communication in Computational Physics **16** (2014), 307–347.

[7] François Bouchut and Michael Westdickenberg, *Gravity driven shallow water models for arbitrary topography*, Communications in Mathematical Sciences **2** (2004), no. 3, 359–389.

[8] François Bouchut, *Nonlinear stability of finite volume methods for hyperbolic conservation laws: And well-balanced schemes for sources*, Springer Science & Business Media, 2004.

[9] Floraine Cordier, Pierre Degond, and Anela Kumbaro, *An Asymptotic–Preserving all–speed scheme for the Euler and Navier–Stokes equations*, Journal of Computational Physics **231** (2012), no. 17, 5685–5704.

[10] Raphaël Danchin, *Low Mach number limit for viscous compressible flows*, ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique **39** (2005), no. 3, 459–475.

[11] Pierre Degond, Jian-Guo Liu, and Marie-Hélène Vignal, *Analysis of an asymptotic preserving scheme for the Euler–Poisson system in the quasineutral limit*, SIAM Journal on Numerical Analysis **46** (2008), no. 3, 1298–1322.
[12] Pierre Degond and Min Tang, *All speed scheme for the low Mach number limit of the isentropic Euler equation*, Communications in Computational Physics **10** (2011), no. 1, 1–31.
[13] Stéphane Dellacherie, *Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number*, Journal of Computational Physics **229** (2010), no. 4, 978–1016.
[14] Stéphane Dellacherie, Pascal Omnes, and Felix Rieper, *The influence of cell geometry on the Godunov scheme applied to the linear wave equation*, Journal of Computational Physics **229** (2010), no. 14, 5315–5338.
[15] Giacomo Dimarco, Raphaël Loubère, and Marie-Hélène Vignal, *Study of a new asymptotic preserving scheme for the Euler system in the low Mach number limit*, Preprint, 2016.
[16] Jan Giesselmann, *Low Mach asymptotic-preserving scheme for the Euler–Korteweg model*, IMA Journal of Numerical Analysis **35** (2015), no. 2, 802–833.
[17] Francis X. Giraldo and Marco Restelli, *High-order semi-implicit time-integrators for a triangular discontinuous Galerkin oceanic shallow water model*, International journal for numerical methods in fluids **63** (2010), no. 9, 1077–1102.
[18] François Golse, Shi Jin, and Charles D. Levermore, *The convergence of numerical transfer schemes in diffusive regimes I: Discrete-ordinate method*, SIAM journal on numerical analysis **36** (1999), no. 5, 1333–1369.
[19] Gene H. Golub and Charles F. Van Loan, *Matrix computations*, vol. 3, JHU Press, 2012.
[20] Robert M. Gray, *Toeplitz and circulant matrices: A review*, Now Publishers Inc., 2006.
[21] Emmanuel Grenier, *Oscillatory perturbations of the Navier–Stokes equations*, Journal de Mathématiques Pures et Appliquées **76** (1997), no. 6, 477–498.
[22] Hervé Guillard and Angelo Murrone, *On the behavior of upwind schemes in the low Mach number limit: II. Godunov type schemes*, Computers & fluids **33** (2004), no. 4, 655–675.
[23] Hervé Guillard and Cécile Viozat, *On the behaviour of upwind schemes in the low Mach number limit*, Computers & fluids **28** (1999), no. 1, 63–86.
[24] Jeffrey Haack, Shi Jin, and Jian-Guo Liu, *An all-speed asymptotic-preserving method for the isentropic Euler and Navier–Stokes equations*, Communications in Computational Physics **12** (2012), no. 4, 955–980.
[25] M. A. Hirsch, *Sur les racines d'une équation fondamentale*, Acta Mathematica **25** (1902), no. 1, 367–370 (French).
[26] Roger A. Horn and Charles R. Johnson, *Matrix analysis*, Cambridge University Press, New York, NY, USA, 1986.
[27] Ilse C. F. Ipsen, *Numerical matrix analysis: Linear systems and least squares*, SIAM, 2009.
[28] Juhi Jang, Fengyan Li, Jing-Mei Qiu, and Tao Xiong, *Analysis of asymptotic preserving DG-IMEX schemes for linear kinetic transport equations in a diffusive scaling*, SIAM Journal on Numerical Analysis **52** (2014), no. 4, 2048–2072.
[29] Shi Jin, *Runge–Kutta methods for hyperbolic conservation laws with stiff relaxation terms*, Journal of Computational Physics **122** (1995), no. 1, 51–67.
[30] ———, *Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations*, SIAM Journal on Scientific Computing **21** (1999), no. 2, 441–454.
[31] ———, *Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: A review*, Lecture Notes for Summer School on Methods and Models of Kinetic Theory(M&MKT), Porto Ercole (Grosseto, Italy) (2010), 177–216.
[32] Shi Jin, Min Tang, and Houde Han, *A uniformly second order numerical method for the one-dimensional discrete-ordinate transport equation and its diffusion limit with interface.*, NHM **4** (2009), no. 1, 35–65.
[33] Klaus Kaiser, Jochen Schütz, Ruth Schöbel, and Sebastian Noelle, *A new stable splitting for the isentropic Euler equations*, IGPM report 442, RWTH Aachen University, Submitted for publication, 2016.
[34] Sergiu Klainerman and Andrew Majda, *Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids*, Communications on Pure and Applied Mathematics **34** (1981), no. 4, 481–524.
[35] ———, *Compressible and incompressible fluids*, Communications on Pure and Applied Mathematics **35** (1982), no. 5, 629–651.
[36] Rupert Klein, *Semi-implicit extension of a Godunov-type scheme based on low Mach number asymptotics I: One-dimensional flow*, Journal of Computational Physics **121** (1995), no. 2, 213–237.
[37] Rupert Klein, Nicola Botta, Thomas Schneider, Claus-Dieter Munz, Sabine Roller, Andreas

Meister, L Hoffmann, and Thomas Sonar, *Asymptotic adaptive methods for multi-scale problems in fluid mechanics*, Practical Asymptotics, Springer, 2001, pp. 261–343.

[38] H-O Kreiss, J Lorenz, and MJ Naughton, *Convergence of the solutions of the compressible to the solutions of the incompressible Navier–Stokes equations*, Advances in Applied Mathematics **12** (1991), no. 2, 187–214.

[39] Edward W. Larsen, J. E. Morel, and Warren F. Miller, *Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes*, Journal of Computational Physics **69** (1987), no. 2, 283–324.

[40] Mohammed Lemou and Luc Mieussens, *A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit*, SIAM Journal on Scientific Computing **31** (2008), no. 1, 334–368.

[41] Qiuhua Liang and Alistair G. L. Borthwick, *Adaptive quadtree simulation of shallow flows with wet–dry fronts over complex topography*, Computers & Fluids **38** (2009), no. 2, 221–234.

[42] Qiuhua Liang and Fabien Marche, *Numerical resolution of well-balanced shallow water equations with complex source terms*, Advances in water resources **32** (2009), no. 6, 873–884.

[43] Jian-Guo Liu and Luc Mieussens, *Analysis of an asymptotic preserving scheme for linear kinetic equations in the diffusion limit*, SIAM Journal on Numerical Analysis **48** (2010), no. 4, 1474–1491.

[44] Andrew Majda, *Compressible fluid flow and systems of conservation laws in several space variables*, vol. 53, Springer Science & Business Media, 2012.

[45] Nader Masmoudi, *Asymptotic problems and compressible-incompressible limit*, Advances in Mathematical Fluid Mechanics, Springer, 2000, pp. 119–158.

[46] ———, *Examples of singular limits in hydrodynamics*, Handbook of Differential Equations: Evolutionary Equations **3** (2007), 195–275.

[47] Guy Métivier and Steve Schochet, *The incompressible limit of the non-isentropic Euler equations*, Archive for rational mechanics and analysis **158** (2001), no. 1, 61–90.

[48] Sebastian Noelle, Georgij Bispen, Koottungal Revi Arun, Mária Lukáčová-Medviďová, and Claus-Dieter Munz, *A weakly asymptotic preserving low Mach number scheme for the Euler equations of gas dynamics*, SIAM Journal on Scientific Computing **36** (2014), no. 6, B989–B1024.

[49] Sebastian Noelle, Rupert Klein, Jochen Schütz, and Hamed Zakerzadeh, *RS-IMEX schemes: Derivation and asymptotic stability*, In preparation, 2016.

[50] Robert-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.

[51] Felix Rieper, *On the dissipation mechanism of upwind–schemes in the low Mach number regime: A comparison between Roe and HLL*, Journal of Computational Physics **229** (2010), no. 2, 221–232.

[52] ———, *A low–Mach number fix for Roe's approximate Riemann solver*, Journal of Computational Physics **230** (2011), no. 13, 5263–5287.

[53] Felix Rieper and Georg Bader, *The influence of cell geometry on the accuracy of upwind schemes in the low Mach number regime*, Journal of Computational Physics **228** (2009), no. 8, 2918–2933.

[54] Benedict D. Rogers, Alistair G. L. Borthwick, and Paul H. Taylor, *Mathematical balancing of flux gradient and source terms prior to using Roes approximate Riemann solver*, Journal of Computational Physics **192** (2003), no. 2, 422–451.

[55] Benedict D. Rogers, Masayuki Fujihara, and Alistair G. L. Borthwick, *Adaptive Q-tree Godunov-type scheme for shallow water equations*, International Journal for Numerical Methods in Fluids **35** (2001), no. 3, 247–280.

[56] Steven Schochet, *The mathematical theory of low Mach number flows*, ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique **39** (2005), no. 3, 441–458.

[57] Jochen Schütz and Klaus Kaiser, *A new stable splitting for singularly perturbed ODEs*, Applied Numerical Mathematics **107** (2016), 18–33.

[58] Jochen Schütz and Sebastian Noelle, *Flux splitting for stiff equations: A notion on stability*, Journal of Scientific Computing (2014), 1–19 (English).

[59] John R. Silvester, *Determinants of block matrices*, The Mathematical Gazette (2000), 460–467.

[60] Lloyd Nicholas Trefethen, *Finite difference and spectral methods for ordinary and partial differential equations*, Cornell University-Department of Computer Science and Center for Applied Mathematics, 1996.

[61] Hamed Zakerzadeh, *Asymptotic analysis of the all-Froude RS-IMEX scheme for the two-dimensional shallow water equations of arbitrary topography*, In preparation, 2016.

[62] ———, *On the Mach–uniformity of the Lagrange–projection scheme*, IGPM report 422, RWTH Aachen University, Submitted for publication, 2016.
[63] Hamed Zakerzadeh and Sebastian Noelle, *A note on the stability of implicit-explicit flux-splittings for stiff systems of hyperbolic conservation laws*, IGPM report 449, RWTH Aachen University, Submitted for publication, 2016.