**ESAIM: PROCEEDINGS AND SURVEYS,** December 2017, Vol. 58, p. 1-26
Stéphane DELLACHERIE, Gloria FACCANONI, Bérénice GREC, Frédéric LAGOUTIERE, Yohan PENEL

# GODUNOV TYPE SCHEME FOR THE LINEAR WAVE EQUATION WITH CORIOLIS SOURCE TERM

Emmanuel Audusse[1], Stéphane Dellacherie[2, 3], Minh Hieu Do[1], Pascal Omnes[1,2] and Yohan Penel[4, 5]

**Abstract.** We propose a method to explain the behaviour of the Godunov finite volume scheme applied to the linear wave equation with Coriolis source term at low Froude number. In particular, we use the Hodge decomposition and we study the properties of the modified equation associated to the Godunov scheme. Based on the structure of the discrete kernel of the linear operator discretized by using the Godunov scheme, we clearly explain the inaccuracy of the classical Godunov scheme at low Froude number and we introduce a way to modify it to recover a correct accuracy.

## 1. Introduction

In this communication, we study some procedures to make finite volume Godunov type schemes accurate when solving perturbations around a steady-state. In what follows, we restrict the analysis to the quasi-1d linear wave equation with Coriolis term. Nevertheless, our future main objective is to derive accurate and stable finite volume collocated schemes for the dimensionless shallow water equations

$$\begin{cases} \mathsf{St}\,\partial_t h + \nabla\cdot(h\bar{\mathbf{u}}) = 0, & \text{(1a)} \\ \mathsf{St}\,\partial_t(h\bar{\mathbf{u}}) + \nabla\cdot(h\bar{\mathbf{u}}\otimes\bar{\mathbf{u}}) + \dfrac{1}{\mathsf{Fr}^2}\nabla\left(\dfrac{h^2}{2}\right) = -\dfrac{1}{\mathsf{Fr}^2}h\nabla b - \dfrac{1}{\mathsf{Ro}}h\bar{\mathbf{u}}^{\perp}, & \text{(1b)} \end{cases}$$

in a rotating frame when the flow is a perturbation around the so-called geostrophic equilibrium. In System (1) unknowns $h$ and $\bar{\mathbf{u}}$ respectively denote the water depth and the velocity of the water column and function $b(x)$ denotes the topography of the considered oceanic basin and is a given function. Dimensionless numbers $\mathsf{St}$, $\mathsf{Fr}$ and $\mathsf{Ro}$ respectively stand for the Strouhal, the Froude and the Rossby numbers defined by

$$\mathsf{St} = \frac{L}{UT}, \quad \mathsf{Fr} = \frac{U}{\sqrt{gH}}, \quad \mathsf{Ro} = \frac{U}{\Omega L}$$

[1] Université Paris 13, LAGA, CNRS UMR 7539, Institut Galilée, 99 Avenue J.-B. Clément, 93430 Villetaneuse Cedex, France
[2] Commissariat à l'Énergie Atomique et aux Énergies Alternatives, CEA, DEN, DM2S-STMF, 91191 Gif-Sur-Yvette, France
[3] Hydro-Québec, TransÉnergie, 75 boulevard René-Lévesque Ouest, Montréal (Qc), H2Z 1A4, Canada (current address)
[4] Team ANGE, CEREMA (Ministry of Ecology, Sustainable Development and Energy) and Inria–Paris
[5] Sorbonne Universités, UPMC Univ. Paris 06 and CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, 75005, Paris, France


Article published online by EDP Sciences and available at http://www.esaim-proc.org or https://doi.org/10.1051/proc/201758001

where the parameters $g$ and $\Omega$ denote the gravity coefficient and the angular velocity of the Earth. Constants $U$, $H$, $L$ and $T$ are some characteristic velocity, vertical and horizontal lengths and time. In the sequel, we shall focus on cases where

$$\mathsf{Ro} = \mathcal{O}(M) \quad \text{and} \quad \mathsf{Fr} = \mathcal{O}(M) \tag{2}$$

with $M$ a small parameter. For large scale oceanographic flows, typical values lead to $M \sim 10^{-2}$ since

$$U \approx 1\,\mathrm{m}\cdot\mathrm{s}^{-1}, \quad L \approx 10^6\,\mathrm{m}, \quad H \approx 10^3\,\mathrm{m}, \quad \Omega \approx 10^{-4}\,\mathrm{rad}\cdot\mathrm{s}^{-1}.$$

In order to exhibit some asymptotic regimes for small Froude and Rossby numbers, we perform an expansion of the unknowns such that

$$f(t,x) = f_0(t,x) + M f_1(t,x) + M^2 f_2(t,x) + \mathcal{O}(M^3) \tag{3}$$

given the orders of magnitude (2). We first focus on long time regimes, i.e. for Strouhal number of order $\mathcal{O}(1)$. At the leading order, solutions of equations (1) satisfy the so-called lake at rest equilibrium

$$\nabla\,(h_0 + b) = 0. \tag{4}$$

At the next order, the flow satisfies the so-called geostrophic equilibrium

$$\nabla h_1 = -\bar{\mathbf{u}}_0^{\perp}. \tag{5}$$

Note that this relation implies

$$\nabla \cdot \bar{\mathbf{u}}_0 = 0. \tag{6}$$

The ability of numerical schemes to well capture the particular solutions (4) and (5) is of great practical interest since it has a direct consequence on the accuracy of the numerical solution when perturbations around these equilibria are considered. A substantial amount of articles in the literature has been devoted to the preservation of the lake at rest equilibrium (4), see in particular [3] and references therein.

The question of the geostrophic equilibrium (5) including the divergence constraint (6) is more complex. It has been studied in a finite element framework by Le Roux [10]. The author considers in his work the linearised version of System (1) and studies the behaviour of several types of finite elements. He shows that spurious modes are created, in particular when the number of degrees of freedom is not the same for height and velocity unknowns. In the finite volume framework, the nonlinear case has been studied in [4, 6, 12]. In particular, Bouchut and coauthors introduce in [4] the *apparent topography method* that allows to adapt to this problem the hydrostatic reconstruction method [2] that was developed to ensure the preservation of the lake at rest equilibrium (4).

Let us now focus on the behaviour of solutions of System (1) for short times, i.e. for Strouhal number of order $\mathcal{O}(M^{-1})$. Here the study is restricted to some flat topography and solutions independent of the $y$ direction. The asymptotic expansion (3) is inserted in System (1). At the leading order, any solution of System (1) satisfies the quasi-1d linear wave equation with Coriolis source term[1]

$$\begin{cases} \partial_t r + a_\star \partial_x u = 0, \\ \partial_t u + a_\star \partial_x r = \omega v, \\ \partial_t v = -\omega u \end{cases} \tag{7}$$

[1]For the sake of simplicity, we note $r = h_1$, $u = u_0$ and $v = v_0$ in (7).

where $a_\star$ and $\omega$ are constants of order one, respectively related to the wave velocity and to the rotating velocity. The stationary state corresponding to Equation (7) is the 1d version of the geostrophic equilibrium (5) and is called *1D geostrophic equilibrium*. It is such that

$$u = 0, \quad a_\star \partial_x r = \omega v. \tag{8}$$

Many works were devoted to the study of the homogeneous wave equations. In particular, in a serie of articles [8,9], Dellacherie and coauthors studied the behaviour of Godunov type schemes for the 2d linear wave equation. Their works are part of a more general study about the use of Godunov type schemes in the context of the incompressible limit for Euler equations, *i.e.* for low Mach number flows, see for example [14, 16, 17]. Similar works are related to low Froude flows [7, 15]. In the present work, we extend the aforementioned approach from Dellacherie and coauthors to take into account the Coriolis source term. First, in Section 2, we analyze the continuous case by using a Hodge type decomposition. Then, in Section 3 and 4, we study three Godunov type numerical schemes to compute approximate solutions of Equation (7):

- The Classical Godunov scheme;
- The *Low Froude* Godunov scheme;
- The *All Froude* Godunov scheme.

For each scheme, we study the kernel of the discrete operator and we compare it to the continuous kernel (8). Then, we study the accuracy of the scheme at low Froude number, *i.e.* when the initial solution is close to the kernel. This is done first for the modified equation associated to the scheme in Section 3 and then in the fully discrete case in Section 4. Moreover we study the stability of each discrete scheme by using Fourier analysis. Finally we present in Section 5 some numerical results to illustrate our purpose.

## 2. Properties of the linear wave equation with Coriolis source term

We first focus on the properties of the linear wave equation on the 1d torus $\mathbb{T}$. To begin with, we introduce the Hilbert space

$$\left(L^2(\mathbb{T})\right)^3 = \left\{ q = (r,u,v) \;\middle|\; \int_{\mathbb{T}} r^2 \,\mathrm{d}x + \int_{\mathbb{T}} \left(u^2+v^2\right) \,\mathrm{d}x < \infty \right\}$$

equipped with the scalar product

$$\langle q_1, q_2 \rangle = \int_{\mathbb{T}} r_1 r_2 \,\mathrm{d}x + \int_{\mathbb{T}} (u_1 u_2 + v_1 v_2) \,\mathrm{d}x.$$

### 2.1. Structure of the kernel of the original model

Let us define the following space

$$\mathcal{E}_{\omega\neq 0} = \left\{ q = (r,u,v) \in \left(L^2(\mathbb{T})\right)^3 \;\middle|\; u = 0,\ \forall\, \phi \in C_c^\infty(\mathbb{T}),\ \int_{\mathbb{T}} a_\star r \partial_x \phi \,\mathrm{d}x = -\int_{\mathbb{T}} \omega v \phi \,\mathrm{d}x \right\}. \tag{9}$$

We then prove this preliminary result:

**Lemma 1.** *The orthogonal of $\mathcal{E}_{\omega\neq 0}$ is*

$$\mathcal{E}_{\omega\neq 0}^\perp = \left\{ q = (r,u,v) \in \left(L^2(\mathbb{T})\right)^3 \;\middle|\; \forall\, \varphi \in C_c^\infty(\mathbb{T}),\ \int_{\mathbb{T}} a_\star v \partial_x \varphi \,\mathrm{d}x = -\int_{\mathbb{T}} \omega r \varphi \,\mathrm{d}x \right\}.$$

*Moreover, we have* $\mathcal{E}_{\omega\neq 0}\oplus\mathcal{E}_{\omega\neq 0}^{\perp}=\left(L^2(\mathbb{T})\right)^3$. *In other words, any* $q\in\left(L^2(\mathbb{T})\right)^3$ *can be uniquely decomposed into*

$$q=\hat{q}+\tilde{q} \tag{10}$$

*where* $\hat{q}\in\mathcal{E}_{\omega\neq 0}$ *and* $\tilde{q}\in\mathcal{E}_{\omega\neq 0}^{\perp}$.

The Hodge decomposition (10) allows us to define the orthogonal projection

$$\mathbb{P}:\left\{\begin{array}{ccc}\left(L^2(\mathbb{T})\right)^3 & \longrightarrow & \mathcal{E}_{\omega\neq 0}\\ q & \longmapsto & \hat{q}\end{array}\right. \tag{11}$$

*Remark* 1. The kernel and its orthogonal set can be described in a simpler way due to the definition of Sobolev spaces, namely

$$\begin{aligned}\mathcal{E}_{\omega\neq 0}&=\left\{q=(r,u,v)\in\left(L^2(\mathbb{T})\right)^3\;\middle|\;r\in H^1(\mathbb{T}),\ u=0,\ v=\frac{a_\star}{\omega}r'\right\},\\ \mathcal{E}_{\omega\neq 0}^{\perp}&=\left\{q=(r,u,v)\in\left(L^2(\mathbb{T})\right)^3\;\middle|\;v\in H^1(\mathbb{T}),\ r=\frac{a_\star}{\omega}v'\right\}.\end{aligned}$$

Moreover, the fact that we consider periodic functions implies that for $\hat{q}\in\mathcal{E}_{\omega\neq 0}$ and $\tilde{q}\in\mathcal{E}_{\omega\neq 0}^{\perp}$, we have

$$\int_{\mathbb{T}}\hat{v}\ \mathrm{d}x=0\quad\text{ and }\quad\int_{\mathbb{T}}\tilde{r}\ \mathrm{d}x=0$$

due to boundary conditions.

*Proof.* Our purpose is to prove that $\mathcal{E}_{\omega\neq 0}^{\perp}=A$ where

$$A=\left\{q=\left(\frac{a_\star}{\omega}v',u,v\right)\;\middle|\;u\in L^2(\mathbb{T}),\ v\in H^1(\mathbb{T})\right\}.$$

Firstly, let us prove that $A\subset\mathcal{E}_{\omega\neq 0}^{\perp}$. For $\tilde{q}\in A$, we have

$$\forall\, q\in\mathcal{E}_{\omega\neq 0},\ \langle\tilde{q},q\rangle=\int_{\mathbb{T}}r\frac{a_\star}{\omega}\tilde{v}'\,\mathrm{d}x+\int_{\mathbb{T}}\frac{a_\star}{\omega}r'\tilde{v}\,\mathrm{d}x=\frac{a_\star}{\omega}\left(\int_{\mathbb{T}}r\tilde{v}'\,\mathrm{d}x+\int_{\mathbb{T}}r'\tilde{v}\,\mathrm{d}x\right).$$

According to [5, Corollary 8.10] with $(r,\tilde{v})\in\left(H^1(\mathbb{T})\right)^2$, we have $r\tilde{v}\in H^1(\mathbb{T})$ and $(r\tilde{v})'=r'\tilde{v}+r\tilde{v}'$. Therefore, we obtain, thanks to periodic boundary conditions on $\mathbb{T}$

$$\int_{\mathbb{T}}r\tilde{v}'\,\mathrm{d}x+\int_{\mathbb{T}}r'\tilde{v}\,\mathrm{d}x=\int_{\mathbb{T}}(r\tilde{v})'\,\mathrm{d}x=0,$$

which leads to $\langle\tilde{q},q\rangle=0$, $\forall\, q\in\mathcal{E}_{\omega\neq 0}$. It means that $\tilde{q}\in\mathcal{E}_{\omega\neq 0}^{\perp}$ .

Secondly, we prove that $\mathcal{E}_{\omega\neq 0}^{\perp}\subset A$. Let $\tilde{q}\in\mathcal{E}_{\omega\neq 0}^{\perp}$. Therefore

$$\forall\, r\in H^1(\mathbb{T}),\ \int_{\mathbb{T}}\tilde{r}r\,\mathrm{d}x+\int_{\mathbb{T}}\frac{a_\star}{\omega}\tilde{v}r'\,\mathrm{d}x=0,$$

which implies

$$\forall\, r\in C_c^\infty(\mathbb{T}),\ \int_{\mathbb{T}}\tilde{v}r'\,\mathrm{d}x=-\frac{\omega}{a_\star}\int_{\mathbb{T}}\tilde{r}r\,\mathrm{d}x.$$

As a result, $\tilde{v} \in H^1(\mathbb{T})$ and $a_\star \tilde{v}' = \omega \tilde{r}$. We come to the conclusion that

$$\mathcal{E}^\perp_{\omega\neq 0} = A = \left\{ q = (r,u,v) \in \left(L^2(\mathbb{T})\right)^3 \;\middle|\; \forall\, \varphi \in C^\infty_c(\mathbb{T}), \int_{\mathbb{T}} a_\star v \partial_x \varphi \,\mathrm{d}x = -\int_{\mathbb{T}} \omega r \varphi \,\mathrm{d}x \right\}.$$

We eventually have to prove that

$$\mathcal{E}_{\omega\neq 0} \oplus \mathcal{E}^\perp_{\omega\neq 0} = \left(L^2(\mathbb{T})\right)^3.$$

We only have to check $\left(L^2(\mathbb{T})\right)^3 \subset \mathcal{E}_{\omega\neq 0} \oplus \mathcal{E}^\perp_{\omega\neq 0}$, because of the fact that $\mathcal{E}_{\omega\neq 0} \oplus \mathcal{E}^\perp_{\omega\neq 0} \subset \left(L^2(\mathbb{T})\right)^3$ is trivial.

For $q = (r,u,v) \in \left(L^2(\mathbb{T})\right)^3$, let us set

$$\begin{aligned} &\hat{r} = \mu(r) - h, && \tilde{r} = r - \mu(r) + h, \\ &\hat{u} = 0, && \tilde{u} = u, \\ &\hat{v} = -\frac{a_\star}{\omega}\partial_x h, && \partial_x \tilde{v} = \frac{\omega}{a_\star}\big(r - \mu(r) + h\big) \;\text{ and } \int_{\mathbb{T}} \tilde{v}\,\mathrm{d}x = \int_{\mathbb{T}} v\,\mathrm{d}x, \end{aligned}$$

where $\mu(r) = \frac{1}{|\mathbb{T}|}\int_{\mathbb{T}} r \;\mathrm{d}x$ and $h \in H^1(\mathbb{T})$ is the unique solution of the variational formulation

$$\forall\; \varphi \in H^1(\mathbb{T}),\; \int_T \partial_x h \partial_x \varphi \;\mathrm{d}x + \frac{\omega^2}{a_\star^2}\int_{\mathbb{T}} \varphi h \;\mathrm{d}x = -\frac{\omega}{a_\star}\int_{\mathbb{T}} v \partial_x \varphi \;\mathrm{d}x - \frac{\omega^2}{a_\star^2}\int_{\mathbb{T}} \big(r - \mu(r)\big)\varphi \;\mathrm{d}x.$$

The existence and uniqueness of $h \in H^1(\mathbb{T})$ results from the Lax-Milgram theorem for $\omega \neq 0$.

We easily check that $\hat{q} \in \mathcal{E}_{\omega\neq 0}$ and $\tilde{q} \in \mathcal{E}^\perp_{\omega\neq 0}$. To reach the conclusion, we have to check that $\hat{q} + \tilde{q} = q$. The equalities $\hat{r} + \tilde{r} = r$ and $\hat{u} + \tilde{u} = u$ are trivially verified. For $v$, we have:

$$\forall\; \Phi \in C^\infty(\mathbb{T}),\; \int_{\mathbb{T}} (v - \hat{v} - \tilde{v})\partial_x \Phi \;\mathrm{d}x = \int_{\mathbb{T}} v \partial_x \Phi \;\mathrm{d}x + \frac{a_\star}{\omega}\int_{\mathbb{T}} \partial_x h \partial_x \Phi \;\mathrm{d}x + \frac{\omega}{a_\star}\int_{\mathbb{T}} \big(r - \mu(r) + h\big)\Phi \;\mathrm{d}x = 0$$

due to the choice of $h$. Using the density of $C^\infty(\mathbb{T})$ in $H^1(\mathbb{T})$, we obtain that $v - (\hat{v} + \tilde{v}) = c$. By using the fact that $\int_{\mathbb{T}} \hat{v} dx = 0$ and $\int_{\mathbb{T}} \tilde{v} dx = \int_{\mathbb{T}} v dx$, we get $c = 0$. Therefore, we have $\hat{v} + \tilde{v} = v$. □

## 2.2. **Behaviour of the solution**

By using Lemma 1, we obtain the following properties for the linear wave equation (7):

**Proposition 1.** *Let $q$ be a solution of (7) on $\mathbb{T}$ with initial condition $q^0$. Then:*

*(i) $\forall\, q^0 \in \mathcal{E}_{\omega\neq 0}$, we have $q(t > 0, \cdot) = q^0 \in \mathcal{E}_{\omega\neq 0}$.*
*(ii) $\forall\, q^0 \in \mathcal{E}^\perp_{\omega\neq 0}$, we have $q(t > 0, \cdot) \in \mathcal{E}^\perp_{\omega\neq 0}$.*

*Proof.* We note that System (7) can be written as

$$\partial_t q + A\partial_x q + Bq = 0 \quad \text{where} \quad A = \begin{pmatrix} 0 & a_\star & 0 \\ a_\star & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\omega \\ 0 & \omega & 0 \end{pmatrix}.$$

Due to the fact that matrix A has 3 real distinct eigenvalues ($\lambda = 0$, $\lambda = -a_\star$ and $\lambda = a_\star$), System (7) is strictly hyperbolic. Therefore this system has a unique solution [1, Th. 2.22]. And for any initial condition $q^0 = (r^0, u^0, v^0)$ in $\mathcal{E}_{\omega\neq 0}$, it is obvious that this unique solution is given by $q(t > 0, \cdot) = q^0$ , which proves *(i)*.

Let $q^0 = (r^0, u^0, v^0) \in \mathcal{E}_{\omega\neq 0}^{\perp}$. We notice that

$$\begin{cases} r = r^0 - a_\star \int_0^t \partial_x u \mathrm{d}\tau, \\ u = u^0 - \int_0^t (a_\star \partial_x r - \omega v)\mathrm{d}\tau, \\ v = v^0 - \int_0^t \omega u \mathrm{d}\tau. \end{cases}$$

Therefore, for all $\hat{q} \in \mathcal{E}_{\omega\neq 0}$, we obtain

$$\begin{aligned} \langle q, \hat{q}\rangle &= \langle q^0, \hat{q}\rangle - a_\star \int_{\mathbb{T}}\int_0^t \partial_x u \hat{r}\,\mathrm{d}x\mathrm{d}\tau - \int_{\mathbb{T}}\int_0^t \omega u \hat{v}\,\mathrm{d}x\mathrm{d}\tau = \langle q^0, \hat{q}\rangle - \int_0^t \int_{\mathbb{T}} a_\star \partial_x u \hat{r}\,\mathrm{d}x\mathrm{d}\tau - \int_0^t \int_{\mathbb{T}} \omega u \hat{v}\,\mathrm{d}x\mathrm{d}\tau \\ &= \langle q^0, \hat{q}\rangle + \int_0^t \int_{\mathbb{T}} a_\star \partial_x \hat{r} u\,\mathrm{d}x\mathrm{d}\tau - \int_0^t \int_{\mathbb{T}} \omega \hat{v} u\,\mathrm{d}x\mathrm{d}\tau = \langle q^0, \hat{q}\rangle = 0. \end{aligned}$$

As a result, we conclude that $q(t > 0, \cdot) \in \mathcal{E}_{\omega\neq 0}^{\perp}$, which proves *(ii)*. □

**Corollary 1.** *Let $q$ be the solution of (7) with initial condition $q^0$. Then, $q$ can be decomposed into*

$$q = \mathbb{P}q^0 + (q - \mathbb{P}q^0) \in \mathcal{E}_{\omega\neq 0} \oplus \mathcal{E}_{\omega\neq 0}^{\perp}.$$

### 2.3. **Evolution of the energy**

Let us define the energy as $E = \langle q, q\rangle$.

**Proposition 2.** *Let $q$ be the solution of (7) on $\mathbb{T}$. Then, the energy is conserved*

$$E(t > 0) = E(t = 0).$$

*Proof.* Because $q$ is the solution of (7), we have

$$\begin{cases} \partial_t r = -a_\star \partial_x u, \\ \partial_t u = \omega v - a_\star \partial_x r, \\ \partial_t v = -\omega u \end{cases}$$

which allows to obtain

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\langle q, q\rangle &= a_\star \int_{\mathbb{T}} r(-\partial_x u)\,\mathrm{d}x + \int_{\mathbb{T}} u\left(\omega v - a_\star \partial_x r\right)\,\mathrm{d}x + \int_{\mathbb{T}} v\left(-\omega u\right)\,\mathrm{d}x \\ &= a_\star \int_{\mathbb{T}} r(-\partial_x u)\,\mathrm{d}x + a_\star \int_{\mathbb{T}} u(-\partial_x r)\,\mathrm{d}x = 0. \end{aligned}$$

Hence we have $E'(t) = 0$ which concludes the proof. □

**Corollary 2.** *For all times $t > 0$, we have $\|q(t, \cdot) - \mathbb{P}q^0\| = \|q^0 - \mathbb{P}q^0\|$.*

## 3. Properties of the first order modified equation associated to the Godunov finite volume scheme

It is well known that the *classical Godunov scheme* is not accurate at low Mach number (or low Froude number). With the homogeneous linear wave equation ($\omega = 0$), the problem appears only in the 2d case over rectangular meshes. The work in [8, 9] clearly points out the main reason of the inaccuracy. Shortly, this is because the *classical Godunov scheme* suffers from the loss of invariance of the well-prepared subspace $\mathcal{E}$ when the numerical diffusion related to the velocity equation is not equal to 0. However in our case with Coriolis source term, the problem appears already in 1d due to the numerical diffusion related to the pressure equation. We shall explain this point by studying the properties of the first-order modified equation associated to 1d Godunov like schemes which is given by

$$\begin{cases} \partial_t r + a_\star \partial_x u - \nu_r \partial^2_{xx} r = 0, \\ \partial_t u + a_\star \partial_x r - \nu_u \partial^2_{xx} u = \omega v, \\ \partial_t v = -\omega u, \end{cases} \tag{12}$$

where

$$\nu_r = \frac{\kappa_r |a_\star| \Delta x}{2}, \quad \nu_u = \frac{\kappa_u |a_\star| \Delta x}{2}, \tag{13}$$

for some mesh size $\Delta x > 0$ and viscosity parameters $\kappa_r > 0$ and $\kappa_u > 0$ (see [8] for more details). The classical Godunov scheme corresponds to $\kappa_r = \kappa_u = 1$. In the sequel, we rewrite (12) under a vector formulation

$$\begin{cases} \partial_t q + L_\nu q = 0, \\ q(t=0,x) = q^0(x) \end{cases} \tag{14}$$

where $L_\nu$ is the following spatial differential operator

$$L_\nu = L - B_\nu, \quad Lq = \begin{pmatrix} a_\star \partial_x u \\ a_\star \partial_x r - \omega v \\ \omega u \end{pmatrix} \quad \text{and} \quad B_\nu q = \begin{pmatrix} \nu_r & 0 & 0 \\ 0 & \nu_u & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \partial^2_{xx} r \\ \partial^2_{xx} u \\ 0 \end{pmatrix}.$$

### 3.1. Evolution of the energy

**Lemma 2.** *Let $q_\nu$ be the solution of System (14) on $\mathbb{T}$. Then:*

*(i) If we define the energy by $E_\nu = \langle q_\nu, q_\nu \rangle = \|r\|^2 + \|u\|^2 + \|v\|^2$, we obtain*

$$E_\nu(t \geq 0) \leq E_\nu(t = 0)$$

*which means that System (14) is dissipative.*

*(ii) If we define the average of energy by $\bar{E}_\nu = \|\bar{r}\|^2 + \|\bar{u}\|^2 + \|\bar{v}\|^2$ with*

$$\bar{r}(t) = \frac{1}{|\mathbb{T}|} \int_{\mathbb{T}} r(t,x)\,\mathrm{d}x, \qquad \bar{u}(t) = \frac{1}{|\mathbb{T}|} \int_{\mathbb{T}} u(t,x)\,\mathrm{d}x, \qquad \bar{v}(t) = \frac{1}{|\mathbb{T}|} \int_{\mathbb{T}} v(t,x)\,\mathrm{d}x,$$

*we obtain*

$$\bar{E}_\nu(t=0) = \bar{E}_\nu(t>0).$$

*(iii) Moreover, we have*

$$\forall\, t > 0,\ \bar{E}_\nu(0) = \bar{E}_\nu(t) \leq E_\nu(t) \leq E_\nu(0).$$

*Proof.* We have

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|q_\nu\|^2(t) = -\langle Lq_\nu, q_\nu\rangle + \langle B_\nu q_\nu, q_\nu\rangle.$$

However,

$$\langle Lq_\nu, q_\nu\rangle = \langle a_\star \partial_x u, r\rangle + \langle a_\star \partial_x r - \omega v, u\rangle + \langle \omega u, v\rangle = 0$$

and

$$\langle B_\nu q_\nu, q_\nu\rangle = \left\langle \nu_r \frac{\partial^2 r}{\partial x^2}, r\right\rangle + \left\langle \nu_u \frac{\partial^2 u}{\partial x^2}, u\right\rangle = -\nu_r\|\partial_x r\|^2 - \nu_u\|\partial_x u\|^2.$$

For this reason, we obtain $E'_\nu(t) \leq 0$ which means that $E_\nu(t \geq 0) \leq E_\nu(t = 0)$.

By integrating the first order modified equation over $\mathbb{T}$ and using periodic boundary conditions, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{r}(t) = 0, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\bar{u}(t) = \omega\bar{v}(t) \qquad \text{and} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\bar{v}(t) = -\omega\bar{u}(t)$$

which leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{r}(t)^2 = 0, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\bar{u}(t)^2 = 2\omega\bar{v}(t)\bar{u}(t) \qquad \text{and} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\bar{v}(t)^2 = -2\omega\bar{u}(t)\bar{v}(t).$$

As a result, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\bar{r}(t)^2 + \bar{u}(t)^2 + \bar{v}(t)^2\right] = 0,$$

which means that $\bar{E}_\nu(t = 0) = \bar{E}_\nu(t > 0)$. It is interesting to note that

$$\begin{aligned} E_\star(t) &:= \int_{\mathbb{T}} (r - \bar{r})^2 \,\mathrm{d}x + \int_{\mathbb{T}} (u - \bar{u})^2 \,\mathrm{d}x + \int_{\mathbb{T}} (v - \bar{v})^2 \,\mathrm{d}x \\ &= \int_{\mathbb{T}} (r^2 + u^2 + v^2) \,\mathrm{d}x - 2\bar{r}\int_{\mathbb{T}} r \,\mathrm{d}x - 2\bar{u}\int_{\mathbb{T}} u \,\mathrm{d}x - 2\bar{v}\int_{\mathbb{T}} v \,\mathrm{d}x + \int_{\mathbb{T}} (\bar{r}^2 + \bar{u}^2 + \bar{v}^2) \,\mathrm{d}x \\ &= \int_{\mathbb{T}} (r^2 + u^2 + v^2) \,\mathrm{d}x - \int_{\mathbb{T}} (\bar{r}^2 + \bar{u}^2 + \bar{v}^2) \,\mathrm{d}x = E_\nu(t) - \bar{E}_\nu(t). \end{aligned}$$

Therefore, we obtain $E'_\star(t) = E'_\nu(t) \leq 0$ and $E_\nu(t) \geq \bar{E}_\nu(t)$ (since $E_\star(t) \geq 0$). □

### 3.2. **Structure of the kernel of the modified equation**

Interestingly, the structure of the kernel of the operator $L_\nu$ is deeply related to the value of $\nu_r$. Indeed, we have:

**Lemma 3.**

*(i) When $\nu_r = 0$, the subspace $\mathcal{E}_{\omega\neq 0}$ is also the kernel of the modified equation*

$$\ker L_{\nu_r=0} = \mathcal{E}_{\omega\neq 0}.$$

*Moreover, $\mathcal{E}^\perp_{\omega\neq 0}$ is invariant by the modified equation.*

*(ii) When $\nu_r \neq 0$, the subspace $\mathcal{E}_{\omega\neq 0}$ is not invariant for the modified equation since*

$$\ker L_{\nu_r\neq 0} = \{q := (r, u, v) \,|\, r = const, u = 0, v = 0\} \subsetneq \mathcal{E}_{\omega\neq 0}.$$

*Proof.* With $\nu_r = 0$, it is easy to see that

$$\ker L_{\nu_r=0} = \mathcal{E}_{\omega\neq 0}.$$

As for the orthogonal space, the proof of Prop. 1 *(ii)* stands for $\nu_r = 0$.

We now focus on the case $\nu_r \neq 0$. Let us suppose that $q = (r, u, v) \in \ker L_\nu$. As $u = 0$, we have $a_\star \partial_x r - \omega v = 0$. Then, from $L_\nu q = 0$, we deduce

$$0 = \langle L_\nu q, q \rangle = \nu_r \|\partial_x r\|^2$$

which implies that $\partial_x r = 0$ or equivalently $r$ is a constant. This leads to $v = 0$ and $q = (const, 0, 0)$. □

The result in Lemma 3 indicates that *the classical Godunov scheme* ($\kappa_r = 1$) does not capture all states $q \in \mathcal{E}_{\omega\neq 0}$ because of the fact that the corresponding kernel is a proper subset of $\mathcal{E}_{\omega\neq 0}$. This gives rise to the loss of invariance. However, when the numerical viscosity on the pressure vanishes ($\nu_r = 0$), we recover all states $q \in \mathcal{E}_{\omega\neq 0}$.

### 3.3. **Behaviour of the solution of the modified equation**

We recall that $M$ is a small parameter. Let us introduce the following definitions:

**Definition 1.** *A state $q^0$ is said to be* well-prepared *if $\|q^0 - \mathbb{P}q^0\| = \mathcal{O}(M)$, where $\mathbb{P}$ is defined by (11).*

**Definition 2.** *The solution $q_\nu$ of System (14) is said to be* accurate at low Froude number at any time *if:*

$$\forall\, C_1 > 0,\, \exists\, C_2 > 0, \|q^0 - \mathbb{P}q^0\| \leq C_1 M \implies \forall\, t \geq 0,\, \|q_\nu - \mathbb{P}q^0\|(t) \leq C_2 M,$$

*where $C_2$ is a positive parameter that does not depend on $M$.*

**Definition 3.** *The solution $q_\nu$ of System (14) is said to be* accurate at low Froude number locally in time *if:*

$$\forall\, C_1 > 0, \forall\, C_2 > 0 : C_2 = \mathcal{O}(1), \exists\, C_3 > 0,\, \|q^0 - \mathbb{P}q^0\| \leq C_1 M \implies \forall\, t \leq C_2,\, \|q_\nu - \mathbb{P}q^0\|(t) \leq C_3 M,$$

*where $C_3 = \mathcal{O}(1)$.*

*Remark* 2. We notice that if the solution is accurate at low Froude number, it is free of spurious acoustic waves (refer to [8] for more details).

We have the following result. We recall that $\nu_\# = \frac{\kappa_\# |a_\star| \Delta x}{2}$.

**Theorem 1.** *Let $q_\nu$ be the solution of System (14). Then:*

- *(i) When $\kappa_r = 0$, the solution is accurate at low Froude number at any time. Moreover, it satisfies $\|q_\nu - \mathbb{P}q^0\|(t) \leq \|q^0 - \mathbb{P}q^0\|$.*
- *(ii) When $\kappa_r = \mathcal{O}(M)$, the solution is accurate at low Froude number locally in time.*
- *(iii) When $\kappa_r = \mathcal{O}(1)$, the solution is accurate at low Froude number locally in time if $\Delta x = \mathcal{O}(M)$.*

*Remark* 3. From Point *(iii)*, we can state that for $\kappa_r = \mathcal{O}(1)$, it is enough to consider a very fine mesh to obtain accurate results. We shall see in the sequel that we actually need to consider fine meshes which is a strong restriction from the computational point of view.

*Proof.* Let $q_\nu^a$ be the solution of

$$\begin{cases} \partial_t q + L_\nu q = 0, \\ q(t = 0, x) = \mathbb{P}q^0(x) \end{cases}$$

and $q_\nu^b$ be the solution of

$$\begin{cases} \partial_t q + L_\nu q = 0, \\ q(t = 0, x) = q^0(x) - \mathbb{P}q^0(x). \end{cases}$$

Then by linearity the solution of (14) is $q_\nu = q_\nu^a + q_\nu^b$. If we suppose that $\|q^0 - \mathbb{P}q^0\| = C_1 M$, then by applying Lemma 2, we obtain

$$\|q_\nu^b\|(t) \leq \|q_\nu^b\|(0) = \|q^0 - \mathbb{P}q^0\| = C_1 M. \tag{15}$$

We also notice that

$$\|q_\nu - \mathbb{P}q^0\|(t) = \|q_\nu^a + q_\nu^b - \mathbb{P}q^0\|(t) \leq \|q_\nu^a - \mathbb{P}q^0\|(t) + \|q_\nu^b\|(t). \tag{16}$$

If $\kappa_r = 0$, then $q_\nu^a = \mathbb{P}q^0$ according to Lemma 3 *(i)*. This proves Point *(i)*.

As for Points *(ii)* and *(iii)*, we set $\hat{q}^0 = (\hat{r}^0, \hat{u}^0, \hat{v}^0) := \mathbb{P}q^0$ and $q_\nu^a = (r_\nu^a, u_\nu^a, v_\nu^a)$. Then, we obtain

$$\begin{cases} \partial_t(r_\nu^a - \hat{r}^0) + a_\star \partial_x(u_\nu^a - \hat{u}^0) - \nu_r \partial_{xx}^2(r_\nu^a - \hat{r}^0) + a_\star \partial_x \hat{u}^0 - \nu_r \partial_{xx}^2 \hat{r}^0 = 0, \\ \partial_t(u_\nu^a - \hat{u}^0) + a_\star \partial_x(r_\nu^a - \hat{r}^0) - \nu_u \partial_{xx}^2(u_\nu^a - \hat{u}^0) + a_\star \partial_x \hat{r}^0 - \nu_u \partial_{xx}^2 \hat{u}^0 = \omega(v_\nu^a - \hat{v}^0) + \omega \hat{v}^0, \\ \partial_t(v_\nu^a - \hat{v}^0) + \omega(u_\nu^a - \hat{u}^0) + \omega \hat{u}^0 = 0. \end{cases} \tag{17}$$

On the other hand, since $\mathbb{P}q^0 \in \mathcal{E}_{\omega \neq 0}$, we have that $\hat{u}^0 = 0$ and $a_\star \partial_x \hat{r}^0 = \omega \hat{v}^0$. Therefore, (17) reduces to

$$\begin{cases} \partial_t(r_\nu^a - \hat{r}^0) + a_\star \partial_x(u_\nu^a - \hat{u}^0) - \nu_r \partial_{xx}^2(r_\nu^a - \hat{r}^0) - \nu_r \partial_{xx}^2 \hat{r}^0 = 0, \\ \partial_t(u_\nu^a - \hat{u}^0) + a_\star \partial_x(r_\nu^a - \hat{r}^0) - \nu_u \partial_{xx}^2(u_\nu^a - \hat{u}^0) = \omega(v_\nu^a - \hat{v}^0), \\ \partial_t(v_\nu^a - \hat{v}^0) = -\omega(u_\nu^a - \hat{u}^0). \end{cases} \tag{18}$$

Multiplying Equation (18) by $q_\nu^a - \hat{q}^0$, integrating over $\mathbb{T}$ and using periodic boundary conditions, we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|q_\nu^a - \mathbb{P}q^0\|^2 = -\nu_r \|\partial_x(r_\nu^a - \hat{r}^0)\|^2 - \nu_u \|\partial_x(u_\nu^a - \hat{u}^0)\|^2 + \nu_r \left\langle \partial_{xx}^2 \hat{r}^0, r_\nu^a - \hat{r}^0 \right\rangle$$

which yields

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|q_\nu^a - \mathbb{P}q^0\|^2 \leq \nu_r \|\partial_{xx}^2 \hat{r}^0\| \cdot \|r_\nu^a - \hat{r}^0\| \leq \nu_r \|\partial_{xx}^2 \hat{r}^0\| \cdot \|q_\nu^a - \mathbb{P}q^0\|.$$

This leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\|q_\nu^a - \mathbb{P}q^0\| \leq \nu_r \|\partial_{xx}^2 \hat{r}^0\|.$$

We deduce from the latter inequality that

$$\|q_\nu^a - \mathbb{P}q^0\|(t) \leq \nu_r t \|\partial_{xx}^2 \hat{r}^0\| \tag{19}$$

since $q_\nu^a(0) = \mathbb{P}q^0$. From (15), (16) and (19), we infer

$$\|q_\nu - \mathbb{P}q^0\|(t) \leq C_1 M + \nu_r t \|\partial_{xx}^2 \hat{r}^0\|.$$

Given (13), we deduce Points *(ii)* and *(iii)* respectively for $\kappa_r = \mathcal{O}(M)$ and $\Delta x = \mathcal{O}(M)$. □

### 3.4. **Fourier analysis**

To go further in the study of the accuracy of the numerical scheme, we perform a Fourier analysis to investigate diffusion and dispersion effects. Let us consider functions of the form

$$q(t, x) = e^{i(\tau t + kx)} \hat{q} \tag{20}$$

where $k$ is the wave number and $\tau$ is the frequency of the wave. These functions can be solutions to the modified equation only under a *dispersion relation* between $\tau$ and $k$ which is commonly written as $\tau = \tau(k)$. In general, this relation lies in the complex set: the real part $\Re(\tau)$ and the imaginary part $\Im(\tau)$ indicate respectively propagation and decay of Fourier modes.

Given a wave number $k$, we only consider mesh sizes satisfying

$$k < \frac{\pi}{\Delta x} \tag{21}$$

so that the associated wave is captured by the scheme.

Functions (20) are solutions to the modified equation (12) if

$$i\tau\hat{q} + A\hat{q} = 0, \qquad \text{where} \quad A(k, \nu_r, \nu_u, a_\star, \omega) = \begin{pmatrix} \nu_r k^2 & a_\star ik & 0 \\ a_\star ik & \nu_u k^2 & -\omega \\ 0 & \omega & 0 \end{pmatrix}. \tag{22}$$

This means that $-i\tau$ is an eigenvalue of $A$. We shall denote by $\lambda$ the eigenvalues of $A$ in the sequel. Hence the decay of Fourier mode $k$ corresponds to $\Re(\lambda) \geq 0$.

**Proposition 3.** *Under Hypothesis (21), the damping of Fourier modes is parametrised by $\kappa_r$ as follows.*

*(i) When $\kappa_r = 0$, the wave associated to the kernel of the wave operator is preserved ($\lambda = 0$).*
*(ii) When $\kappa_r = \mathcal{O}(M)$, the wave resulting from $\lambda(\nu_r = 0) = 0$ is damped at an $\mathcal{O}(M)$ speed.*
*(iii) When $\kappa_r = \mathcal{O}(1)$ and $\Delta x = \mathcal{O}(1)$, all Fourier modes are strongly damped at an $\mathcal{O}(1)$ speed.*

*Proof.* The linear system (22) reads in terms of eigenvalues $\lambda$

$$\nu_r k^2\, r + ika_\star u = \lambda\, r, \tag{23a}$$
$$ika_\star r + \nu_u k^2\, u - \omega\, v = \lambda\, u, \tag{23b}$$
$$\omega\, u = \lambda\, v. \tag{23c}$$

The characteristic polynomial of Matrix $A$ is

$$\chi(\lambda, \nu_r) := \lambda^3 - k^2(\nu_r + \nu_u)\lambda^2 + (\omega^2 + k^2 a_\star^2 + k^4 \nu_r \nu_u)\lambda - k^2\omega^2\nu_r = 0. \tag{24}$$

It is a third order polynomial whose highest order coefficient is equal to one. It thus has either one real root and two complex conjugate roots (denoted respectively by $\lambda_0$, $\lambda_c$ and $\bar{\lambda}_c$) or three real roots (denoted respectively by $\lambda_0$, $\lambda_+$ and $\lambda_-$).

It is possible to determine its three roots when $\nu_r = 0$:

$$\lambda_0(\nu_r = 0) = 0,$$
$$\lambda_c(\nu_r = 0) = \frac{1}{2}\left[k^2\nu_u + i\sqrt{4(\omega^2 + k^2 a_\star^2) - k^4\nu_u^2}\right].$$

We mention that the term under the square root is actually positive under Hyp. (21) (see (13) for the definition of $\nu_u$). Point *(i)* is proven.

We remark that $\partial_\lambda \chi$ does not vanish as soon as

$$k^2\Delta x^2\left((\kappa_r - \kappa_u)^2 + \kappa_r\kappa_u\right) < 12\left(1 + \frac{\omega^2}{a_\star^2 k^2}\right). \tag{25}$$

Due to Hyp. (21), this inequality always holds for $\kappa_r$ and $\kappa_u$ in $[0,1]$. Hence by means of the implicit function theorem, we can define a function $\nu_r \mapsto \lambda_0(\nu_r)$ for $\nu_r$ small enough. Since coefficients multiplying $\lambda^k$ in (24) are affine functions in $\nu_r$, we infer that $\lambda_0$ is continuous and analytic with respect to $\nu_r$ [13]. This shows that

$$\lambda_0(\nu_r) \underset{\nu_r \to 0}{\sim} \lambda_0'(\nu_r = 0)\nu_r = -\frac{\partial_{\nu_r}\chi(0,0)}{\partial_\lambda \chi(0,0)}\nu_r = \frac{k^2\omega^2}{k^2 a_\star^2 + \omega^2}\nu_r.$$

In particular, we deduce that if $\kappa_r = \mathcal{O}(M)$, then $\lambda_0(\nu_r) = \mathcal{O}(M)$. This proves Point *(ii)*.

Let us now provide other properties of the eigenvalues. We substitute (23c) into (23b) and then multiply (23a) by $\bar{r}$ and (23b) by $\bar{u}$ to obtain

$$\frac{1}{\lambda}\omega^2|u|^2 + \lambda(|r|^2 + |u|^2) = k^2(\nu_r|r|^2 + \nu_u|u|^2) + ika_\star(u\bar{r} + r\bar{u}). \tag{26}$$

On the one hand, the real part of (26)

$$\Re(\lambda)\left[\frac{\omega^2|u|^2}{|\lambda|^2} + |r|^2 + |u|^2\right] = k^2\left(\nu_r|r|^2 + \nu_u|u|^2\right),$$

shows that all eigenvalues have positive real parts (unless $k = 0$ for which eigenvalues are pure imaginary), which ensures the decay for all Fourier modes.

The three roots of (24) satisfy

$$\lambda_1 + \lambda_2 + \lambda_3 = k^2(\nu_r + \nu_u), \tag{27a}$$
$$\lambda_1\lambda_2 + (\lambda_1 + \lambda_2)\lambda_3 = \omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u, \tag{27b}$$
$$\lambda_1\lambda_2\lambda_3 = k^2\omega^2\nu_r. \tag{27c}$$

Substituting $\lambda_1\lambda_2$ from (27c) into (27b), we get

$$(\lambda_1 + \lambda_2)\lambda_3 + \frac{k^2\omega^2\nu_r}{\lambda_3} = \omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u. \tag{28}$$

Let us first focus on the case of a single real eigenvalue: we take $\lambda_1 = \lambda_c$, $\lambda_2 = \bar{\lambda}_c$ and $\lambda_3 = \lambda_0$. Eq. (28) yields

$$\lambda_0 \geq \frac{k^2\omega^2\nu_r}{\omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u} \tag{29}$$

since it has been proven that $\Re(\lambda_c) \geq 0$.

We also notice that $\chi(0,\nu_r) = -k^2\omega^2\nu_r < 0$ and $\chi(k^2\nu_r,\nu_r) = k^4 a_\star^2\nu_r > 0$. Hence, since there is a single real eigenvalue, this implies that $\lambda_0 \leq k^2\nu_r$ and we have

$$\frac{\omega^2}{\omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u}k^2\nu_r \leq \lambda_0 \leq k^2\nu_r. \tag{30}$$

As for the complex conjugate roots, we get from (27a) and (28) that $\mu := 2\Re(\lambda_c)$ verifies

$$f(\mu) := \mu^2 - k^2(\nu_r + \nu_u)\mu - \frac{k^2\omega^2\nu_r}{k^2(\nu_r + \nu_u) - \mu} + \omega^2 + k^2 a_\star^2 + k^4\nu_r\nu_u = 0. \tag{31}$$

We remark that $f(\mu) \geq g(\mu)$ where

$$g(\mu) := -k^2(\nu_r+\nu_u)\mu - \frac{k^2\omega^2\nu_r}{k^2(\nu_r+\nu_u)-\mu} + \omega^2 + k^2a_\star^2 + k^4\nu_r\nu_u. \tag{32}$$

Since $f(0) = g(0) > 0$, this implies that any root $\mu$ of (31) is larger than the smallest positive root of (32).

Equation $g(\mu) = 0$ can be written as

$$k^2(\nu_r+\nu_u)\mu^2 - \left[k^4(\nu_r+\nu_u)^2 + (\omega^2+k^2a_\star^2+k^4\nu_r\nu_u)\right]\mu + (\omega^2+k^2a_\star^2+k^4\nu_r\nu_u)k^2(\nu_r+\nu_u) - k^2\omega^2\nu_r = 0. \tag{33}$$

Due to the fact that

$$\Delta = \left[k^4(\nu_r+\nu_u)^2 - (\omega^2+k^2a_\star^2+k^4\nu_r\nu_u)\right]^2 + 4k^4(\nu_r+\nu_u)\nu_r\omega^2 > 0,$$

Equation (33) has two real positive solutions so that

$$2\Re(\lambda_c) \geq \frac{(\omega^2+k^2a_\star^2+k^4\nu_r\nu_u)k^2(\nu_r+\nu_u) - k^2\omega^2\nu_r}{k^4(\nu_r+\nu_u)^2 + (\omega^2+k^2a_\star^2+k^4\nu_r\nu_u)}. \tag{34}$$

In the case of three real roots, (29) holds for each of them by symmetry as they are all positive. Lower bounds (29) and (34) ensure that real parts of all eigenvalues are of order 1 when $\nu_r$ is of order 1. This proves Point *(iii)*. □

## 4. Analysis of fully discrete Godunov schemes

There are two main possible time strategies for Godunov type schemes applied to the linear wave equation with Coriolis source term. The first one is a classical splitting discretisation where one deals with the problem without source term in a first step and then the Coriolis source term is considered in a second step, which then consists in solving an ordinary differential equation. It is well known that this splitting strategy is not well adapted to preserve stationary states and then to compute small perturbations around them [3, 11], see also Appendix A. Thus we focus on the analysis of the second strategy that consists in computing acoustic and Coriolis effects in a single step. As a matter of fact, there are many ways to take into account the Coriolis source term. For example, we can discretise this term using explicit, implicit and even Crank-Nicolson strategies. Hence, we introduce two new parameters $\theta_1$ and $\theta_2$ to parametrise the strategy.

### 4.1. **Study of the discrete kernel of the one step Godunov scheme**

We consider a homogeneous cartesian mesh $(x_i)_{1\leq i\leq N}$. The one step fully discrete Godunov scheme is given by

$$\begin{cases} \dfrac{r_i^{n+1}-r_i^n}{\Delta t} + a_\star \dfrac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} - \nu_r \dfrac{r_{i+1}^n - 2r_i^n + r_{i-1}^n}{\Delta x^2} = 0, \\ \dfrac{u_i^{n+1}-u_i^n}{\Delta t} + a_\star \dfrac{r_{i+1}^n - r_{i-1}^n}{2\Delta x} - \nu_u \dfrac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} = \omega\left[\theta_1 v_i^n + (1-\theta_1)v_i^{n+1}\right], \\ \dfrac{v_i^{n+1}-v_i^n}{\Delta t} = -\omega\left[\theta_2 u_i^n + (1-\theta_2)u_i^{n+1}\right] \end{cases} \tag{35}$$

for $i \in \{1, \dots, N\}$ and $0 \leq \theta_1, \theta_2 \leq 1$. Periodic boundary conditions read

$$q_0^{n+1} = q_N^{n+1},\ q_{N+1}^{n+1} = q_1^{n+1}. \tag{36}$$

We now investigate the kernel of the fully discrete one step scheme. It is strongly related to the value of the numerical viscosity $\kappa_r$. In particular we have the following result:

**Lemma 4.**

*(i) When $\nu_r = 0$, the kernel of the one step scheme is*

$$\mathcal{E}^h_{\omega \neq 0} := \ker L_{\nu_r = 0, h} = \left\{ q = (r, u, v) \in \mathbb{R}^{3N} \;\middle|\; u_i = 0,\ \frac{a_\star}{2\Delta x}(r_{i+1} - r_{i-1}) = \omega v_i \right\}.$$

*(ii) When $\nu_r \neq 0$, the kernel of the one step scheme is*

$$\ker L_{\nu_r \neq 0, h} = \left\{ q = (r, u, v) \in \mathbb{R}^{3N} \;\middle|\; \exists C \in \mathbb{R} : r_i = C, u_i = 0, v_i = 0 \right\}.$$

*Proof.* A stationary state verifies $r_i^{n+1} = r_i^n$, $u_i^{n+1} = u_i^n$ and $v_i^{n+1} = v_i^n$. Therefore, we easily obtain from (35) that

$$\begin{cases} a_\star \frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} - \nu_r \frac{r_{i+1}^n - 2r_i^n + r_{i-1}^n}{\Delta x^2} = 0, & \text{(37a)} \\ a_\star \frac{r_{i+1}^n - r_{i-1}^n}{2\Delta x} - \nu_u \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} = \omega v_i^n, & \text{(37b)} \\ 0 = -\omega u_i^n. & \text{(37c)} \end{cases}$$

Point *(i)* is straightforward: we get from (37c) that $u_i^n = 0$, and then (37a) is trivially satisfied since $\nu_r = 0$. Then, (37b) yields that

$$\frac{a_\star}{2\Delta x}(r_{i+1}^n - r_{i-1}^n) = \omega v_i^n.$$

Now we consider the case $\nu_r \neq 0$. According to (37c), $u_i^n = 0$ for all $i$. Together with (37a) and $\nu_r \neq 0$, we get $r_{i+1}^n - r_i^n = r_i^n - r_{i-1}^n$. By induction we get $r_{N+1}^n - r_N^n = r_N^n - r_{N-1}^n = \ldots = r_2^n - r_1^n = r_1^n - r_0^n = c$ where $c$ is a constant. This implies $r_N^n = r_0^n + Nc$. On the other hand, periodic conditions require to have $r_N^n = r_0^n$. Therefore, we get $c = 0$ and $r_i^n =$ constant. This leads to $v_i^n = 0$ by using (37b). Point *(ii)* is proven. □

## 4.2. **Stability of the discrete one step Godunov scheme**

For $0 \leq \theta_1, \theta_2 \leq 1$, let us denote

$$\Theta_1 = 1 - \theta_1 - \theta_2, \quad \Theta_2 = \theta_1\theta_2 + (1 - \theta_1)(1 - \theta_2) \in [0, 1], \quad \Theta_3 = (1 - 2\theta_1)(1 - 2\theta_2) \in [-1, 1].$$

**Lemma 5.** *For $\kappa_r = 0$ and $\kappa_u > 0$, we have:*

*(i) When $\theta_1 + \theta_2 > 1$, the one step scheme (35) is unstable.*
*(ii) When $\theta_1 + \theta_2 \leq 1$, we consider two cases:*
*(a) If $\frac{\kappa_u^2 a_\star^2}{\omega^2 \Delta x^2} \leq \Theta_3$, the one step scheme (35) is stable provided that*

$$\Delta t \leq \Delta t_a := \frac{\kappa_u \Delta x}{2|a_\star|} \frac{1}{\left(1 - \frac{\omega \Delta x}{|a_\star|}\sqrt{\Theta_1}\right)_+}; \tag{38a}$$

*(b) If* $\frac{\kappa_u^2 a_\star^2}{\omega^2 \Delta x^2} > \Theta_3$*, the one step scheme* (35) *is stable provided that*

$$\Delta t \leq \min\{\Delta t_a, \Delta t_b\} \quad \textit{where} \quad \Delta t_b := \frac{\Delta x}{\kappa_u |a_\star|} \times \begin{cases} \frac{2\kappa_u^2 a_\star^2}{\omega^2 \Delta x^2 \Theta_3} \left[1 - \sqrt{1 - \frac{\omega^2 \Delta x^2}{\kappa_u^2 a_\star^2} \Theta_3}\right], & \textit{if } \Theta_3 \neq 0, \\ 1, & \textit{otherwise.} \end{cases} \tag{38b}$$

*Remark* 4. The standard CFL condition for the homogeneous case ($\omega = 0$) reads [8]

$$\Delta t \leq \Delta t_0 := \frac{\Delta x}{|a_\star|} \min\left\{\frac{\kappa_u}{2}, \frac{1}{\kappa_u}\right\}.$$

Inequality (38a) clearly shows that taking Coriolis forces into account requires a less restrictive CFL condition. It is also the case for (38b) when $\Theta_3 \geq 0$ thanks to the convexity of the function $x \mapsto 1 - \sqrt{1-x}$. We also notice that for the Crank-Nicolson scheme $\theta_1 = \theta_2 = \frac{1}{2}$, we recover the standard bound $\Delta t_0$.

*Remark* 5. An asymptotic expansion for $\Delta x \ll 1$ in the bound $\Delta t_a$ and $\Delta t_b$ in (38a-38b) yields

$$\Delta t_a = \frac{\kappa_u \Delta x}{2|a_\star|} + \mathcal{O}(\Delta x^2), \quad \Delta t_b = \frac{\Delta x}{\kappa_u |a_\star|} + \mathcal{O}(\Delta x^3)$$

and then one still recovers the classical bound $\Delta t_0$ for the homogeneous problem.

*Remark* 6. For large values of the Coriolis parameter $\omega$, the constraint (38a) is always satisfied ($\Delta t_a = +\infty$) while for the second constraint (38b), it depends on the sign of $\Theta_3$:

- If $\Theta_3 \geq 0$, there is no constraint upon $\Delta t$ for $\omega$ large enough;
- If $\Theta_3 < 0$, the asymptotic bound reads

$$\Delta t_b \approx \frac{2}{\omega}.$$

We then recover the standard stability condition for the ODE system solved by means of a $\theta$-scheme (46b).

*Remark* 7. Figure 1 specifies the stability area. In the red zone, the scheme is unstable according to Point *(i)*. In the green zone, the scheme is stable under a CFL-like constraint (characterised by $\Delta t_a$ or $\min(\Delta t_a, \Delta t_b)$) that is less restrictive than the homogeneous bound $\Delta t_0$ while in the blue zone, the scheme is stable provided $\Delta t$ is smaller than $\min(\Delta t_a, \Delta t_b) \leq \Delta t_0$.

*Proof.* We perform a Von Neumann analysis to investigate the stability condition for Scheme (35). Let us denote

$$\sigma = \frac{\Delta t}{\Delta x}, \ \gamma = \omega \Delta t \ \text{ and } \ s = \sin\left(\frac{k\Delta x}{2}\right).$$

We now substitute

$$q_j^n = \begin{pmatrix} r_j^n \\ u_j^n \\ v_j^n \end{pmatrix} = \begin{pmatrix} R_n \\ U_n \\ V_n \end{pmatrix} e^{ikj\Delta x}$$

into (35) in order to obtain

$$A q_j^{n+1} = B q_j^n \tag{39}$$

where the matrices $A$ and $B$ are given by

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -(1-\theta_1)\gamma \\ 0 & (1-\theta_2)\gamma & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 - 2\kappa_r |a_\star| \sigma s^2 & -a_\star \sigma i \sin(k\Delta x) & 0 \\ -a_\star \sigma i \sin(k\Delta x) & 1 - 2\kappa_u |a_\star| \sigma s^2 & \theta_1 \gamma \\ 0 & -\theta_2 \gamma & 1 \end{pmatrix}.$$

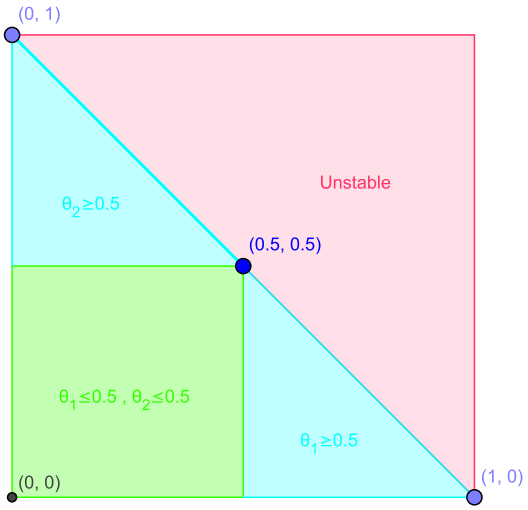



FIGURE 1. Region of stability condition.

In addition, we have

$$A^{-1} = \frac{1}{\Lambda(\theta_1,\theta_2)} \begin{pmatrix} \Lambda(\theta_1,\theta_2) & 0 & 0 \\ 0 & 1 & \gamma(1-\theta_1) \\ 0 & -\gamma(1-\theta_2) & 1 \end{pmatrix}$$

with

$$\Lambda(\theta_1,\theta_2) = 1 + \gamma^2(1-\theta_1)(1-\theta_2). \tag{40}$$

Therefore, we can rewrite (39) as the following equation

$$q_j^{n+1} = C q_j^n$$

where the amplification matrix $C$ is given by

$$C = A^{-1}B = \frac{1}{\Lambda(\theta_1,\theta_2)} \begin{pmatrix} (1-2\kappa_r|a_\star|\sigma s^2)\Lambda(\theta_1,\theta_2) & -a_\star\sigma i\sin(k\Delta x)\Lambda(\theta_1,\theta_2) & 0 \\ -a_\star\sigma i\sin(k\Delta x) & 1-\gamma^2\theta_2(1-\theta_1)-2\kappa_u|a_\star|\sigma s^2 & \gamma \\ \gamma(1-\theta_2)a_\star\sigma i\sin(k\Delta x) & -\gamma[1-(1-\theta_2)2\kappa_u|a_\star|\sigma s^2] & 1-\gamma^2\theta_1(1-\theta_2) \end{pmatrix}, \tag{41}$$

whose characteristic polynomial will be denoted by $\mathcal{P}(\lambda)$. We now consider the modes which are constant in space ($k = 0$). In this case, the amplification matrix in $(u, v)$ is given by

$$\frac{1}{1+\gamma^2(1-\theta_1)(1-\theta_2)} \begin{pmatrix} 1-\gamma^2\theta_2(1-\theta_1) & \gamma \\ -\gamma & 1-\gamma^2\theta_1(1-\theta_2) \end{pmatrix}.$$

Therefore the characteristic equation $\mathcal{P}(\lambda) = 0$ reduces to

$$\lambda^2 - \frac{2-\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)}{\Lambda(\theta_1,\theta_2)}\lambda + \frac{1+\gamma^2\theta_1\theta_2}{\Lambda(\theta_1,\theta_2)} = 0, \tag{42}$$

and the condition $|\lambda_1\lambda_2| \leq 1$ is equivalent to

$$1 + \gamma^2\theta_1\theta_2 \leq 1 + \gamma^2(1-\theta_1)(1-\theta_2),$$

that is fulfilled if and only if

$$\gamma^2[(\theta_1+\theta_2)-1] \leq 0,$$

which leads to the condition $\theta_1 + \theta_2 \leq 1$. This proves Point *(i)*.

Now we consider the case of interest $\kappa_r = 0$ (*cf.* Lemma 4). The characteristic polynomial $\mathcal{P}(\lambda)$ reduces to

$$\mathcal{P}_0(\lambda) = (1-\lambda)\left[\lambda^2 - \frac{2-\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)}\lambda + \frac{1+\gamma^2\theta_1\theta_2-2\kappa_u|a_\star|\sigma s^2+4a_\star^2\sigma^2s^2(1-s^2)}{\Lambda(\theta_1,\theta_2)}\right]. \tag{43}$$

One root of this polynomial is $\lambda_0 = 1$ and the two other roots $\lambda_\pm$ are the solutions of the following second degree equation

$$\lambda^2 + \xi\lambda + \zeta = 0 \tag{44}$$

with

$$\xi = -\frac{2-\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)-2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)} \quad \text{and} \quad \zeta = \frac{1+\gamma^2\theta_1\theta_2-2\kappa_u|a_\star|\sigma s^2+4a_\star^2\sigma^2s^2(1-s^2)}{\Lambda(\theta_1,\theta_2)}.$$

In order to ensure that the roots of (44) are in the unit circle ($|\lambda_\pm| \leq 1$), the coefficients $\xi$ and $\zeta$ must satisfy

$$|\zeta| \leq 1 \quad \text{and} \quad |\xi| \leq 1+\zeta.$$

- Firstly, the condition $\zeta \leq 1$ is equivalent to

$$\frac{1+\gamma^2\theta_1\theta_2-2\kappa_u|a_\star|\sigma s^2+4a_\star^2\sigma^2s^2(1-s^2)}{\Lambda(\theta_1,\theta_2)} \leq 1$$

  which leads to

$$f_1(s^2) := -\gamma^2\Theta_1 - 2\kappa_u|a_\star|\sigma s^2 + 4a_\star^2\sigma^2s^2(1-s^2) \leq 0.$$

  With $s$ varying in $[-1,1]$, the previous condition holds provided $\max_{[0,1]} f_1 \leq 0$. As Function $f_1$ is maximal over $\mathbb{R}$ at $X_1 := \frac{1}{2}\left(1 - \frac{\kappa_u}{2|a_\star|\sigma}\right)$, we deduce that

$$\max_{[0,1]} f_1 = \begin{cases} f_1(0), & \text{if } X_1 \leq 0, \\ f_1(X_1), & \text{otherwise.} \end{cases}$$

  If $X_1 \leq 0$ which is equivalent to $\sigma \leq \frac{\kappa_u}{2|a_\star|}$, the condition $f_1(0) \leq 0$ is always satisfied. If $X_1 > 0$, $f_1(X_1) \leq 0$ reads

$$\left(|a_\star|\sigma - \frac{\kappa_u}{2}\right)^2 \leq \gamma^2\Theta_1 \iff \left(\frac{|a_\star|}{\Delta x} - \omega\sqrt{\Theta_1}\right)\Delta t \leq \frac{\kappa_u}{2}.$$

  Hence $\Delta t \leq \Delta t_a$.
- Next, the condition $\zeta \geq -1$ can be written as

$$f_2(s^2) := \gamma^2\Theta_2 + 2(1-\kappa_u|a_\star|\sigma s^2) + 4a_\star^2\sigma^2s^2(1-s^2) \geq 0.$$

  We shall see below that this constraint is weaker than another one ($f_3(s^2) \geq 0$) and needs not be taken into account.

- Let us now turn to the condition upon $\xi$. The first case $-\xi \leq 1+\zeta$ reads

$$2-\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)-2\kappa_u|a_\star|\sigma s^2 \leq 2+\gamma^2[1-(\theta_1+\theta_2)+2\theta_1\theta_2]-2\kappa_u|a_\star|\sigma s^2+4a_\star^2\sigma^2 s^2(1-s^2)$$

  which comes down to
$$-\gamma^2-4a_\star^2\sigma^2 s^2(1-s^2)\leq 0.$$
  The latter inequality always holds and does not imply an additional constraint upon $\Delta t$.
- Finally, we consider the case $\xi \leq 1+\zeta$. This leads to

$$-2+\gamma^2(\theta_1+\theta_2-2\theta_1\theta_2)+2\kappa_u|a_\star|\sigma s^2 \leq 2+\gamma^2[1-(\theta_1+\theta_2)+2\theta_1\theta_2]-2\kappa_u|a_\star|\sigma s^2+4a_\star^2\sigma^2 s^2(1-s^2).$$

  It follows that
$$f_3(s^2) := \gamma^2\Theta_3+4(1-\kappa_u|a_\star|\sigma s^2)+4a_\star^2\sigma^2 s^2(1-s^2)\geq 0.$$
  From $\Theta_3 = 2\Theta_2-1$, we infer that $2f_2(s^2)-f_3(s^2)\geq 0$ over $[0,1]$. This implies that the condition $f_2(s^2)\geq 0$ is a consequence of $f_3(s^2)\geq 0$.
  Function $f_3$ is maximal over $\mathbb{R}$ at $X_3 := \frac{1}{2}\left(1-\frac{\kappa_u}{|a_\star|\sigma}\right) \leq \frac{1}{2}$. The minimum over $[0,1]$ is reached for $s^2=1$ and the condition $f_3(s^2)\geq 0$ reduces to
$$0\leq f_3(1) = \omega^2\Theta_3\Delta t^2-\frac{4\kappa_u|a_\star|}{\Delta x}\Delta t+4 =: Q_3(\Delta t).$$
  The resolution of the second order equation $Q_3(\Delta t)=0$ leads to the stability condition (38b) depending on the sign of $\omega^2\Theta_3\Delta x^2-\kappa_u^2 a_\star^2$.

□

**Lemma 6** (Stability of the All Froude Godunov scheme)**.** *The CFL condition (38a-38b) obtained for $\kappa_r = 0$ still ensures the stability of the All Froude Godunov scheme, i.e. for the choice $\kappa_r = \mathcal{O}(M)$.*

The proof is obtained by using a classical continuity argument. The key point is to prove that the modulus of the eigenvalue $\lambda_0$ is increasing when $\kappa_r \to 0^+$. The detailed proof of Lemma 6 is given in Appendix B.

## 5. Numerical results

Let us fix the parameters $a_\star = 1$, $\omega = 1$ , $M = 10^{-3}$ and consider the initial condition

$$q_i^0 = \hat{q}_i^0 + M\frac{\tilde{q}_i^0}{\|\tilde{q}_i^0\|} \quad \text{with} \quad \hat{q}_i^0 = \begin{pmatrix} \sin(\omega x_i) \\ 0 \\ a_\star\cos(\omega x_i)\frac{\sin(\omega\Delta x)}{\omega\Delta x}\end{pmatrix} \in \mathcal{E}^h_{\omega\neq 0},\ \tilde{q}_i^0 = \begin{pmatrix} a_\star\cos(\omega x_i)\frac{\sin(\omega\Delta x)}{\omega\Delta x} \\ 1 \\ \sin(\omega x_i)\end{pmatrix} \in \mathcal{E}^{h,\perp}_{\omega\neq 0},$$

that is close to the kernel $\mathcal{E}^h_{\omega\neq 0}$ (see Lemma 4) up to a perturbation of order $M$.

We solve the 1D linear wave equation (7) by means of the schemes we analyzed in the previous sections, namely the *low Froude* scheme (35) for $\kappa_r = 0$, the *all Froude scheme* (35) for $\kappa_r = \mathcal{O}(M)$, and the *classical Godunov* scheme (35) for $\kappa_r = 1$. In a first step, we take $\theta_1 = 1$ and $\theta_2 = 0$.

We observe on Figure 2 that the two schemes designed for the low Froude regime have the correct behaviour as the Froude number goes to 0, unlike the classical Godunov scheme which is not accurate as stated before.

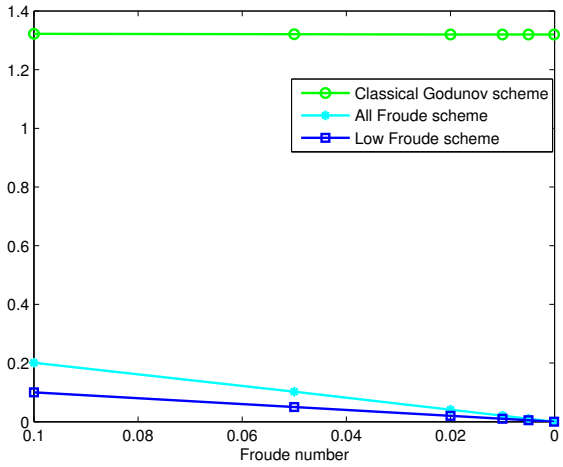



FIGURE 2. Evolution of $\max_t \|q_h - \mathbb{P}q_h^0\|(t)$ for $t = \mathcal{O}(1)$ when the Froude number goes to 0 for the Low Froude Godunov, the All Froude Godunov and the Classical Godunov schemes.

We now investigate the accuracy with time at a fixed Froude number. As it was stated in Theorem 1 *(i)*, we see on Figures 3(a)-(c) that the two aforementioned schemes are accurate for times $t = \mathcal{O}(1)$ since the numerical solutions remain close to the projection of the initial data onto the kernel (the norm of the difference is of order $10^{-3}$). However for large times the all Froude scheme turns out be inaccurate as the corresponding solution is moving away from the kernel. It illustrates the result from Theorem 1 *(ii)*.

Next, we now focus on Theorem 1 *(iii)* by means of Figures 4(a)-(b), where we see that the total deviation of the Classical Godunov scheme is of order $M$ when the mesh is sufficiently refined ($\Delta x = \mathcal{O}(M)$). Note that even in this case, the behaviour of the low/all Froude Godunov schemes is better than that of the classical Godunov scheme.

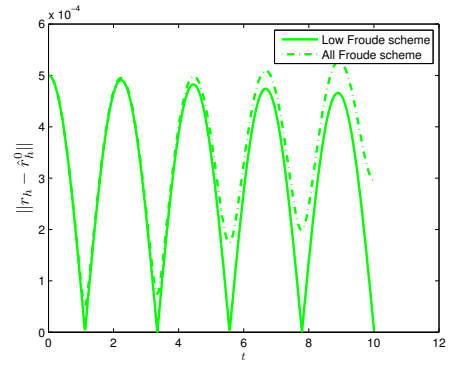
In Figures 5 and 6, we change the value $\theta_1$ and $\theta_2$ of the Low/All Froude schemes. These figures indicate that the total deviation depends on the value of $\theta_1 + \theta_2$.

In the second test case, we consider the initial condition given by

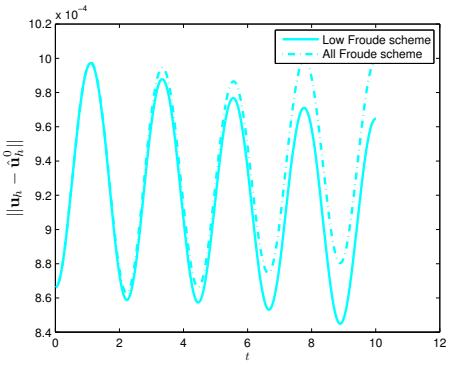
$$\begin{cases} r_i^0 = \chi_{[-\frac{1}{2},\frac{1}{2}]}(x_i), \\ u_i^0 = 1, \\ v_i^0 = 1. \end{cases} \tag{45}$$

In this test, we choose $\omega = 1$, $\Delta x = 0.01$ and $a_\star$ such that the Rossby deformation is equal to $R_d := \frac{a_\star}{\omega} = \Delta x$ and $\kappa_u = 1$. In Figure 7(a), we choose $\theta_1 = 0.5$ and $\theta_2 = 0$. Hence, in this case we have $\Theta_1 = 0.5$ and $\Theta_3 = 0$ which leads to $\Delta t_a = \frac{0.5}{1-\sqrt{0.5}}$, $\Delta t_b = 1$ and $\Delta t_0 = 0.5$. Therefore, the new time step $\Delta t = \min\{\Delta t_a, \Delta t_b\} = 1$ is less restrictive than the classical time step $\Delta t_0 = 0.5$. Figure 7(a) shows that the new time step is optimal since when $\Delta t = 0.999 < 1$ the *Low Froude scheme* is stable while when $\Delta t = 1.001$ the *Low Froude scheme* is unstable.

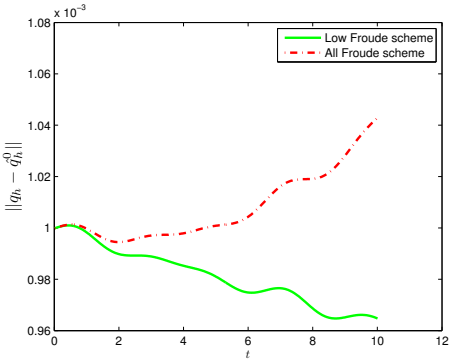
On the other hand, if we take $\theta_1 = 0$ and $\theta_2 = 0$, then $\Theta_1 = 1$ and $\Theta_3 = 1$. Due to the fact that $\frac{\kappa_u^2 a_\star^2}{\omega^2 \Delta x^2} \leq \Theta_3$, the constraint over the time step for the Low Froude scheme is prescribed by $\Delta t_a$. However as $a_\star = \omega \Delta x \sqrt{\Theta_1}$, $\Delta t_a = +\infty$ and the Low Froude scheme is always stable without regard to the time step $\Delta t$. Figure 7(b) confirms this statement by showing that the Low Froude scheme is stable even for $\Delta t = 10$.



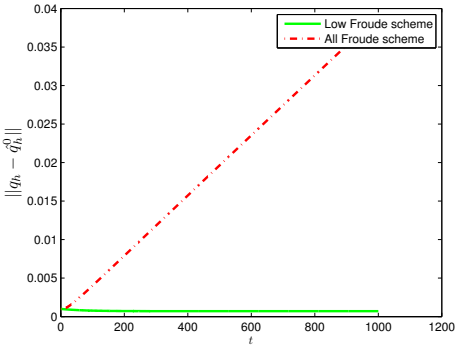

(A) Deviation on the pressure




(B) Deviation on the velocity

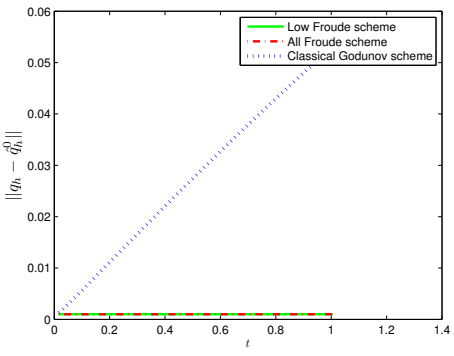



(C) Total deviation (short time)

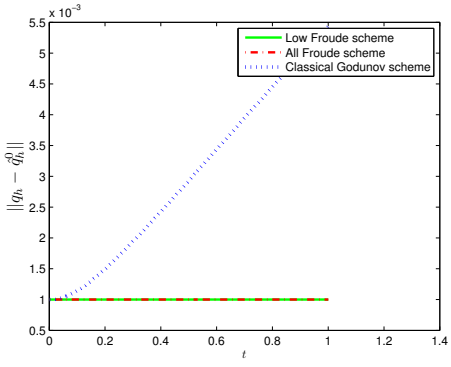



(D) Total deviation (long time)

FIGURE 3. Comparisons of schemes: proximity to the discrete kernel as time increases.

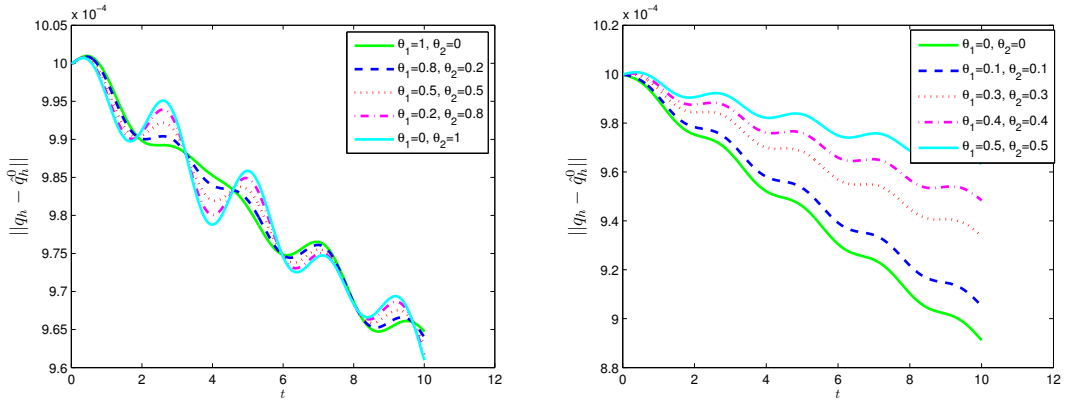



(A) Total deviation ($\Delta x = 2\pi \times 10^{-2}$ )




(B) Total deviation ($\Delta x = 2\pi \times 10^{-3}$)

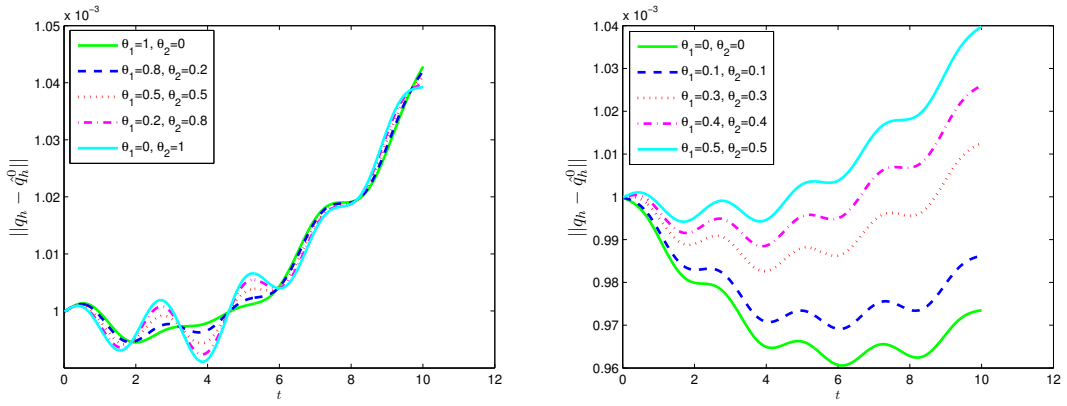
FIGURE 4. Comparisons of schemes: proximity to the discrete kernel as time increases.




(A) Total deviation with $\theta_1 + \theta_2 = 1$.

(B) Total deviation with various values of $\theta_1 + \theta_2$.

FIGURE 5. Comparisons of Low Froude schemes: proximity to the discrete kernel as time increases.




(A) Total deviation with $\theta_1 + \theta_2 = 1$.

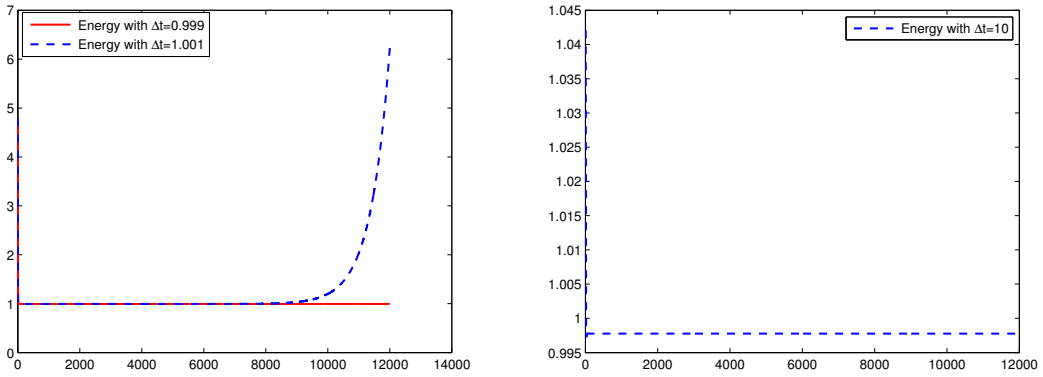
(B) Total deviation with various values of $\theta_1 + \theta_2$.

FIGURE 6. Comparisons of All Froude schemes: proximity to the discrete kernel as time increases.

In Figure 8, we take $\Delta t = \Delta t_0 = 0.5$ and change the value of $\theta_1$ and $\theta_2$. This figure indicates that the behaviour of the energy of the Low Froude scheme depends on the value of $\theta_1$ and $\theta_2$. The choice $\theta_1 = \theta_2 = \frac{1}{2}$ (Crank-Nicolson approximation for the Coriolis term) is able to preserve the energy exaclty although this choice requires a more restrictive constraint upon the time step than for $\theta_1, \theta_2 \leq \frac{1}{2}$.

## 6. CONCLUSION

It is well known that the *classical Godunov scheme* applied to the linear wave equation is not accurate at low Froude number on cartesian meshes in dimension 2 [8]. In this work, we have shown that, when a Coriolis source term is involved, the *classical Godunov scheme* is not accurate at low Froude number even in dimension 1.




(A) $\theta_1 = 0.5, \theta_2 = 0$ (B) $\theta_1 = 0, \theta_2 = 0$

FIGURE 7. Influence of the time step upon the Low Froude scheme

This is because the stationary space of the *classical Godunov* discrete operator is not a good approximation of the invariant subspace $\mathcal{E}_{\omega\neq 0}$. The loss of invariance of $\mathcal{E}_{\omega\neq 0}$ is explained by studying the associated modified equation. It is strongly related to the numerical diffusion $\kappa_r$ on the pressure equation. In particular when we set $\kappa_r = 0$, the inaccuracy problem does not occur. As a result, we derived two modified schemes by decreasing the value of the numerical diffusion $\kappa_r$ on the pressure equation. From this, we deduce that:

- The *Low Froude Godunov scheme* ($\kappa_r = 0$) is accurate at low Froude number.
- The *All Froude Godunov scheme* ($\kappa_r = M$) is accurate at low Froude number locally in time.

We then proved that both schemes are stable under suitable constraints upon the time step. These stability conditions turn out to be less restrictive than classical ones for a suitable treatment of the Coriolis source term.

In forthcoming works, we shall extend our analysis to the two-dimensional case and to the nonlinear framework in order to derive and analyse accurate and stable numerical schemes for System (1). In particular we aim at comparing our work with previous approaches from [4, 6, 12].

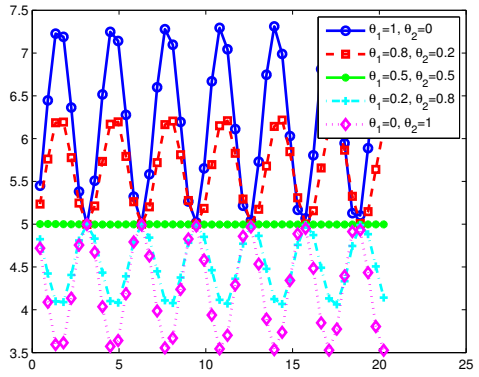
## A. Analysis of splitting scheme

Let us define a two-step Godunov scheme using a splitting strategy to take into account the Coriolis source term. The first step is related to the acoustic term

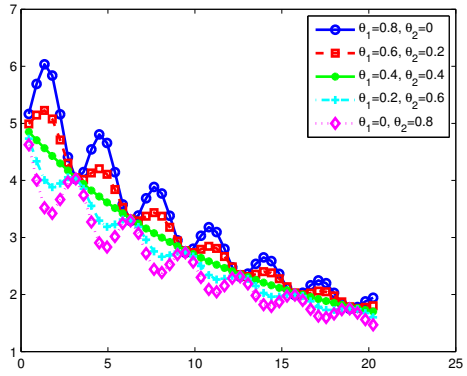
$$\begin{cases} r_i^\star - r_i^n + a_\star \Delta t \frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} - \nu_r \Delta t \frac{r_{i+1}^n - 2r_i^n + r_{i-1}^n}{\Delta x^2} = 0, \\ u_i^\star - u_i^n + a_\star \Delta t \frac{r_{i+1}^n - r_{i-1}^n}{2\Delta x} - \nu_u \Delta t \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} = 0, \\ v_i^\star - v_i^n = 0, \end{cases} \tag{46a}$$

and we use a $\theta$-scheme to deal with the Coriolis term in the second step

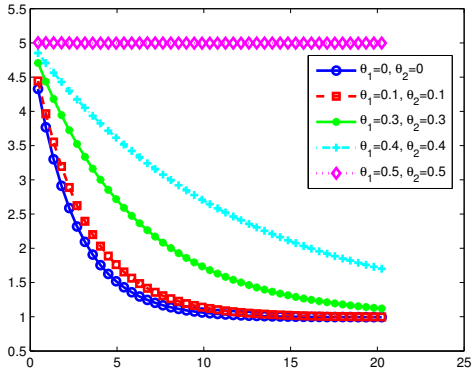
$$\begin{cases} r_i^{n+1} = r_i^\star, \\ u_i^{n+1} - u_i^\star = \omega \Delta t \left[\theta_1 v_i^\star + (1-\theta_1) v_i^{n+1}\right], \\ v_i^{n+1} - v_i^\star = -\omega \Delta t \left[\theta_2 u_i^\star + (1-\theta_2) u_i^{n+1}\right], \end{cases} \tag{46b}$$




(A) Energy with $\theta_1 + \theta_2 = 1$.




(B) Energy with $\theta_1 + \theta_2 = 0.8$.




(C) Energy with various values of $\theta_1 + \theta_2$.

FIGURE 8. Comparisons of Low Froude schemes depending on the time step

for $0 \leq \theta_1, \theta_2 \leq 1$.

**Lemma 7.**

*(i) For $\nu_r = 0$, the splitting scheme preserves steady states only if $\theta_2 = 0$.*
*(ii) For $\nu_r \neq 0$, steady states are not preserved without regard to the value of $\theta_1$ and $\theta_2$.*

*Proof.* Let us assume that the numerical solution at time $t^n = n\Delta t$ belongs to the discrete kernel $\mathcal{E}^h_{\omega \neq 0}$

$$\forall i \in \mathbb{Z}, \quad u_i^n = 0 \quad \text{and} \quad a_\star \frac{r_{i+1}^n - r_{i-1}^n}{2\Delta x} = \omega v_i^n.$$

We shall show that at the next time step the numerical solution does not lie in the discrete kernel anymore. After the first step, we easily obtain

$$\begin{cases} r_i^\star = r_i^n + \nu_r \Delta t \frac{r_{i+1}^n - 2r_i^n + r_{i-1}^n}{\Delta x^2}, \\ u_i^\star = u_i^n - a_\star \Delta t \frac{r_{i+1}^n - r_{i-1}^n}{2\Delta x} = -\omega \Delta t v_i^n, \\ v_i^\star = v_i^n. \end{cases}$$

Then the second step leads to

$$\begin{cases} r_i^{n+1} = r_i^n + \nu_r \Delta t \frac{r_{i+1}^n - 2r_i^n + r_{i-1}^n}{\Delta x^2}, \\ u_i^{n+1} = \omega \Delta t (1-\theta_1)(v_i^{n+1} - v_i^n), \\ v_i^{n+1} + \omega \Delta t (1-\theta_2) u_i^{n+1} = \left[1 + (\omega \Delta t)^2 \theta_2\right] v_i^n. \end{cases}$$

As a result, we have

$$v_i^{n+1} + (\omega \Delta t)^2 (1-\theta_1)(1-\theta_2)(v_i^{n+1} - v_i^n) = \left[1 + (\omega \Delta t)^2 \theta_2\right] v_i^n$$

from which it follows that

$$\left[1 + (\omega \Delta t)^2 (1-\theta_1)(1-\theta_2)\right] v_i^{n+1} = \left[1 + (\omega \Delta t)^2 (1-\theta_1)(1-\theta_2) + (\omega \Delta t)^2 \theta_2\right] v_i^n.$$

Therefore, $v_i^{n+1} = v_i^n$ (and $u_i^{n+1} = 0$) iff $\theta_2 = 0$. The kernel is recovered if $\nu_r = 0$ as $r_i^{n+1} = r_i^n$. □

Let us now note that the choice $\theta_2 = 0$ is not really a splitting method since it can be written as a one-step method

$$\begin{cases} r_i^{n+1} - r_i^n + a_\star \Delta t \frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} - \nu_r \Delta t \frac{r_{i+1}^n - 2r_i^n + r_{i-1}^n}{\Delta x^2} = 0, \\ u_i^{n+1} - u_i^n + a_\star \Delta t \frac{r_{i+1}^n - r_{i-1}^n}{2\Delta x} - \nu_u \Delta t \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{\Delta x^2} = \omega \Delta t [\theta_1 v_i^n + (1-\theta_1) v_i^{n+1}], \\ v_i^{n+1} - v_i^n = -\omega \Delta t u_i^{n+1}. \end{cases}$$

## B. Stability of the one step All-Froude Godunov scheme

Here we detail the proof of Lemma 6.

*Proof.* The characteristic polynomial $\mathcal{P}(\lambda)$ of the amplification matrix (41) is given by

$$\mathcal{P}(\lambda) = (1 - \lambda - 2\kappa_r |a_\star| \sigma s^2) \left( \lambda^2 - \frac{2 - \gamma^2(\theta_1 + \theta_2 - 2\theta_1\theta_2) - 2\kappa_u |a_\star| \sigma s^2}{\Lambda(\theta_1, \theta_2)} \lambda + \frac{1 + \gamma^2 \theta_1 \theta_2 - 2\kappa_u |a_\star| \sigma s^2}{\Lambda(\theta_1, \theta_2)} \right) + (1-\lambda) \frac{4 a_\star^2 \sigma^2 s^2 (1 - s^2)}{\Lambda(\theta_1, \theta_2)},$$

which can be decomposed as

$$\mathcal{P}(\lambda) = \mathcal{P}_0(\lambda) + \kappa_r \mathcal{P}_1(\lambda), \tag{47}$$

where $\mathcal{P}_0$ is given by (43) and

$$\mathcal{P}_1(\lambda) = -2|a_\star| \sigma s^2 \left( \lambda^2 - \frac{2 - \gamma^2(\theta_1 + \theta_2 - 2\theta_1\theta_2) - 2\kappa_u |a_\star| \sigma s^2}{\Lambda(\theta_1, \theta_2)} \lambda + \frac{1 + \gamma^2 \theta_1 \theta_2 - 2\kappa_u |a_\star| \sigma s^2}{\Lambda(\theta_1, \theta_2)} \right).$$

Since the roots of polynomial $\mathcal{P}_0$ are simple – see the proof of Lemma 5 – a classical continuity argument [13] allows us to write the roots of the polynomial $\mathcal{P}$ by using an asymptotic expansion

$$\lambda = \lambda^{(0)} + \kappa_r \lambda^{(1)} + \mathcal{O}(\kappa_r^2) \tag{48}$$

where $\lambda^{(0)}$ is a root of $\mathcal{P}_0$. The stability of the scheme is obtained if the modulus of all roots of $\mathcal{P}$ is smaller than one. If $\lambda^{(0)} = \lambda_\pm$, the results is obvious since one can ensure $|\lambda_\pm| < 1$ by considering

$$\Delta t \leq K \min\{\Delta t_a, \Delta t_b\},$$

with $K < 1$ small enough and $\Delta t_a$, $\Delta t_b$ given in (38a-38b). The case $\lambda^{(0)} = \lambda_0 = 1$ is a bit more tricky. By inserting the asymptotic expansion (48) into relation (47), we obtain

$$\mathcal{P}(\lambda) = \kappa_r \left[\lambda_1 \mathcal{P}_0'(\lambda_0) + \mathcal{P}_1(\lambda_0)\right] + \mathcal{O}(\kappa_r^2).$$

The condition $\mathcal{P}(\lambda) = 0$ thus implies

$$\lambda_1 = -\frac{\mathcal{P}_1(\lambda_0)}{\mathcal{P}_0'(\lambda_0)}.$$

Easy computations lead to

$$\begin{aligned}\mathcal{P}_1(\lambda_0) &= -2|a_\star|\sigma s^2 \left(1 - \frac{2 - \gamma^2(\theta_1 + \theta_2 - 2\theta_1\theta_2) - 2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)} + \frac{1 + \gamma^2\theta_1\theta_2 - 2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)}\right) \\ &= -\frac{2|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)}\gamma^2 < 0.\end{aligned}$$

On the other hand, since $\mathcal{P}_0(\lambda) = (1-\lambda)\widetilde{\mathcal{P}_0}(\lambda)$, we have

$$\begin{aligned}\mathcal{P}_0'(1) = -\widetilde{\mathcal{P}_0}(1) &= -1 + \frac{2 - \gamma^2(\theta_1 + \theta_2 - 2\theta_1\theta_2) - 2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)} - \frac{1 + \gamma^2\theta_1\theta_2 - 2\kappa_u|a_\star|\sigma s^2}{\Lambda(\theta_1,\theta_2)} - \frac{4a_\star^2\sigma^2 s^2(1-s^2)}{\Lambda(\theta_1,\theta_2)} \\ &= -\frac{\gamma^2 + 4a_\star^2\sigma^2 s^2(1-s^2)}{\Lambda(\theta_1,\theta_2)} < 0.\end{aligned}$$

It follows that

$$-2|a_\star|\sigma < \lambda_1 < 0,$$

and the scheme is stable. □

## References


[1] S. Alinhac. *Hyperbolic partial differential equations*. Springer, 2009.

[2] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, and B. Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM J. Sci. Comput.*, 25(6):2050–2065, 2004.

[3] F. Bouchut. *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*. Frontiers in Mathematics. Birkhäuser Verlag, 2004.

[4] F. Bouchut, J. Le Sommer, and V. Zeitlin. Frontal geostrophic adjustment and nonlinear wave phenomena in one-dimensional rotating shallow water. II. High-resolution numerical simulations. *J. Fluid Mech.*, 514:35–63, 2004.

[5] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer, 2011.

[6] M.J. Castro, J.A. López, and C. Parés. Finite volume simulation of the geostrophic adjustment in a rotating shallow-water system. *SIAM J. Sci. Comput.*, 31(1):444–477, 2008.

[7] F. Couderc, A. Duran, and J.-P. Vila. An explicit asymptotic preserving low Froude scheme for the multilayer shallow water model with density stratification. arXiv:1607.00200, 2016.

[8] S. Dellacherie. Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number. *J. Comput. Phys.*, 229(4):978–1016, 2010.
[9] S. Dellacherie, P. Omnes, and F. Rieper. The influence of cell geometry on the Godunov scheme applied to the linear wave equation. *J. Comput. Phys.*, 229(14):5315–5338, 2010.
[10] D.Y. Le Roux. Spurious inertial oscillations in shallow-water models. *J. Comput. Phys.*, 231(24):7959–7987, 2012.
[11] R.J. LeVeque. *Finite volume methods for hyperbolic problems*, volume 31. Cambridge Univ. Press, 2002.
[12] M. Lukacova-Medvidova, S. Noelle, and M. Kraft. Well-balanced finite volume evolution Galerkin methods for the shallow water equations. *J. Comput. Phys.*, 221(1):122–147, 2007.
[13] M. Marden. *Geometry of polynomials*, volume 3. American Mathematical Society, 1949.
[14] Y. Moguen, T. Kousksou, P. Bruel, J. Vierendeels, and E. Dick. Pressure–velocity coupling allowing acoustic calculation in low Mach number flow. *J. Comput. Phys.*, 231(16):5522–5541, 2012.
[15] M. Parisot and J.-P. Vila. Numerical scheme for multilayer shallow-water model in the low-Froude number regime. *C. R. Acad. Sci. Ser. I Math.*, 352(11):953–957, 2014.
[16] F. Rieper. A low-Mach number fix for Roe's approximate Riemann solver. *J. Comput. Phys.*, 230(13):5263–5287, 2011.
[17] S. Vater and R. Klein. Stability of a Cartesian grid projection method for zero Froude number shallow water flows. *Numer. Math.*, 113(1):123–161, 2009.