UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Schémas compacts hermitiens sur la Sphère - Applications en climatologie et océanographie numérique

THÈSE

présentée et soutenue publiquement le 3 Juillet 2018

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention mathématiques appliquées)

par

Matthieu Brachet

Composition du jury

Président :	Michael Ghil	Professeur à l'ENS
Rapporteurs :	Eric Blayo	Professeur à l'Université de Grenoble
	Thomas Dubos	Professeur à L'École Polytechnique
Examinateurs :	Didier Clamond	Professeur à l'Université de Nice
	Jean-Pierre Croisille	Professeur à l'Université de Lorraine (Directeur de Thèse)
	Véronique Martin	Maitre de conférence à l'Université de Picardie
	Dong Ye	Professeur à l'Université de Lorraine

Institut Elie Cartan de Lorraine — UMR 7502

Mis en page avec la classe thesul.

Remerciements

Quand j'ai commencé ma thèse, on m'a dit que c'était un travail très solitaire. Je ne peux pas dire ça. J'ai eu la chance de toujours être bien entouré.

Je tiens tout d'abord à exprimer ma gratitude et mes remerciements à Jean-Pierre Croisille pour avoir encadré mon travail de thèse. Il serait étrange de dire que sans lui ce travail n'aurait pas eu lieu, je le remercie pour les nombreuses discussions, ses conseils et sa patience. Il m'a permis de mener ce travail dans d'excellentes conditions pendant ces quatre années. A ses cotés, j'ai énormément appris autant en termes de sciences et de rigueur que humainement.

J'adresse mes remerciements à Eric Blayo et à Thomas Dubos pour avoir accepté la lourde tâche de rapporteur. Ma gratitude est aussi adressée à Michael Ghil pour la présidence du jury de ma thèse. J'exprime ma reconnaissance à Didier Clamond, Véronique Martin et Dong Ye pour avoir accepté de faire partie du jury de ma thèse.

Je souhaite remercier tout particulièrement Jean-Paul Chehab pour nos discussions, ses conseils et ses remarques. Après mon stage de master 2, j'ai eu la chance de pouvoir continuer à travailler à ses côtés et c'est toujours un plaisir pour moi. Je l'en remercie. Ma gratitude est aussi adressée à l'ensemble de l'Institut Elie Cartan de Lorraine. En particulier je remercie Xavier Antoine, David Dos Santos Ferreira et Julien Lequeurre pour nos discussions. Ma gratitude va aussi à Olivier Botella, Antoine Henrot, Vladimir Latocha et Didier Schmitt. Je remercie également Claude Coppin, Hélène Jouve et Paola Schneider pour leur accompagnement administratif. Bien sûr je remercie Laurence Quirot et sa bonne humeur permanente ainsi que Didier Gemmerlé (sans qui j'aurais eu de sérieux problèmes informatiques) et Elodie Cunat (qui a toujours l'information recherchée). Sachez que j'ai apprécié chaque moment de nos discussions.

A la fin de ma première année j'ai eu l'opportunité de participer au CEMRACS. C'était une chance et je suis très heureux d'y avoir participé. A cette occasion, je remercie l'ensemble de l'équipe Hydromorpho : Nora Aissiouene, Tarik Amtout, Romain Hild, Christophe Prud'homme, Antoine Rousseau et Stéphanie Salmon. Vous avez rendu le projet particulièrement enrichissant. Le CEMRACS fut aussi l'occasion de rencontres : Andrea, Clémentine, Guillaume, Hélène, Charlotte, Ranine, ... (la liste est bien trop longue, merci à tous ! sincèrement). Je vous croise encore, nous discutons souvent et j'apprécie toujours autant ces moments.

J'ai commencé ces remerciements en parlant de mon entourage. Je ne peux pas oublier de remercier mes parents. Ils m'ont permis de faire des études (longtemps, très longtemps) et m'ont toujours accompagné, que ce soit dans les bons ou les mauvais moments. Ils sont toujours à mes cotés et me soutiennent. Nous ne nous disons pas ce genre de choses mais merci beaucoup et merci d'être là. Je n'oublierai pas non plus mes frères, Samuel et Valentin, qui trouveront toujours une bonne raison de m'empêcher de travailler.

Ma thèse ne se serait pas déroulée comme elle s'est déroulée sans ces gens qui m'ont supporté quotidiennement (alors qu'ils avaient le choix !). Je parle bien sûr de mes collègues et amis. Un très grand merci à mes coburaux : Clément et Benjamin à Metz (sans qui cette thèse aurait été très différente), ainsi que Dimitry, Florian et Yang à Nancy (pour le 513). En écrivant ce texte je pense aussi évidement à Coralie (que je retrouverai à 16h je suppose), Clémence (mais tu fais des maths !), Tom, Maxime, Allan, ... Un grand merci à l'ensemble des doctorants de l'IECL.

Il y a ceux que j'ai rencontré grâce aux mathématiques et il y a ceux qui préfèrent sûrement lorsque j'en parle un peu moins. Jennifer et Mathieu, je n'aurais jamais cru que vous croiser entre deux rayons puisse aboutir à cette amitié. Magaly, je t'attend toujours pour boire un monaco. David et Hélène, ma L2 est loin mais nous discutons toujours et c'est un plaisir. Émilie, Julien et Mira, j'ai fait mes études avec vous, merci de les avoir faites avec moi ;). Un grand merci à Guillaume, Jasmine, Typhaine, Vincent et Anne-Sophie qui finalement étaient une bonne raison d'aller dans les Vosges. Un merci un peu spécial à Camille. Je remercie également Alexis, certes je t'ai rencontré grâce aux maths mais je m'en fous je te mets ici.

Je remercie également tous les étudiants avec lesquels j'ai travaillé, qui ont supporté mon incontestable désorganisation et ma numérotation aléatoire des chapitres et propositions.

Enfin, j'adresse mes remerciements à l'équipe du LAMFA qui, à travers les enseignements, m'a donné un goût particulier et une certaine curiosité pour les mathématiques.

"Ce n'est qu'en essayant continuellement que l'on finit par réussir...
En d'autres termes... Plus ça rate et plus on a de chances que ça marche..."
Devise Shadoks, J. Rouxel.

Je dédie cette thèse à ma Grand-Mère,
qui n'a pas pu voir ce travail aboutir.
A mes parents, à ma famille, à mes amis.

Table des matières

Introduction générale

Simuler numériquement la dynamique des fluides atmosphérique et océanographique par des méthodes de haute précision représente un enjeu fondamental pour la prévision climatique à grande échelle. Les questions écologiques, sociales et politiques du réchauffement climatique rendent ce problème central. On renvoie aux ouvrages récents de G. K. Vallis [82] de B. Cushman-Roisin et *al.* [26] ainsi qu'au livre de M. Ghil et *al.* [39]. Il s'agit d'un problème délicat. Ceci est dû en particulier au couplage entre phénomènes convectifs et thermodynamiques présents dans les équations de Navier-Stokes. De plus, le contexte sphérique et la variété des échelles de temps et d'espace rajoutent encore à la complexité.

Notre objectif concerne la résolution des équations Shallow Water sphériques (équations de Saint-Venant). Ces dernières représentent le modèle le plus simple pour les mouvements d'une atmosphère de faible épaisseur sur la sphère en rotation. Bien qu'il s'agisse d'une simplification, ce modèle prend en compte plusieurs difficultés attachées à la propagation atmosphérique. En particulier, il permet d'analyser les principales ondes de propagation (ondes de Poincaré, ondes de Rossby).

Dans ce travail, nous considérons la conception d'un schéma pour la résolution des équations Shallow Water sur la sphère. Ce problème va au delà du cadre plan où il est souvent traité sous différentes hypothèses concernant la force de Coriolis (par exemple, l'hypothèse du $\beta-$plan où la force de Coriolis dépend seulement de la coordonnée y). Des références classiques pour les méthodes numériques en dynamique des fluides sont [7, 30, 88].

Dans de nombreux travaux récents, les équations Shallow Water sphériques sont résolues par des méthodes numériques sur grille. Des grilles particulières sont la grille icosaèdrale, la grille Yin-Yang et la grille Cubed-Sphere. Dans [78], le schéma volumes finis Dynamico utilise une grille icosaèdrale. Une autre méthode de volumes finis utilisant une grille Yin-Yang est introduite dans [61]. Différentes méthodes de Galerkin Discontinu [58] ont également été considérées sur les maillages suivants : maillage icosaèdral [40], maillage Yin-Yang [43] et Cubed-Sphere [56, 67]. D'autres travaux utilisent la théorie des éléments finis mimétiques [31]. Notons également une activité importante utilisant des approximations sans grille. Les équations SWE sont résolues par méthodes particulaires dans [13]. Par ailleurs, une méthode utilisant les fonctions de base radiale est considérée dans [33, 35].

Dans [24, 25] est introduit un schéma aux différences finies pour le calcul du gradient d'un champ scalaire sphérique et de la divergence d'un champ vectoriel sphérique donné aux points de la Cubed-Sphere [74]. Ce schéma est basé sur une approximation hermitienne le long de grands cercles correspondant à des lignes de coordonnées sur la grille.

Dans cette thèse, nous étudions différents aspects de cette approche. Nous commençons par analyser en détail la grille Cubed-Sphere, en particulier ses propriétés de symétrie. Dans un second temps, nous montrons comment la structure en grands cercles de la grille permet de définir de façon naturelle des opérateurs différentiels discrets. On en déduit un algorithme de référence pour la discrétisation des problèmes sur la sphère. Nous nous sommes attachés à effectuer de nombreux tests de la littérature en climatologie numérique et à étudier les résultats obtenus en détail. Comme nous le verrons dans les chapitres 5 et 6, les résultats sont comparables aux meilleurs schémas conservatifs d'ordre 4 disponibles actuellement.

Plan de la thèse :

Chapitre 1 : Schémas aux différences.

L'objectif de ce chapitre est d'introduire les notations et les schémas aux différences finies 1D et 2D qui seront utilisés sur la sphère. Nous détaillons des schémas d'approximations de la dérivée première à l'aide de méthodes de différences finies classiques ou hermitiennes. Nous introduisons aussi les opérateurs de filtrage. Les propriétés spectrales de ces outils sont étudiées.

Chapitre 2 : Analyse numérique.

La discrétisation d'équations aux dérivées partielles d'évolution est introduite et étudiée. En particulier, nous analysons les propriétés de précision, stabilité et conservation pour l'équation d'advection en dimension 1 et l'équation Shallow Water linéarisée en dimension 2. Des tests sont aussi effectués sur l'équation de Burgers. L'opérateur de filtrage est analysé sur ces problèmes d'évolution.

Chapitre 3 : Grille Cubed-Sphere.

Nous introduisons le maillage Cubed-Sphere. Ce dernier est construit à partir de grands cercles. Un produit scalaire est analysé sur les fonctions de grille. Il permet d'obtenir l'orthogonalité d'un grand nombre d'harmoniques sphériques sur la grille. Des formules de quadrature issues de ce produit scalaire sont analysées.

Chapitre 4 : Approximation des opérateurs différentiels sur la Cubed-Sphere.

On utilise la structure en grands cercles du maillage pour construire des opérateurs gradient, divergence et vorticité discrets sur la Cubed-Sphere. Nous analysons la consistance de ces opérateurs et effectuons des expériences numériques. L'opérateur de filtrage en dimension 1 est étendu à la Cubed-Sphere.

Chapitre 5 : Équations d'advection sphériques.

Des expériences numériques sont effectuées sur l'équation d'advection sphérique linéaire et non-linéaire. La précision du schéma est analysée ainsi que l'influence du filtrage sur des tests de convection de type rotation solide ou de type tourbillon. Sur l'équation d'advection non linéaire, nous observons le comportement du schéma en présence d'un choc et la conservation d'une solution stationnaire. Nous nous restreignons à un schéma linéaire sans opérateur de capture de chocs.

Chapitre 6 : Équations Shallow Water sphériques.

Nous évaluons les performances du schéma sur le système d'équations Shallow Water et son linéarisé. Des tests sont faits sur des solutions stationnaires et des problèmes évoluant dans le temps. Ces derniers sont issus de la littérature classique. Nous analysons le comportement de la solution calculée par l'algorithme et les propriétés de conservation observées numériquement.

Table des figures

Rapport de Thèse

Chapitre 1

Schémas aux différences

1.1 Opérateurs aux différences en dimension 1

1.1.1 Notations

On considère $\Omega = [a, b]$, $a < b$, un intervalle de $\mathbb{R}$ de longueur $L = b - a$. Nous utilisons les lettres latines pour noter les fonctions continues : $f(x)$, $u(x)$, ... $x \in \Omega$ à valeur complexe. Pour u et v, des fonctions définies sur Ω, le produit scalaire $L^2(\Omega)$ est défini par

$$(u, v) = \int_\Omega u(x)\bar{v}(x)dx = \int_a^b u(x)\bar{v}(x)dx. \tag{1.1}$$

Pour u et v à valeurs réelles, on a

$$(u, v) = \int_a^b u(x)v(x)dx. \tag{1.2}$$

La norme sur $L^2(\Omega)$ est donnée par

$$\|u\|_{L^2(\Omega)} = \sqrt{(u, u)}. \tag{1.3}$$

Pour $u \in \mathcal{C}(\bar{\Omega}) \cap L^\infty(\Omega)$, on note

$$\|u\|_\infty = \sup_{x\in\Omega} |u(x)|. \tag{1.4}$$

Une fonction $u : x \in \mathbb{R} \mapsto u(x) \in \mathbb{R}$ est *périodique* de période L si

$$u(x + L) = u(x), \forall x \in \mathbb{R}. \tag{1.5}$$

En particulier, on a $u(a) = u(b)$.

On considère une grille régulière sur Ω constituée de $N + 1$ points :

$$a = x_0 < x_1 < \ldots < x_{N-1} < x_N = b, \tag{1.6}$$

où les valeurs x_j sont définies par :

$$x_j = a + jh,\ j = 0, 1, \ldots, N \text{ et } h = \frac{b-a}{N} \text{ le pas d'espace.} \tag{1.7}$$

Les points $x_0 = a$ et $x_N = a + L = b$ sont les points de bord du domaine et les points $(x_j)_{1\le j\le N-1}$ désignent les points intérieurs.

Nous distinguons trois types de données aux points de grille x_j, $0 \le j \le N$:

1. Une *fonction de grille* est une fonction définie uniquement aux points $(x_j)_{0\le j\le N}$. Les fonctions de grilles sont notées en fonte gothique : $\mathfrak{u}$, $\mathfrak{v}$, ... On note

$$\mathfrak{u} = (\mathfrak{u}(x_0), \mathfrak{u}(x_1), \mathfrak{u}(x_2), ..., \mathfrak{u}(x_N)). \tag{1.8}$$

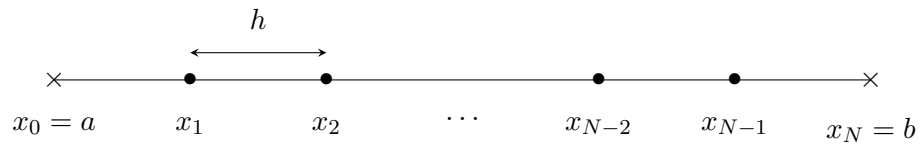


FIGURE 1.1 – Grille différences finies en dimension 1. Les symboles $\times$ désignent les points de bord, les symboles $\bullet$ désignent les points intérieurs.

De plus, l^2_h désigne l'espace des fonctions de grille, $h > 0$ fixé. On munit cet espace du produit scalaire $(\cdot,\cdot)_h$ et de la norme associée :

$$(\mathfrak{u},\mathfrak{v})_h = h\sum_{j=0}^{N}\mathfrak{u}(x_j)\bar{\mathfrak{v}}(x_j),\ |\mathfrak{u}|_h^2 = h\sum_{j=0}^{N}|\mathfrak{u}(x_j)|^2. \tag{1.9}$$

On définit aussi la norme $\|\cdot\|_\infty$ pour les fonctions de grille :

$$\|\mathfrak{u}\|_\infty = \max_{0\le j\le N}|\mathfrak{u}(x_j)|. \tag{1.10}$$

On notera

$$\mathfrak{u}_j = \mathfrak{u}(x_j) \text{ pour tout } 0 \le j \le N. \tag{1.11}$$

On note $l^2_{h,\text{pér}}$ l'espace des fonctions de grilles périodiques. Si $\mathfrak{u} \in l^2_{h,\text{pér}}$ alors $\mathfrak{u}(x_0) = \mathfrak{u}(x_N)$ et on a

$$\mathfrak{u} = (\mathfrak{u}(x_0), \mathfrak{u}(x_1), ..., \mathfrak{u}(x_{N-1})). \tag{1.12}$$

Le produit scalaire et la norme associée dans $l^2_{h,\text{pér}}$ sont

$$(\mathfrak{u},\mathfrak{v})_{h,\text{pér}} = h\sum_{j=0}^{N-1}\mathfrak{u}(x_j)\bar{\mathfrak{v}}(x_j),\ |\mathfrak{u}|^2_{h,\text{pér}} = h\sum_{j=0}^{N-1}|\mathfrak{u}(x_j)|^2 \text{ avec } \mathfrak{u},\mathfrak{v} \in l^2_{h,\text{pér}}. \tag{1.13}$$

Pour les fonctions de grille périodiques, on note :

$$\|\mathfrak{u}\|_\infty = \max_{0\le j\le N-1}|\mathfrak{u}(x_j)|. \tag{1.14}$$

2. Les lettres latines capitales désignent les vecteurs de $\mathbb{R}^{N+1}$ et les matrices de $\mathbb{M}_{N+1}(\mathbb{R})$. Par exemple, le vecteur $U \in \mathbb{R}^{N+1}$ associé à la fonction de grille $\mathfrak{u} \in l^2_h$ est

$$U = \begin{bmatrix}\mathfrak{u}_0\\ \mathfrak{u}_1\\ \vdots\\ \mathfrak{u}_N\end{bmatrix} = \begin{bmatrix}\mathfrak{u}(x_0)\\ \mathfrak{u}(x_1)\\ \vdots\\ \mathfrak{u}(x_N)\end{bmatrix}. \tag{1.15}$$

La norme euclidienne sur $\mathbb{R}^{N+1}$ est notée $|U|$. Elle induit une norme pour les matrices $A \in \mathbb{M}_{N+1}(\mathbb{R})$ définie par

$$|A|_2 = \sup_{U\neq 0}\frac{|AU|}{|U|}. \tag{1.16}$$

Si A est symétrique, l'identité suivante est vérifiée

$$|A|_2 = \rho(A) = \max\{|\lambda| \text{ tels que } \lambda \in \text{Sp}(A)\}, \tag{1.17}$$

$\rho(A)$ est le *rayon spectral* de A. De plus, on note également

$$|U|_\infty = \max_{1\leq j\leq N+1} |U_j|. \tag{1.18}$$

La norme sur $\mathbb{M}_{N+1}(\mathbb{R})$ subordonnée à $|\cdot|_\infty$ est définie pour $A = (a_{i,j})_{1\leq i,j\leq N+1} \in \mathbb{M}_{N+1}(\mathbb{R})$ par

$$|A|_\infty = \sup_{U\neq 0} \frac{|AU|_\infty}{|U|_\infty} = \max_{1\leq i\leq N+1} \sum_{j=1}^{N+1} |a_{i,j}|. \tag{1.19}$$

3. Soit $u : x \in \Omega \mapsto u(x)$, on définit la *fonction de grille* u^* associée à u par :

$$u_j^* = u^*(x_j) \text{ pour } 0 \leq j \leq N, \tag{1.20}$$

u^* est la restriction de u aux points de la grille. Si u est une fonction $L-$périodique, alors u^* est définie par

$$u_j^* = u^*(x_j) \text{ pour } 0 \leq j \leq N-1. \tag{1.21}$$

Nous distinguons l_h^2, l'espace des fonctions de grilles d'une part, et d'autre part l'espace vectoriel $\mathbb{R}^{N+1}$ même si ces deux espaces sont isomorphes.

Cette distinction permet de faire une claire différence entre :

- les opérateurs aux différences finies, qui agissent sur les fonctions de grille,
- les matrices, qui agissent sur les vecteurs.

Les fonctions de grilles contiennent toutes les échelles nécessaires dans le contexte physique alors que les vecteurs sont sans dimension. De plus, le raisonnement au niveau discret est plus naturel avec les fonctions de grilles. Il s'effectue d'une façon abstraite à l'aide d'opérateurs aux différences. En revanche, le codage est effectué dans le cadre de l'espace vectoriel $\mathbb{R}^{N+1}$.

On note que plusieurs normes infinies ont été définies :

- l'écriture $\|\cdot\|_\infty$ désigne à la fois la norme pour une fonction u définie sur Ω ou une fonction de grille $\mathfrak{u}$. Le contexte permettra de distinguer les deux cas de figure.
- La notation $|\cdot|_\infty$ désigne la norme d'une matrice A ou d'un vecteur U. La distinction se fera en fonction du contexte.

1.1.2 Transformée de Fourier discrète

Quitte à opérer une translation sur x, on peut supposer $a = 0$ et $b = L$. Le pas du maillage est $h = \frac{L}{N}$. Soit, pour tout k vérifiant $-N/2+1 \leq k \leq N/2$, la fonction $u^k : x \mapsto u^k(x) \in \mathbb{C}$ périodique de période L définie par

$$u^k(x) = \frac{1}{\sqrt{L}} \exp\left(\frac{2i\pi kx}{L}\right). \tag{1.22}$$

Les fonctions $(u^k)_{-N/2+1\leq k\leq N/2}$ forment une famille libre et orthonormée de $N+1$ fonctions. C'est à dire

$$(u^k, u^{k'})_{L^2([a,b])} = \delta_{k,k'} \text{ avec } -N/2+1 \leq k, k' \leq N/2. \tag{1.23}$$

On définit les *fonctions de base* $\mathfrak{u}^k$ de $l^2_{h,\text{pér}}$ par

$$\mathfrak{u}^k = \sqrt{h}(u^k)^* \tag{1.24}$$

donc

$$\mathfrak{u}_j^k = \sqrt{h}u^k(x_j) = \frac{1}{\sqrt{N}} \exp\left(\frac{2i\pi jkh}{L}\right) = \frac{1}{\sqrt{N}} \exp\left(\frac{2i\pi jk}{N}\right) \text{ avec } 0 \leq j \leq N-1. \tag{1.25}$$

Proposition 1.1. *Les fonctions* $(\mathfrak{u}^k)_{-N/2+1\leq k\leq N/2}$ *forment une base orthonormée de* $l^2_{h,pér}$.

Démonstration. On montre que $(\mathfrak{u}^k)_{-N/2+1\leq k\leq N/2}$ satisfait

$$(\mathfrak{u}^k,\mathfrak{u}^{k'})_{h,\text{pér}}=\delta_{k,k'} \tag{1.26}$$

pour tous $-N/2+1\leq k,k'\leq N/2$.

Considérons d'abord k et k' sont deux entiers distincts tels que $-N/2+1\leq k,k'\leq N/2$. $\mathfrak{u}^k,\mathfrak{u}^{k'}\in l^2_{h,\text{pér}}$ et

$$\begin{aligned}(\mathfrak{u}^k,\mathfrak{u}^{k'})_{h,\text{pér}}&=\frac{1}{N}\sum_{j=0}^{N-1}\exp\left(\frac{2i\pi jk}{N}\right)\exp\left(-\frac{2i\pi jk'}{N}\right)\\&=\frac{1}{N}\sum_{j=0}^{N-1}\exp\left(ij(k-k')\frac{2\pi}{N}\right)\\&=\frac{1}{N}\frac{1-\exp\left(i2\pi(k-k')\right)}{1-\exp\left(i(k-k')\frac{2\pi}{N}\right)}\\&=0.\end{aligned}$$

De plus, si $k=k'$, on a :

$$\begin{aligned}(\mathfrak{u}^k,\mathfrak{u}^{k'})_{h,\text{pér}}&=\frac{1}{N}\sum_{j=0}^{N-1}\exp\left(\frac{2i\pi jk}{N}\right)\exp\left(-\frac{2i\pi jk'}{N}\right)\\&=\frac{1}{N}\sum_{j=0}^{N-1}1\\&=1,\end{aligned}$$

d'où le résultat. □

Pour tout $\mathfrak{v}\in l^2_{h,\text{pér}}$, on note $(\hat{\mathfrak{v}}_k)_{-N/2\leq k\leq N/2}$ les composantes de $\mathfrak{v}$ sur la base $(\mathfrak{u}^k)_{-N/2+1\leq k\leq N/2}$:

$$\mathfrak{v}=\sum_{-N/2+1}^{N/2}\hat{\mathfrak{v}}_k\mathfrak{u}^k. \tag{1.27}$$

En effectuant le produit scalaire par $\mathfrak{u}^{k'}$ on obtient

$$\begin{aligned}(\mathfrak{v},\mathfrak{u}^{k'})_{h,\text{pér}}&=\sum_{k=-N/2+1}^{N/2}\hat{\mathfrak{v}}_k\underbrace{(\mathfrak{u}^k,\mathfrak{u}^{k'})_{h,\text{pér}}}_{\delta_{k,k'}}\\&=\hat{\mathfrak{v}}_{k'}.\end{aligned}$$

Définition 1.1. *On définit la transformée de Fourier discrète de* $\mathfrak{v}$ *par* $(\hat{\mathfrak{v}}_k)_{-N/2+1\leq k\leq N/2}$ *où*

$$\hat{\mathfrak{v}}_k=(\mathfrak{v},\mathfrak{u}^k)_{h,pér}. \tag{1.28}$$

Proposition 1.2. ***(Relation de Parseval discrète)*** *Pour tout* $\mathfrak{v}\in l^2_{h,pér}$, *on a*

$$|\mathfrak{v}|^2_{h,pér}=\sum_{k=-N/2+1}^{N/2}|\hat{\mathfrak{v}}_k|^2 \tag{1.29}$$

Démonstration. On sait que

$$|\mathfrak{v}|^2_{h,\text{pér}} = \sum_{j=0}^{N-1} \mathfrak{v}_j \bar{\mathfrak{v}}_j. \tag{1.30}$$

Or par décomposition, on a

$$\begin{aligned} \mathfrak{v}_j &= \sum_{k=-N/2+1}^{N/2} \hat{v}_k \mathfrak{u}_j^k \\ \bar{\mathfrak{v}}_j &= \sum_{k'=-N/2+1}^{N/2} \bar{\hat{\mathfrak{v}}}_{k'} \bar{\mathfrak{u}}_j^{k'}. \end{aligned} \tag{1.31}$$

Alors, on obtient

$$\begin{aligned} |\mathfrak{v}|^2_{h,\text{pér}} &= h \sum_{j=0}^{N-1} \sum_{k=-N/2+1}^{N/2} \sum_{k'=-N/2+1}^{N/2} \hat{\mathfrak{v}}_j \mathfrak{u}_j^k \bar{\hat{\mathfrak{v}}}_j \bar{\mathfrak{u}}_j^{k'} \\ &= h \sum_{k=-N/2+1}^{N/2} \sum_{k'=-N/2+1}^{N/2} \hat{\mathfrak{v}}_k \bar{\hat{\mathfrak{v}}}_{k'} \sum_{j=0}^{N-1} \mathfrak{u}_j^k \bar{\mathfrak{u}}_j^{k'} \\ &= \sum_{k=-N/2+1}^{N/2} \sum_{k'=-N/2+1}^{N/2} \hat{\mathfrak{v}}_k \bar{\hat{\mathfrak{v}}}_{k'} (\mathfrak{u}^k, \mathfrak{u}^{k'})_{h,\text{pér}}. \end{aligned}$$

Or on sait que $(\mathfrak{u}^k, \mathfrak{u}^{k'})_{h,\text{pér}} = \delta_{k,k'}$ donc

$$\begin{aligned} |\mathfrak{v}|^2_{h,\text{pér}} &= \sum_{k=-N/2+1}^{N/2} \sum_{k'=-N/2+1}^{N/2} \hat{\mathfrak{v}}_k \bar{\hat{\mathfrak{v}}}_{k'} \delta_{k,k'} \\ &= \sum_{k=-N/2+1}^{N/2} \hat{\mathfrak{v}}_k \bar{\hat{\mathfrak{v}}}_k \\ &= \sum_{k=-N/2+1}^{N/2} |\hat{\mathfrak{v}}_k|^2. \end{aligned}$$

Ce qui conclut la démonstration. □

1.1.3 Opérateur de translation périodique

On se donne $\mathfrak{u} \in l^2_{h,\text{pér}}$ une fonction de grille périodique.

Définition 1.2. *L'opérateur τ_p, $p \in \mathbb{Z}$, est défini, pour $\mathfrak{u}$ fonction de grille périodique, par*

$$(\tau_p \mathfrak{u})_j = \mathfrak{u}_{j+p} \text{ avec } 0 \leq j \leq N-1. \tag{1.32}$$

L'opérateur linéaire τ_p agit sur les fonctions périodiques $u : \mathbb{R} \mapsto u(x) \in \mathbb{R}$ par :

$$(\tau_p u)^*_j = \tau_p u(x_j) = u(x_{j+p}) = u^*_{j+p}. \tag{1.33}$$

En particulier, lorsque $p = 1$, on note τ l'*opérateur de translation à droite* :

$$\tau = \tau_1. \tag{1.34}$$

De plus, il est clair que l'on a

$$\begin{array}{rcl} \text{i) } \tau^0 & = & \mathrm{Id} \\ \text{ii) } \tau^p & = & \underbrace{\tau \circ \tau \circ \tau \circ \cdots \circ \tau}_{p \text{ fois.}} = \tau_p. \end{array} \tag{1.35}$$

En particulier, $\tau^N = \tau_N = \mathrm{Id}$, donc τ est inversible et

$$\tau^{-1} = \tau^{N-1}. \tag{1.36}$$

L'analyse des opérateurs périodiques repose sur la diagonalisation de τ. C'est l'objet de la proposition suivante. On note

$$\omega = \exp\left[\frac{2i\pi}{N}\right] \tag{1.37}$$

ainsi que

$$\omega^k = \exp\left[\frac{2ik\pi}{N}\right] \tag{1.38}$$

pour $-N/ + 1 \leq k \leq N/2$, les racines $N-$ièmes de l'unité.

Proposition 1.3.

- *Les valeurs propres de τ sont $\omega^k \in \mathbb{C}$, $-N/2 + 1 \leq k \leq N/2$,*
- *L'espace propre associé à ω^k est Vect$(\mathfrak{u}^k)$. En particulier, τ est diagonalisable et sa décomposition spectrale est*

$$\tau = \sum_{k=-N/2+1}^{N/2} \omega^k p_k \tag{1.39}$$

où p_k est le projecteur orthogonal sur Vect$(\mathfrak{u}^k)$. Pour tout $\mathfrak{v} \in l^2_{h,pér}$ et $-N/2 + 1 \leq k \leq N/2$, on a

$$p_k \mathfrak{v} = (\mathfrak{v}, \mathfrak{u}^k)_{h,pér} \mathfrak{u}^k. \tag{1.40}$$

Démonstration. Soient j, k tels que $0 \leq j \leq N-1$ et $-\frac{N}{2} + 1 \leq k \leq \frac{N}{2}$. On a

$$\begin{aligned} (\tau \mathfrak{u}^k)_j = \mathfrak{u}^k_{j+1} &= \frac{1}{\sqrt{N}} \exp\left[\frac{2i(j+1)\pi k}{N}\right] \\ &= \frac{1}{\sqrt{N}} \exp\left[\frac{2ij\pi k}{N}\right] \exp\left[\frac{2i\pi k}{N}\right] \\ &= \omega^k \mathfrak{u}^k_j. \end{aligned}$$

L'opérateur τ possède N valeurs propres distinctes, il est diagonalisable et sa décomposition spectrale s'obtient par la formule

$$\tau = \sum_{k=-N/2+1}^{N/2} \omega^k p_k.$$

□

Soit $P \in \mathbb{C}[X]$ un polynôme. Les valeurs propres et les fonctions propres de $P(\tau)$ sont données par la proposition suivante :

Proposition 1.4.

- *Les valeurs propres de $P(\tau)$ sont $P(\omega^k) \in \mathbb{C}$, $-N/2 + 1 \leq k \leq N/2$,*
- *L'espace propre associé à $P(\omega^k)$ est Vect$(\mathfrak{u}^k)$. $P(\tau)$ est diagonalisable, sa décomposition spectrale est*

$$P(\tau) = \sum_{k=-N/2+1}^{N/2} P(\omega^k) p_k \tag{1.41}$$

où p_k est le projecteur donné par (1.40).

Démonstration. P est un polynôme de $\mathbb{C}[X]$ donc il existe un nombre fini d'éléments de $\mathbb{C}$ notés a_0, a_1, a_2, ... tels que

$$P(X) = \sum_n a_n X^n. \tag{1.42}$$

Soient j, k tels que $0 \leq j \leq N-1$ et $-\frac{N}{2}+1 \leq k \leq \frac{N}{2}$. On a

$$\begin{aligned}(P(\tau)\mathfrak{u}^k)_j &= \left(\sum_n a_n \tau^n \mathfrak{u}^k\right)_j \\ &= \sum_n a_n \mathfrak{u}^k_{j+n}.\end{aligned}$$

Or par construction de $\mathfrak{u}^k$, on a

$$\begin{aligned}(P(\tau)\mathfrak{u}^k)_j &= \sum_n a_n (\omega^k)^n \mathfrak{u}^k_j \\ &= P(\omega^k)\mathfrak{u}^k_j,\end{aligned}$$

donc $P(\omega^k)$ est valeur propre de $P(\tau)$. De plus, les espaces vectoriels $\mathrm{Vect}(\mathfrak{u}^k)$ sont en somme directe, donc $P(\tau)$ est diagonalisable et la formule est (1.41) vérifiée. □

Remarque 1.1. *Noter que cette proposition est vraie non seulement pour τ mais aussi pour tout opérateur diagonalisable.*

On définit vec_1 l'opérateur appliquant un élément de $l^2_{h,\text{pér}}$ sur un vecteur de $\mathbb{R}^N$:

Définition 1.3. *Soient $\mathfrak{u} \in l^2_{h,pér}$ une fonction de grille et $(\mathbf{e}_j)_{1\leq j\leq N}$ la base canonique de $\mathbb{R}^N$. On définit l'opérateur vec_1 par :*

$$\begin{array}{rccc} vec_1 : l^2_{h,pér} & \to & \mathbb{R}^N \\ \mathfrak{u} & \mapsto & vec_1(\mathfrak{u}) \end{array} \tag{1.43}$$

avec

$$vec_1(\mathfrak{u}) = \sum_{j=1}^{N} \mathfrak{u}_{j-1}\mathbf{e}_j \tag{1.44}$$

Lorsque cela ne porte à aucune confusion, nous noterons vec au lieu de vec_1. La distinction se fera aisément en fonction du contexte.

Pour toute fonction de grille $\mathfrak{u} \in l^2_{h,\text{pér}}$ on note

$$U = \mathrm{vec}(\mathfrak{u}) = \mathrm{vec}_1(\mathfrak{u}) = \begin{bmatrix} \mathfrak{u}_0 \\ \mathfrak{u}_1 \\ \vdots \\ \mathfrak{u}_{N-1} \end{bmatrix} \in \mathbb{R}^N \tag{1.45}$$

On s'intéresse à présent à l'interprétation matricielle de τ. On note $T \in \mathbb{M}_N(\mathbb{R})$ la matrice donnée par

$$T = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & (0) & \\ & & \ddots & \ddots & \\ & (0) & & 0 & 1 \\ 1 & & & & 0 \end{bmatrix}. \tag{1.46}$$

La matrice T agit sur un vecteur $U = \begin{bmatrix} U_1 & U_2 & \cdots & U_N \end{bmatrix}^T \in \mathbb{R}^N$ par

$$(TU)_j = U_{j+1} \text{ avec } 1 \leq j \leq N. \tag{1.47}$$

C'est à dire

$$\mathrm{vec}(\tau \mathfrak{u}) = T(\mathrm{vec}(\mathfrak{u})). \tag{1.48}$$

Les propriétés spectrales de T se déduisent de celles de τ :

Corollaire 1.1. • *Les valeurs propres de T sont les valeurs $(\omega^k)_{-N/2+1\leq k\leq N/2}$. L'espace vectoriel associé est Vect(U^k) avec*

$$U^k = vec(\mathfrak{u}^k). \tag{1.49}$$

$\mathfrak{u}^k$ et ω sont donnés par la proposition 1.3.

- *Si $P \in \mathbb{R}[X]$ alors les valeurs propres de $P(T)$ sont $P(\omega^k)$ avec $-N/2+1 \leq k \leq N/2$ et l'espace propre associé est* (1.49).

Les vecteurs $(U^k)_{-N/2+1\leq k\leq N/2}$ forment une base orthonormée de $\mathbb{R}^N$ pour le produit scalaire usuel.

1.1.4 Opérateurs aux différences discrets

Dans cette section on s'intéresse à une approximation symétrique de $\partial_x : u \in \mathcal{C}^1 \mapsto \partial_x u \in \mathcal{C}^0$ de la forme

$$\delta_{2J,x} = \sum_{p=1}^{J} a_p \frac{\tau_p - \tau_{-p}}{2ph} \tag{1.50}$$

C'est le type d'opérateur dont nous aurons besoin dans la suite pour le calcul approché d'opérateurs sphériques.

Cette question est classique en analyse numérique. Elle est abordée dans des références classiques telles que [4, 18, 55]. En mathématiques, on cite en particulier [18, 55] ou en mécanique des fluides numériques [48].

Nous commençons par rappeler la *consistance* d'un opérateur différentiel [76]. On s'intéresse à l'équation aux dérivées partielles

$$\mathcal{L}u = f \tag{1.51}$$

où f est une fonction donnée. L'équation est discrétisée en

$$L_h \mathbf{u} = f^*. \tag{1.52}$$

La consistance est définie par

Définition 1.4. *le couple (L_h, R_h) est consistant avec $\mathcal{L}$ à l'ordre α si pour toute fonction u régulière, on a*

$$L_h(u^*) - R_h(\mathcal{L}u) = \mathcal{O}(h^\alpha) \tag{1.53}$$

avec R_h un opérateur d'interpolation approchant l'identité. $L_h(u^*) - R_h(\mathcal{L}u)$ est l'erreur de troncature.

- **Exemple 1 :** L'opérateur aux différences centré usuel est

$$\delta_x = \frac{\tau_1 - \tau_{-1}}{2h}. \tag{1.54}$$

Appliqué à la fonction de grille $\mathfrak{u}$, cet opérateur s'exprime par

$$\delta_x \mathfrak{u}_j = \frac{\mathfrak{u}_{j+1} - \mathfrak{u}_{j-1}}{2h} \text{ pour } 0 \leq j \leq N-1. \tag{1.55}$$

Il s'agit d'un opérateur qui approche la dérivée première. On a :

Proposition 1.5. *Soit $u : x \in \Omega \mapsto u(x) \in \mathbb{R}$ et u^* la fonction de grille correspondante. Si $u \in \mathcal{C}^3(\Omega)$ alors*

$$\delta_x u_j^* - \partial_x u(x_j) = \frac{h^2}{6}\partial_x^{(3)} u(\alpha) \text{ avec } \alpha \in]x_{j-1}, x_{j+1}[, \tag{1.56}$$

Démonstration. Pour u de classe $\mathcal{C}^3$, on considère les développements de Taylor-Lagrange :

$$u(x_j + h) = u(x_j) + h\partial_x u(x_j) + \frac{h^2}{2}\partial_x^{(2)} u(x_j) + \frac{h^3}{6}\partial_x^{(3)} u(\eta) \text{ avec } \eta \in]x_j, x_{j+1}[\tag{1.57}$$

et également :

$$u(x_j - h) = u(x_j) - h\partial_x u(x_j) + \frac{h^2}{2}\partial_x^{(2)} u(x_j) - \frac{h^3}{6}\partial_x^{(3)} u(\xi) \text{ avec } \xi \in]x_{j-1}, x_j[. \tag{1.58}$$

Alors par différence, on obtient :

$$\delta_x u_j^* = \partial_x u(x_j) + \frac{h^2}{12}\left[\partial_x^{(3)} u(\xi) + \partial_x^{(3)} u(\eta)\right] \text{ avec } \xi, \eta \in]x_{j-1}, x_{j+1}[. \tag{1.59}$$

Comme u est de classe $\mathcal{C}^3$, $\partial_x^{(3)} u$ est continue. Donc il existe $\alpha \in]\xi, \eta[\subset]x_{j-1}, x_{j+1}[$ tel que

$$2\partial_x^{(3)} u(\alpha) = \partial_x^{(3)} u(\xi) + \partial_x^{(3)} u(\eta) \tag{1.60}$$

d'où on obtient

$$\delta_x u_j^* - \partial_x u(x_j) = \frac{h^2}{6}\partial_x^{(3)} u(\alpha) \text{ avec } \alpha \in]x_{j-1}, x_{j+1}[. \tag{1.61}$$

□

Il résulte de la proposition 1.5 que (1.54) est consistant à l'ordre 2 au sens de la définition 1.4 avec $R_h = \mathrm{Id}$.

- **Exemple 2 :** Une manière d'augmenter la consistance de l'exemple 1 est de modifier l'opérateur R_h [18]. Pour cela on introduit σ_x l'opérateur de Simpson

$$\sigma_x = \frac{1}{6}\tau + \frac{4}{6}\mathrm{Id} + \frac{1}{6}\tau^{-1}. \tag{1.62}$$

Si $\mathfrak{u} \in l^2_{h,\text{pér}}$ est une fonction de grille alors

$$\sigma_x \mathfrak{u}_j = \frac{1}{6}\mathfrak{u}_{j+1} + \frac{4}{6}\mathfrak{u}_j + \frac{1}{6}\mathfrak{u}_{j-1}. \tag{1.63}$$

La proposition suivante est alors vérifiée :

Proposition 1.6. *Soit $u : x \in \Omega \mapsto u(x) \in \mathbb{R}$ et u^* la fonction de grille associée. Si $u \in \mathcal{C}^5(\Omega)$ alors*

$$\|\sigma_x \partial_x u_j^* - \delta_x u_j^*\|_\infty \leq \frac{h^4}{45}\|\partial_x^{(5)} u\|_\infty. \tag{1.64}$$

Démonstration. Soit f une fonction régulière de Ω dans $\mathbb{R}$. Pour $0 \leq j \leq N-1$, on pose

$$\mathfrak{e}_j = \frac{h}{3}\left[f(x_{j-1}) + 4f(x_j) + f(x_{j+1})\right] - \int_{x_{j-1}}^{x_{j-1}} f(x)dx = 2h\sigma_x f_j^* - \int_{x_{j-1}}^{x_{j-1}} f(x)dx. \tag{1.65}$$

Il s'agit de l'erreur de quadrature de la formule de Simpson. D'une part, par développement de Taylor, on a

$$2h\sigma_x f_j^* = 2hf(x_j) + \frac{h}{3}\partial_x^{(2)} f(x_j) + \frac{h^5}{72}\left(\partial_x^{(4)} f(\xi) + \partial_x^{(4)} f(\eta)\right) \tag{1.66}$$

avec $\xi \in]x_{j-1}, x_j[$ et $\eta \in]x_j, x_{j+1}[$. D'autre part, au voisinage $]x_{j-1}, x_{j+1}[$ de x_j et en posant $t = x - x_j$, on a

$$f(x) = f(x_j) + t\partial_x f(x_j) + \frac{t^2}{2}\partial_x^{(2)} f(x_j) + \frac{t^3}{6}\partial_x^{(3)} f(x_j) + \frac{t^4}{24}\partial_x^{(4)} f(\omega) \tag{1.67}$$

avec $\omega \in]x_{j-1}, x_{j+1}[$. On intègre cette dernière équation d'où

$$\begin{aligned}\int_{x_{i-1}}^{x_{i+1}} f(x)dx &= \int_{-h}^{h}\left(f(x_j) + t\partial_x f(x_j) + \frac{t^2}{2}\partial_x^{(2)} f(x_j) + \frac{t^3}{6}\partial_x^{(3)} f(x_j) + \frac{t^4}{24}\partial_x^{(4)} f(\omega)\right)dt \\ &= 2hf(x_j) + \frac{h}{3}\partial_x^{(2)} f(x_j) + \frac{h^5}{60}\partial_x^{(4)} f(\omega).\end{aligned}$$

Alors, par comparaison, on trouve

$$\begin{aligned}|\mathfrak{e}_j| &= |\frac{h^5}{72}\left(\partial_x^{(4)} f(\xi) + \partial_x^{(4)} f(\eta)\right) - \frac{h^5}{60}\partial_x^{(4)} f(\omega)| \\ &\leq \frac{2h^5}{45}\max_x |\partial_x^{(4)} f(x)|.\end{aligned}$$

Pour conclure, on applique la formule précédente avec $f = \partial_x u$. On note que

$$\int_{x_{i-1}}^{x_{i+1}} f(x)dx = u(x_{i+1}) - u(x_{i-1}), \tag{1.68}$$

de là, on obtient

$$|\sigma\left(\partial_x u^*\right)_j - \delta_x u_j^*| \leq \frac{h^5}{45}\max_x |\partial_x^{(4)} f(x)| \tag{1.69}$$

et on peut conclure directement. □

De la proposition 1.6, il découle que le couple $(L_h, R_h) = (\delta_x, \sigma_x)$ est consistant avec la dérivée première à l'ordre 4.

On s'intéresse à présent aux opérateurs centrés et antisymétriques dont le stencil est composé de $2J+1$ points de la forme (1.50). L'intérêt de cette famille d'opérateurs réside en l'absence de dissipation numérique. On choisit dans ce travail de maximiser l'ordre de consistance de ces opérateurs mais d'autres options sont possibles. On citera en particulier les schémas peu dispersifs [12] ou les schémas du type "dispersion relation preserving" [77].

Théorème 1.1. *Soit $f : \Omega \to \mathbb{R}$ alors l'identité discrète*

$$\delta_{2J,x}\mathfrak{u} = f^* \tag{1.70}$$

est consistante à l'ordre $2J$ avec

$$\partial_x u = f \tag{1.71}$$

(avec $R_h = Id$ et $L_h = \delta_{2J,x}$) si et seulement si les coefficients $(a_p)_{1\leq p\leq J}$ sont solutions du système

$$\begin{cases} \displaystyle\sum_{p=1}^{J} a_p &= 1 \\ \displaystyle\sum_{p=1}^{J} p^{2k} a_p &= 0 \text{ pour tous } 1 \leq k \leq J-1. \end{cases} \tag{1.72}$$

L'erreur de troncature est de la forme :

$$(\partial_x u)_j^* - \delta_{2J,x} u_j^* = h^{2J} \frac{2J}{2(2J+1)!} \left(\sum_{p=1}^{J} a_p p^{2J} \partial_x^{(2J+1)} u(\alpha_p) \right) \tag{1.73}$$

avec $\alpha_p \in]x_{j-p}, x_{j+p}[$ *et* $u \in \mathcal{C}^{2J+1}$.

Démonstration. Soit $u : x \in \Omega \mapsto u(x) \in \mathbb{R}$ une fonction de classe $\mathcal{C}^{2P+1}(\Omega)$ et u^* la fonction de grille correspondante.

On considère les développements de Taylor :

$$\begin{aligned} u(x_j + ph) &= u(x_j) + ph\partial_x u(x_j) + \cdots + \frac{(ph)^k}{k!}\partial_x^{(k)} u(x_j) + \cdots + \frac{(ph)^{2J+1}}{(2J+1)!}\partial_x^{(2J+1)} u(\xi_j) \\ u(x_j - ph) &= u(x_j) - ph\partial_x u(x_j) + \cdots + \frac{(-ph)^k}{k!}\partial_x^{(k)} u(x_j) + \cdots + \frac{(-ph)^{2J+1}}{(2J+1)!}\partial_x^{(2J+1)} u(\eta_j) \end{aligned} \tag{1.74}$$

avec $\xi_j \in]x_j, x_j + ph[$ et $\eta_j \in]x_j - ph, x_j[$. En combinant ces deux égalités, on obtient

$$\frac{\tau_p u_j^* - \tau_{-p} u_j^*}{2ph} = \partial_x u(x_j) + \cdots + \frac{(ph)^{k-1}(1-(-1)^k)}{2 \cdot k!}\partial_x^{(k)}(x_j) + \cdots + \frac{(ph)^{2J}}{2(2J+1)!}\left(\partial_x^{(2J+1)} u(\xi_j) + \partial_x^{(2J+1)} u(\eta_j)\right) \tag{1.75}$$

Donc la combinaison linéaire de ces termes pondérée par les coefficients $(a_p)_{1 \le p \le J}$ est consistante avec la dérivée première à l'ordre $2J$ si et seulement si :

$$\begin{cases} \displaystyle\sum_{p=1}^{J} a_p &= 1 \\ \displaystyle\sum_{p=1}^{J} p^{k-1} \frac{(1-(-1)^k)}{k!} a_p &= 0 \text{ pour tous } 1 \le k \le 2J-1. \end{cases} \tag{1.76}$$

La seconde égalité est vérifiée pour tout k pair. Ce système se simplifie en :

$$\begin{cases} \displaystyle\sum_{p=1}^{J} a_p &= 1 \\ \displaystyle\sum_{p=1}^{P} p^{2k} a_p &= 0 \text{ pour tout } 1 \le k \le J-1. \end{cases} \tag{1.77}$$

L'erreur de troncature prend la forme

$$\partial_x u(x_j) - \delta_{2J,x} u_j^* = h^{2J} \sum_{p=1}^{J} a_p \frac{p^{2J}}{2(2J+1)!} \left(\partial_x^{(2J+1)}(\xi_j) + \partial_x^{(2J+1)} u(\eta_j)\right). \tag{1.78}$$

pour conclure, on utilise le théorème des valeurs intermédiaires. Comme $u \in \mathcal{C}^{2J+1}$, il existe $\alpha_p \in]x_{j-p}, x_{j+p}[$ tel que

$$\partial_x^{(2J+1)}(\xi_j) + \partial_x^{(2J+1)} u(\eta_j) = 2\partial_x^{(2J+1)} u(\alpha_p). \tag{1.79}$$

Ce qui conclut la preuve. □

Corollaire 1.2. *Si les coefficients* $(a_p)_{1 \le p \le J}$ *sont solutions du système* (1.72) *alors il existe* $C > 0$ *indépendant de* h *et de* u *tel que pour toute fonction* u *régulière on a*

$$\|(\partial_x u)^* - \delta_{2J,x}(u^*)\|_\infty \le C h^{2J} \|(\partial_x^{(2J+1)} u)^*\|_\infty. \tag{1.80}$$

Proposition 1.7. *Le système* (1.72) *admet une unique solution. De plus, pour tout* $1 \leq p \leq J$*, on a*

$$a_p = (-1)^{p-1} \left(\frac{J!}{p}\right)^2 \frac{\displaystyle\prod_{1\leq i<j\leq J-1} (\beta_j^p - \beta_i^p)}{\displaystyle\prod_{1\leq i<j\leq J} (j^2 - i^2)}, \tag{1.81}$$

avec

$$\beta_j^p = \begin{cases} j^2 & si \quad j < p \\ (j+1)^2 & si \quad j \geq p. \end{cases} \tag{1.82}$$

Enfin, les coefficients sont de signes alternés, c'est à dire $a_p a_{p+1} \leq 0$ *pour tout* p*.*

Démonstration. Le système (1.72) s'écrit matriciellement

$$\underbrace{\begin{bmatrix} (1^2)^0 & (2^2)^0 & (3^2)^0 & (4^2)^0 & \cdots & (J^2)^0 \\ (1^2)^1 & (2^2)^1 & (3^2)^1 & (4^2)^1 & \cdots & (J^2)^1 \\ (1^2)^2 & (2^2)^2 & (3^2)^2 & (4^2)^2 & \cdots & (J^2)^2 \\ \vdots & & & \vdots & & \vdots \\ (1^2)^{J-1} & (2^2)^{J-1} & (3^2)^{J-1} & (4^2)^{J-1} & \cdots & (J^2)^{J-1} \end{bmatrix}}_{=A} \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_J \end{bmatrix}}_{=a} = \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{=e_1}. \tag{1.83}$$

Ce système admet une solution si $\det(A) \neq 0$. Or on reconnaît un déterminant de Vandermonde. Donc

$$\begin{aligned} \det(A) &= \mathrm{VDM}(\alpha_1, \cdots, \alpha_J) \text{ où VDM désigne un déterminant de Vandermonde [32]} \\ &= \prod_{1\leq i<j\leq J} (\alpha_j - \alpha_i) \text{ avec } \alpha_l = l^2 \\ &= \prod_{1\leq i<j\leq J} (j^2 - i^2) \\ &> 0 \text{ car } j^2 > i^2 \text{ lorsque } j > i. \end{aligned}$$

Le système (1.72) admet donc une unique solution. Calculons cette solution.

D'après la formule de Cramer, on a

$$a_p = \frac{\det(A_p)}{\det(A)} \tag{1.84}$$

avec

$$A_p = \begin{bmatrix} (1^2)^0 & (2^2)^0 & \cdots & ((p-1)^2)^0 & 1 & ((p+1)^2)^0 & \cdots & (J^2)^0 \\ (1^2)^1 & (2^2)^1 & \cdots & ((p-1)^2)^1 & 0 & ((p+1)^2)^1 & \cdots & (J^2)^1 \\ (1^2)^2 & (2^2)^2 & \cdots & ((p-1)^2)^2 & 0 & ((p+1)^2)^2 & \cdots & (J^2)^2 \\ \vdots & \vdots & & & \vdots & \vdots & & \vdots \\ (1^2)^{J-1} & (2^2)^{J-1} & \cdots & ((p-1)^2)^{J-1} & 0 & ((p+1)^2)^{J-1} & \cdots & (J^2)^{J-1} \end{bmatrix}. \tag{1.85}$$

En développant le déterminant de A_p par rapport à la $p-$ième colonne, on trouve

$$\begin{aligned}
\det(A_p) &= (-1)^{p-1}\begin{vmatrix} (1^2)^1 & (2^2)^1 & \cdots & ((p-1)^2)^1 & ((p+1)^2)^1 & \cdots & (J^2)^1 \\ (1^2)^2 & (2^2)^2 & \cdots & ((p-1)^2)^2 & ((p+1)^2)^2 & \cdots & (J^2)^2 \\ \vdots & \vdots & & & \vdots & & \vdots \\ (1^2)^{J-1} & (2^2)^{J-1} & \cdots & ((p-1)^2)^{J-1} & ((p+1)^2)^{J-1} & \cdots & (J^2)^{J-1} \end{vmatrix} \\
&= (-1)^{p-1}\left(\frac{J!}{p}\right)^2\begin{vmatrix} (1^2)^0 & (2^2)^0 & \cdots & ((p-1)^2)^0 & ((p+1)^2)^0 & \cdots & (J^2)^0 \\ (1^2)^1 & (2^2)^1 & \cdots & ((p-1)^2)^1 & ((p+1)^2)^1 & \cdots & (J^2)^1 \\ (1^2)^2 & (2^2)^2 & \cdots & ((p-1)^2)^2 & ((p+1)^2)^2 & \cdots & (J^2)^2 \\ \vdots & \vdots & & & \vdots & & \vdots \\ (1^2)^{J-2} & (2^2)^{J-2} & \cdots & ((p-1)^2)^{J-2} & ((p+1)^2)^{J-2} & \cdots & (J^2)^{J-2} \end{vmatrix} \\
&= (-1)^{p-1}\left(\frac{J!}{p}\right)^2 \mathrm{VDM}(1, 2^2, 3^2, \cdots, (p-1)^2, (p+1)^2, \cdots J^2) \\
&= (-1)^{p-1}\left(\frac{J!}{p}\right)^2 \prod_{1\le i<j\le J-1}(\beta_j^p - \beta_i^p)
\end{aligned}$$

avec

$$\beta_j^p = \begin{cases} j^2 & \text{si } j < p \\ (j+1)^2 & \text{si } p \ge j. \end{cases} \tag{1.86}$$

VDM désigne le déterminant de Vandermonde [32] déterminé par

$$\mathrm{VDM}(\beta_1^p, \cdots, \beta_n^p) = \prod_{1\le i<j\le n}(\beta_j^p - \beta_i^p). \tag{1.87}$$

On note que $\beta_j^p > \beta_i^p$ lorsque $j > i$, donc $\det(A_p)$ est du signe de $(-1)^{p-1}$. On trouve donc

$$a_p = (-1)^{p-1}\left(\frac{J!}{p}\right)^2 \frac{\displaystyle\prod_{1\le i<j\le J-1}(\beta_j^p - \beta_i^p)}{\displaystyle\prod_{1\le i<j\le J}(j^2 - i^2)}, \tag{1.88}$$

ainsi que l'alternance des signes car

$$\left(\frac{J!}{p}\right)^2 \frac{\displaystyle\prod_{1\le i<j\le J-1}(\beta_j^p - \beta_i^p)}{\displaystyle\prod_{1\le i<j\le J}(j^2 - i^2)} > 0, \tag{1.89}$$

donc $a_p a_{p+1} < 0$. Ce qui conclut la preuve. □

A présent, dans cette section, nous supposons que les coefficients $(a_p)_{1\le p\le J}$ sont issus de la résolution du système (1.72). On a déjà vu que $\tau^{-1} = \tau_{-1} = \tau^{N-1}$, donc à partir des coefficients $(a_p)_{1\le p\le J}$ on construit le polynôme Q_{2J} tel que

$$\begin{aligned}
\delta_{2J,x} &= \sum_{p=1}^{J} a_p \frac{\tau_p - \tau_{-p}}{2ph} \\
&= \sum_{p=1}^{J} \frac{a_p}{2ph}\left(\tau^p - \tau^{N-p}\right) \\
&= \frac{1}{h} Q_{2J}(\tau),
\end{aligned}$$

où le polynôme Q_{2J} de $\mathbb{R}_{N-1}[X]$ est donné par

$$Q_{2J}(X) = \sum_{p=1}^{J} \frac{a_p}{2p} \left(X^p - X^{N-p}\right). \tag{1.90}$$

Exemples : on donne quelques exemples d'opérateurs d'approximation de la dérivée première. Soit $u : x \in \Omega \mapsto u(x) \in \mathbb{R}$ une fonction régulière.

- Si $u \in \mathcal{C}^3(\Omega)$ alors l'opérateur :

$$\delta_{2,x} = \delta_x = \frac{\tau_1 - \tau_{-1}}{2h} = \frac{1}{h} Q_2(\tau) \tag{1.91}$$

  est un opérateur d'approximation de la dérivée première à l'ordre 2 ($J = 1$) avec

$$Q_2(X) = \frac{1}{2}(X - X^{N-1}). \tag{1.92}$$

- Si $u \in \mathcal{C}^5(\Omega)$ alors l'opérateur :

$$\delta_{4,x} = \frac{4}{3}\frac{\tau_1 - \tau_{-1}}{2h} - \frac{1}{3}\frac{\tau_2 - \tau_{-2}}{4h} = \frac{1}{h} Q_4(\tau) \tag{1.93}$$

  est un opérateur d'approximation de la dérivée première à l'ordre 4 ($J = 2$), avec

$$Q_4(X) = \frac{2}{3}(X - X^{N-1}) - \frac{1}{12}(X^2 - X^{N-2}). \tag{1.94}$$

- Si $u \in \mathcal{C}^7(\Omega)$ alors l'opérateur :

$$\delta_{6,x} = \frac{3}{2}\frac{\tau_1 - \tau_{-1}}{2h} - \frac{3}{5}\frac{\tau_2 - \tau_{-2}}{4h} + \frac{1}{10}\frac{\tau_3 - \tau_{-3}}{6h} = \frac{1}{h} Q_6(\tau) \tag{1.95}$$

  est un opérateur d'approximation de la dérivée première à l'ordre 6 ($J = 3$), avec

$$Q_6(X) = \frac{3}{4}(X - X^{N-1}) - \frac{3}{20}(X^2 - X^{N-2}) + \frac{1}{60}(X^3 - X^{N-3}). \tag{1.96}$$

- Si $u \in \mathcal{C}^9(\Omega)$ alors l'opérateur :

$$\delta_{8,x} = \frac{8}{5}\frac{\tau_1 - \tau_{-1}}{2h} - \frac{4}{5}\frac{\tau_2 - \tau_{-2}}{4h} + \frac{8}{35}\frac{\tau_3 - \tau_{-3}}{6h} - \frac{1}{35}\frac{\tau_4 - \tau_{-4}}{8h} = \frac{1}{h} Q_8(\tau) \tag{1.97}$$

  est un opérateur d'approximation de la dérivée première à l'ordre 8 ($J = 4$), avec

$$Q_8(X) = \frac{4}{5}(X - X^{N-1}) - \frac{1}{5}(X^2 - X^{N-2}) + \frac{4}{105}(X^3 - X^{N-3}) - \frac{1}{180}(X^4 - X^{N-4}). \tag{1.98}$$

L'opérateur $\delta_{2J,x}$ agit sur $l^2_{h,\text{pér}}$. On s'intéresse à présent aux propriétés spectrales de cet opérateur.

Proposition 1.8. *Les valeurs propres de $\delta_{2J,x}$ sont*

$$\frac{1}{h} Q_{2J}(\omega^k) \tag{1.99}$$

où ω^k est la $k-$ième racine de l'unité, avec $-N/2 + 1 \leq k \leq N/2$. L'espace propre associé à chaque valeur propre $\dfrac{1}{h} Q_{2J}(\omega^k)$ est Vect$(\mathfrak{u}^k)$.

Démonstration. Ce résultat se déduit immédiatement de la proposition 1.4 avec le polynôme $Q_{2J}(X)$. □

Proposition 1.9. *Soit $k \in \mathbb{N}$ fixé, on a*

$$\frac{2i\pi k}{N} - Q_{2J}\left(\exp\left(\frac{2i\pi k}{N}\right)\right) = \mathcal{O}(h^{2J+1}) \tag{1.100}$$

avec $h = L/N$. De plus, si on pose $\theta = 2k\pi/N$ alors

$$-iQ_{2J}\left(\exp(i\theta)\right) = \sin\theta \sum_{p=1}^{J} \frac{a_p}{p} U_p(\cos\theta) \in \mathbb{R}, \tag{1.101}$$

où U_p est un polynôme de Tchebychev de seconde espèce. De plus

$$Q_{2J}\left(\exp(i0)\right) = Q_{2J}\left(\exp(i\pi)\right) = 0. \tag{1.102}$$

Démonstration. D'une part, d'après le théorème 1.1 on a

$$(\partial_x u^k)^* - \delta_{2J,x}(u^k)^* = \mathcal{O}(h^{2J}). \tag{1.103}$$

D'autre part, on rappelle que la fonction u^k et la fonction de grille $\mathfrak{u}^k$ sont liées par

$$u^k = \sqrt{h}\mathfrak{u}^k. \tag{1.104}$$

Donc on a

$$\begin{aligned}
(\partial_x u^k)^* - \delta_{2J,x}(u^k)^* &= \frac{2i\pi k}{L}(u^k)^* - \frac{1}{h}Q_{2J}(\omega^k)(u^k)^* \\
&= \frac{1}{h}\left(\frac{2i\pi hk}{L} - Q_{2J}(\omega^k)\right)(u^k)^* \\
&= \frac{1}{h}\left(\frac{2i\pi k}{N} - Q_{2J}\left(\exp\left(\frac{2i\pi k}{N}\right)\right)\right)(u^k)^*.
\end{aligned}$$

La fonction u^k est bornée, donc

$$\frac{2i\pi k}{N} - Q_{2J}\left(\exp\left(\frac{2i\pi k}{N}\right)\right) = \mathcal{O}(h^{2J+1}). \tag{1.105}$$

De plus par construction on a

$$\begin{aligned}
Q_{2J}\left(\exp(i\theta)\right) &= \sum_{p=1}^{J} a_p \frac{\exp(ip\theta) - \exp(i(N-p)\theta)}{2p} \text{par périodicité des opérateurs.} \\
&= \sum_{p=1}^{J} a_p \frac{\exp(ip\theta) - \exp(-ip\theta)}{2p} \\
&= \sum_{p=1}^{J} a_p \frac{i\sin(p\theta)}{p} \\
&= i\sin\theta \sum_{p=1}^{J} \frac{a_p}{p} U_p(\cos\theta) \in i\mathbb{R}
\end{aligned}$$

où U_p désigne un polynôme de Tchebychev de seconde espèce. Enfin, on a

$$\begin{aligned} Q_{2J}\left(\exp(i0)\right) &= \sum_{p=1}^{J} a_p \frac{\exp(ip0) - \exp(i(N-p)0)}{2p} \\ &= \sum_{p=1}^{J} a_p \frac{1-1}{2p} \\ &= 0, \end{aligned}$$

ainsi que

$$\begin{aligned} Q_{2J}\left(\exp(i\pi)\right) &= i \sin \pi \sum_{p=1}^{J} \frac{a_p}{p} U_p(\cos \pi) \\ &= i0 \sum_{p=1}^{J} \frac{a_p}{p} U_p(-1) \\ &= 0, \end{aligned}$$

ce qui conclut la démonstration. □

La fonction

$$\theta \in \mathbb{R} \mapsto -iQ_{2J}(\exp(i\theta)) \in \mathbb{R} \tag{1.106}$$

est $2\pi-$périodique et impaire. De plus, en tenant compte de la proposition 1.9 on compare $\theta \mapsto \theta$ et $\theta \mapsto -iQ_{2J}(\exp(i\theta))$ sur l'intervalle $[0, \pi]$. On représente quelques exemples de ces fonctions pour différentes valeurs de $2J$ sur la Figure 1.2. On constate que plus l'ordre $2J$ est élevé, mieux $-iQ_{2J}(\exp(i\theta))$ approche θ pour θ proche de 0. De plus, les courbes coïncident en $\theta = 0$.

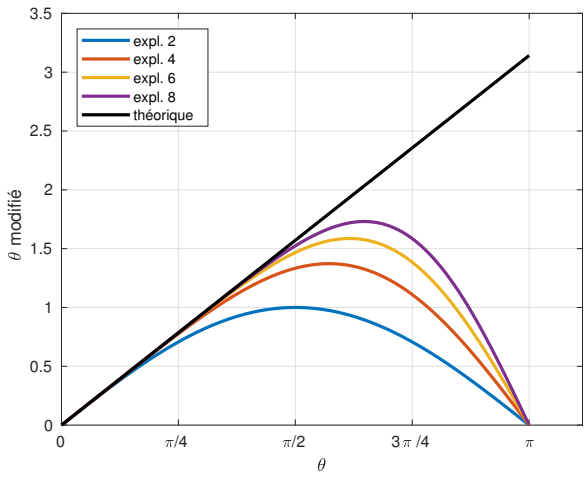


FIGURE 1.2 – Représentation de $-iQ_{2J}\left(\exp(i\theta)\right)$ en fonction de θ pour les schémas d'approximation explicites $\delta_{x,2J}$ d'ordres $2J = 2$, 4, 6 et 8.

On s'intéresse à présent à la version matricielle de l'opérateur $\delta_{2J,x}$. On définit $D_{2J} \in \mathbb{M}_N(\mathbb{R})$ la matrice associée à l'opérateur $\delta_{2J,x}$:

$$D_{2J} = \frac{1}{h} Q_{2J}(T). \tag{1.107}$$

Pour toute fonction de grille $\mathfrak{u} \in l^2_{h,\text{pér}}$, la relation suivante est vérifiée

$$\text{vec}(\delta_{2J,x}\mathfrak{u}) = D_{2J} \cdot \text{vec}(\mathfrak{u}). \tag{1.108}$$

- **Exemple 1 :** A l'ordre $2J = 2$, le schéma centré d'ordre 2 : $\delta_{2,x}$ est associée à la matrice

$$D_2 = \frac{1}{2h}\begin{bmatrix} 0 & 1 & & & & -1 \\ -1 & 0 & 1 & & (0) & \\ & -1 & 0 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & (0) & & -1 & 0 & 1 \\ 1 & & & & -1 & 0 \end{bmatrix} \tag{1.109}$$

- **Exemple 2 :** A l'ordre $2J = 4$, le schéma centré d'ordre 4 : $\delta_{4,x}$ est associée à la matrice

$$D_4 = \frac{1}{2h}\begin{bmatrix} 0 & 4/3 & -1/6 & & & & 1/6 & -4/3 \\ -4/3 & 0 & 4/3 & -1/6 & & & & 1/6 \\ 1/6 & -4/3 & 0 & 4/3 & -1/6 & & (0) & \\ & 1/6 & -4/3 & 0 & 4/3 & -1/6 & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & (0) & & 1/6 & -4/3 & 0 & 4/3 & -1/6 \\ -1/6 & & & & 1/6 & -4/3 & 0 & 4/3 \\ 4/3 & -1/6 & & & & 1/6 & -4/3 & 0 \end{bmatrix}. \tag{1.110}$$

Les propriétés spectrales de D_{2J} se déduisent de celles de $\delta_{2J,x}$ vues dans la proposition 1.8.

Proposition 1.10. *Les valeurs propres de D_{2J} sont les valeurs*

$$\frac{1}{h}Q_{2J}(\omega^k) \text{ avec } -N/2+1 \leq k \leq N/2, \tag{1.111}$$

l'espace propre associé est Vect(U^k) avec $U^k = vec(\mathfrak{u}^k)$.

De plus, on note la propriété de symétrie de D_{2J} :

Proposition 1.11. *La matrice D_{2J} est antisymétrique.*

Démonstration. Montrons que $D_{2J}^T = -D_{2J}$:

$$\begin{aligned} D_{2J}^T &= \frac{1}{h}Q_{2J}(T)^T \\ &= \frac{1}{h}\left(\sum_{p=1}^{J}\frac{a_j}{2p}(T^p - T^{N-p})\right)^T \\ &= \frac{1}{h}\sum_{p=1}^{P}\frac{a_p}{2p}((T^p)^T - (T^{N-p})^T) \\ &= \frac{1}{h}\sum_{p=1}^{J}\frac{a_p}{2p}(T^{N-p} - T^p) \\ &= -\frac{1}{h}\sum_{p=1}^{J}\frac{a_p}{2p}(T^p - T^{N-p}) \\ &= -D_{2J} \end{aligned}$$

d'où le résultat. □

1.1.5 Opérateurs Hermitiens périodiques 1D

Dans ce travail, nous n'utilisons pas de schémas de la forme (1.50). Les schémas hermitiens [60] donnent de meilleurs résultats en ce qui concerne l'analyse fréquentielle. Ils sont couramment utilisés sur la résolution d'équations hyperboliques [17] et permettent de bons résultats pour la capture des chocs [50]. De bons résultats ont aussi été donnés pour l'approximation de la dérivée seconde [1, 52] ou en mécanique des fluides [8, 9]. Dans ce travail, nous ne considérons que les schémas à 3 points implicites pour lesquels nous souhaitons optimiser l'ordre de convergence. D'autres travaux visent à minimiser d'autres formes d'erreurs [53, 54]. Nous définissons l'opérateur σ_x par :

$$(\sigma_x \mathfrak{u})_j = (1-2\beta)\mathfrak{u}_j + \beta\left(\mathfrak{u}_{j+1} + \mathfrak{u}_{j-1}\right) \text{ avec } 0 \leq j \leq N-1. \tag{1.112}$$

On considère l'opérateur δ_x de la forme (1.50), c'est à dire

$$\delta_x = \sum_{p=1}^{J} a_p \frac{\tau^p - \tau^{-p}}{2ph}. \tag{1.113}$$

Théorème 1.2. *Soit $f : \Omega \to \mathbb{R}$ alors l'identité discrète*

$$\delta_x \mathfrak{u} = f^* \tag{1.114}$$

est consistante à l'ordre $2J+2$ avec

$$\partial_x u = f \tag{1.115}$$

(avec $R_h = \sigma_x$ et $L_h = \delta_x$) si et seulement si les coefficients $(a_p)_{1\leq p\leq J}$ et β sont solutions du système

$$\left\{ \begin{array}{rcl} \displaystyle\sum_{p=1}^{J} a_p & = & 1 \\ \displaystyle\sum_{p=1}^{J} a_p \frac{p^{2n}}{2n+1} & = & 2\beta \text{ pour } n = 1, 2, ...J. \end{array} \right. \tag{1.116}$$

Si u est une fonction de $\mathcal{C}^{2J+3}$, alors pour tout $0 \leq j \leq N-1$, on a

$$(\delta_x u^*)_j - (\sigma_x(\partial_x u)^*)_j = h^{2J+2}\left[\frac{1}{(2J+3)!}\left(\sum_{p=1}^{J} a_p p^{2J+2}\partial_x^{(2J+3)}u(\theta_p)\right) - \left(\frac{2\beta}{(2J+2)!}\partial_x^{(2J+3)}u(\rho_p)\right)\right] \tag{1.117}$$

avec $\rho_p, \theta_p \in]x_{j-p}, x_{j+p}[$.

Démonstration. Soit $u : x \in \Omega \mapsto u(x) \in \mathbb{R}$ une fonction de classe $\mathcal{C}^{2J+3}(\Omega)$ et u^* la fonction de grille correspondante. On considère les développements de Taylor :

$$\begin{array}{rcl} u(x_j + ph) & = & u(x_j) + ph\partial_x u(x_j) + \cdots + \dfrac{(ph)^k}{k!}\partial_x^{(k)}u(x_j) + \cdots + \dfrac{(ph)^{2J+3}}{(2J+3)!}\partial_x^{(2J+3)}u(\xi) \\ u(x_j - ph) & = & u(x_j) - ph\partial_x u(x_j) + \cdots + \dfrac{(-ph)^k}{k!}\partial_x^{(k)}u(x_j) + \cdots + \dfrac{(-ph)^{2J+3}}{(2J+3)!}\partial_x^{(2J+3)}u(\eta) \end{array} \tag{1.118}$$

avec $\xi \in]x_j, x_j + ph[$ et $\eta \in]x_j - ph, x_j[$. En combinant ces deux égalités, on a

$$\frac{\tau_p u_j^* - \tau_{-p}u_j^*}{2ph} = \partial_x u(x_j) + \cdots + \frac{(ph)^{k-1}(1-(-1)^k)}{2\cdot k!}\partial_x^{(k)}u(x_j) + \cdots + \frac{(ph)^{2J+2}}{2(2J+3)!}\left(\partial_x^{(2J+3)}u(\xi) + \partial_x^{(2J+3)}u(\eta)\right). \tag{1.119}$$

D'autre part, on a

$$(1-2\beta)\partial_x u_j^* + \beta\left(\tau_1\partial_x u_j^* + \tau_{-1}\partial_x u_j^*\right) = \partial_x u_j^* + \sum_{k=1}^{2J+1} \beta\frac{h^k}{k!}\left(1+(-1)^k\right)\partial_x^{(k+1)}u(x_j) + \beta\frac{h^{2J+2}}{(2J+2)!}\left(\partial_x^{(2J+3)}u(\varrho_p) + \partial_x^{(2J+3)}u(\sigma_p)\right) \quad (1.120)$$

avec $\varrho_p \in]x_j, x_j + ph[$ et $\sigma_p \in]x_j - h, x_j[$.

On remarque directement que (1.120) et $\sum_{p=0}^{J} a_p$(1.119) coïncident pour les puissances de h impaires. Pour les autres valeurs, l'égalité est vrai si les coefficients $(a_p)_{1\leq p\leq J}$ et β sont solutions de (1.116). L'erreur de troncature prend directement la forme

$$(\delta_x u^*)_j - (\sigma_x\partial_x u^*)_j = h^{2J+2}\left(\sum_{p=1}^{J} a_p\frac{p^{2J+2}}{2(2J+3)!}\left(\partial_x^{(2J+3)}u(\xi_p) + \partial_x^{(2J+3)}u(\eta_p)\right) - \frac{\beta}{(2J+2)!}\left(\partial_x^{(2J+3)}u(\varrho_p) + \partial_x^{(2J+3)}(\sigma_p)\right)\right). \quad (1.121)$$

On conclut en utilisant le théorème des valeurs intermédiaires. □

Dans cette section, nous supposons que le système (1.116) admet une solution. Ce sera le cas pour tous les exemples considérés.

Proposition 1.12. *Les valeurs propres de σ_x sont*

$$R(\omega^k) \text{ avec } R(X) = (1-2\beta) + \beta(X + X^{N-1}) \in \mathbb{R}_{N-1}[X] \text{ et } -N/2+1 \leq k \leq N/2. \quad (1.122)$$

L'espace propre associé est Vect$(\mathfrak{u}^k)$.

Démonstration. Il s'agit d'une conséquence de la proposition 1.4 appliqué au polynôme R. □

Dans la suite, nous aurons besoin d'inverser l'opérateur σ_x. La propriété suivante permet de s'assurer de son inversibilité.

Corollaire 1.3. *L'opérateur σ_x est inversible si*

$$|\beta| < \frac{1}{2}. \quad (1.123)$$

Définition 1.5. *On suppose que β et $(a_p)_{1\leq p\leq J}$ sont solutions de (1.116) et que $|\beta| < 1/2$, on définit l'opérateur hermitien d'approximation de la dérivée première $\delta_{2J+2,x}^H$ par*

$$\delta_{2J+2,x}^H = \sigma_x^{-1}\circ\delta_x. \quad (1.124)$$

Théorème 1.3. *Il existe $C > 0$ indépendant de h tel que pour tout $u : x \in \Omega \mapsto u(x) \in \mathbb{R}$ fonction de classe $\mathcal{C}^{(2J+3)}$ on a*

$$\|(\partial_x u)^* - \delta_{2J+2,x}^H(u^*)\|_\infty \leq Ch^{2J+2}\|\partial_x^{(2J+3)}u\|_\infty. \quad (1.125)$$

Démonstration. Par calcul immédiat, on a :

$$\begin{aligned}\|(\partial_x u)^* - \delta_{2J+2,x}^H(u^*)\|_\infty &= \|\sigma_x^{-1}\circ(\sigma_x u'^* - \delta_x u^*)\|_\infty \\ &\leq \|\sigma_x^{-1}\|_\infty\|\sigma_x u'^* - \delta_x u^*\|_\infty \\ &\leq Ch^{2J+2}\|\partial_x^{(2J+3)}u\|_\infty\end{aligned} \quad (1.126)$$

en utilisant σ_x inversible et l'équation (1.117). □

Quelques exemples de schémas hermitiens sont donnés.
Exemples : Soit $u : x \in \Omega \mapsto u(x) \in \mathbb{R}$ périodique alors

- Si $u \in \mathcal{C}^5(\Omega)$ alors
$$\begin{cases} \sigma_x = \dfrac{4}{6}\mathrm{Id} + \dfrac{1}{6}\left(\tau_1 + \tau_{-1}\right) \\ \delta_x = \dfrac{\tau_1 - \tau_{-1}}{2h} \end{cases} \tag{1.127}$$
et $\delta^H_{4,x} = \sigma_x^{-1} \circ \delta_x$ définit un opérateur d'approximation de la dérivée première à l'ordre 4. En particulier, comme $\|\sigma_x^{-1}\|_\infty \leq 3$, on montre à partir de la proposition 1.6 que :
$$\|(\partial_x u)^* - \delta^H_{4,x}(u^*)\|_\infty \leq \frac{h^4}{15}\|\partial_x^{(4)} u\|_\infty. \tag{1.128}$$

- Si $u \in \mathcal{C}^7(\Omega)$ alors
$$\begin{cases} \sigma_x = \dfrac{9}{15}\mathrm{Id} + \dfrac{1}{5}\left(\tau_1 + \tau_{-1}\right) \\ \delta_x = \dfrac{14}{15}\dfrac{\tau_1 - \tau_{-1}}{2h} + \dfrac{1}{15}\dfrac{\tau_2 - \tau_{-2}}{4h} \end{cases} \tag{1.129}$$
et $\delta^H_x = \sigma_x^{-1} \circ \delta_{4,x}$ définit un opérateur d'approximation de la dérivée première à l'ordre 6,

- Si $u \in \mathcal{C}^9(\Omega)$ alors
$$\begin{cases} \sigma_x = \dfrac{4}{7}\mathrm{Id} + \dfrac{3}{14}\left(\tau_1 + \tau_{-1}\right) \\ \delta_x = \dfrac{25}{28}\dfrac{\tau_1 - \tau_{-1}}{2h} + \dfrac{4}{35}\dfrac{\tau_2 - \tau_{-2}}{4h} - \dfrac{1}{140}\dfrac{\tau_3 - \tau_{-3}}{6h} \end{cases} \tag{1.130}$$
et $\delta^H_{8,x} = \sigma_x^{-1} \circ \delta_x$ définit un opérateur d'approximation de la dérivée première à l'ordre 8.

Les opérateurs σ_x et δ_x s'expriment en fonction de τ donc ils commutent, tout comme σ_x^{-1} et δ_x :
$$\delta^H_{2J+2,x} = \sigma_x^{-1} \circ \delta_x = \delta_x \circ \sigma_x^{-1}. \tag{1.131}$$

Notons la fraction rationnelle $Q^H_{2J+2} \in \mathbb{R}(X)$ telle que
$$Q^H_{2J+2}(X) = \frac{Q_{2J}(X)}{R(X)} = \frac{\displaystyle\sum_{p=1}^{J} \frac{a_p}{2p}(X^p - X^{N-p})}{(1-2\beta) + \beta(X + X^{N-1})}. \tag{1.132}$$

Exemples :

- $Q^H_4(X)$ est donné par
$$Q^H_4(X) = \frac{X - X^{N-1}}{\dfrac{4}{3} + \dfrac{1}{3}(X + X^{N-1})}, \tag{1.133}$$

- $Q^H_6(X)$ est donné par
$$Q^H_6(X) = \frac{\dfrac{14}{15}(X - X^{N-1}) + \dfrac{1}{15}(X^2 - X^{N-2})}{\dfrac{9}{15} + \dfrac{1}{5}(X + X^{N-1})}, \tag{1.134}$$

- $Q^H_8(X)$ est donné par
$$Q^H_8(X) = \frac{\dfrac{25}{56}(X - X^{N-1}) + \dfrac{4}{140}(X^2 - X^{N-2}) - \dfrac{1}{840}(X^3 - X^{N-3})}{\dfrac{4}{7} + \dfrac{3}{14}(X + X^{N-1})}. \tag{1.135}$$

La fraction rationnelle Q^H_{2J+2} permet d'exprimer $\delta^H_{2J+2,x}$ en fonction de τ

$$\delta^H_{2J+2,x} = \frac{1}{h} Q^H_{2J+2}(\tau). \tag{1.136}$$

L'opérateur $\delta^H_{2J+2,x}$ agit sur les fonctions de $l^2_{h,\text{pér}}$. Les propriétés spectrales de $\delta^H_{2J,x}$ s'expriment à l'aide de la fraction rationnelle Q^H_{2J+2}.

Proposition 1.13. *Les valeurs propres de $\delta^H_{2J+2,x}$ sont*

$$\frac{1}{h} Q^H_{2J+2}(\omega^k) \tag{1.137}$$

où $-N/2+1 \leq k \leq N/2$. L'espace propre associé à $\dfrac{1}{h} Q^H_{2J+2}(\omega^k)$ est donné par Vect$(\mathfrak{u}^k)$.

Proposition 1.14. *Si les coefficient $(a_p)_{1\leq p\leq J}$ et β sont solutions du système (1.116), alors pour $k \in \mathbb{N}$ fixé et $h = L/N$ on a*

$$\frac{2i\pi k}{N} - Q^H_{2J+2}\left(\exp\left(\frac{2i\pi k}{N}\right)\right) = \mathcal{O}(h^{2J+3}). \tag{1.138}$$

De plus, en posant $\theta = 2k\pi/N$, on a

$$-iQ^H_{2J+2}(\exp(i\theta)) = \frac{\sin(\theta)}{1+2\beta(\cos(\theta)-1)} \sum_{p=1}^{J} \frac{a_p}{p} U_p(\cos\theta) \tag{1.139}$$

où $(U_p)_{1\leq p\leq J}$ désigne des polynômes de Tchebychev de seconde espèce.

Démonstration. D'une part, d'après le théorème 1.2 on a

$$\sigma_x(\partial_x u^k)^* - \delta_x(u^k)^* = \mathcal{O}(h^{2J+2}), \tag{1.140}$$

donc en inversant l'opérateur σ_x, on a

$$(\partial_x u^k)^* - \delta^H_{2J+2,x}(u^k)^* = \mathcal{O}(h^{2J+2}). \tag{1.141}$$

D'autres part, on rappelle que la fonction u^k et la fonction de grille $\mathfrak{u}^k$ sont liées par

$$u^k = \sqrt{h}\mathfrak{u}^k, \tag{1.142}$$

donc on a directement

$$\begin{aligned}
(\partial_x u^k)^* - \delta^H_{2J+2,x}(u^k)^* &= \frac{2i\pi k}{L}(u^k)^* - \frac{1}{h} Q^H_{2J+2}(\omega^k)(u^k)^* \\
&= \frac{1}{h}\left(\frac{2i\pi hk}{L} - Q^H_{2J+2}(\omega^k)\right)(u^k)^* \\
&= \frac{1}{h}\left(\frac{2i\pi k}{N} - Q^H_{2J+2}\left(\exp\left(\frac{2i\pi k}{N}\right)\right)\right)(u^k)^*.
\end{aligned}$$

La fonction u^k est bornée, donc

$$\frac{2i\pi k}{N} - Q^H_{2J+2}\left(\exp\left(\frac{2i\pi k}{N}\right)\right) = \mathcal{O}(h^{2J+1}). \tag{1.143}$$

De plus par construction on a

$$\begin{aligned} Q_{2J}\left(\exp(i\theta)\right) &= \sum_{p=1}^{J} a_p \frac{\exp(ip\theta) - \exp(i(N-p)\theta)}{2p} \text{par périodicité des opérateurs.} \\ &= \sum_{p=1}^{J} a_p \frac{\exp(ip\theta) - \exp(-ip\theta)}{2p} \\ &= \sum_{p=1}^{J} a_p \frac{i\sin(p\theta)}{p}. \end{aligned}$$

D'où

$$Q_{2J}^{H}\left(\exp(i\theta)\right) = i\frac{\sin(\theta)}{1+2\beta(\cos(\theta)-1)} \sum_{p=1}^{J} \frac{a_p}{p} U_p(\cos\theta) \in i\mathbb{R} \tag{1.144}$$

où U_p désigne un polynôme de Tchebychev de seconde espèce. De plus, on a $\sin(0) = \sin(\pi) = 0$, donc

$$Q_{2J}^{H}\left(\exp(i0)\right) = Q_{2J}^{H}\left(\exp(i\pi)\right) = 0, \tag{1.145}$$

ce qui conclut la proposition. □

Par 2π-périodicité et imparité de $\theta \in \mathbb{R} \mapsto -iQ_{2J+2}^{H}(e^{i\theta})$, et d'après la proposition 1.14, on souhaite comparer $\theta \in \mathbb{R} \mapsto -iQ_{2J+2}^{H}(e^{i\theta})$ et $\theta \in \mathbb{R} \mapsto \theta$. Quelques exemples, pour les valeurs de $2J+2 = 4$, 6 et 8, sont tracés dans la Figure 1.3. On compare ces fonctions avec celles obtenues pour les schémas $\delta_{2J,x}$ aux ordre 2, 4, 6 et 8. On constate que les schémas hermitiens $\delta_{2J+2,x}^{H}$ représentent mieux θ que les schémas $\delta_{2J,x}$. La représentation spectrale de la dérivée est meilleure en utilisant le schéma $\delta_{2J+2,x}^{H}$ que le schéma $\delta_{2J,x}$. De plus, plus l'ordre du schéma est élevé, mieux la fonction $\theta \in \mathbb{R} \mapsto \theta$ est approchée.

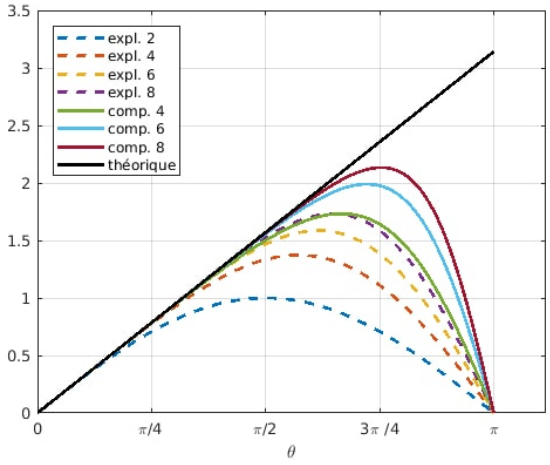


FIGURE 1.3 – Représentation de $-iQ_{2J+2}^{H}\left(e^{i\theta}\right)$ en fonction de θ pour les schémas d'approximation hermitien $\delta_{2J+2,x}^{H}$ d'ordres 2, 4, 6 et 8. Les courbes en pointillés représentent les fonctions $-iQ_{2J}\left(e^{i\theta}\right)$ associées aux opérateurs d'approximations $\delta_{2J,x}$.

Considérons à présent la version matricielle de l'opérateur $\delta_{2J+2,x}^{H}$. On pose $P_\sigma \in \mathbb{M}_N(\mathbb{R})$ la matrice

associée à l'opérateur σ_x. Cette matrice s'exprime par

$$P_\sigma = R(T) \tag{1.146}$$

$$= \begin{bmatrix} 1-2\beta & \beta & & & \beta \\ \beta & 1-2\beta & \beta & (0) & \\ & \ddots & \ddots & \ddots & \\ & (0) & \beta & 1-2\beta & \beta \\ \beta & & & \beta & 1-2\beta \end{bmatrix} \in \mathbb{M}_N(\mathbb{R}). \tag{1.147}$$

La relation suivante est immédiate

$$\mathrm{vec}(\sigma_x \mathfrak{u}) = P_\sigma \cdot \mathrm{vec}(\mathfrak{u}). \tag{1.148}$$

Les valeurs propres de P_σ sont connues et données par le corollaire 1.1.

Proposition 1.15. *Les valeurs propres de P_σ sont*

$$R(\omega^k) \tag{1.149}$$

avec $-N/2+1 \leq k \leq N/2$. L'espace propre associé à $R(\omega^k)$ est Vect(U^k) avec $U^k = vec(\mathfrak{u}^k)$.

La matrice P_σ est inversible si $|\beta| < 1/2$ et trivialement symétrique. De la même manière, on a déjà vu que

$$D_{2J} = \frac{1}{h} Q_{2J}(T) \in \mathbb{M}_N(\mathbb{R}) \tag{1.150}$$

donc

$$\mathrm{vec}(\delta_x \mathfrak{u}) = D_{2J} \cdot \mathrm{vec}(\mathfrak{u}). \tag{1.151}$$

Le calcul de $\delta^H_{2J+2,x}\mathfrak{u}$ se fait par la résolution du système

$$P_\sigma \cdot \mathrm{vec}(\delta^H_{2J+2,x}\mathfrak{u}) = D_{2J} \cdot \mathrm{vec}(\mathfrak{u}) \tag{1.152}$$

et on a

$$\mathrm{vec}(\delta_{2J+2,x}\mathfrak{u}) = P_\sigma^{-1} D_{2J} \cdot \mathrm{vec}(\mathfrak{u}). \tag{1.153}$$

Proposition 1.16. *Les matrices D_{2J}, P_σ et P_σ^{-1} commutent.*

Démonstration. Les matrices D_{2J} et P_σ s'expriment en fonction de T d'où la propriété. □

Proposition 1.17. *La matrice $P_\sigma^{-1} D_{2J}$ est antisymétrique.*

Démonstration. Par calcul immédiat, on a :

$$(P_\sigma^{-1} D_{2J})^T = D_{2J}^T P_\sigma^{-T} = -D_{2J} P_\sigma^{-1} = -P_\sigma^{-1} D_{2J}, \tag{1.154}$$

car D_{2J} est antisymétrique et P_σ est symétrique (donc P_σ^{-1} aussi). □

Le calcul de $\delta^H_{2J+2,x}\mathfrak{u}$ se fait par résolution d'un système linéaire. Ce système peut être résolu grâce à la formule de Shermann-Morisson-Woodbury couplé à un solveur tridiagonal comme l'algorithme de Thomas ou grâce à un solveur basé sur la transformée de Fourier rapide.

Proposition 1.18. ***(Formule de Shermann-Morisson-Woodbury)*** *Soient $A, B \in \mathbb{M}_N(\mathbb{R})$ deux matrices inversibles telles que*

$$A = B + RS^T, \tag{1.155}$$

avec R et S deux matrices de $\mathbb{M}_{N,n}(\mathbb{R})$ avec $n \leq N$. Alors l'inverse de A peut s'écrire

$$A^{-1} = B^{-1} - B^{-1} R \left(Id + S^T B^{-1} R\right)^{-1} S^T B^{-1}. \tag{1.156}$$

Démonstration. Il suffit de vérifier que

$$\left(B + RS^T\right)\left(B^{-1} - B^{-1}R\left(Id + S^TB^{-1}R\right)^{-1}S^TB^{-1}\right) = Id. \tag{1.157}$$

□

Dans le cas où $n \lll N$, A est une petite perturbation de la matrice B de la forme

$$A = B + \delta B \tag{1.158}$$

avec rang(δB) "petit". Si on peut facilement calculer l'inverse de B la formule de Shermann-Morisson-Woodbury (1.156) donne un algorithme efficace de résolution du système

$$AX = b. \tag{1.159}$$

Algorithme 1 : Algorithme de Shermann-Morisson-Woodbury

1: Calcul de $V_1 = B^{-1}b$,
2: Calcul de $V_2 = S^TV_1$,
3: Calcul de $V_3 = (Id + S^TB^{-1}R)^{-1}V_2$ (résolution d'un système de petite taille),
4: Calcul de $V_4 = RV_3$,
5: Calcul de $V_5 = B^{-1}V_4$,
6: Calcul de $X = V_1 - V_5$.

Proposition 1.19. *Soit P_σ et $\tilde{P}_\sigma$ les matrices données par*

$$P_\sigma = \begin{bmatrix} 1-2\beta & \beta & & & \beta \\ \beta & 1-2\beta & \beta & (0) & \\ & \ddots & \ddots & \ddots & \\ & (0) & \beta & 1-2\beta & \beta \\ \beta & & & \beta & 1-2\beta \end{bmatrix} \text{ et } \tilde{P}_\sigma = \begin{bmatrix} 1-2\beta & \beta & & & \\ \beta & 1-2\beta & \beta & (0) & \\ & \ddots & \ddots & \ddots & \\ & (0) & \beta & 1-2\beta & \beta \\ & & & \beta & 1-2\beta \end{bmatrix}. \tag{1.160}$$

Alors

$$P_\sigma = \tilde{P}_\sigma + RS^T \tag{1.161}$$

avec

$$R = \beta \begin{bmatrix} 1 & 0 \\ 0 & \vdots \\ \vdots & \vdots \\ \vdots & 0 \\ 0 & 1 \end{bmatrix} \text{ et } S = \begin{bmatrix} 0 & 1 \\ \vdots & 0 \\ \vdots & \vdots \\ 0 & \vdots \\ 1 & 0 \end{bmatrix} \tag{1.162}$$

Il découle l'algorithme de calcul de $\text{vec}(\delta^H_{2J+2,x}\mathfrak{u}) = P_\sigma^{-1} \cdot D_{2J} \cdot \text{vec}(\mathfrak{u})$ donné par

Algorithme 2 : Calcul Hermitien

1: Calcul de $b = D_{2J}\text{vec}(\mathfrak{u})$,
2: Calcul de $V_1 = \tilde{P}_\sigma^{-1}b$,
3: Calcul de $V_2 = S^TV_1$,
4: Calcul de $V_3 = (Id + S^T\tilde{P}_\sigma^{-1}R)^{-1}V_2$ (résolution d'un système de taille 2×2),
5: Calcul de $V_4 = RV_3$,
6: Calcul de $V_5 = \tilde{P}_\sigma^{-1}V_4$,
7: Calcul de $\text{vec}(\delta^H_{2J+2,x}\mathfrak{u}) = V_1 - V_5$.

L'utilisation de l'algorithme 2 pour la résolution du système (1.152) permet de se ramener à la résolution de systèmes tridiagonaux. La résolution de ces derniers se fait en utilisant l'algorithme de Thomas [20, 72]. Le coût global de l'algorithme de résolution est alors $\mathcal{O}(N)$. On aurait aussi pu utiliser un solveur rapide de type transformée de Fourier rapide [83]. Mais le coût en calcul d'un tel algorithme est de l'ordre de $\mathcal{O}(N\log(N))$, ce qui est plus élevé.

1.1.6 Opérateurs de filtrage

Pour améliorer les propriétés de stabilité d'un schéma centré en espace, il est connu que l'utilisation d'un opérateur de type "filtrage", à chaque pas de temps, est bénéfique. Un opérateur de type filtrage est un opérateur de la forme

$$\mathcal{F}_{2J,x} = \sum_{p=0}^{J} a_p \frac{\tau^p + \tau^{-p}}{2}. \tag{1.163}$$

Pour $\mathfrak{u} \in l^2_{h,\text{pér}}$ fonction de grille périodique, on a

$$\mathcal{F}_{2J,x}(\mathfrak{u})_j = \sum_{p=0}^{J} a_p \frac{\mathfrak{u}_{j+p} + \mathfrak{u}_{j-p}}{2} \text{ avec } 0 \leq j \leq N-1. \tag{1.164}$$

Soit $S_{2J} \in \mathbb{R}_{N-1}[X]$ défini par

$$S_{2J}(X) = \sum_{p=0}^{J} a_p \frac{X^p + X^{N-p}}{2}, \tag{1.165}$$

on a $\mathcal{F}_{2J,x}(\mathfrak{u})_j = S_{2J}(\tau)(\mathfrak{u})_j$ pour tout $0 \leq j \leq N-1$.

L'opérateur $\mathcal{F}_{2J,x}$ est un opérateur d'interpolation au sens de la définition 1.4. C'est à dire qu'il est consistant avec l'identité. D'autre part, il joue le rôle d'une dissipation numérique. Cette contrainte se traduit par le fait que $\mathfrak{u}^{N/2}$, qui vérifie

$$\mathfrak{u}^{N/2}_j = (-1)^j \text{ pour } 0 \leq j \leq N-1, \tag{1.166}$$

est inclus dans le noyau de $\mathcal{F}_{2J,x}$: $\mathcal{F}_{2J,x}(\mathfrak{u}^{N/2}) = \mathfrak{o}$. On note que l'on a alors

$$\ker(\mathcal{F}_{2J,x}) \subset \ker(\delta_{2J,x}) \text{ et } \ker(\mathcal{F}_{2J,x}) \subset \ker(\delta^H_{2J,x}). \tag{1.167}$$

En pratique, cela correspond à annuler le mode oscillant à la fréquence du maillage. A chaque itération en temps, on souhaite supprimer ce mode en projetant la solution sur l'orthogonal du mode oscillant :

$$\text{Vect}(\mathfrak{u}^{N/2})^{\perp}. \tag{1.168}$$

De plus, un opérateur de filtrage doit laisser $\mathfrak{u}^0$ inchangé, c'est à dire

$$\mathcal{F}_{2J,x}(\mathfrak{u}^0) = \mathfrak{u}^0. \tag{1.169}$$

Comme $\mathfrak{u}^0_j = 1$ pour $0 \leq j \leq N-1$, cette condition correspond à avoir

$$\mathcal{F}_{2J,x}(\mathfrak{u}^0) = \sum_{p=0}^{J} a_p = 1. \tag{1.170}$$

Nous nous concentrons sur les filtres maximisant l'ordre de précision [73]. D'autres filtres visent à minimiser la perturbation [12]. Dans ce cadre, nous définissons l'opérateur de filtrage par

Définition 1.6. *L'opérateur $\mathcal{F}_{2J,x}$ est appelé filtre centré s'il s'agit d'un opérateur aux différences finies de la forme* (1.163) *et si les coefficients $(a_p)_{1\leq p\leq J}$ vérifient :*

- *Consistance du filtrage :*

$$\sum_{p=0}^{J} a_p = 1, \tag{1.171}$$

- *Suppression du mode* $\mathfrak{u}^{N/2}$ *:*

$$\sum_{p=0}^{J} a_p (-1)^p = 0, \tag{1.172}$$

- *Précision de l'opérateur*

$$\sum_{p=0}^{J} a_p p^{2k} = 0 \text{ pour } 1 \leq k \leq J-1. \tag{1.173}$$

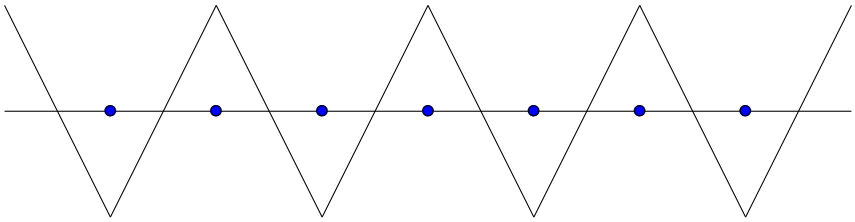


FIGURE 1.4 – Représentation graphique du monde $\mathfrak{u}^{N/2} \in l^2_{h,\text{pér}}$.

Proposition 1.20. *Si les* J *coefficients* $(a_p)_{1\leq p\leq J}$ *vérifient* (1.171), (1.172) *et* (1.173), *c'est à dire*

$$\begin{cases} \displaystyle\sum_{p=0}^{J} a_p &= 1 \\ \displaystyle\sum_{p=0}^{J} a_p (-1)^p &= 0 \\ \displaystyle\sum_{p=0}^{J} a_p p^{2k} &= 0 \text{ avec } 1 \leq k \leq J-1, \end{cases} \tag{1.174}$$

alors l'opérateur de filtrage $\mathcal{F}_{x,2J}$ *est consistant avec l'identité.*

Soit u *une fonction régulière et* u^* *la fonction de grille associée. L'erreur de troncature de l'opérateur de filtrage* $\mathcal{F}_{x,2J}$ *est donnée par*

$$\mathcal{F}_{x,2J}(u^*)_j - u^*_j = \frac{h^{2J}}{(2J)!} \left(\sum_{p=1}^{J} a_p p^{2J} \partial_x^{(2J)} u(\alpha_p) \right) \tag{1.175}$$

avec $\alpha_p \in]x_{j-p}, x_{j+p}[$.

Démonstration. u est une fonction régulière, donc d'après la formule de Taylor-Lagrange il existe $\xi \in]x, x+h[$ tel que

$$u(x+h) = \sum_{k=0}^{2J-1} \frac{h^k}{k!} \partial_x^{(k)} u(x) + \frac{1}{(2J)!} \partial_x^{(2J)} u(\xi). \tag{1.176}$$

D'après cette formule et le théorème des valeurs intermédiaires, on a

$$\begin{aligned} \frac{\tau_p \mathfrak{u}_j + \tau_{-p} \mathfrak{u}_j}{2} &= \frac{u(x_j + ph) - u(x_j - ph)}{2} \\ &= \sum_{k=0}^{2J-1} \frac{h^k + (-h)^k}{2k!} \partial_x^{(k)} u(x_j) + \frac{h^{2J}}{(2J)!} \partial_x^{(2J)} u(\alpha_p) \end{aligned}$$

avec $\alpha_p \in]x_j - ph, x_j + ph[$. Alors en multipliant cette relation par a_p et en sommant sur p, on obtient :

$$\begin{aligned}
\sum_{p=1}^{J} a_p \frac{\tau_p \mathfrak{u}_j + \tau_{-p}\mathfrak{u}_j}{2} &= \sum_{p=1}^{J}\sum_{k=0}^{2J-1} a_p \frac{(ph)^k + (-ph)^k}{2k!}\partial_x^{(k)}u(x_j) + \sum_{p=1}^{J} a_p \frac{(ph)^{2J}}{(2J)!}\partial_x^{(2J)}u(\alpha_p) \\
&= \sum_{k=0}^{2J-1} \frac{h^k}{2(k!)} \left(\sum_{p=0}^{J}(p^k + (-p)^k)\right)\partial_x^{(k)}u(x) + \sum_{p=0}^{J} a_p \frac{(ph)^{2J}}{(2J)!}\partial_x^{(2J)}u(\alpha_p) \\
&= \underbrace{\left(\sum_{p=0}^{J} a_p\right)}_{=1} u(x_j) + \sum_{k=1}^{2J-1} \frac{h^k}{2(k!)} \underbrace{\left(\sum_{p=0}^{J}(p^k + (-p)^k)\right)}_{=0 \text{ si } k \text{ impair.}} \partial_x^{(k)}u(x_j) + \frac{h^{2J}}{(2J)!}\sum_{p=0}^{J} a_p p^{2J}\partial_x^{(2J)}u(\alpha_p) \\
&= u(x_j) + \sum_{k=1}^{J} \frac{h^{2k}}{k!} \underbrace{\left(\sum_{p=0}^{J} a_p p^{2k}\right)}_{=0 \text{ d'après (1.173)}} + \frac{h^{2J}}{(2J)!}\left(\sum_{p=0}^{J} a_p p^{2J}\partial_x^{(2J)}u(\alpha_p)\right) \\
&= u(x_j) + \frac{h^{2J}}{(2J)!}\left(\sum_{p=0}^{J} a_p p^{2J}\partial_x^{(2J)}u(\alpha_p)\right).
\end{aligned}$$

On retrouve ainsi l'erreur de consistance souhaitée et le résultat est prouvé. □

Corollaire 1.4. *Si les coefficients $(a_p)_{1\le p\le J}$ sont solution de* (1.174) *alors il existe $C > 0$ indépendant de h tel que pour toute fonction u régulière on a*

$$\|u^* - \mathcal{F}_{x,2J}(u^*)\|_\infty \le Ch^{2J}\|(\partial_x^{(2J)}u)^*\|_\infty. \tag{1.177}$$

L'existence et l'unicité des opérateurs de filtrage $\mathcal{F}_{x,2J}$ est donnée par la proposition suivante :

Proposition 1.21. *Le système* (1.174) *admet une unique solution.*

Démonstration. Le système (1.174) s'écrit en matriciel :

$$\underbrace{\begin{bmatrix}
1 & 1 & 1 & 1 & \cdots & 1 \\
1 & -1 & 1 & -1 & \cdots & (-1)^J \\
0 & 1 & 2^2 & 3^2 & \cdots & J^2 \\
0 & 1 & 2^4 & 3^4 & \cdots & J^4 \\
0 & 1 & 2^6 & 3^6 & \cdots & J^6 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 1 & (2^2)^{J-1} & (3^2)^{J-1} & \cdots & (J^2)^{J-1}
\end{bmatrix}}_{=A}
\underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ \vdots \\ a_J \end{bmatrix}}_{=a} =
\underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{=b}. \tag{1.178}$$

En remplaçant la deuxième ligne par une combinaison de la première et de la deuxième ligne : $L_2 \leftarrow L_1 - L_2$, on obtient :

$$\underbrace{\begin{bmatrix}
1 & 1 & 1 & 1 & \cdots & 1 \\
0 & 2 & 0 & 2 & \cdots & 1-(-1)^J \\
0 & 1 & 2^2 & 3^2 & \cdots & J^2 \\
0 & 1 & 2^4 & 3^4 & \cdots & J^4 \\
0 & 1 & 2^6 & 3^6 & \cdots & J^6 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 1 & (2^2)^{J-1} & (3^2)^{J-1} & \cdots & (J^2)^{J-1}
\end{bmatrix}}_{=\tilde{A}}
\underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ \vdots \\ a_J \end{bmatrix}}_{=a} =
\underbrace{\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{=\tilde{b}}. \tag{1.179}$$

Pour montrer que le système (1.174), on montre que la matrice $\tilde{A} \in \mathbb{M}_J(\mathbb{R})$ est inversible. En effet :

$$\begin{aligned}
\det(\tilde{A}) &= \begin{vmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 0 & 2 & 0 & 2 & \cdots & 1-(-1)^J \\ 0 & 1 & 2^2 & 3^2 & \cdots & J^2 \\ 0 & 1 & 2^4 & 3^4 & \cdots & J^4 \\ 0 & 1 & 2^6 & 3^6 & \cdots & J^6 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & (2^2)^{J-1} & (3^2)^{J-1} & \cdots & (J^2)^{J-1} \end{vmatrix} \\
&= \begin{vmatrix} 2 & 0 & 2 & \cdots & 1-(-1)^J \\ 1 & 2^2 & 3^2 & \cdots & J^2 \\ 1 & 2^4 & 3^4 & \cdots & J^4 \\ 1 & 2^6 & 3^6 & \cdots & J^6 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & (2^2)^{J-1} & (3^2)^{J-1} & \cdots & (J^2)^{J-1} \end{vmatrix} \\
&= 2 \sum_{p=0}^{\lfloor \frac{J-1}{2} \rfloor} \Delta_{2p+1}.
\end{aligned}$$

On va montrer que pour tout $1 \leq p \leq J$, $\Delta_p > 0$. Ainsi, $\det(\tilde{A})$ est la somme de nombres strictement positifs, donc $\det(\tilde{A}) > 0$. Δ_p est donné par

$$\begin{aligned}
\Delta_p &= \begin{vmatrix} 1 & 2^2 & 3^2 & \cdots & (p-1)^2 & (p+1)^2 & \cdots & J^2 \\ 1 & (2^2)^2 & (3^2)^2 & \cdots & ((p-1)^2)^2 & ((p+1)^2)^2 & \cdots & (J^2)^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & (2^2)^{J-1} & (3^2)^{J-1} & \cdots & ((p-1)^2)^{J-1} & ((p+1)^2)^{J-1} & \cdots & (J^2)^{J-1} \end{vmatrix} \\
&= \left(\frac{J!}{p}\right)^2 \begin{vmatrix} 1 & 1^2 & 1^2 & \cdots & 1^2 & 1^2 & \cdots & 1^2 \\ 1 & 2^2 & 3^2 & \cdots & (p-1)^2 & (p+1)^2 & \cdots & J^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & (2^2)^{J-2} & (3^2)^{J-2} & \cdots & ((p-1)^2)^{J-2} & ((p+1)^2)^{J-2} & \cdots & (J^2)^{J-2} \end{vmatrix} \\
&= \left(\frac{J!}{p}\right)^2 \mathrm{VDM}(1, 2^2, 3^2, \cdots (p-1)^2, (p+1)^2, \cdots, J^2)
\end{aligned}$$

où VDM désigne un déterminant de Vandermonde [32] déterminé par

$$\mathrm{VDM}(\alpha_1, \cdots, \alpha_n) = \prod_{1 \leq i < j \leq n} (\alpha_j - \alpha_i). \tag{1.180}$$

Dans notre cadre, on a

$$\Delta_p = \left(\frac{J!}{p}\right)^2 \mathrm{VDM}(\alpha_1, \alpha_2, \cdots \alpha_{J-1}) \tag{1.181}$$

avec α_j donné par

$$\alpha_j = \begin{cases} j^2 & \text{si } j < p \\ (j+1)^2 & \text{si } j \geq p. \end{cases} \tag{1.182}$$

Ainsi, si $j > i$, on observe que $\alpha_j - \alpha_i > 0$, donc

$$\begin{aligned}
\Delta_p &= \left(\frac{J!}{p}\right)^2 \mathrm{VDM}(1, 2^2, 3^2, \cdots (p-1)^2, (p+1)^2, \cdots, J^2) \\
&= \left(\frac{J!}{p}\right)^2 \prod_{1 \leq i < j \leq J-1} \underbrace{(\alpha_j - \alpha_i)}_{>0} > 0.
\end{aligned}$$

$\det(\tilde{A})$ est une somme de termes strictement positifs donc $\det(\tilde{A}) > 0$ et en particulier, $\det(\tilde{A}) \neq 0$ et le système admet une unique solution. □

Exemples : on donne les filtres pour $J = 1, 2, 3, 4$ ou 5.

- Soit $u \in \mathcal{C}^2(\Omega)$ alors
$$\mathcal{F}_{2,x} = \frac{1}{2}\mathrm{Id} + \frac{1}{2}\frac{\tau_1 + \tau_{-1}}{2} \tag{1.183}$$
définit un opérateur de filtrage à l'ordre 2,

- soit $u \in \mathcal{C}^4(\Omega)$ alors
$$\mathcal{F}_{4,x} = \frac{10}{16}\mathrm{Id} + \frac{8}{16}\frac{\tau_1 + \tau_{-1}}{2} - \frac{2}{16}\frac{\tau_2 + \tau_{-2}}{2} \tag{1.184}$$
définit un opérateur de filtrage à l'ordre 4,

- soit $u \in \mathcal{C}^6(\Omega)$ alors
$$\mathcal{F}_{6,x} = \frac{44}{64}\mathrm{Id} + \frac{30}{64}\frac{\tau_1 + \tau_{-1}}{2} - \frac{12}{64}\frac{\tau_2 + \tau_{-2}}{2} + \frac{2}{64}\frac{\tau_3 + \tau_{-3}}{2} \tag{1.185}$$
définit un opérateur de filtrage à l'ordre 6,

- soit $u \in \mathcal{C}^8(\Omega)$ alors
$$\mathcal{F}_{8,x} = \frac{186}{256}\mathrm{Id} + \frac{112}{256}\frac{\tau_1 + \tau_{-1}}{2} - \frac{56}{256}\frac{\tau_2 + \tau_{-2}}{2} + \frac{16}{256}\frac{\tau_3 + \tau_{-3}}{2} - \frac{2}{256}\frac{\tau_4 + \tau_{-4}}{2} \tag{1.186}$$
définit un opérateur de filtrage à l'ordre 8,

- soit $u \in \mathcal{C}^{10}(\Omega)$ alors
$$\mathcal{F}_{10,x} = \frac{772}{1024}\mathrm{Id} + \frac{420}{1024}\frac{\tau_1 + \tau_{-1}}{2} - \frac{240}{1024}\frac{\tau_2 + \tau_{-2}}{2} + \frac{90}{1024}\frac{\tau_3 + \tau_{-3}}{2} - \frac{20}{1024}\frac{\tau_4 + \tau_{-4}}{2} + \frac{2}{1024}\frac{\tau_5 + \tau_{-5}}{2} \tag{1.187}$$
définit un opérateur de filtrage à l'ordre 10.

La Table 1.1 résume les valeurs de a_p associées à $\mathcal{F}_{2J,x}$ pour un ordre de précision $2J$ fixé.

Ordre de précision : 2J	a_0	a_1	a_2	a_3	a_4	a_5
2	1/2	1/2				
4	10/16	8/16	−2/16			
6	44/64	30/64	−12/64	2/64		
8	186/256	112/256	−56/256	16/256	−2/256	
10	772/1024	420/1024	−240/1024	90/1024	−20/1024	2/1024

TABLE 1.1 – Exemples de filtres de la forme (1.163) et leurs ordres de précision.

Les valeurs propres et fonctions propres de $\mathcal{F}_{x,2J}$ sont issues de la proposition 1.4 appliquée au polynôme $S_{2J} \in \mathbb{R}_J[X]$ défini par (1.165).

Proposition 1.22. *Les valeurs propres de $\mathcal{F}_{2J,x}$ sont*
$$S_{2J}(\omega^k) \tag{1.188}$$
où $-N/2 + 1 \leq k \leq N/2$. L'espace propre associé à $S_{2J}(\omega^k)$ est donné par Vect$(\mathfrak{u}^k)$.

Théorème 1.4. *Soit* $\beta : \theta \in [0,\pi] \mapsto \beta(\theta) = \sum_{p=0}^{J} a_p \cos(p\theta)$ *le symbole de l'opérateur de filtrage* $\mathcal{F}_{2J,x}$, *alors*

- *il existe un unique polynôme* $P \in \mathbb{R}_J(X)$ *tel que*

$$\beta(\theta) = P(\cos(\theta)), \tag{1.189}$$

ce polynôme est déterminé par

$$P(X) = 1 - \frac{1}{2^J}(1-X)^J. \tag{1.190}$$

- *Pour tout* $\theta \in [0,\pi]$, *on a*

$$0 \leq \beta(\theta) \leq 1. \tag{1.191}$$

Démonstration. • Les polynômes de Tchebytchev de première espèce, notés T_p , sont définis par la relation

$$T_p(\cos(\theta)) = \cos(p\theta). \tag{1.192}$$

On a $T_p \in \mathbb{R}_p[X]$ et les polynômes T_p forment une suite de polynômes de degrés croissants. Ils forment une base de $\mathbb{R}_p[X]$ et on a

$$\begin{aligned}\beta(\theta) &= \sum_{p=0}^{J} a_p \cos(p\theta) \\ &= \sum_{p=0}^{J} a_p T_p(\cos(\theta)).\end{aligned}$$

Ainsi il existe un polynôme $P = \sum_{p=0}^{J} a_p T_p$ de degré J tel que

$$\beta(\theta) = P(\cos(\theta)). \tag{1.193}$$

Comme les polynômes $(T_p)_{0\leq p\leq J}$ forment une base de $\mathbb{R}_J[X]$, ce polynôme est unique.

Les fonctions $\theta \mapsto \cos(n\theta)$ forment une famille libre pour $0 \leq n \leq J$. On pose E_J l'espace engendré par ces fonctions, $\dim(E_J) = J$. Alors $\beta \in E_J$ et si on pose $Q(X) = 1 - \frac{1}{2^J}(1-X)^J$, la fonction $\theta \mapsto Q(\cos(\theta))$ est une fonction de E_J. Montrons que ces fonctions coïncident. Soit la fonction de E_J donnée par

$$f(\theta) = \beta(\theta) - Q(\cos(\theta)). \tag{1.194}$$

Alors pour commencer, le résultat suivant est vérifié :

$$f^{(n)}(0) = 0 \text{ pour tout } 0 \leq n \leq J-1. \tag{1.195}$$

En effet, pour $n = 0$, on a

$$\begin{aligned}f(0) &= \beta(0) - \left(1 - \frac{1}{2^J}(1-\cos(0))^J\right) \\ &= \sum_{p=0}^{J} a_p - 1 \\ &= 0.\end{aligned}$$

Pour traiter le cas $1 \leq n \leq 2J$, nous utilisons la formule de Fàa Di Bruno [19], ainsi

$$(Q \circ \cos)^{(2n)}(\theta) = \sum_{k=0}^{2n} P^{(k)}(\cos(\theta)) B_{2n,k}(-\sin\theta, -\cos\theta, \sin\theta, \cos\theta, ...), \tag{1.196}$$

où $B_{2n,k}$ désignent les polynômes de Bell [19]. En particulier pour $\theta = 0$

$$(Q \circ \cos)^{(2n)}(0) = \sum_{k=0}^{2n} Q^{(k)}(1) B_{2n,k}(0, -1, 0, 1, 0, -1, ...). \tag{1.197}$$

et la dérivée $k-$ième de Q est donnée par

$$Q^{(k)}(X) = -\frac{(-1)^k}{2^J} \frac{J!}{(J-n)!}(1-X)^{J-n} \text{ pour tout } k > 1. \tag{1.198}$$

Donc $Q^{(k)}(1) = 0$ et $(Q \circ \cos)^{(n)}(0) = 0$. De là, il vient directement

$$\begin{aligned} f^{(n)}(0) &= \beta^{(n)}(0) - (Q \circ \cos)^{(n)}(0) \\ &= \sum_{p=1}^{J} a_p p^{2n} - 0 \\ &= 0 \text{ pour } 1 \leq n \leq J-1, \end{aligned}$$

donc en particulier

$$f^{(2n)}(0) = 0. \tag{1.199}$$

De plus,

$$\begin{aligned} f(\pi) &= \beta(\pi) - \left(1 - \frac{1}{2^J}(1 - \cos(\pi))^J\right) \\ &= \sum_{p=1}^{J} a_p (-1)^p \\ &= 0. \end{aligned}$$

Comme $f \in E_J$, il existe $(f_p)_{0 \leq p \leq J}$ tels que

$$f(\theta) = \sum_{p=0}^{J} f_p \cos(p\theta). \tag{1.200}$$

De plus, $f^{(2k)}(0) = 0$ pour tout k et $f(\pi) = 0$, alors les coefficients $(f_p)_{0 \leq p \leq J}$ sont solutions de

$$\begin{cases} \displaystyle\sum_{p=0}^{J} f_p &= 0 \\ \displaystyle\sum_{p=0}^{J} f_p (-1)^p &= 0 \\ \displaystyle\sum_{p=0}^{J} f_p p^{2k} &= 0 \text{ pour } 1 \leq k \leq J-1. \end{cases} \tag{1.201}$$

On a déjà vu que ce système est inversible donc pour tout $0 \leq p \leq J$, $f_p = 0$ et $f(\theta) = 0$ pour tout θ. Ainsi

$$\beta(\theta) = Q(\cos\theta). \tag{1.202}$$

On a bien

$$P(X) = Q(X) = 1 - \frac{1}{2^J}(1-X)^J. \tag{1.203}$$

- On a vu que

$$\beta(\theta) = P(\cos(\theta)). \tag{1.204}$$

Or par dérivation $P'(X) = \frac{J}{2^J}(1-X)^{J-1}$ qui ne change pas de signe pour $X \in [-1,1]$, donc P est monotone sur $[-1,1]$.

De plus, $\cos(\theta) \in [-1,1]$ pour $\theta \in [0,\pi]$, donc $\beta(\theta) \subset [(P(-1),P(1))] = [0,1]$. De là, on déduit que pour tout $\theta \in [0,\pi]$, on a

$$0 \leq \beta(\theta) \leq 1. \tag{1.205}$$

□

Remarque 1.2. *On a vu dans la proposition 1.22 que les valeurs propres de $\mathcal{F}_{2J,x}$ sont de la forme*

$$\begin{aligned} S_{2J}(\omega^k) &= S_{2J}\left(\exp\left(\frac{2i\pi k}{N}\right)\right) \\ &= \beta\left(\frac{2\pi k}{N}\right) \text{ d'après la formule d'Euler,} \\ &\in [0,1] \text{ d'après le théorème 1.4,} \end{aligned}$$

avec $-N/2+1 \leq k \leq N/2$. On en déduit que l'opérateur de filtrage $\mathcal{F}_{2J,x}$ est bien dissipatif.

La fonction $\theta \mapsto \beta(\theta)$ permet de considérer le comportement du filtre sur les différents fonctions propres. Par parité et $2\pi-$périodicité de β, on peut observer la fonction β sur $[0,\pi]$. On représente cette fonction dans la Figure 1.5 pour les filtres d'ordres $2J = 10, 8, ..., 2$. Les fréquences proches de 0 sont peu affectées par $\mathcal{F}_{2J,x}$ alors que celles proches de π sont de plus en plus atténuées à mesure que l'on s'approche de π jusqu'à être supprimées. L'importance de ce filtrage sera vue au chapitre 2 dans l'analyse numérique.

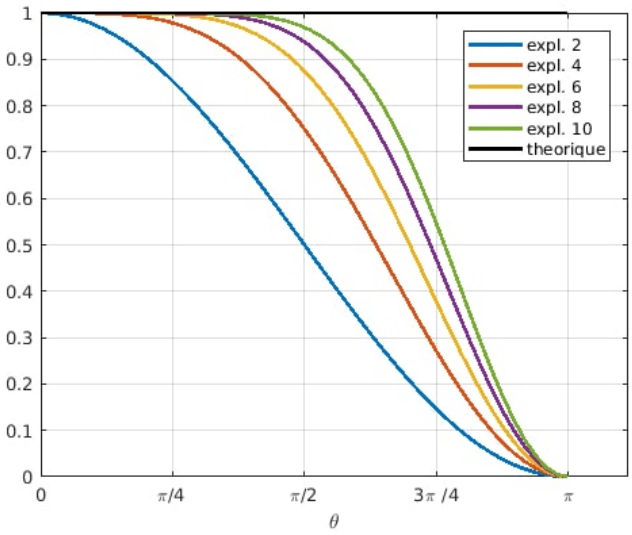


FIGURE 1.5 – Fonction d'amplification $\theta \mapsto \beta(\theta)$ pour les filtres explicites d'ordre 2, 4, 6, 8 et 10.

Pour plus de clarté, dans la Table 1.2, on représente la fréquence $\theta_{0.95}$ telle que

$$\begin{aligned} \beta(\theta) &\geq 0.95 \quad \text{si} \quad 0 \leq \theta \leq \theta_{0.95} \\ \beta(\theta) &\leq 0.95 \quad \text{si} \quad \theta_{0.95} \leq \theta \leq \pi. \end{aligned} \tag{1.206}$$

Ordre du filtre : 2J	$\theta_{0.95}(\mathbf{J})$
10	1.6695
8	1.5165
6	1.3045
4	0.9851
2	0.4510

TABLE 1.2 – Valeur de $\theta_{0.95}(J)$ pour quelques valeurs de J.

Comme la fonction $\theta \mapsto \beta(\theta)$ est strictement décroissante et continue sur $[0, \pi]$, c'est une bijection de $[0, \pi]$ dans $[0, 1]$ et on a

$$\theta_{0.95}(J) = \arccos\left[1 - 2(0.05)^{1/J}\right]. \tag{1.207}$$

La fonction $J \mapsto \theta_{0.95}(J)$ est strictement croissante donc lorsque J croît, $\theta_{0.95}(J)$ croît et les fonctions sont moins affectées par le filtre $\mathcal{F}_{2J,x}$. De plus, on a la limite

$$\lim_{J \to +\infty} \theta_{0.95}(J) = \pi. \tag{1.208}$$

Cependant, il faut noter que plus J augmente, plus les fonctions manipulées sont supposées régulières et ce qui précède n'est vrai que pour des fonctions très régulières.

Remarque 1.3. *Le filtre $\mathcal{F}_{2J,x}$ est linéaire et agit sur les composantes des fonctions de grilles. Il existe $M_{2J} \in \mathbb{M}_N(\mathbb{R})$ la matrice associée au filtrage des données en dimension 1 telle que*

$$vec(\mathcal{F}_{2J,x}\mathfrak{u}) = M_{2J} \cdot vec(\mathfrak{u}). \tag{1.209}$$

La matrice M_{2J} est donnée par

$$M_{2J} = S_{2J}(T), \tag{1.210}$$

de plus, cette matrice est symétrique.

1.2 Opérateurs aux différences en dimension 2

Comme nous le verrons dans le chapitre 3, la Cubed-Sphere est divisée en panels assimilables à des carrés en dimension 2. Pour cette raison, il est utile de considérer les opérateurs aux différences en dimension 2. Tous les opérateurs introduits dans la section précédente en dimension 1 peuvent être étendus à un carré périodique.

1.2.1 Notations

En dimension 2, les notations sont analogues à celles utilisées en dimension 1. Si a et b sont des réels positifs, nous notons $\Omega = [a, b]^2$. Chaque côté du carré est de longueur $L = b - a$. Soient u et v dans $L^2(\Omega, \mathbb{C})$. Le produit scalaire est

$$(u, v) = \int_\Omega u(x, y)\bar{v}(x, y)dxdy. \tag{1.211}$$

La norme associée est :

$$\|u\|_{L^2(\Omega)} = \sqrt{(u, u)}. \tag{1.212}$$

On note également

$$\|u\|_{L^\infty(\Omega)} = \max_{(x,y)\in\Omega} |u(x, y)|. \tag{1.213}$$

Ces deux normes sont notées $\|u\|_{L^2}$ et $\|u\|_{L^\infty}$.

Dans le domaine Ω, la grille est constituée des points $(x_i, y_j)_{0\leq i,j\leq N}$ où $N \geq 1$ avec $a = x_0 < x_1 < \ldots < x_N = b$ et $a = y_0 < y_1 < \ldots < y_N = b$. Le pas d'espace h est fixe et donné par $h = \frac{L}{N}$. Les points de grilles sont (x_i, y_j) avec

$$\left\{ \begin{array}{rcl} x_i & = & a + ih \\ y_j & = & a + jh \end{array} \right. \text{ avec } 0 \leq i, j \leq N. \tag{1.214}$$

Les points $(x_i, y_j)_{0\leq i,j\leq N}$ sont de deux types (voir Fig. 1.6) :

- Les points de bords (x_i, y_j) avec

$$i \in \{0, N\} \text{ ou } j \in \{0, N\}, \tag{1.215}$$

- les points intérieurs (x_i, y_j) avec

$$1 \leq i, j \leq N - 1. \tag{1.216}$$

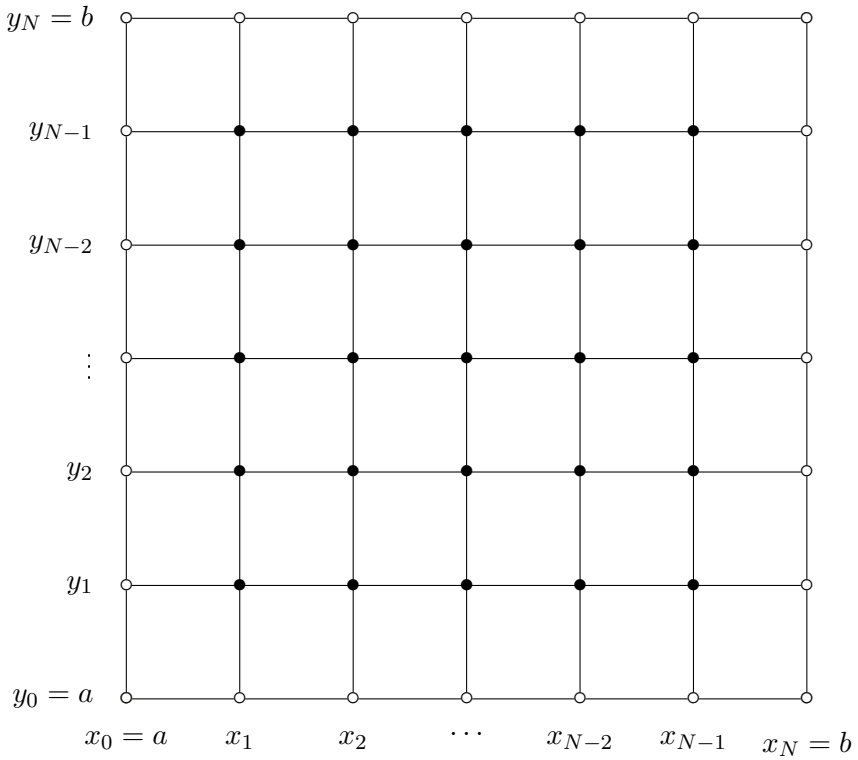


FIGURE 1.6 – Grille en dimension 2. Les symboles $\circ$ désignent les points de bords, les symboles $\bullet$ désignent les points intérieurs de la grille.

Une fonction $u : (x, y) \in \mathbb{R} \mapsto u(x, y) \in \mathbb{C}$ est *$L-$périodique* dans les directions x et y si

$$\begin{array}{rcl} u(x, y + L) & = & u(x, y) \\ u(x + L, y) & = & u(x, y) \end{array} \text{ pour tous } (x, y) \in \mathbb{R}^2. \tag{1.217}$$

Les notions de fonctions de grilles sont analogues à celles vues en dimension 1 :

1. Une *fonction de grille* est une fonction définie aux points de la grille $(x_i, y_j)_{0\leq i,j\leq N}$. Nous notons ces fonctions en fonte gothique comme $\mathfrak{u}$ ou $\mathfrak{v}$. On note :

$$\mathfrak{u}_{i,j} = \mathfrak{u}(x_i, y_j) \text{ et } \mathfrak{u} = (\mathfrak{u}_{i,j})_{0\leq i,j\leq N}. \tag{1.218}$$

On note L^2_h l'espace des fonctions de grilles. Cet espace est équipé d'un produit scalaire et de la norme associée :

$$(\mathfrak{u},\mathfrak{v})_h = h^2 \sum_{i,j=0}^{N} \mathfrak{u}(x_i,y_j)\bar{\mathfrak{v}}(x_i,y_j) \text{ et } |\mathfrak{u}|_h = \sqrt{(\mathfrak{u},\mathfrak{u})_h}. \tag{1.219}$$

De plus, pour $\mathfrak{u}$ fonction de grille, on note

$$|\mathfrak{u}|_\infty = \max_{0\le i,j\le N} |\mathfrak{u}_{i,j}|. \tag{1.220}$$

2. Soit $u : (x,y) \in \Omega \mapsto u(x,y) \in \mathbb{R}$, nous définissons la fonction de grille associée, notée u^*, comme la restriction de u à la grille :

$$u^*_{i,j} = u(x_i,y_j) \text{ pour tous } 0 \le i,j \le N. \tag{1.221}$$

3. Cas périodique : $\mathfrak{u}$ est périodique si $\mathfrak{u}_{i,0} = \mathfrak{u}_{i,N}$ et $\mathfrak{u}_{0,j} = \mathfrak{u}_{N,j}$ pour tous $0 \le i,j \le N$. On note $L^2_{h,\text{pér}}$ l'espace des fonctions de grilles périodiques. Cet espace est doté du produit scalaire et de la norme

$$(\mathfrak{u},\mathfrak{v})_{h,\text{pér}} = h^2 \sum_{i,j=0}^{N-1} \mathfrak{u}_{i,j}\bar{\mathfrak{v}}_{i,j} \text{ et } |\mathfrak{u}|_{h,\text{pér}} = \sqrt{(\mathfrak{u},\mathfrak{u})_{h,\text{pér}}}. \tag{1.222}$$

Pour u périodique selon x et y, on a $u^*_{i,0} = u^*_{i,N}$ et $u^*_{0,j} = u^*_{N,j}$ pour tous $0 \le i,j \le N$.

4. Une fonction de grille $\mathfrak{u} \in L^2_h$est associée au vecteur $U \in \mathbb{R}^{(N+1)^2}$ dont les composantes sont les valeurs de $\mathfrak{u}$ dans l'ordre anti-lexicographique :

$$U = \begin{bmatrix} \mathfrak{u}_{0,0} \\ \mathfrak{u}_{1,0} \\ \vdots \\ \mathfrak{u}_{N,0} \\ \mathfrak{u}_{0,1} \\ \mathfrak{u}_{1,1} \\ \vdots \\ \mathfrak{u}_{N,1} \\ \vdots \\ \mathfrak{u}_{N,N} \end{bmatrix}. \tag{1.223}$$

On note ces vecteurs par des lettres capitales.

1.2.2 Opérateurs aux différences en géométrie cartésienne

On considère dans cette section les fonctions de grille périodiques $u \in L^2_{h,\text{pér}}$. Nous définissons les opérateurs agissant sur les fonctions de grilles comme suit.

Définition 1.7. *Les opérateurs τ_x et τ_y dans les directions x et y sont définis par*

$$\begin{cases} \tau_x \mathfrak{u}_{i,j} &= \mathfrak{u}_{i+1,j} \\ \tau_y \mathfrak{u}_{i,j} &= \mathfrak{u}_{i,j+1} \end{cases} \tag{1.224}$$

avec $\mathfrak{u}$ une fonction de grille et $1 \le i,j \le N$. Il s'agit d'opérateurs de translations dans les directions x et y.

Les opérateurs obtenus en dimension 1 sont définis en dimension 2 grâce à ces deux opérateurs de translation. Les opérateurs centrés permettant d'approcher les dérivées partielles d'ordre 1 sont donnés par

$$\begin{cases} \delta_{2J,x} = \dfrac{1}{h} Q_{2J}(\tau_x) \\ \delta_{2J,y} = \dfrac{1}{h} Q_{2J}(\tau_y). \end{cases} \tag{1.225}$$

De même, on définit les opérateurs dans chaque direction par

$$\left\{ \begin{array}{rcl} \sigma_x & = & R(\tau_x) \\ \sigma_y & = & R(\tau_y) \end{array} \right. \tag{1.226}$$

où R est donné par $R(X) = (1-2\beta) + \beta(X + X^{N-1})$. Chacun des opérateurs σ_x et σ_y est inversible si $|\beta| < 1/2$. L'opérateur hermitien en dimension 1 : δ_x^H est étendu en dimension 2 grâce à la relation suivante

$$\left\{ \begin{array}{rcl} \delta_{2J+2,x}^H & = & \sigma_x^{-1} \circ \delta_{2J,x} \\ \delta_{2J+2,y}^H & = & \sigma_y^{-1} \circ \delta_{2J,y}. \end{array} \right. \tag{1.227}$$

Théorème 1.5. *Il existe C_x et C_y des constantes positives indépendantes de h telles que pour toute fonction $u : x \in \Omega \mapsto u(x) \in \mathbb{R}$ de $\mathcal{C}^5(\Omega)$ on a*

$$\begin{array}{rcl} \|(\partial_x u)^* - \delta_{2P+2,x}^H u^*\|_\infty & \leq & C_x h^{2J+2} \|\partial_x^{(2J+3)} u\|_\infty \\ \|(\partial_y u)^* - \delta_{2P+2,y}^H u^*\|_\infty & \leq & C_y h^{2J+2} \|\partial_y^{(2J+3)} u\|_\infty. \end{array} \tag{1.228}$$

Démonstration. Conséquence du théorème 1.3 appliqué à chaque direction x et y. □

1.2.3 Écriture matricielle des opérateurs aux différences en dimension 2

Dans cette section, nous précisons les notations vectorielles et matricielles utiles en dimension 2. La *base canonique* de $\mathbb{R}^N$, notée $(e_i)_{0 \leq i \leq N-1}$ est donnée par

$$(e_i) = \delta_{i,j} = \left\{ \begin{array}{ll} 1 & \text{si } j = i, \\ 0 & \text{sinon.} \end{array} \right. \tag{1.229}$$

$\delta_{i,j}$ est le symbole de Kronecker.

Définition 1.8. *Soit A une matrice $m \times n$ et B une matrice $p \times q$, avec $m, n, p, q \in \mathbb{N}^\star$. Le produit de Kronecker $A \otimes B$ est la matrice $mp \times nq$ définie par*

$$A \otimes B = \begin{bmatrix} a_{1,1}B & \cdots & a_{1,n}B \\ \vdots & \ddots & \vdots \\ a_{n,1}B & \cdots & a_{n,n}B \end{bmatrix}. \tag{1.230}$$

Le produit de Kronecker possède les propriétés suivantes [83] :

Proposition 1.23. *Soient A, B, C et D des matrices complexes.*

- *Pour tout $\alpha \in \mathbb{C}$, on a*
$$\alpha(A \otimes B) = (\alpha A) \otimes B = A \otimes (\alpha B), \tag{1.231}$$
- *si AC et BD sont bien définis alors*
$$(A \otimes B)(C \otimes D) = AC \otimes BD, \tag{1.232}$$
- *par transposition, on a*
$$(A \otimes B)^T = A^T \otimes B^T. \tag{1.233}$$

- *si* $A \in \mathbb{M}_n(\mathbb{C})$ *et* $B \in \mathbb{M}_p(\mathbb{C})$ *sont inversibles, alors* $A \otimes B$ *est inversible et*

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}. \tag{1.234}$$

- *si* $A \in \mathbb{M}_n(\mathbb{C})$ *et* $B \in \mathbb{M}_p(\mathbb{C})$ *alors*

$$\det(A \otimes B) = \det(A)^p \det(B)^n. \tag{1.235}$$

Proposition 1.24. *Soit* $A \in \mathbb{M}_n(\mathbb{C})$ *et* $p \in \mathbb{N}^*$*, alors les spectres des matrices* $(A \otimes I_p)$ *et* $(I_p \otimes A)$ *sont connus et on a*

- $A \otimes I_p \in \mathbb{M}_{np}(\mathbb{C})$ *et*

$$Sp(A \otimes I_p) = Sp(A), \tag{1.236}$$

- $I_p \otimes A \in \mathbb{M}_{np}(\mathbb{C})$ *et*

$$Sp(I_p \otimes A) = Sp(A), \tag{1.237}$$

où I_p *désigne la matrice identité de* $\mathbb{M}_p(\mathbb{C})$.

Démonstration. Soit $\lambda \in \mathrm{Sp}(A \otimes \mathrm{I}_p)$ alors λ est solution de

$$\det(A \otimes \mathrm{I}_p - \lambda \mathrm{I}_{np}) = 0 \tag{1.238}$$

Donc d'après la proposition 1.23 et comme $\mathrm{I}_n \otimes \mathrm{I}_p = \mathrm{I}_{np}$, on a

$$\begin{aligned} \det((A \otimes \mathrm{I}_p) - \lambda \mathrm{I}_{np}) &= \det(A \otimes \mathrm{I}_p - \lambda \mathrm{I}_n \otimes \mathrm{I}_p) \\ &= \det(A - \lambda \mathrm{I}_n)^p \det(\mathrm{I}_p)^n \\ &= \det(A - \lambda \mathrm{I}_n)^p \end{aligned}$$

Ainsi, on a

$$\lambda \in \mathrm{Sp}(A \otimes \mathrm{I}_p) \Leftrightarrow \lambda \in \mathrm{Sp}(A).$$

De là, il découle l'égalité souhaitée sur les ensembles :

$$\mathrm{Sp}(A \otimes \mathrm{I}_p) = \mathrm{Sp}(A). \tag{1.239}$$

La seconde égalité est obtenue de la même manière.

□

On note vec_2 l'opérateur suivant :

Définition 1.9. *L'opérateur* vec_2 *est défini par*

$$\begin{array}{rccc} vec_2 : L^2_h & \longrightarrow & \mathbb{C}^{N^2} \\ \mathfrak{v} & \longrightarrow & V = vec_2(\mathfrak{v}) \end{array} \tag{1.240}$$

avec

$$vec_2(\mathfrak{v}) = \sum_{i,j=0}^{N-1} (e_j \otimes e_i)\, \mathfrak{v}_{i,j}. \tag{1.241}$$

L'opérateur vec_2 applique une fonction de grille $\mathfrak{v} \in L^2_{h,\mathrm{pér}}$ en un vecteur $V \in \mathbb{C}^{N^2}$ dont les composantes sont déduites de $\mathfrak{v}$ dans l'ordre antilexicographique. Autrement dit $V = \mathrm{vec}_2(\mathfrak{v})$ est donné par

$$V = [\mathfrak{v}_{0,0}, \mathfrak{v}_{1,0}, \mathfrak{v}_{2,0}, \cdots, \mathfrak{v}_{N-1,0}, \mathfrak{v}_{0,1}, \mathfrak{v}_{1,1}, \cdots, \mathfrak{v}_{N-1,1}, \mathfrak{v}_{N-1,N-1}]^T. \tag{1.242}$$

On note vec au lieu de vec_2 quand le contexte est clair.

Proposition 1.25. *Soit* $\mathfrak{u}$ *une fonction de grille. Alors les opérateurs de dérivées centrés* (1.225) *s'écrivent sous forme matricielle*

$$\left\{\begin{array}{rcl} vec(\delta_{2J,x}\mathfrak{u}) & = & (I_N \otimes D_{2J})vec(\mathfrak{u}) \\ vec(\delta_{2J,y}\mathfrak{u}) & = & (D_{2J} \otimes I_N)vec(\mathfrak{u}) \\ vec(\sigma_x\mathfrak{u}) & = & (I_N \otimes P_\sigma)vec(\mathfrak{u}) \\ vec(\sigma_y\mathfrak{u}) & = & (P_\sigma \otimes I_N)vec(\mathfrak{u}) \end{array}\right. \tag{1.243}$$

où D_{2J} *est issu de* (1.107) *et* P_σ *de* (1.147).

Démonstration. Par définition de vec_2, on a

$$\begin{aligned} \mathrm{vec}(\delta_{2J,x}\mathfrak{u}) &= \sum_{i,j=0}^{J} (e_j \otimes e_i)\delta_{2J,x}\mathfrak{u}_{i,j} \\ &= \sum_{i,j=0}^{N-1}\sum_{p=1}^{J}(e_j \otimes e_i)\frac{a_p}{2ph}(\mathfrak{u}_{i+p,j} - \mathfrak{u}_{i-p,j}) \\ &= \sum_{p=1}^{J}\frac{a_p}{2ph}\left(\sum_{i,j=0}^{N-1}(e_j \otimes e_i)\mathfrak{u}_{i+p,j} - (e_j \otimes e_i)\mathfrak{u}_{i-p,j}\right) \\ &= \sum_{p=1}^{J}\frac{a_p}{2ph}\sum_{i,j=0}^{N-1} e_j \otimes (e_i\mathfrak{u}_{i+p,j} - e_i\mathfrak{u}_{i-p,j}) \\ &= \sum_{p=1}^{J}\frac{a_p}{2ph}\left(\mathrm{I}_N \otimes \left(T^p - T^{-p}\right)\right)\mathrm{vec}(\mathfrak{u}) \\ &= (\mathrm{I}_N \otimes D_{2J})\mathrm{vec}(\mathfrak{u}). \end{aligned}$$

Les autres égalités se montrent de la même manière. □

On déduit également $\mathrm{vec}(\delta^H_{2J+2,x}\mathfrak{u})$ et $\mathrm{vec}(\delta^H_{2J+2,y}\mathfrak{u})$. On considère à présent les matrices associées à $\delta^H_{2J+2,x}\mathfrak{u}$ et $\delta^H_{2J+2,y}\mathfrak{u}$.

Théorème 1.6. *Soit* $\mathfrak{u}$ *une fonction de grille. Alors, si on pose*

$$\left\{\begin{array}{rcl} U & = & vec(\mathfrak{u}) \\ U_x & = & vec(\delta^H_{2J+2,x}\mathfrak{u}) \\ U_y & = & vec(\delta^H_{2J+2,y}\mathfrak{u}) \end{array}\right. \tag{1.244}$$

les matrices associées à $\delta^H_{2J+2,x}\mathfrak{u}$ *et* $\delta^H_{2J+2,y}\mathfrak{u}$ *sont telles que*

$$\left\{\begin{array}{rcl} U_x & = & (I_N \otimes P_\sigma^{-1}D_{2J})U \\ U_y & = & (P_\sigma^{-1}D_{2J} \otimes I_N)U. \end{array}\right. \tag{1.245}$$

Démonstration. L'égalité suivante sont vérifiée

$$(P_\sigma \otimes I_N)U_x = (D_{2J} \otimes \mathrm{I}_N)U. \tag{1.246}$$

Donc on obtient

$$\begin{aligned} U_y &= (P_\sigma \otimes I_N)^{-1}(D_{2J} \otimes \mathrm{I}_N)U \\ &= (P_\sigma^{-1}D_{2J} \otimes \mathrm{I}_N^{-1}\mathrm{I}_N)U \\ &= (P_\sigma^{-1}D_{2J} \otimes \mathrm{I}_N)U. \end{aligned}$$

La seconde égalité est obtenue de la même manière. □

Les matrices $(Id \otimes P^{-1}D_{2P})$ et $(P^{-1}D_{2P} \otimes Id)$ sont antisymétriques.

Proposition 1.26. *Les valeurs propres de $\delta_{2J+2,x}$ et de $\delta_{2J+2,y}$ sont*

$$\frac{1}{h}Q^H_{2J+2}(\omega^k) \tag{1.247}$$

pour $-N/2+1 \leq k \leq N/2$.

Démonstration. Il s'agit d'une conséquence de la proposition 1.24. □

1.2.4 Opérateur de filtrage

Dans cette partie, nous utilisons les notations de la section 1.2.1 en contexte périodique. Définissons les opérateurs de filtrage bidimensionnel dans les directions x et y par

$$\left\{\begin{array}{rcl} \mathcal{F}_{2J,x} & = & \displaystyle\sum_{p=0}^{J} a_p \frac{\tau_x^p + \tau_x^{-p}}{2} = S_{2J}(\tau_x), \\ \mathcal{F}_{2J,y} & = & \displaystyle\sum_{p=0}^{J} a_p \frac{\tau_y^p + \tau_y^{-p}}{2} = S_{2J}(\tau_y), \end{array}\right. \tag{1.248}$$

où S_{2J} est donné par (1.165). Les coefficients $(a_p)_{0\leq p\leq J}$ vérifient le système (1.174). Les opérateurs $\mathcal{F}_{2J,x}$ et $\mathcal{F}_{2J,y}$ commutent :

$$\mathcal{F}_{2J,x} \circ \mathcal{F}_{2J,y} = \mathcal{F}_{2J,y} \circ \mathcal{F}_{2J,x}. \tag{1.249}$$

La proposition 1.20 permet de vérifier la consistance des opérateurs de filtrages. Si $u : (x,y) \in \Omega \mapsto u(x,y) \in \mathbb{R}$ est une fonction de $\mathcal{C}^{2J}$ alors :

$$\left\{\begin{array}{rcl} \|u^* - \mathcal{F}_{2J,x}u^*\|_\infty & \leq & C_x h^{2J} \|\partial_x^{(2J)} u\|_\infty \\ \|u^* - \mathcal{F}_{2J,y}u^*\|_\infty & \leq & C_y h^{2J} \|\partial_y^{(2J)} u\|_\infty \end{array}\right. \tag{1.250}$$

où C_x et C_y sont des constantes indépendantes de u et de h. En composant les opérateurs, l'opérateur $\mathcal{F}_{2J,x} \circ \mathcal{F}_{2J,y}$ désigne aussi un opérateur de filtrage au sens où

$$\|u^* - (\mathcal{F}_{2J,x} \circ \mathcal{F}_{2J,y})u^*\|_\infty = Ch^{2J} \max(\|\partial_x^{(2J)} u\|_\infty, \|\partial_y^{(2J)} u\|_\infty). \tag{1.251}$$

où C est un réel indépendant de u et de h.

L'écriture matricielle de l'opérateur de filtrage est donnée par la proposition suivante :

Proposition 1.27. *Soit $\mathfrak{u} \in L^2_{h,pér}$ une fonction de grille. Alors les opérateurs de filtrages vérifient*

$$\left\{\begin{array}{rcl} vec(\mathcal{F}_{2J,x}\mathbf{u}) & = & (I_N \otimes M_{2J}) vec(\mathbf{u}) \\ vec(\mathcal{F}_{2J,y}\mathbf{u}) & = & (M_{2J} \otimes I_N) vec(\mathbf{u}). \end{array}\right. \tag{1.252}$$

Comme la matrice M_{2J} est symétrique, il est immédiat que $\mathrm{I}_N \otimes M_{2J}$ et $M_{2J} \otimes \mathrm{I}_N$ sont symétriques.

Chapitre 2

Analyse numérique des schémas compacts

2.1 Introduction

Dans ce chapitre, nous considérons le cadre général de la procédure d'approximation utilisée à partir du chapitre 3. On considère des équations aux dérivées partielles d'évolution de la forme

$$\frac{\partial q}{\partial t} = J(q) \tag{2.1}$$

dans le contexte périodique. Dans l'équation (2.1), q représente la fonction inconnue et $q \mapsto J(q)$ représente un opérateur différentiel spatial. Les exemples considérés sont

1. L'équation de transport 1D :
$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0, \text{ avec } c \in \mathbb{R}, \tag{2.2}$$

2. l'équation de transport 2D :
$$\frac{\partial u}{\partial t} + c_x\frac{\partial u}{\partial x} + c_y\frac{\partial u}{\partial y} = 0, \text{ avec } c_x, c_y \in \mathbb{R}, \tag{2.3}$$

3. L'équation des ondes avec force de Coriolis :
$$\left\{\begin{array}{rcl} \dfrac{\partial \eta}{\partial t} + H\left(\dfrac{\partial u}{\partial x} + \dfrac{\partial v}{\partial y}\right) &=& 0 \\ \dfrac{\partial u}{\partial t} + g\dfrac{\partial \eta}{\partial x} - fv &=& 0 \\ \dfrac{\partial v}{\partial t} + g\dfrac{\partial \eta}{\partial y} + fu &=& 0, \end{array}\right. \tag{2.4}$$
avec H, g et f dans $\mathbb{R}$.

4. L'équation Shallow Water
$$\left\{\begin{array}{rcl} \dfrac{\partial h}{\partial t} + \left(\dfrac{\partial hu}{\partial x} + \dfrac{\partial hv}{\partial y}\right) &=& 0 \\ \dfrac{\partial u}{\partial t} + u\dfrac{\partial u}{\partial x} + v\dfrac{\partial u}{\partial y} + g\dfrac{\partial h}{\partial x} - fv &=& 0 \\ \dfrac{\partial v}{\partial t} + u\dfrac{\partial v}{\partial x} + v\dfrac{\partial v}{\partial y} + g\dfrac{\partial h}{\partial y} + fu &=& 0, \end{array}\right. \tag{2.5}$$
avec g et f des réels.

A chacune de ces équations on ajoute les conditions utiles au caractère bien posé des problèmes.

La propriété générale de ces équations est l'hyperbolicité qui garantie l'existence d'un régime de type ondes linéaires et ondes non-linéaires.

Dans ce chapitre, nous étudions les propriétés classiques du schéma utilisé : la consistance, la stabilité, la convergence et la conservation. Bien que ces propriétés soient classiques, nous n'avons pas trouvé ces résultats dans la littérature. L'approche envisagée est celle d'une approximation centrée en espace de l'opérateur $J(q)$. On note $J_\Delta(q)$ cette approximation spatiale. La situation typique est celle de l'équation de transport 1D périodique dans laquelle on a

$$\frac{du}{dt} = J(u) \tag{2.6}$$

avec $J(u) = -c\partial_x u$ et $c \in \mathbb{R}$. La donnée $u(t,\cdot)^*$ est approchée par $\mathfrak{u}(t)$ fonction de grille périodique (voir chapitre 1) et J est approché par J_Δ avec

$$J_\Delta(\mathfrak{u}) = -c\delta^H_{4,x}\mathfrak{u}. \tag{2.7}$$

L'inconnue semi-discrète $t \mapsto \mathfrak{u}(t)$ est solution de

$$\frac{d\mathfrak{u}}{dt} = J_\Delta(\mathfrak{u}). \tag{2.8}$$

Dans ce qui suit, on s'occupe principalement de différents points d'analyse numérique (consistance, stabilité, convergence) dans différentes situations relatives aux équations précédentes. Bien que le contexte de cette thèse soit la géométrie sphérique, nous considérons ici le cadre plan et périodique.

2.2 Schémas de Runge-Kutta explicites

2.2.1 Le schéma de Runge-Kutta RK4

Pour résoudre une équation ordinaire de la forme

$$\left\{\begin{array}{rcl} \dfrac{dq}{dt} & = & J_\Delta(q) \\ q(t=0) & = & q_0, \end{array}\right. \tag{2.9}$$

où $t \in \mathbb{R}^+$, on utilise une méthode de Runge-Kutta à $p \in \mathbb{N}^\star$ étapes [15, 27]. Les méthodes de Runge-Kutta sont de la forme

$$\left\{\begin{array}{rcl} K^{(i)} & = & J_\Delta\left(q^n + \Delta t \displaystyle\sum_{j=1}^{p} A_{i,j}K^{(j)}\right) \text{ pour } 1 \leq i \leq p \\ q^{n+1} & = & q^n + \Delta t \displaystyle\sum_{j=1}^{p} b_j K^{(j)}, \end{array}\right. \tag{2.10}$$

on note le pas de temps $\Delta t > 0$, et q^n est une approximation de $q(n\Delta t)$. Les coefficients $(A_{i,j})_{1\leq i,j\leq p}$ et $(b_j)_{1\leq j\leq p}$ sont des réels donnés. Une méthode de Runge-Kutta est dite explicite si $A_{i,j} = 0$ lorsque $j \geq i$. C'est le cas des méthodes considérées dans ce mémoire. L'état q^n étant connu, l'état q^{n+1} s'écrit en fonction de q^n, et non de q^n, q^{n-1}, q^{n-2}, etc., via une relation de la forme

$$q^{n+1} = Q(q^n). \tag{2.11}$$

Le schéma de résolution temporelle de référence que nous considérons est le schéma de Runge-Kutta d'ordre 4 (RK4). Le schéma de Runge-Kutta d'ordre 4 est le schéma de référence d'ordre 4 pour les

problèmes non linéaires convectifs. :

$$\begin{cases} K^{(1)} &= J_\Delta(q^n) \\ K^{(2)} &= J_\Delta\left(q^n + \dfrac{\Delta t}{2}K^{(1)}\right) \\ K^{(3)} &= J_\Delta\left(q^n + \dfrac{\Delta t}{2}K^{(2)}\right) \\ K^{(4)} &= J_\Delta\left(q^n + \Delta t K^{(3)}\right) \end{cases} \tag{2.12}$$

$$q^{n+1} = q^n + \frac{\Delta t}{6}\left(K^{(1)} + 2K^{(2)} + 2K^{(3)} + K^{(4)}\right). \tag{2.13}$$

Il est pratique de représenter une méthode de Runge-Kutta à l'aide d'un tableau de Butcher [15] de la forme de la Table 2.1. Le vecteur $c \in \mathbb{R}^p$ désigne la position de l'approximation, $A \in \mathbb{M}_p(\mathbb{R})$ représente la dépendance entre l'étape donnée et les autres étapes, $b \in \mathbb{R}^p$ est issue de la règle de quadrature associée à la méthode de Runge-Kutta.

c	A
	b^T

TABLE 2.1 – Tableau de Butcher d'une méthode de Runge-Kutta. $A \in \mathbb{M}_p(\mathbb{R})$, $b, c \in \mathbb{R}^p$.

Pour la méthode de Runge-Kutta d'ordre 4 détaillée ici, le tableau de Butcher prend la forme donnée dans la Table 2.2. La matrice A ainsi que les vecteurs b et c associés au tableau de Butcher

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

TABLE 2.2 – Tableau de Butcher de la méthode de Runge Kutta d'ordre 4.

pour RK4 sont

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad b = \begin{bmatrix} 1/6 \\ 1/3 \\ 1/3 \\ 1/6 \end{bmatrix} \qquad c = \begin{bmatrix} 0 \\ 1/2 \\ 1/2 \\ 1 \end{bmatrix}. \tag{2.14}$$

Dans (2.11), en supposant que $q^n = q(n\Delta t)$ alors si $q^{n+1} = Q(q^n)$ satisfait

$$q^{n+1} - q((n+1)\Delta t) = \mathcal{O}(\Delta t^{p+1}) \tag{2.15}$$

on dit que la méthode est d'ordre p.

Proposition 2.1. *La méthode de Runge-Kutta :* (2.12) *et* (2.13) *est d'ordre 4.*

En ce qui concerne la preuve de cette proposition, nous renvoyons à [27].

2.2.2 Stabilité d'un schéma en temps

On considère une équation différentielle ordinaire de la forme

$$\frac{dq}{dt} = J_\Delta(q), \tag{2.16}$$

q^n approchant $q(n\Delta t)$ donné par un schéma en temps de la forme (2.11). Nous considèrons les deux notions de stabilité suivantes :

1. La *stabilité au sens de Von Neumann.* Soit $(q^n)_{0\leq n\leq N}$ (avec $N\Delta t = T < \infty$ un temps fini) une approximation de la solution de (2.16) pour $0 \leq t \leq T$. Le schéma est stable au sens de Von Neumann si il existe $C(T) > 0$ tel que

$$\sup_{N>0, N\Delta t=T} \max_{0\leq n\leq N} |q^n| < C(T). \tag{2.17}$$

2. On considère $q' = J(q)$ discrétisée par un schéma en temps, $q(t) \in \mathbb{C}^p$ pour tout t. Cette fois, Δt est fixé et $n \to +\infty$. On reprend le schéma en temps indépendamment des propriétés de l'équation différentielle considérée. La *stabilité asymptotique* se caractérise par le comportement de la solution numérique q^n sur un intervalle non borné à Δt fixé. On distingue deux types de stabilités asymptotiques en supposant q^n calculable pour tout n :

   - la suite $(|q^n|)_{n\in\mathbb{N}}$ est bornée,
   - La suite $|q^n| \to 0$ lorsque $n \to +\infty$.

   Noter que $|\cdot|$ est la norme usuelle sur $\mathbb{C}^p$.

Ces deux définitions de stabilité sont illustrées dans le contexte particulier de l'équation de Dahlquist qui joue un rôle essentiel :

$$\left\{ \begin{array}{rcl} q' & = & \lambda q \\ q(0) & = & q_0. \end{array} \right. \tag{2.18}$$

avec $\lambda \in \mathbb{C}$ et q_0 non nul. La solution de cette équation est explicitement donnée par

$$q(t) = e^{\lambda t} q_0 \tag{2.19}$$

On remarque que

$$\lim_{+\infty} q(t) = 0 \text{ si et seulement si } \mathrm{Re}(\lambda) < 0 \tag{2.20}$$

ainsi que

$$\lim_{+\infty} |q(t)| = +\infty \text{ si et seulement si } \mathrm{Re}(\lambda) > 0. \tag{2.21}$$

Nous considérons les méthodes de Runge-Kutta explicites donc A est triangulaire inférieure stricte. Si l'on note $K = [K^{(1)}, K^{(2)}, ...]^T$. Dans le cadre de l'équation de Dahlquist (2.18) [42], K est lié à q^n par

$$K^{(i)} = \lambda \left(q^n + \Delta t \sum_{j=1}^{i-1} A_{i,j} K^{(k)} \right) \tag{2.22}$$

pour tout i. Cette égalité s'écrit sous forme matricielle :

$$K = \lambda q^n + \lambda \Delta t A K. \tag{2.23}$$

La matrice $\mathrm{I} - \Delta t \lambda A$ est triangulaire inférieure avec des 1 sur la diagonale. Il s'agit d'une matrice inversible et on a $K = \lambda(\mathrm{I} - \Delta t \lambda A)^{-1} q^n$. Alors

$$q^{n+1} = q^n + \Delta t b^T \cdot K,$$

donc on obtient

$$q^{n+1} = q^n + \lambda \Delta t b^T \cdot (\mathrm{I} - \lambda \Delta t A)^{-1} q^n. \tag{2.24}$$

On considère une méthode de Runge-Kutta appliquée à la résolution de l'équation (2.18). En utilisant les notations du tableau de Butcher 2.1, on a

$$q^{n+1} = R(\lambda \Delta t) q^n, \tag{2.25}$$

où R est donné par

$$R(z) = 1 + z b^T \cdot \left((\mathrm{I} - zA)^{-1} \mathbf{1} \right). \tag{2.26}$$

on désigne par $\mathbf{1}$ le vecteur ne contenant que des 1.

Définition 2.1. *On appelle fonction de stabilité d'une méthode de Runge-Kutta la fonction* $R : z \in \mathbb{C} \mapsto R(z) \in \mathbb{C}$ *donnée par*

$$R(z) = 1 + zb^T \cdot \left((I - zA)^{-1}\mathbf{1}\right). \tag{2.27}$$

Pour la méthode de Runge-Kutta d'ordre 4, on a

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} \tag{2.28}$$

alors le résultat de stabilité est le suivant :

Proposition 2.2. *La méthode de Runge-Kutta d'ordre 4 (RK4) est asymptotiquement stable pour l'équation* (2.18) *si et seulement si*

$$|1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}| \leq 1, \tag{2.29}$$

avec $z = \lambda\Delta t$.

La condition $|R(z)|^2 = 1$ définit implicitement une courbe décrivant la frontière d'un compact. Par le théorème des fonctions implicites, cette frontière est régulière et fermée. La zone de stabilité de RK4 est constituée de l'intérieur de cette courbe. On définit la zone de stabilité de RK4, notée $\mathcal{D}_{\text{RK4}}$, par

$$\mathcal{D}_{\text{RK4}} = \left\{ z \in \mathbb{C} \text{ tels que } |1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}| \leq 1 \right\}. \tag{2.30}$$

On représente $\mathcal{D}_{\text{RK4}}$ sur la Figure 2.1.

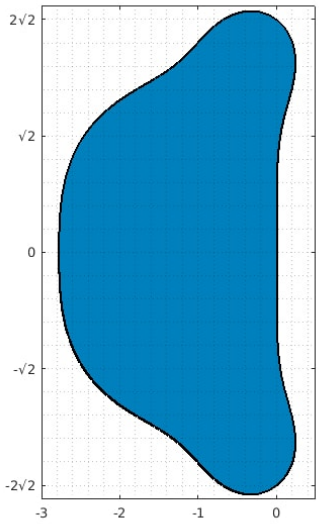


FIGURE 2.1 – Zone de stabilité de la méthode de Runge-Kutta d'ordre 4 : $\mathcal{D}_{\text{RK4}}$.

En particulier, comme détaillé dans [49], on a

$$\mathcal{D}_{\text{RK4}} \cap i\mathbb{R} = i\left[-2\sqrt{2}, 2\sqrt{2}\right]. \tag{2.31}$$

Lemme 2.1. *Soit* $(e_n)_{n\in\mathbb{N}}$ *une suite réelle. On suppose qu'il existe* $a \in [0,1]$ *et* $b > 0$ *tels que*

$$e_{n+1} \leq ae_n + b \tag{2.32}$$

pour tout $n \in \mathbb{N}$, *alors on a*

$$e_n \leq a^n e_0 + nb. \tag{2.33}$$

Démonstration. Montrons la propriété par récurrence. Initialement, on a directement

$$e_1 \leq a e_0. \tag{2.34}$$

Supposons qu'il existe $n \in \mathbb{N}$ tel que

$$e_n \leq a^n e_0 + nb. \tag{2.35}$$

Au rang $n+1$, on a

$$\begin{aligned} e_{n+1} &\leq a e_n + b \\ &\leq a\left(a^n e_0 + nb\right) + b \text{ par hypothèse de récurrence,} \\ &\leq a^{n+1} e_0 + anb + b \\ &\leq a^{n+1} e_0 + (n+1)b \text{ car } 0 \leq a \leq 1. \end{aligned}$$

D'après ce raisonnement par récurrence, pour tout $n \in \mathbb{N}$,

$$e_n \leq a^n e_0 + nb, \tag{2.36}$$

ce qui conclut la preuve. □

Proposition 2.3. *Supposons que $\Delta t = T/N$, $N \in \mathbb{N}$, est choisi tel que $|R(\lambda \Delta t)| \leq 1$ (RK4 est asymptotiquement stable). Si (q^n) est la solution approchée par RK4 de (2.18) et $q(t)$ la solution exacte en $t \in [0,T]$, alors il existe $C > 0$ indépendant de q et de Δt tel que*

$$e^n = |q^n - q(t^n)| \leq CT\Delta t^4 \max_{0 \leq t \leq T} |q(t)|, \tag{2.37}$$

pour $0 \leq n \leq N$.

Démonstration. Soit $n \leq N$, alors

$$\begin{aligned} e_{n+1} &= |q^{n+1} - q(t^{n+1})| \\ &= |R(\lambda\Delta t) q^n - e^{\lambda \Delta t} q(t^n)| \\ &= |R(\lambda\Delta t) q^n - R(\lambda\Delta t) q(t^n) + R(\lambda\Delta t) q(t^n) - e^{\lambda \Delta t} q(t^n)| \\ &\leq |R(\lambda\Delta t)| e_n + |R(\lambda\Delta t) - e^{\lambda\Delta t}||q(t^n)|. \end{aligned}$$

Or, d'après la formule de Taylor-Lagrange, il existe ξ tel que

$$R(\lambda\Delta t) - e^{\lambda\Delta t} = \frac{\lambda^5 \Delta t^5}{120} e^{\xi}. \tag{2.38}$$

Comme $\mathcal{D}_{\mathrm{RK4}}$ est un compact de $\mathbb{C}$, on a

$$|R(\lambda\Delta t) - e^{\lambda\Delta t}| \leq \frac{|\lambda|^5 \Delta t^5}{720} \max_{z \in \mathcal{D}_{\mathrm{RK4}}} |e^z| = C\Delta t^5.$$

De plus, par hypothèse $|R(\lambda\Delta t)| \leq 1$, donc d'après le lemme 2.1, pour tout $n \leq N$, on a

$$\begin{aligned} e_n &\leq |R(\lambda\Delta t)|^n \underbrace{e_0}_{=0} + \underbrace{n\Delta t}_{\leq T} \Delta t^4 C \max_{0 \leq t \leq T} |q(t)| \\ &\leq CT\Delta t^4 \max_{0 \leq t \leq T} |q(t)|. \end{aligned}$$

Ce qui prouve la formule souhaitée. □

2.2.3 Schémas de Runge-Kutta pour les systèmes d'équations différentielles

La notion de stabilité appliquée à l'équation de Dahlquist (2.18) se généralise aux systèmes d'équations linéaires

$$\left\{ \begin{array}{rcl} \mathbf{q}' & = & J\mathbf{q} \\ \mathbf{q}(0) & = & \mathbf{q}_0 \end{array} \right. \tag{2.39}$$

où $\mathbf{q} : t > 0 \mapsto \begin{bmatrix} q_1(t) & q_2(t) & \cdots & q_j(t) & \cdots & q_N(t) \end{bmatrix}^T \in \mathbb{R}^N$. q_j désigne une fonction de $\mathbb{R}^+$ dans $\mathbb{C}$, $J \in \mathbb{M}_N(\mathbb{C})$ désigne une matrice carrée. On se restreint au cas où J est une matrice diagonalisable. Cette dernière propriété donne

$$\mathbf{q}(t) = e^{Jt}\mathbf{q}_0 \text{ avec } t \in \mathbb{R}^+. \tag{2.40}$$

Comme J est diagonalisable, il existe $P \in \mathbb{M}_N(\mathbb{R})$ une matrice de passage et $\Lambda \in \mathbb{M}_N(\mathbb{R})$ matrice diagonale (contenant les valeurs propres de J) telles que

$$J = P^{-1}\Lambda P. \tag{2.41}$$

De là, il vient que

$$e^{Jt} = P^{-1}e^{\Lambda t}P, \tag{2.42}$$

donc la solution de (2.39) vérifie l'égalité

$$\mathbf{q}(t) = e^{Jt}\mathbf{q}_0 = P^{-1}e^{\Lambda t}P\mathbf{q}_0. \tag{2.43}$$

A présent, considérons que l'équation (2.39) est discrétisée via la méthode de Runge Kutta d'ordre 4 de l'algorithme 3. On obtient alors

$$\mathbf{q}^{n+1} = R(\Delta t J)\mathbf{q}^n. \tag{2.44}$$

Comme J est diagonalisable, on a même

$$\mathbf{q}^{n+1} = P^{-1}R(\Delta t\Lambda)P\mathbf{q}^n. \tag{2.45}$$

On déduit la relation liant $\mathbf{q}^n$ à la condition initiale $\mathbf{q}_0$:

$$\mathbf{q}^n = R(\Delta t J)\mathbf{q}_0 = P^{-1}R(\Delta t\Lambda)P\mathbf{q}_0. \tag{2.46}$$

Le schémas utilisé est dit linéairement stable si $(\|\mathbf{q}^n\|)_{n\in\mathbb{N}}$ est une suite bornée. Comme J est diagonalisable, cela revient à dire que $|R(\Delta t\lambda)| \leq 1$ pour tout $\lambda \in \mathrm{Sp}(J)$.

Proposition 2.4. *Le schéma RK4 est stable sous la condition*

$$\forall \lambda \in Sp(J),\ |R(\lambda\Delta t)| \leq 1, \tag{2.47}$$

c'est à dire

$$\rho(R(\Delta t J)) \leq 1. \tag{2.48}$$

Cette condition est équivalente à

$$Sp(R(\Delta t J)) \subset \mathcal{D}_{RK4}. \tag{2.49}$$

L'algorithme de RK4 s'écrit :

Algorithme 3 : RK4

$q^0 = q_0$ connu,
for $n = 0, 1, \ldots$ **do**
  $K^{(1)} = J_\Delta\left(q^n\right)$,
  $K^{(2)} = J_\Delta\left(q^n + \dfrac{\Delta t}{2} K^{(1)}\right)$,
  $K^{(3)} = J_\Delta\left(q^n + \dfrac{\Delta t}{2} K^{(2)}\right)$,
  $K^{(4)} = J_\Delta\left(q^n + \Delta t K^{(3)}\right)$,
  $q^{n+1} = q^n + \dfrac{\Delta t}{6}\left(K^{(1)} + 2K^{(2)} + 2K^{(3)} + K^{(4)}\right)$.
end for

2.3 Equation d'advection en dimension 1

Dans cette section, on considère l'équation de transport à vitesse constante $c > 0$,

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0 \text{ avec } 0 \leq x \leq L,\ t > 0. \tag{2.50}$$

La donnée initiale est $u(t = 0, x) = u_0(x)$. On se place dans le cadre des fonctions $L-$périodiques donc $u(t, 0) = u(t, L)$ pour tout $t \geq 0$. La solution exacte de cette équation est

$$u(t, x) = u_0(x - ct) \text{ avec } x \in \mathbb{R},\ t > 0. \tag{2.51}$$

On introduit le développement de Fourier de u_0. Pour tout $x \in [0, L]$, on a

$$u_0(x) = \sum_{k\in\mathbb{Z}} \hat{u}_0^k \exp\left(\frac{2i\pi k}{L} x\right) \tag{2.52}$$

avec $(\hat{u}_0^k)_{k\in\mathbb{Z}}$ les coefficients de Fourier donnés par

$$\hat{u}_0^k = \frac{1}{L}\int_0^L u_0(\tau) \exp\left(-\frac{2i\pi k}{L}\tau\right) d\tau. \tag{2.53}$$

Ainsi, le développement en série de Fourier de u est donné par

$$u(t, x) = u_0(x - ct) \tag{2.54}$$

$$= \sum_{k\in\mathbb{Z}} \exp\left(\frac{2i\pi k}{L}(x - ct)\right) \hat{u}_0^k \tag{2.55}$$

$$= \frac{1}{L}\sum_{k\in\mathbb{Z}} \exp\left(\frac{2i\pi k}{L}(x - ct)\right) \cdot \left(\int_0^L u_0(\tau) \exp\left(-\frac{2i\pi k}{L}\tau\right) d\tau\right). \tag{2.56}$$

Par ailleurs, on se s'intéresse à présent à la conservation de l'énergie. On se retreint au cas $u_0 \in \mathcal{C}^1(\mathbb{R}) \cap L^\infty(\mathbb{R}) \cap L^2(\mathbb{R})$. Quand on multiplie (2.50) par u et qu'on intègre, on obtient

$$\begin{aligned}
\frac{d}{dt}\|u(t,\cdot)\|_{L^2}^2 &= \frac{d}{dt}\int_0^L |u(t,x)|^2 dx \\
&= 2\int_0^L u(t,x)\frac{\partial u}{\partial t}(t,x)dx \\
&= -2c\int_0^L u(t,x)\frac{\partial u}{\partial x}(t,x)dx \\
&= -c\int_0^L \frac{\partial}{\partial x}|u(t,x)|^2 dx \\
&= 0.
\end{aligned}$$

Donc l'énergie $t \in \mathbb{R}^+ \mapsto \|u(t,\cdot)\|_{L^2}^2$ est conservée au fil du temps. Pour interpréter la conservation de l'énergie, on remarque que :

$$\begin{aligned}\hat{u}^k &= \frac{1}{L}\int_0^L u_0(\tau - ct)\exp\left(-\frac{2i\pi k}{L}\tau\right)d\tau \text{ en utilisant (2.51)}\\ &= \frac{1}{L}\int_0^L u_0(\sigma)\exp\left(-\frac{2i\pi k}{L}\sigma\right)d\sigma \cdot \exp\left(-\frac{2i\pi k}{L}ct\right) \text{ par } L-\text{périodicité et en posant } \sigma = \tau - ct,\\ &= \hat{u}_0^k \exp\left(-\frac{2i\pi k}{L}ct\right).\end{aligned}$$

En particulier, on note que

$$|\hat{u}^k|^2 = |\hat{u}_0^k|, \tag{2.57}$$

donc le module de chaque mode est conservé. Cette relation et l'égalité de Parseval permettent d'écrire :

$$\begin{aligned}\|u(t,\cdot)\|_{L^2}^2 &= \sum_{k\in\mathbb{Z}} |\hat{u}^k(t)|^2 \text{ par égalité de Parseval,}\\ &= \sum_{k\in\mathbb{Z}} |\hat{u}_0^k|^2 \text{ par (2.57),}\\ &= \|u_0\|_{L^2}^2 \text{ par égalité de Parseval.}\end{aligned}$$

Et cette relation permet d'interpréter la conservation de l'énergie à l'aide des coefficients de Fourier.

2.3.1 Discrétisation en espace et en temps

L'équation d'advection (2.50) sert d'exemple prototype au schéma qui sera utilisé sur la sphère. Cette équation est discrétisée en espace et en temps en utilisant la méthode des lignes qui consiste à discrétiser dans un premier temps en espace et dans un second temps, discrétiser l'équation en temps. On approche l'opérateur $-c(\partial_x u)^*$ par $-c\delta_{4,x}^H u^*$. Cette approche est inspirée de l'aéroacoustique dans laquelle on discrétise en espace puis en temps. Lors de la discrétisation en temps, on ajoute une étape de filtrage. On cherche alors à déterminer $t > 0 \mapsto \mathfrak{u}(t)$ une approximation de $t > 0 \mapsto u(t,\cdot)^*$ et solution de

$$\left\{\begin{array}{rcl}\dfrac{d\mathfrak{u}}{dt} &=& -c\delta_{4,x}^H \mathfrak{u},\\ \mathfrak{u}_{|t=0} &=& u_0^*.\end{array}\right. \tag{2.58}$$

La version matricielle est déduite de (2.58) par l'opérateur vec. On note

$$U = \mathrm{vec}(\mathfrak{u}(t)) = \begin{bmatrix}\mathfrak{u}_1(t)\\ \mathfrak{u}_2(t)\\ \vdots\\ \mathfrak{u}_N(t)\end{bmatrix} \in \mathbb{R}^N \text{ et } U_0 = \mathrm{vec}(u_0^*) = \begin{bmatrix}u_0(x_0)\\ u_0(x_1)\\ \vdots\\ u_0(x_{N-1})\end{bmatrix} \in \mathbb{R}^N. \tag{2.59}$$

En appliquant vec à (2.58), on obtient

$$\left\{\begin{array}{rcl}\dfrac{dU}{dt} &=& -cP_\sigma^{-1}D_2 U\\ U_{|t=0} &=& U_0,\end{array}\right. \tag{2.60}$$

où $P_\sigma \in \mathbb{M}_N(\mathbb{R})$ est donné par (1.147) avec $\beta = 1/6$ et $D_2 \in \mathbb{M}_N(\mathbb{R})$ est donné par (1.109). La solution de (2.60) est

$$U(t) = \exp\left[-cP_\sigma^{-1}D_2 t\right] U_0. \tag{2.61}$$

On a vu dans la proposition 1.13 que $P_\sigma^{-1}D_2 \in \mathbb{M}_N(\mathbb{R})$ admet N valeurs propres distinctes. $P_\sigma^{-1}D_2$ est diagonalisable, il existe $V \in \mathbb{M}_N(\mathbb{R})$ inversible et $\Lambda \in \mathbb{M}_N(\mathbb{R})$ diagonale telle que

$$P_\sigma^{-1}D_2 = V\Lambda\bar{V}^T. \tag{2.62}$$

Les matrices V et Λ sont données par

$$\Lambda = \begin{bmatrix} \lambda_{-N/2+1} & & & \\ & \lambda_{-N/2+2} & & (0) \\ & & \ddots & \\ & (0) & & \lambda_{N/2-1} \\ & & & & \lambda_{N/2} \end{bmatrix} \text{ et } V = \operatorname{col}\left(U^{-N/2+1}, U^{-N/2+2}, \cdots, U^{N/2-1}, U^{N/2}\right), \tag{2.63}$$

où pour tout $-N/2+1 \leq k \leq N/2$, U^k vérifie

$$U^k = \operatorname{vec}_1(\mathfrak{u}^k) \tag{2.64}$$

et les valeurs propres associées $\lambda_k = \dfrac{1}{h} Q_4^H(\omega^k)$ sont les valeurs propres de $\delta_{4,x}^H$ données par la proposition 1.13. On rappelle que Q_4^H est donné par (1.133).

Donc l'équation (2.61) se réécrit

$$U(t) = V \exp\left[-c\Lambda t\right] \bar{V}^T U_0. \tag{2.65}$$

Ainsi, on a

$$U_j(t) = \sum_{k=1}^{N} \sum_{p=1}^{N} V_{j,k} \exp\left(-c\Lambda_{k,k} t\right) \bar{V}_{p,k} (U_0)_p. \tag{2.66}$$

Comme $\mathfrak{u}(t) = \operatorname{vec}^{-1}(U(t))$, en remplaçant chaque composante par sa valeur, on obtient

$$\mathfrak{u}_j(t) = \sum_{k=-N/2+1}^{N/2} \exp\left(\frac{2ik\pi}{L}\left(x_j - ct\frac{L\lambda_k}{2i\pi k}\right)\right) \underbrace{\frac{h}{L} \sum_{p=0}^{N-1} \exp\left(-\frac{2i\pi k}{L} x_p\right) u_0(x_p)}_{\approx \hat{u}_0^k}. \tag{2.67}$$

En comparant les équations (2.56) et (2.67), on constate que (2.67) représente une série de Fourier tronquée de (2.56) en tenant compte de $\lambda_k = \dfrac{1}{h} Q_4^H(\omega^k)$. On a en effet

$$Q_4^H(\omega^k) = \frac{2i\pi k}{L} + \mathcal{O}\left(h^5\right) \tag{2.68}$$

d'après la proposition 1.14, donc

$$\frac{L\lambda_k}{2i\pi k} = 1 + \mathcal{O}(h^4). \tag{2.69}$$

De plus, la solution vectorielle $U(t)$ de (2.60) vérifie

$$\begin{aligned} \frac{d}{dt}\|U(t)\|_2^2 &= 2U(t)^T \cdot \frac{dU(t)}{dt} \\ &= -2cU(t)^T \cdot \left(P_\sigma^{-1} D_2 U(t)\right) \\ &= 0 \text{ car } P_\sigma^{-1} D_2 \text{ est antisymétrique (proposition 1.17).} \end{aligned}$$

Ainsi, l'énergie $t \mapsto \|\mathfrak{u}(t)\|_{h,\text{pér}}^2 = h^2 \|U(t)\|_2^2$ est conservée.

Le schéma semi-discrétisé (2.58) est dispersif mais pas dissipatif. La dissipation du schéma complètement discrétisé est liée à la discrétisation en temps.

L'approximation de l'opérateur en espace ∂_x par l'opérateur $\delta_{4,x}^H$ introduit une erreur dans le calcul de la solution. L'estimation suivante de l'erreur est obtenue :

Proposition 2.5. *L'erreur entre $u(t,\cdot)^*$ et $\mathfrak{u}(t)$ est donnée pour $t \in [0,T]$ en norme $\|\cdot\|_{h,pér}$ par l'estimation suivante :*

$$\|\mathfrak{u}(t) - u(t,\cdot)^*\|_{h,pér} \leq \tilde{C}c\sqrt{L}th^4\|\partial_x^{(5)}u(t,\cdot)\|_{\infty,[0,T]\times[0,L]} \tag{2.70}$$

où $\tilde{C} > 0$ est une constante indépendante de h et de u avec $\tilde{C} < 0.3210$.

Démonstration. On pose la fonction de grille d'erreur $\mathfrak{e}(t) = u(t,\cdot)^* - \mathfrak{u}(t)$ et $\tau(t) = \delta_{4,x}^H u(t,\cdot)^* - (\partial_x u(t,\cdot))^*$, alors

$$\begin{aligned}\frac{d\mathfrak{e}}{dt}(t) &= \left(\frac{\partial u}{\partial t}(t,\cdot)\right)^* - \frac{d\mathfrak{u}}{dt}(t)\\ &= -c\left(\frac{\partial u}{\partial x}(t,\cdot)\right)^* + c\delta_{4,x}^H\mathfrak{u}(t)\\ &= c\tau(t) - c\delta_{4,x}^H\mathfrak{e}(t).\end{aligned}$$

On a vu que $\delta_{4,x}^H$ est un opérateur antisymétrique d'où

$$(\delta_{4,x}^H\mathfrak{e}(t), \mathfrak{e}(t))_{h,\text{pér}} = 0. \tag{2.71}$$

De ce dernier résultat, on déduit :

$$(\frac{d\mathfrak{e}}{dt}(t), \mathfrak{e}(t))_{h,\text{pér}} = c(\tau(t), \mathfrak{e}(t))_{h,\text{pér}} \tag{2.72}$$

Or, pour tout $\alpha > 0$, et pour toutes fonctions de grille $\mathfrak{b}_1, \mathfrak{b}_2 \in l^2_{h,\text{pér}}$, on a

$$\alpha\|\mathfrak{b}_1\|^2_{h,\text{pér}} + \frac{1}{\alpha}\|\mathfrak{b}_2\|^2_{h,\text{pér}} \geq 2|(\mathfrak{b}_1, \mathfrak{b}_2)_{h,\text{pér}}|^2. \tag{2.73}$$

En utilisant les propriétés de la dérivation, on a

$$\begin{aligned}2\frac{d}{dt}\|\mathfrak{e}(t)\|^2_{h,\text{pér}} &= 2c(\tau(t), \mathfrak{e}(t))_{h,\text{pér}}\\ &\leq 2c|(\tau(t), \mathfrak{e}(t))_{h,\text{pér}}|\\ &\leq c\left[\alpha\|\tau(t)\|^2_{h,\text{pér}} + \frac{1}{\alpha}\|\mathfrak{e}(t)\|^2_\infty\right]\\ &\leq c\alpha\|\tau(t)\|^2_{h,\text{pér}} + \frac{c}{\alpha}\|\mathfrak{e}(t)\|^2_{h,\text{pér}}\end{aligned}$$

pour tout $\alpha > 0$. D'après le lemme de Gronwall, si $y'(t) \leq ay(t) + b$ alors $y(t) \leq y_0e^{at} - b/a$. Donc on a

$$\|\mathfrak{e}(t)\|^2_{h,\text{pér}} \leq \alpha^2 \max_{t\in[0,T]}\|\tau\|^2_{h,\text{pér}}\left(\exp\left(\frac{ct}{\alpha}\right) - 1\right). \tag{2.74}$$

Évaluons $\|\tau(t)\|^2_{h,\text{pér}}$:

$$\begin{aligned}\|\tau(t)\|^2_{h,\text{pér}} = h\sum_{j=0}^{N-1}|\tau_j(t)|^2 &\leq hN\|\tau(t)\|^2_\infty\\ &\leq L\|\tau(t)\|^2_\infty\\ &\leq LC^2h^8\|\partial_x^{(5)}u(t,\cdot)\|^2_\infty \text{ d'après le théorème 1.3.}\end{aligned}$$

De là, on déduit

$$\|\mathfrak{e}(t)\|^2_{h,\text{pér}} \leq \alpha^2LC^2h^8\|\partial_x^{(5)}u(t,\cdot)\|^2_{\infty,[0,T]\times[0,L]}\max_{t\in[0,T]}\left(\exp\left(\frac{ct}{\alpha}\right) - 1\right). \tag{2.75}$$

Pour $t > 0$ fixé, on pose $\beta = \alpha/(ct) \in \mathbb{R}^+$, on a

$$\|\mathfrak{e}(t)\|^2_{h,\text{pér}} \leq LC^2h^8\|\partial_x^{(5)}u(t,\cdot)\|^2_{\infty,[0,T]\times[0,L]}c^2t^2\beta^2\left(\exp\left(\frac{1}{\beta}\right)-1\right) \tag{2.76}$$

On minimise à présent la fonction a définie par

$$a : \beta \in]0,+\infty[\mapsto \beta^2\left(\exp\left(\frac{1}{\beta}\right)-1\right). \tag{2.77}$$

La fonction a est continue, positive et de plus :

$$\lim_{\beta\to+\infty} a(\beta) = \lim_{\beta\to 0^+} a(\beta) = +\infty \tag{2.78}$$

donc a admet un minimum $m > 0$. De là, il découle que

$$\|\mathfrak{e}(t)\|^2_{h,\text{pér}} \leq LC^2mc^2t^2h^8\|\partial_x^{(5)}u(t,\cdot)\|^2_{\infty,[0,T]\times[0,L]}.$$

L'estimation est obtenue en prenant la racine carré de cette équation. De plus, on a vu que $C = 1/15$ convient. Et p, vérifie que $m < 1.545$, donc $\tilde{C} = \sqrt{mC} < \sqrt{\dfrac{1.545}{15}} < 0.3210$. □

La présence du temps t dans le terme d'erreur permet de mettre en évidence la détérioration linéaire de l'erreur au fil du temps.

L'équation semi discrétisée de l'équation de transport (2.50) est :

$$\left\{\begin{array}{rcl} \dfrac{d\mathfrak{u}}{dt} + c\delta^H_{4,x}\mathfrak{u} &=& 0 \\ \mathfrak{u}(0) &=& u_0^* \end{array}\right. \text{ avec } t \in [0,T]. \tag{2.79}$$

Elle est résolue en utilisant la méthode de Runge-Kutta d'ordre 4. L'algorithme de résolution est donné par l'algorithme 4 dans lequel on a ajouté une étape de filtrage au schéma RK4 Comme indiqué ci-dessous, les expériences montrent qu'un opérateur de filtrage [84] est utile.

Algorithme 4 : Schémas en temps RK4 avec étape de filtrage pour l'équation (2.79)

$\mathfrak{u}^0 = u_0^*$ connu,
for $n = 0, 1, \ldots$ **do**
  $K^{(1)} = -c\delta^H_{4,x}(\mathfrak{u}^n)$,
  $K^{(2)} = -c\delta^H_{4,x}\left(\mathfrak{u}^n + \dfrac{\Delta t}{2}K^{(1)}\right)$,
  $K^{(3)} = -c\delta^H_{4,x}\left(\mathfrak{u}^n + \dfrac{\Delta t}{2}K^{(2)}\right)$,
  $K^{(4)} = -c\delta^H_{4,x}\left(\mathfrak{u}^n + \Delta t K^{(3)}\right)$,
  $\mathfrak{u}^{n+1} = \mathcal{F}_{2J,x}\left(\mathfrak{u}^n + \dfrac{\Delta t}{6}\left(K^{(1)} + 2K^{(2)} + 2K^{(3)} + K^{(4)}\right)\right)$.
end for

Dans cet algorithme, $\Delta t > 0$ désigne le pas de temps. $\mathfrak{u}^n$ est une approximation de $\mathfrak{u}(n\Delta t)$.

Comme on combine un opérateur d'ordre 4 en temps et en espace, il est clair que pour $J \geq 2$, le schéma global est d'ordre 4 en espace et en temps. C'est à dire

$$\max_{0\leq n\leq N}\|\mathfrak{u}^n - u(t^n,\cdot)^*\|_\infty = \mathcal{O}\left(h^4,\Delta t^4\right). \tag{2.80}$$

2.3.2 Étude de stabilité

On s'intéresse dans cette section à la stabilité asymptotique matricielle [15, 48]. On exprime $U^{n+1} = \mathrm{vec}(\mathfrak{u}^{n+1})$ en fonction de $U^n = \mathrm{vec}(\mathfrak{u}^n)$ obtenus par l'algorithme 4. On montre que

$$\begin{aligned} U^{n+1} &= M_{2J}\left(1 - c\Delta t P_\sigma^{-1}D_2 + \frac{(c\Delta t P_\sigma^{-1}D_2)^2}{2} - \frac{(c\Delta t P_\sigma^{-1}D_2)^3}{6} + \frac{(c\Delta t P_\sigma^{-1}D_2)^4}{24}\right)U^n \\ &= S_{2J,x}(T)R(-c\Delta t P_\sigma^{-1}D_2)U^n \text{ avec } R \text{ exprimé par l'équation (2.28).} \\ &= S_{2J,x}(T)R\left(-\lambda Q_4^H(T)\right)U^n \end{aligned}$$

où $M_{2J} = S_{2J}(T)$ est la matrice associée au filtrage $\mathcal{F}_{2J,x}$ donnée par (1.210) et (1.165) en notant T la matrice de translation (1.46). $P_\sigma^{-1}D_2 = \dfrac{1}{h}Q_4^H(T)$ est donné par (1.132). On pose $\lambda = c\Delta t/h$.

D'après la proposition 2.4, l'algorithme 4 est asymptotiquement stable si et seulement si

$$\mathrm{Sp}\left(S_{2J}(T)R(-\lambda Q_4^H(T))\right) \subset \mathcal{D}_{\mathrm{RK4}}. \tag{2.81}$$

La fonction $\lambda \mapsto \mathrm{Sp}\left(S_{2J}(T)R\left(-\lambda Q_4^H(T)\right)\right)$ dépend continûment de λ et tend vers l'infini lorsque λ tend vers l'infini. En remarquant que $\mathcal{D}_{\mathrm{RK4}}$ est un compact connexe de $\mathbb{C}$, il existe λ_{2J} tel que si $\lambda > \lambda_{2J}$, alors

$$\mathrm{Sp}\left(S_{2J}(T)R\left(-\lambda Q_4^H(T)\right)\right) \not\subset \mathcal{D}_{\mathrm{RK4}}. \tag{2.82}$$

En l'absence d'opérateur de filtrage, on note λ_∞ tel que si $\lambda > \lambda_\infty$ alors

$$\mathrm{Sp}\left(R\left(-\lambda Q_4^H(T)\right)\right) \not\subset \mathcal{D}_{\mathrm{RK4}}. \tag{2.83}$$

Le schéma est stable sous la condition

$$\lambda = \frac{c\Delta t}{h} \leq \lambda_{2J}. \tag{2.84}$$

Il s'agit d'une condition de Courant–Friedrichs–Lewy [23] (notée condition CFL). On a vu que les matrices $S(T)$ et $Q_4^H(T)$ sont diagonalisables, donc en considérant V la matrice donnée par (2.63) et en notant

$$\Omega = \begin{bmatrix} \omega^{-N/2+1} & & & \\ & \omega^{-N/2+2} & & (0) \\ & & \ddots & \\ & (0) & & \omega^{N/2-1} \\ & & & & \omega^{N/2} \end{bmatrix} \in \mathbb{M}_N(\mathbb{C}), \tag{2.85}$$

on déduit

$$\begin{aligned} S_{2J}(T) &= V S_{2J}(\Omega)\bar{V}^T \\ Q_4^H(T) &= V Q_4^H(\Omega)\bar{V}^T \end{aligned} \tag{2.86}$$

d'où

$$S_{2J}(T)R(-\lambda Q_4^H(T)) = V\left[S_{2J}(\Omega)R(-\lambda Q_4^H(\Omega))\right]\bar{V}^T. \tag{2.87}$$

Les valeurs propres de $S_{2J}(T)R(-\lambda Q_4^H(T))$ sont donc

$$S(\omega^k)R(-\lambda Q_4^H(\omega^k)) \text{ pour } -N/2+1 \leq k \leq N/2. \tag{2.88}$$

Considérons dans un premier temps le cas sans filtrage ($\mathcal{F} = \mathrm{Id}$). On a alors

$$\mathrm{Sp}\left(R\left(-\lambda Q_4^H(T)\right)\right) = \left\{R\left(-\lambda Q_4^H(\omega^k)\right) \text{ avec } -N/2+1 \leq k \leq N/2\right\}. \tag{2.89}$$

Or, pour tout $-N/2+1 \leq k \leq N/2$, l'application

$$\lambda \mapsto R\left(-\lambda Q_4^H(\omega^k)\right) \tag{2.90}$$

est polynomiale en λ. Donc

$$\left\{\lambda \in \mathbb{R}^+ \text{ tels que } \mathrm{Sp}\left(R\left(-\lambda Q_4^H(T)\right)\right) \subset \mathcal{D}_{\mathrm{RK4}}\right\} = \left\{\lambda \in \mathbb{R}^+ \text{ tels que } \max_{0\leq\theta\leq\pi}\left(|R(-\lambda Q_4^H(e^{i\theta})| \leq 1\right)\right\} \tag{2.91}$$

est borné et on définit λ_∞ par :

$$\lambda_\infty = \max\left\{\lambda \in \mathbb{R}^+ \text{ tels que } \max_{0\leq\theta\leq\pi}\left(|R(-\lambda Q_4^H(e^{i\theta})| \leq 1\right)\right\}. \tag{2.92}$$

Proposition 2.6. *En l'absence de filtrage, le schéma est asymptotiquement stable sous la condition*

$$\lambda \leq \lambda_\infty = 2\sqrt{\frac{2}{3}}. \tag{2.93}$$

Démonstration. Pour tout $-N/2+1 \leq k \leq N/2$, on a

$$\omega^k = \exp\left(i\frac{2\pi k}{N}\right). \tag{2.94}$$

donc, on a

$$\begin{aligned} Q_4^H(\omega^k) &= \frac{\exp\left(i\frac{2\pi k}{N}\right) - \exp\left(-i\frac{2\pi k}{N}\right)}{\frac{4}{6} + \frac{1}{6}\left(\exp\left(i\frac{2\pi k}{N}\right) + \exp\left(-i\frac{2\pi k}{N}\right)\right)} \\ &= i\frac{\sin\left(\frac{2\pi k}{N}\right)}{\frac{2}{3} + \frac{1}{3}\cos\left(\frac{2\pi k}{N}\right)} \\ &= ig\left(\frac{2\pi k}{N}\right) \end{aligned}$$

avec

$$g(x) = \frac{\sin(x)}{2/3 + 1/3\cos(x)}. \tag{2.95}$$

D'après la proposition 2.4, l'algorithme 4 sans filtrage ($\mathcal{F} = \mathrm{Id}$) est asymptotiquement stable si

$$|R(-\lambda Q_4^H(\omega^k))| \leq 1. \tag{2.96}$$

D'après (2.31), comme $Q_4^H(\omega^k) \in i\mathbb{R}$, ce dernier résultat est équivalent à avoir

$$|\lambda Q_4^H(\omega^k)| = |\lambda g\left(\frac{2\pi k}{N}\right)| \leq 2\sqrt{2}. \tag{2.97}$$

Or $\max_{0\leq x\leq\pi} g(x) = \sqrt{3} = g\left(\frac{2\pi}{3}\right)$ donc

$$\begin{aligned} |\lambda g\left(\frac{2k\pi}{N}\right)| \leq 2\sqrt{2} &\Leftrightarrow \lambda \leq \frac{2\sqrt{2}}{\max_{0\leq x\leq\pi} g(x)} \\ &\Leftrightarrow \lambda \leq 2\sqrt{\frac{2}{3}} \end{aligned}$$

et cette inégalité est une égalité lorsque $3k = N$. □

Ce dernier résultat donne condition sur λ pour que le schéma soit stable, y compris lorsqu'un opérateur de filtrage est présent. Comme l'opérateur de filtrage est symétrique, ses valeurs propres sont réelles et on a :

$$\lambda_{2J} = \max\left\{\lambda \in \mathbb{R}^+ \text{ tels que } \max_{0\leq\theta\leq\pi}\left(|S_{2J}(e^{i\theta})||R(-\lambda Q_4^H(\theta)|\right) \leq 1\right\} \tag{2.98}$$

où S_{2J} est issue de l'opérateur de filtrage d'ordre $2J$ et donné par (1.165).

D'après la proposition 1.4, on obtient le théorème suivant

Théorème 2.1. *Quel que soit le filtre $\mathcal{F}_{2J,x}$ d'ordre $2J$, on a*

$$\lambda_{2J} \geq \lambda_\infty. \tag{2.99}$$

Démonstration. Pour tout $\theta \in [0,\pi]$ et pour tout $\lambda > 0$, on a

$$|S_{2J}(e^{i\theta})||R(-\lambda Q_4^H(e^{i\theta}))| \leq |R(-\lambda Q_4^H(e^{i\theta}))| \tag{2.100}$$

car $|S_{2J}(e^{i\theta})| \leq 1$ d'après la proposition 1.4. Donc en particulier, pour le maximum :

$$\max_{0\leq\theta\leq\pi}\left\{|S_{2J}(e^{i\theta})||R(-\lambda Q_4^H(e^{i\theta}))|\right\} \leq \max_{0\leq\theta\leq\pi}\left\{|R(-\lambda Q_4^H(e^{i\theta}))|\right\}. \tag{2.101}$$

Ainsi, l'inclusion suivante est vérifiée :

$$\left\{\lambda \in \mathbb{R}^+ \text{ tels que } \max_{0\leq\theta\leq\pi}\left(|R(-\lambda Q_4^H(e^{i\theta}))|\right) \leq 1\right\} \subset ...$$
$$... \left\{\lambda \in \mathbb{R}^+ \text{ tels que } \max_{0\leq\theta\leq\pi}\left(|S_{2J}(e^{i\theta})||R(-\lambda Q_4^H(e^{i\theta}))|\right) \leq 1\right\}. \tag{2.102}$$

On conclut en prenant le maximum de cette inclusion. □

En évaluant numériquement la valeur de λ_{2J} par un algorithme de dichotomie, on obtient la Table 2.3. On constate que λ_{2J} augmente lorsque $2J$ diminue. La condition de stabilité est moins restrictive pour un ordre de filtre bas. Ce résultat était attendu puisqu'un filtre d'ordre bas est plus sélectif qu'un filtre d'ordre élevé cependant les ondes sont plus atténuées qu'avec un filtre d'ordre élevé.

Ordre du filtre $\mathcal{F}_{2J,x}$: $2J$	λ_{2J}
Pas de filtrage	$2\sqrt{2/3} \approx 1.6329$
10	1.6883
8	1.7114
6	1.7485
4	1.8156
2	1.9749

TABLE 2.3 – Valeurs de λ_{2J} pour différentes valeurs de l'ordre du filtre $2J$.

2.3.3 Dissipation et dispersion numérique

On effectue dans cette section l'étude de dissipation et de dispersion pour un schéma linéaire appliqué à l'équation de transport (2.50). Cette étude est similaire à celles présentes dans [28, 29].

La solution de l'équation de transport (2.50) périodique est donnée pour tout $x \in [0, L]$ et $t > 0$ par

$$u(t,x) = u_0(x - ct). \tag{2.103}$$

Si la fonction initiale u_0 est une onde de la forme

$$u_0(x) = \exp\left(i\frac{2\pi k}{L}x\right) \tag{2.104}$$

avec $-N/2+1 \leq k \leq N/2$, alors la solution est

$$u(t,x) = \exp\left(i\frac{2\pi k}{L}(x-ct)\right). \tag{2.105}$$

Elle vérifie en $x_j = jh = jL/N$ et $t^n = n\Delta t$:

$$u(t^{n+1}, x_j) = e^{-i\lambda\theta}u(t^n, x_j) \tag{2.106}$$

où $\lambda = c\Delta t/h$ et $\theta = 2\pi k/N$.

D'autre part, l'application du schéma de discrétisation spatiale linéaire $\delta^H_{4,x}$ et du schéma d'intégration temporel donné dans l'algorithme 4 à (2.50) est linéaire. Ils donnent une relation de la forme

$$\mathfrak{u}_j^{n+1} = G(\lambda,\theta)\mathfrak{u}_j^n \tag{2.107}$$

où $\mathfrak{u}_j^n$ est calculé par le schéma lorsque $\mathfrak{u}^0 = u_0^*$. La fonction $(\lambda,\theta) \in \mathbb{R}^+ \times \mathbb{R} \mapsto G(\lambda,\theta)$ est la fonction d'amplification du schéma numérique et est donnée par

$$G(\lambda,\theta) = S_{2J}(e^{i\theta})R(-\lambda Q_4^H(e^{i\theta})) \tag{2.108}$$

où R est donné par (2.28), S_{2J} correspond au filtre $\mathcal{F}_{x,2J}$ et est donné par (1.165), Q_4^H est la fraction rationnelle du schéma hermitien $\delta^H_{4,x}$. Elle est donnée par l'équation (1.133).

Par comparaison avec (2.106), on définit la vitesse numérique de phase du schéma par $c(\lambda,\theta) = c_R(\lambda,\theta) + ic_I(\lambda,\theta)$ telle que

$$G(\lambda,\theta) = \exp\left(-i\frac{c(\lambda,\theta)}{c}\lambda\theta\right). \tag{2.109}$$

On définit $\varepsilon(\lambda,\theta)$ par

$$\begin{aligned}\varepsilon(\lambda,\theta) &= \frac{G(\lambda,\theta)}{e^{-i\lambda\theta}} \\ &= \exp\left(\frac{c_I(\lambda,\theta)}{c}\lambda\theta\right)\exp\left(-i\lambda\theta\left(\frac{c_R(\lambda,\theta)}{c}-1\right)\right) \\ &= |G(\lambda,\theta)|\exp\left(-i\lambda\theta\left(\frac{c_R(\lambda,\theta)}{c}-1\right)\right).\end{aligned}$$

La *fonction de dissipation* ε_D et *la fonction de dispersion* ε_Φ sont déduites de $\varepsilon(\lambda,\theta)$.

Définition 2.2. *Soit λ fixé.*

- *La fonction de dissipation ε_D est définie par*

$$\begin{array}{rcl}\varepsilon_D :]-\pi,\pi[& \to & \mathbb{R} \\ \theta & \mapsto & |\varepsilon(\lambda,\theta)| = |G(\lambda,\theta)|.\end{array} \tag{2.110}$$

  On note que

$$\varepsilon_D(\theta) = \exp\left(\frac{c_I(\lambda,\theta)}{c}\lambda\theta\right). \tag{2.111}$$

- *La fonction de dispersion ε_Φ est définie par*

$$\begin{array}{rcl}\varepsilon_\Phi :]-\pi,\pi[& \to & \mathbb{R} \\ \theta & \mapsto & c_R(\lambda,\theta)/c.\end{array} \tag{2.112}$$

Remarque 2.1. *On observe que ε_Φ s'écrit sous la forme*

$$\varepsilon_\Phi(\lambda,\theta) = 1 - \frac{1}{\lambda\theta} arg\left(\frac{G(\lambda,\theta)}{|G(\lambda,\theta)|e^{i\lambda\theta}}\right), \tag{2.113}$$

ainsi que la relation

$$\varepsilon(\lambda,\theta) = \varepsilon_D(\lambda,\theta)\exp\left(-i\lambda\theta\left(\varepsilon_\Phi(\lambda,\theta)-1\right)\right). \tag{2.114}$$

On s'intéresse à l'influence du filtrage sur ε_D et ε_Φ.

La fonction de dissipation ε_D mesure la dissipation du schéma numérique. On note en particulier que le schéma est asymptotiquement stable si et seulement si pour tout θ, on a

$$\varepsilon_D(\lambda,\theta) \leq 1 \tag{2.115}$$

ce qui implique directement $c_I(\lambda,\theta) \leq 0$. Lorsque $\varepsilon_D(\lambda,\theta) = 1$, le schéma n'est pas dissipatif. Si $\varepsilon_D(\lambda,\theta) < 1$, le schéma est dissipatif.

La fonction de dispersion ε_Φ mesure l'erreur de phase du schéma numérique. Si $\varepsilon_\Phi(\lambda,\theta) > 1$ alors $c_R(\lambda,\theta) > c$ et le schéma est en avance de phase. Inversement si $\varepsilon_\Phi(\lambda,\theta) < 1$, il est en retard de phase. Lorsque $\varepsilon_\Phi(\lambda,\theta) = 1$, le schéma n'est pas dispersif. L'opérateur de filtrage n'a aucune influence sur la fonction de dispersion. En effet, il s'agit de la multiplication par un scalaire de la fonction d'amplification $G(\lambda,\theta)$, ce qui est confirmé par la figure 2.2

Sur la Figure 2.2, on représente la fonction de dissipation ε_D et la fonction de dispersion ε_Φ pour différentes valeurs de λ. Le but est de comparer l'influence de l'opérateur de filtrage par rapport à l'absence d'opérateur de filtrage quand on utilise RK4 avec le schéma compact d'ordre 4 : $\delta^H_{4,x}$.

A λ fixé, on ne trace ε_D et ε_Φ que pour $\theta \in [0,\pi]$ pour des raisons de parité. Lorsque $\lambda = 0.05$, le schéma sans filtre est peu dissipatif. Le pas de temps Δt est petit donc il y a peu d'action du schéma RK4. Le schéma est dispersif. En revanche, l'introduction du filtrage d'ordre 10 rend le schéma dissipatif. Lorsque $\lambda = 1$, l'action de RK4 est plus importante le schéma est un peu plus dissipatif lorsque le filtrage est absent. Lors du choix $\lambda = \lambda_\infty \approx 1.6330$, on se place à la limite de stabilité du schéma sans filtre. Dans ce cadre, deux zones pour lesquelles le schéma est dissipatif aparaissent. Lorsque l'on utilise le filtre d'ordre 10, le schéma est plus dissipatif, en particulier autour de $5\pi/8$. Lorsque $\lambda = 1.6883$, le schéma sans filtrage est instable. En revanche le schéma avec filtrage d'ordre 10 est à la limite de la stabilité. Il présente une bonne amplification et les ondes sont amorties lorsque $\theta \geq \pi/4$.

2.3.4 Relations de conservation

L'équation de transport (2.50) est une équation de conservation. En effet il est immédiat que la périodicité de $x \mapsto u(t,x)$ entraine

Proposition 2.7. *Si u est une solution périodique de (2.50) alors pour tout $t > 0$, on a*

$$\int_0^1 u(t,x)dx = \int_0^1 u_0(x)dx. \tag{2.116}$$

Autrement dit, la masse totale de u est conservée au cours du temps. Dans la pratique, la résolution par un schéma numérique peut entraîner une perte de conservation. La contrepartie discrète de la conservation de la masse est la proposition suivante :

Proposition 2.8. *La suite $(\mathfrak{u}^n)$, calculée par l'algorithme 4, satisfait*

$$(\mathfrak{u}^{n+1}, \mathfrak{1})_{h,pér} = (\mathfrak{u}^n, \mathfrak{1})_{h,pér}. \tag{2.117}$$

pour tout $n \in \mathbb{N}$, où $\mathfrak{1}$ est la fonction de grille constante égale à 1.

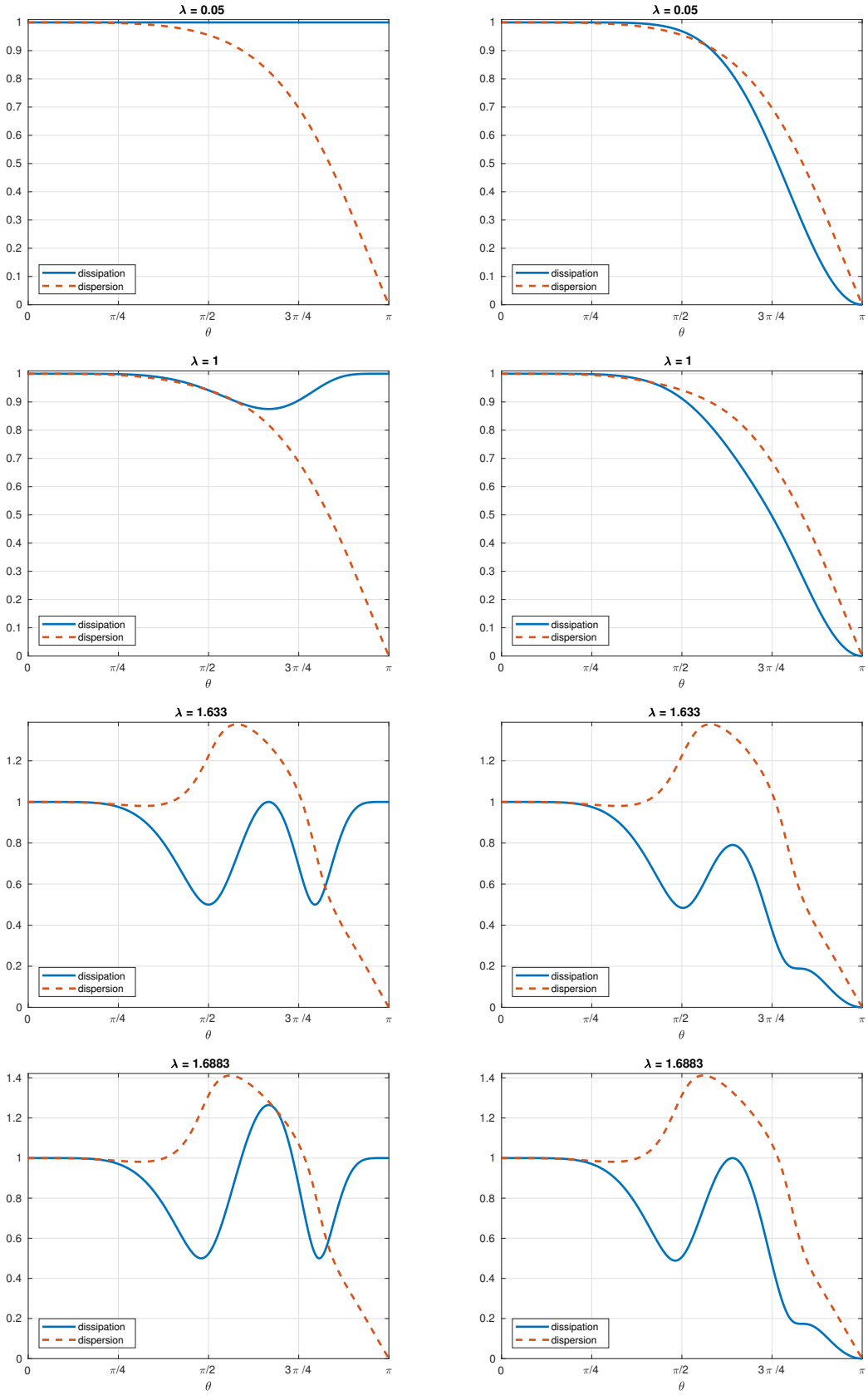


FIGURE 2.2 – Fonctions de dissipation et de dispersion associées à l'algorithme 4 de résolution de l'équation (2.50) sans un filtre (gauche) et avec filtre d'ordre 10 (droite) pour différentes valeurs de $\lambda = c\Delta t/h$.

Démonstration. Soit $\mathfrak{b} \in l^2_{h,\text{pér}}$ une fonction de grille quelconque et $b = \text{vec}(\mathfrak{b}) \in \mathbb{R}^N$, on a

$$\begin{aligned}(\delta^H_{4,x}\mathfrak{b}, \mathfrak{1})_{h,\text{pér}} &= \left(Q^H_4(T)b\right)^T \cdot \mathbf{1} \\ &= \frac{1}{h} b^T \cdot (\bar{Q}^H_4(T)\mathbf{1}) \\ &= \frac{1}{h} b^T \cdot \mathbf{0} \text{ par antisymétrie de } Q^H_4(T) \\ &= 0.\end{aligned}$$

De cette égalité, on déduit

$$(K^{(i)}, \mathfrak{1})_{h,\text{pér}} = 0 \tag{2.118}$$

pour tout $i \in \{1, 2, 3, 4\}$ dans l'algorithme 4.

De là, on déduit :

$$\begin{aligned}(\mathfrak{u}^{n+1}, \mathfrak{1})_{h,\text{pér}} &= h\left(S_{2J}(T)\left(U^n + \frac{\Delta t}{6}(K^{(1)} + 2K^{(2)} + 2K^{(3)} + K^{(4)})\right)\right)^T \cdot \mathbf{1} \\ &= h\left(U^n + \frac{\Delta t}{6}(K^{(1)} + 2K^{(2)} + 2K^{(3)} + K^{(4)})\right)^T \cdot \mathbf{1} \text{ par symétrie de } S_{2J}(T) \\ &= (\mathfrak{u}^n, \mathfrak{1})_{h,\text{pér}}.\end{aligned}$$

avec $U^n = \text{vec}(\mathfrak{u}^n)$. □

2.3.5 Résultats numériques

Dans cette section, on évalue numériquement les performances du schéma numérique dans les deux cas suivants.

Condition initiale régulière

On considère la condition initiale u_0 donnée par

$$u_0(x) = \frac{1}{\sqrt{2}}\left[\cos(2\pi x)\sin(4\pi x) + \sin(2\pi x)\right] \text{ avec } x \in \Omega = [0, 1], \tag{2.119}$$

avec $c = 0.2$. On compare la solution exacte avec la solution numérique associée. L'erreur relative

$$e^n_l = \frac{\|\mathfrak{u}^n - u(t^n, \cdot)^*\|_l}{\|u(t^n, \cdot)^*\|_l}, \text{ avec } l \in \{2, \infty\}, \tag{2.120}$$

calculée au temps t^n. Les valeurs obtenues sont données dans la Table 2.4. Les résultats permettent de confirmer la convergence à l'ordre 4 attendue. L'erreur est tracée au cours du temps dans la Figure

N	**norme 2**	**norme ∞**
50	1.1158(−2)	1.2630(−2)
100	7.1441(−4)	8.0641(−4)
500	1.1484(−6)	1.2998(−6)
1000	7.1839(−8)	8.1303(−8)
ordre estimé	3.9917	3.9913

TABLE 2.4 – Equation de convection avec la condition initiale (2.119). Table de convergence de l'algorithme 4 avec le filtre d'ordre 10. Le temps final est $T = 10$ et $c\Delta t/h = 1.5$.

2.3 en $\|\cdot\|_2$ et $\|\cdot\|_\infty$. Comme cela a été vu dans la proposition 2.5, on observe la dépendance linéaire en t de l'erreur.

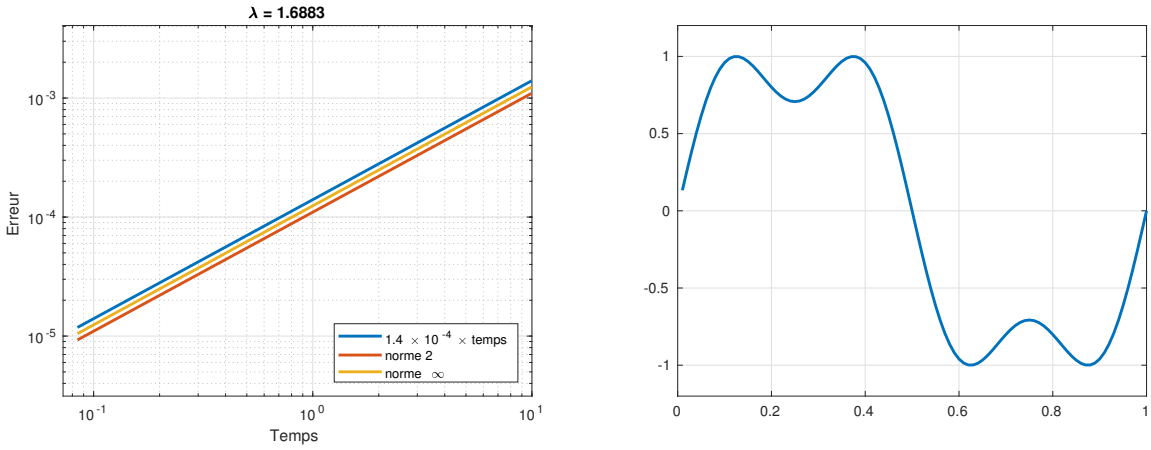


FIGURE 2.3 – Equation de convection avec la condition initiale (2.119). A gauche, historique de l'erreur de l'algorithme 4 avec le filtre d'ordre 10 pour la condition initiale (2.119). A droite, la condition initiale (2.119). On choisit $N = 100$ points de grille et $c\Delta t/h = 1.6883$ (118 pas de temps).

Condition initiale de type créneau

On considère à présent la donnée initiale

$$u_0(x) = \begin{cases} 1 & \text{si } 0.25 \leq x \leq 0.75, \\ -1 & \text{sinon,} \end{cases} \quad \text{pour } x \in \Omega = [0,1], \tag{2.121}$$

avec $c = 0.2$. Les résultats numériques pour la convection de la donnée initiale (2.121) sont donnés dans la figure 2.4 au temps final $t = 10$, c'est à dire au bout de deux périodes.

Des oscillations parasites de nature dispersives apparaissent. Ces oscillations peuvent être atténuées par l'utilisation d'un filtre. On compare sur la Figure 2.4 la solution au temps $t = 10$ avec les solution obtenues en utilisant différentes fonctions de filtrage dans l'algorithme 4.

On constate que le filtre d'ordre 2 permet de supprimer les ondes parasites mais est trop dissipatif. Les filtres d'ordres 4, 6, 8 et 10 donnent des résultats moins dissipatifs tout en atténuant les oscillations dispersives.

Il est bien connu que l'ammélioration du comportement dissipatif d'un schéma centré de type (2.79) est un problème délicat. On renvoie à ce sujet aux méthodes d'hyperviscosités [22], aux méthodes non linéaires de type WENO [71] ou aux méthodes utilisant des senseurs non linéaires [87]. Nous n'allons pas plus loin dans cette direction car les expériences numériques effectuées pour des problèmes sphériques en climatologie numérique n'ont pas nécessité de tels traitements.

2.4 Équation Shallow Water linéarisée avec Coriolis constant

On considère dans cette partie l'équation des ondes avec force de Coriolis sur un carré périodique. Le paramètre $f > 0$ constant représentant la force de Coriolis. Cette équation est la linéarisation de l'équation Shallow Water au voisinage d'un état de repos avec l'hypothèse d'une force de Coriolis constante. On parle généralement de $f-$plan [11]. Ce système présente de nombreuses propriétés propres aux équations Shallow Water linéarisées. Il permet aussi de représenter les ondes d'inertie-gravité mais pas les ondes de Rossby [11, 41].

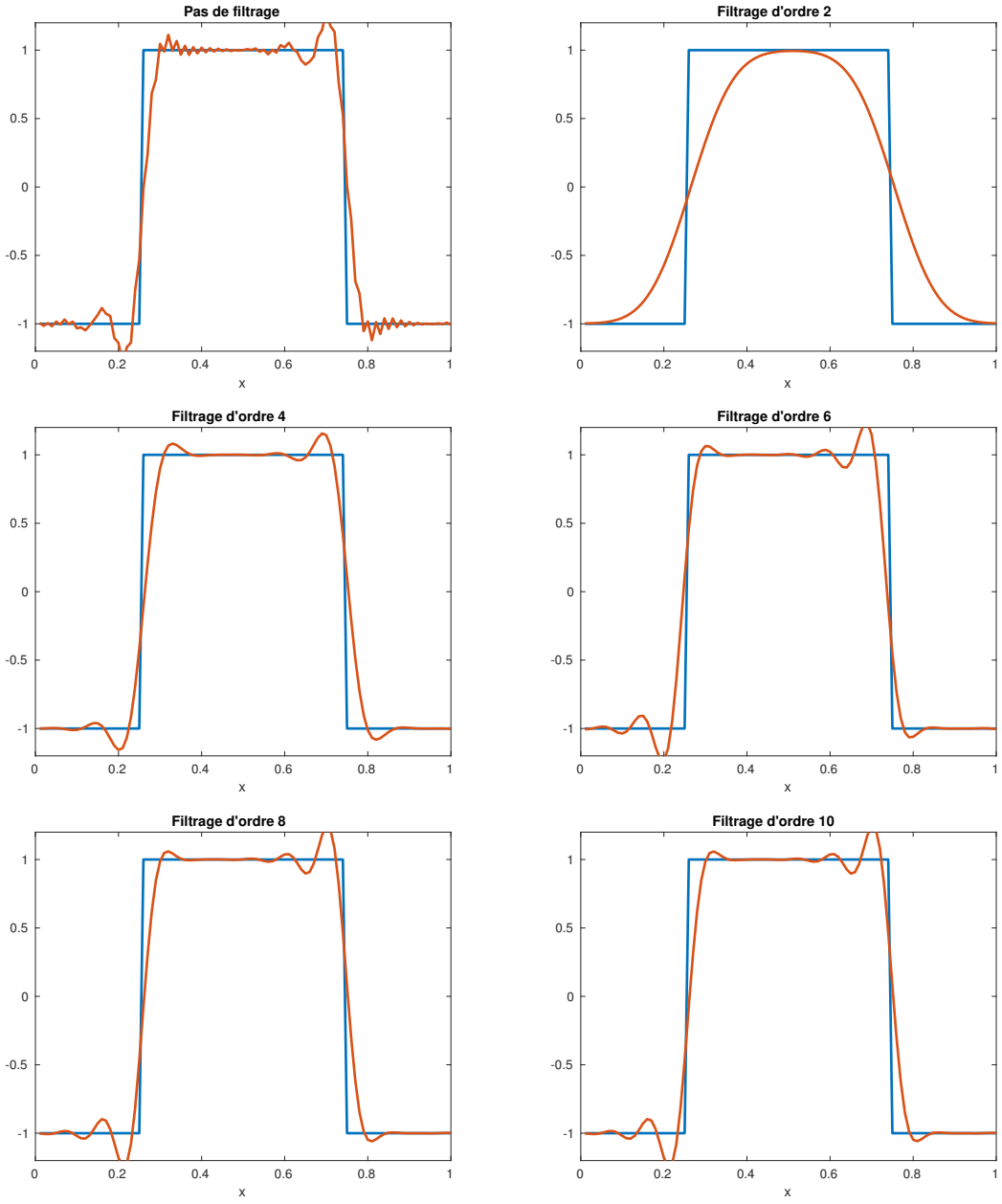


FIGURE 2.4 – Comparaison de la solution exacte (bleu) avec la solution obtenue par l'algorithme 4 (rouge) au temps $t = 10$ pour la résolution de l'équation (2.50) avec différents filtres. $\lambda = c\Delta t/h = 1.5$ et $N = 100$. Les résultats sont obtenus en 133 itérations et au bout de 2 périodes.

Pour $(x,y) \in \Omega = [0,1]^2$ et $t > 0$, le problème s'écrit

$$\left\{ \begin{array}{rcl} \dfrac{\partial \eta}{\partial t} + H\left(\dfrac{\partial u}{\partial x} + \dfrac{\partial v}{\partial y}\right) & = & 0 \\ \dfrac{\partial u}{\partial t} + g\dfrac{\partial \eta}{\partial x} - fv & = & 0 \\ \dfrac{\partial v}{\partial t} + g\dfrac{\partial \eta}{\partial y} + fu & = & 0. \end{array} \right. \tag{2.122}$$

La constante $H > 0$ est une hauteur de fluide de référence et $g > 0$ est la constante de gravité. On ajoute, à cette équation, une condition initiale $1-$périodique de la forme

$$\left\{ \begin{array}{rcl} \eta(0,x,y) & = & \eta_0(x,y) \\ u(0,x,y) & = & u_0(x,y) \quad \text{avec } (x,y) \in \Omega. \\ v(0,x,y) & = & v_0(x,y) \end{array} \right. \tag{2.123}$$

Proposition 2.9. *Si $\eta(t,x,y)$, $u(t,x,y)$ et $v(t,x,y)$ sont trois fonctions de $\mathbb{R}^+ \times \Omega$ dans $\mathbb{R}$ solutions de (2.122), alors les relations de conservation suivantes sont vérifiées :*

- ***Conservation de la masse :***

$$\frac{d}{dt}\int_\Omega h(t,x,y)dxdy = 0, \tag{2.124}$$

- ***Conservation de l'énergie :***

$$\frac{d}{dt}\int_\Omega \left(\frac{1}{2}gh(t,x,y)^2 + \frac{1}{2}H\left(u(t,x,y)^2 + v(t,x,y)^2\right)\right)dxdy = 0. \tag{2.125}$$

Démonstration. Soient η, u et v trois fonctions de $\mathbb{R}^+ \times \Omega$ dans $\mathbb{R}$ solutions de (2.122) alors :

- **Conservation de la masse :**

$$\begin{aligned} \frac{d}{dt}\int_{[0,1]^2} \eta(t,x,y)dxdy &= -H\int_{[0,1]^2} \frac{\partial u}{\partial x}(t,x,y) + \frac{\partial v}{\partial y}(t,x,y)dxdy \\ &\quad - H\int_0^1 \left(u(t,1,y) - u(t,0,y)\right)dy - H\int_0^1 \left(u(t,x,1) - u(t,x,0)\right)dx \\ &= 0 \text{ par périodicité de } u \text{ et } v. \end{aligned}$$

- **Conservation de l'énergie :** en multipliant la première équation de (2.122) par η et en intégrant, on a

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\int_{[0,1]^2} \eta(t,x,y)^2dxdy &= \int_{[0,1]^2} \eta(t,x,y)\frac{\partial \eta}{\partial t}(t,x,y)dxdy \\ &= -H\int_{[0,1]^2} \eta(t,x,y)\left(\frac{\partial u}{\partial x}(t,x,y) + \frac{\partial v}{\partial y}(t,x,y)\right)dxdy. \end{aligned}$$

D'autre part, on a

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\int_{[0,1]^2} u(t,x,y)^2dxdy &= \int_{[0,1]^2} u(t,x,y)\frac{\partial u}{\partial t}(t,x,y)dxdy \\ &= \int_{[0,1]^2} fu(t,x,y)v(t,x,y) - gu(t,x,y)\frac{\partial \eta}{\partial x}(t,x,y)dxdy, \end{aligned}$$

ainsi que

$$\frac{1}{2}\frac{d}{dt}\int_{[0,1]^2} v(t,x,y)^2dxdy = \int_{[0,1]^2} -fu(t,x,y)v(t,x,y) - gu(t,x,y)\frac{\partial \eta}{\partial y}(t,x,y)dxdy.$$

En combinant ces trois dernières relations, on obtient

$$\frac{d}{dt}\int_{[0,1]^2}\left(\frac{1}{2}gh(t,x,y)^2+\frac{1}{2}H\left(u(t,x,y)^2+v(t,x,y)^2\right)\right)= \\ -gH\int_{[0,1]^2}\frac{\partial}{\partial x}\left(\eta(t,x,y)u(t,x,y)\right)+\frac{\partial}{\partial y}\left(\eta(t,x,y)v(t,x,y)\right)dxdy=0$$

en utilisant la périodicité de η, u et v.

□

2.4.1 Schéma centré en dimension 2

De façon analogue à l'équation de transport (2.50), on utilise la méthode des lignes. La version semi-discrétisée de (2.122) est

$$\left\{\begin{array}{rcl}\dfrac{d\eta}{dt}+H(\delta^H_{x,4}\mathfrak{u}+\delta^H_{y,4}\mathfrak{v}) &=& 0\\ \dfrac{d\mathfrak{u}}{dt}+g\delta^H_{x,4}\eta-f\mathfrak{v} &=& 0\\ \dfrac{d\mathfrak{v}}{dt}+g\delta^H_{y,4}\eta+f\mathfrak{u} &=& 0.\end{array}\right. \tag{2.126}$$

On cherche alors $t\mapsto\eta(t)\in L^2_{h,\text{pér}}$, $t\mapsto\mathfrak{u}(t)\in L^2_{h,\text{pér}}$ et $t\mapsto\mathfrak{v}(t)\in L^2_{h,\text{pér}}$ des approximations de η, u et v les solutions de (2.122) aux points du maillage.

Ce sont effectivement des approximations comme le montre la proposition suivante :

Proposition 2.10. *Les fonctions de grilles η, $\mathfrak{u}$ et $\mathfrak{v}$ solutions de* (2.126) *convergent vers η, u et v solutions de* (2.122) *sur $[0,T]$ au sens où il existe $\tilde{C}>0$ indépendant de t, η, u et v tel que*

$$E(t)\leq gHtCh^4 \tag{2.127}$$

avec $C=\tilde{C}\max_{t\in[0,T]}\left(\frac{1}{\sqrt{H}}\|\partial_x^{(5)}u\|_\infty,\frac{1}{\sqrt{H}}\|\partial_y^{(5)}v\|_\infty,\frac{1}{\sqrt{g}}\|\partial_x^{(5)}\eta\|_\infty,\frac{1}{\sqrt{g}}\|\partial_y^{(5)}\eta\|_\infty\right)$ *et en notant les termes d'erreur $\mathfrak{e}_\eta=\eta^*-\eta$, $\mathfrak{e}_u=u^*-\mathfrak{u}$ et $\mathfrak{e}_v=v^*-\mathfrak{v}$ ainsi que*

$$E(t)=\sqrt{g\|\mathfrak{e}_\eta\|^2_{h,pér}+H\|\mathfrak{e}_u\|^2_{h,pér}+H\|\mathfrak{e}_v\|^2_{h,pér}}. \tag{2.128}$$

Démonstration. On définit les termes de troncature $\tau_x\eta=\delta^H_{4,x}\eta-(\partial_x\eta)^*$, $\tau_y\eta=\delta^H_{4,y}\eta-(\partial_y\eta)^*$, $\tau_xu=\delta^H_{4,x}\mathfrak{u}-(\partial_xu)^*$ et $\tau_yv=\delta^H_{4,y}\mathfrak{v}-(\partial_yv)^*$.

Les termes d'erreurs sont solutions de

$$\left\{\begin{array}{rcl}\dfrac{d}{dt}\mathfrak{e}_\eta &=& H\left(\delta^H_{4,x}\mathfrak{e}_u+\delta^H_{4,y}\mathfrak{e}_v\right)+H\left(\tau_xu+\tau_yv\right)\\ \dfrac{d}{dt}\mathfrak{e}_u &=& -g\delta^H_{4,x}\mathfrak{e}_\eta+f\mathfrak{e}_v-g\tau_x\eta\\ \dfrac{d}{dt}\mathfrak{e}_v &=& -g\delta^H_{4,y}\mathfrak{e}_\eta-f\mathfrak{e}_u-g\tau_y\eta\end{array}\right. \tag{2.129}$$

Par produit scalaire avec $\mathfrak{e}_\eta$ de la première équation on obtient :

$$\frac{1}{2}\frac{d}{dt}\|\mathfrak{e}_\eta\|^2_{h,\text{pér}}=(\frac{d}{dt}\mathfrak{e}_\eta,\mathfrak{e}_\eta)_{h,\text{pér}}=-H(\delta^H_{4,x}\mathfrak{e}_u+\delta^H_{4,y}\mathfrak{e}_v,\mathfrak{e}_\eta)_{h,\text{pér}}+H(\tau_xu+\tau_yv,\mathfrak{e}_\eta)_{h,\text{pér}} \tag{2.130}$$

En effectuant les produits scalaires de la seconde et troisième équations par $\mathfrak{e}_u$ et $\mathfrak{e}_v$, on obtient :

$$(\frac{d}{dt}\mathfrak{e}_u,\mathfrak{e}_u)_{h,\text{pér}}=-g(\delta^H_{4,x}\mathfrak{e}_\eta,\mathfrak{e}_u)_{h,\text{pér}}+f(\mathfrak{e}_v,\mathfrak{e}_u)_{h,\text{pér}}-g(\tau_x\eta,\mathfrak{e}_u)_{h,\text{pér}} \tag{2.131}$$

ainsi que

$$(\frac{d}{dt}\mathfrak{e}_v, \mathfrak{e}_v)_{h,\text{pér}} = -g(\delta^H_{4,y}\mathfrak{e}_\eta, \mathfrak{e}_v)_{h,\text{pér}} - f(\mathfrak{e}_u, \mathfrak{e}_v)_{h,\text{pér}} - g(\tau_y\eta, \mathfrak{e}_v)_{h,\text{pér}}. \tag{2.132}$$

Alors, en sommant ces deux équations et, par antisymétrie de $\delta^H_{4,x}$ et de $\delta^H_{4,y}$, on a

$$\frac{1}{2}\frac{d}{dt}\left(\|\mathfrak{e}_u\|^2_{h,\text{pér}} + \|\mathfrak{e}_v\|^2_{h,\text{pér}}\right) = (d_t\mathfrak{e}_u, \mathfrak{e}_u)_{h,\text{pér}}) + (d_t\mathfrak{e}_v, \mathfrak{e}_v)_{h,\text{pér}} \tag{2.133}$$

d'où

$$\frac{1}{2}\frac{d}{dt}\left(\|\mathfrak{e}_u\|^2_{h,\text{pér}} + \|\mathfrak{e}_v\|^2_{h,\text{pér}}\right) = g(\delta^H_{4,x}\mathfrak{e}_u + \delta^H_{4,y}\mathfrak{e}_v, \mathfrak{e}_\eta)_{h,\text{pér}} - g(\tau_x\eta, \mathfrak{e}_u)_{h,\text{pér}} - g(\tau_y\eta, \mathfrak{e}_v)_{h,\text{pér}}. \tag{2.134}$$

En sommant $g\times$(2.130) $+H\times$(2.134), on obtient :

$$\frac{d}{dt}E^2(t) = 2gH\left((\tau_x u + \tau_y v, \mathfrak{e}_\eta)_{h,\text{pér}} - (\tau_x\eta, \mathfrak{e}_u)_{h,\text{pér}} - (\tau_y\eta, \mathfrak{e}_v)_{h,\text{pér}}\right). \tag{2.135}$$

En majorant le terme de droite par sa valeur absolue, on obtient

$$\frac{d}{dt}E^2(t) \leq 2gH\left(|(\tau_x u + \tau_y v, \mathfrak{e}_\eta)_{h,\text{pér}}| + |(\tau_x\eta, \mathfrak{e}_u)_{h,\text{pér}}| + |(\tau_y\eta, \mathfrak{e}_v)_{h,\text{pér}}|\right). \tag{2.136}$$

Alors, pour tout $\alpha > 0$, on a

$$\frac{d}{dt}E^2(t) \leq 2gH\left(\frac{\alpha}{H}\|\tau_x u + \tau_y v\|^2_{h,\text{pér}} + \frac{H}{\alpha}\|\mathfrak{e}_\eta\|^2_{h,\text{pér}} + \frac{\alpha}{g}(\|\tau_x\eta\|^2_{h,\text{pér}} + \|\tau_y\eta\|^2_{h,\text{pér}}) + \frac{g}{\alpha}(\|\mathfrak{e}_u\|^2_{h,\text{pér}} + \|\mathfrak{e}_v\|^2_{h,\text{pér}})\right). \tag{2.137}$$

Ainsi, si on pose

$$\Upsilon(t) = \frac{1}{H}\|\tau_x u + \tau_y v\|_{h,\text{pér}} + \frac{1}{g}\|\tau_x\eta\|^2_{h,\text{pér}} + \frac{1}{g}\|\tau_y\eta\|^2_{h,\text{pér}} \tag{2.138}$$

on obtient

$$\frac{d}{dt}E^2(t) \leq 2gh\left(\alpha\Upsilon(t) + \frac{1}{g}E^2(t)\right). \tag{2.139}$$

Par consistance des opérateurs $\delta^H_{4,x}$ avec ∂_x et $\delta^H_{4,y}$ avec ∂_y, il existe $C_1 > 0$ indépendant des fonctions η, u v et h tels que pour tout $t \in [0, T]$

$$\Upsilon(t) \leq C_1 h^8 \max_{t\in[0,T]}\left(\frac{1}{\sqrt{H}}\|\partial^{(5)}_x u\|_\infty, \frac{1}{\sqrt{H}}\|\partial^{(5)}_y v\|_\infty, \frac{1}{\sqrt{g}}\|\partial^{(5)}_x\eta\|_\infty, \frac{1}{\sqrt{g}}\|\partial^{(5)}_y\eta\|_\infty\right) = h^8\Upsilon_\infty. \tag{2.140}$$

Ainsi, on a

$$\frac{d}{dt}E^2(t) \leq 2gh\left(\alpha h^8\Upsilon_\infty + \frac{1}{\alpha}E^2(t)\right), \tag{2.141}$$

et d'après le lemme de Gronwall,

$$E^2(t) \leq \alpha^2 h^8\Upsilon_\infty\left(\exp\left(\frac{2gHt}{\alpha}\right) - 1\right). \tag{2.142}$$

Pour tout $t > 0$ fixé, on pose $\beta = \alpha/(2gHt)$, l'équation (2.141) se réécrit

$$E^2(t) \leq 4g^2H^2t^2h^8\Upsilon_\infty a(\beta) \tag{2.143}$$

avec a la fonction donnée par

$$a : \beta > 0 \mapsto \beta^2\left(\exp\left(\frac{1}{\beta}\right) - 1\right). \tag{2.144}$$

Or, la fonction $a(\beta)$ est continue, positive et

$$\lim_{\beta\to+\infty} a(\beta) = \lim_{\beta\to 0} a(\beta) = +\infty. \tag{2.145}$$

Donc a admet un minimum $m > 0$, on vérifie que $m < 1.545$. Ainsi, on a

$$E^2(t) \leq g^2H^2t^2m\Upsilon_\infty h^8. \tag{2.146}$$

On conclut en prenant la racine carrée de cette dernière relation. □

De plus, le système semi-discrétisé (2.126) vérifie des relations de conservation très semblables à celles de l'équation d'origine (2.122).

Proposition 2.11. *Si η, $\mathfrak{u}$ et $\mathfrak{v}$ sont solutions de* (2.126) *alors les relations de conservation suivantes sont vérifiées :*

- ***Conservation de la masse :***
$$\frac{d}{dt}(\eta, \mathfrak{1})_{h,pér} = 0. \tag{2.147}$$
- ***Conservation de l'énergie :***
$$\frac{d}{dt}\left(\frac{1}{2}g\|\eta\|^2_{h,pér} + \frac{1}{2}H(\|\mathfrak{u}\|^2_{h,pér} + \|\mathfrak{v}\|^2_{h,pér})\right) = 0. \tag{2.148}$$

Démonstration. • **Conservation de la masse :** par consistance des opérateurs $\delta^H_{4,x}$ et $\delta^H_{4,y}$ avec ∂_x et ∂_y, on note que $\delta^H_{4,x}\mathfrak{1} = \mathfrak{o}$ et $\delta^H_{4,y}\mathfrak{1} = \mathfrak{o}$, où $\mathfrak{1}$ (resp. $\mathfrak{o}$) est la fonction de grille égale à 1 (resp. 0). De là, il découle :

$$\begin{aligned}\frac{d}{dt}(\eta, \mathfrak{1})_{h,\text{pér}} &= (d_t\eta, \mathfrak{1})_{h,\text{pér}} \\ &= -H(\delta_{4,x}\mathfrak{u}, \mathfrak{1})_{h,\text{pér}} - H(\delta_{4,y}\mathfrak{v}, \mathfrak{1})_{h,\text{pér}} \\ &= H(\mathfrak{u}, \delta_{4,x}\mathfrak{1})_{h,\text{pér}} + H(\mathfrak{v}, \delta_{4,y}\mathfrak{1})_{h,\text{pér}} \\ &= 0\end{aligned}$$

où on a utilisé l'anti-symétrie des opérateurs.

- **Conservation de l'énergie :** On a
$$(\frac{d}{dt}\eta, \eta)_{h,\text{pér}} = \frac{1}{2}\frac{d}{dt}\|\eta\|^2_{h,\text{pér}} = -H(\delta_{4,x}\mathfrak{u}, \eta)_{h,\text{pér}} - H(\delta_{4,y}\mathfrak{v}, \eta)_{h,\text{pér}}. \tag{2.149}$$
De même, on a :
$$\frac{1}{2}\frac{d}{dt}(\|\mathfrak{u}\|^2_{h,\text{pér}} + \|\mathfrak{v}\|^2_{h,\text{pér}}) = g(\delta^H_{4,x}\mathfrak{u} + \delta^H_{4,y}\mathfrak{v}, \eta)_{h,\text{pér}}. \tag{2.150}$$
Par combinaison, on obtient :
$$\frac{d}{dt}\left(\frac{1}{2}g\|\eta\|^2_{h,\text{pér}} + \frac{1}{2}H(\|\mathfrak{u}\|^2_{h,\text{pér}} + \|\mathfrak{v}\|^2_{h,\text{pér}})\right) = 0. \tag{2.151}$$

□

A partir d'ici, on pose $F_h : (L^2_{h,\text{pér}})^3 \to (L^2_{h,\text{pér}})^3$ l'application linéaire définie par

$$F_h : \begin{pmatrix}\eta\\ \mathfrak{u}\\ \mathfrak{v}\end{pmatrix} \mapsto \begin{pmatrix}-H(\delta^H_{4,x}\mathfrak{u} + \delta^H_{4,y}\mathfrak{v})\\ -g\delta^H_{4,x}\eta + f\mathfrak{v}\\ -g\delta^H_{4,y}\eta - f\mathfrak{u}\end{pmatrix}. \tag{2.152}$$

Alors, le problème (2.126) s'écrit

$$\frac{d}{dt}\begin{pmatrix}\eta\\ \mathfrak{u}\\ \mathfrak{v}\end{pmatrix} = F_h\begin{pmatrix}\eta\\ \mathfrak{u}\\ \mathfrak{v}\end{pmatrix}. \tag{2.153}$$

On note $\Delta t > 0$ le pas de temps et η^n, $\mathfrak{u}^n$ et $\mathfrak{v}^n$ les approximations de $\eta(n\Delta t)$, $\mathfrak{u}(n\Delta t)$ et $\mathfrak{v}(n\Delta t)$ obtenues par un algorithme de type RK4 filtré. On note $\mathcal{F}_{2J} = \mathcal{F}_{2J,x} \circ \mathcal{F}_{2J,y}$ l'opérateur de filtrage tel que $\mathcal{F}_{2J,x}$ (resp. $\mathcal{F}_{2J,y}$) est un opérateur de filtrage dans la direction de x (resp. y). La méthode est détaillée dans l'algorithme 5.

Algorithme 5 : Schéma en temps RK4 avec étape de filtrage pour le système périodique (2.126)

$\eta^0 = \eta_0^*$, $\mathfrak{u}^0 = u_0^*$ et $\mathfrak{v}^0 = v_0^*$ connus,
for $n = 0, 1, \ldots$ **do**
$$\left(K_\eta^{(1)}, K_u^{(1)}, K_v^{(1)}\right) = F_h(\eta^n, \mathfrak{u}^n, \mathfrak{v}^n),$$
$$\left(K_\eta^{(2)}, K_u^{(2)}, K_v^{(2)}\right) = F_h\left(\eta^n + \frac{\Delta t}{2}K_\eta^{(1)}, \mathfrak{u}^n + \frac{\Delta t}{2}K_u^{(1)}, \mathfrak{v}^n + \frac{\Delta t}{2}K_v^{(1)}\right),$$
$$\left(K_\eta^{(3)}, K_u^{(3)}, K_v^{(3)}\right) = F_h\left(\eta^n + \frac{\Delta t}{2}K_\eta^{(2)}, \mathfrak{u}^n + \frac{\Delta t}{2}K_u^{(2)}, \mathfrak{v}^n + \frac{\Delta t}{2}K_v^{(2)}\right),$$
$$\left(K_\eta^{(4)}, K_u^{(4)}, K_v^{(4)}\right) = F_h\left(\eta^n + \Delta t K_\eta^{(3)}, \mathfrak{u}^n + \Delta t K_u^{(3)}, \mathfrak{v}^n + \Delta t K_v^{(3)}\right),$$
$$\eta^{n+1} = \mathcal{F}_{2J}\left(\eta^n + \frac{\Delta t}{6}\left(K_\eta^{(1)} + 2K_\eta^{(2)} + 2K_\eta^{(3)} + K_\eta^{(4)}\right)\right),$$
$$\mathfrak{u}^{n+1} = \mathcal{F}_{2J}\left(\mathfrak{u}^n + \frac{\Delta t}{6}\left(K_u^{(1)} + 2K_u^{(2)} + 2K_u^{(3)} + K_u^{(4)}\right)\right),$$
$$\mathfrak{v}^{n+1} = \mathcal{F}_{2J}\left(\mathfrak{v}^n + \frac{\Delta t}{6}\left(K_v^{(1)} + 2K_v^{(2)} + 2K_v^{(3)} + K_v^{(4)}\right)\right).$$
end for

Pour assurer la précision de la méthode utilisée, il faut que l'algorithme (5) soit stable. On considère dans un premier temps l'algorithme sans filtrage, c'est à dire $\mathcal{F}_{2J} = \mathrm{Id}$. On a déjà vu que l'algorithme est stable si et seulement si

$$\mathrm{Sp}(\Delta t F_h) \subset \mathcal{D}_{RK4}. \tag{2.154}$$

Cela donne lieu à la proposition suivante.

Proposition 2.12. *Si $\lambda \in Sp(F_h)$ alors $\lambda = 0$ ou $\lambda = \pm if$ ou $\lambda^2 \in Sp\left(gH(\delta_{4,x}^H \circ \delta_{4,x}^H + \delta_{4,y}^H \circ \delta_{4,y}^H) - f^2\right)$.*

Démonstration. Si $\lambda \in \mathrm{Sp}(F_h)$ alors il existe η, $\mathfrak{u}$ et $\mathfrak{v}$ non tous nuls dans $L^2_{h,\text{pér}}$ tels que

$$\left\{\begin{array}{rcll} \lambda\eta + H\delta_{4,x}^H\mathfrak{u} + H\delta_{4,y}^H\mathfrak{v} &=& 0 & \text{(a)} \\ g\delta_{4,x}^H\eta + \lambda\mathfrak{u} - f\mathfrak{v} &=& 0 & \text{(b)} \\ g\delta_{4,y}^H\eta + f\mathfrak{u} + \lambda\mathfrak{v} &=& 0 & \text{(c)} \end{array}\right. . \tag{2.155}$$

En considérant $f \times (2.155.a) + H\delta_{4,y}^H \times (2.155.b)$, on montre que

$$(\lambda f + gH\delta_{4,x}^H \circ \delta_{4,y}^H)\eta + H(f\delta_{4,x}^H + \lambda\delta_{4,y}^H)\mathfrak{u} = 0. \tag{2.156}$$

De même, avec $\lambda \times (2.155.a) - H\delta_{4,x}^H(2.155.c)$, on a

$$(\lambda^2 - gH\delta_{4,y}^H \circ \delta_{4,y}^H)\eta + H(\lambda\delta_{4,x}^H - f\delta_{4,y}^H)\mathfrak{u} = 0. \tag{2.157}$$

En effectuant les opérations $(f\delta_{4,x}^H + \lambda\delta_{4,y}^H) \times (2.157) - (\lambda\delta_{4,x}^H - f\delta_{4,y}^H) \times (2.156)$, on montre que

$$\lambda\left(gH(\delta_{4,x}^H \circ \delta_{4,x}^H + \delta_{4,y}^H \circ \delta_{4,y}^H - f^2 - \lambda^2\right)\delta_{4,y}^H\eta = 0. \tag{2.158}$$

De la même manière, on obtient l'égalité :

$$\lambda\left(gH(\delta_{4,x}^H \circ \delta_{4,x}^H + \delta_{4,y}^H \circ \delta_{4,y}^H - f^2 - \lambda^2\right)\delta_{4,x}^H\eta = 0. \tag{2.159}$$

Il y a alors plusieurs possibilités. Soit $\lambda = 0$, soit

$$\lambda^2 \in \mathrm{Sp}\left(gH(\delta_{4,x}^H \circ \delta_{4,x}^H + \delta_{4,y}^H \circ \delta_{4,y}^H) - f^2\right), \tag{2.160}$$

soit $\delta^H_{4,x}\eta = \mathfrak{o} = \delta^H_{4,y}\eta$. Dans ce dernier cas de figure, d'après (2.155.b) et (2.155.c), on trouve

$$\begin{cases} (\lambda^2 + f^2)\mathfrak{u} = 0 \\ (\lambda^2 + f^2)\mathfrak{v} = 0. \end{cases} \tag{2.161}$$

Si $\mathfrak{u} = \mathfrak{v} = \mathfrak{o}$, alors $\eta = \mathfrak{o}$ d'après (2.155.a), ce qui est impossible. Si $\mathfrak{u}$ ou $\mathfrak{v}$ est non nul, alors nécessairement on a $\lambda = \pm if$ ce qui conclut la preuve. □

Les valeurs propres de l'opérateur $gH(\delta^H_{4,x} \circ \delta^H_{4,x} + \delta^H_{4,y} \circ \delta^H_{4,y}) - f^2$ sont connues. D'après les propositions 1.13 et 1.24, elles sont données par

$$\frac{gH}{h^2}\left(Q^H_4(\omega^{k_1})^2 + Q^H_4(\omega^{k_2})^2\right) - f^2 \tag{2.162}$$

avec $-N/2 + 1 \leq k_1, k_2 \leq N/2$. On note le lien entre ces valeurs propres et la relation de dispersion liée aux équations (2.122). Il s'agit en effet d'une approximation discrète de la relation de dispersion des ondes d'inertie .gravité [11, 36, 59]. Il découle le théorème de stabilité suivant :

Théorème 2.2. *En l'absence de filtrage, le schéma énoncé par l'algorithme 5 est stable sous la condition*

$$\Delta t \leq \Delta t_\infty = \frac{2\sqrt{2}}{\sqrt{\frac{6gH}{h^2} + f^2}}. \tag{2.163}$$

Démonstration. Le schéma donné par l'algorithme 5 est stable si

$$\mathrm{Sp}(\Delta t F_h) \subset \mathcal{D}_{\mathrm{RK4}}. \tag{2.164}$$

On a vu dans la proposition 2.12 que $\lambda \in \mathrm{Sp}(F_h)$ implique $\lambda \in i\mathbb{R}$. En effet, les cas $\lambda = 0$ et $\lambda = \pm if$ sont clairs. De plus, si $\lambda \in \mathrm{Sp}\left(gH(\delta^H_{4,x} \circ \delta^H_{4,x} + \delta^H_{4,y} \circ \delta^H_{4,y}) - f^2\right)$, alors il existe $-N/2 + 1 \leq k_1, k_2 \leq N/2$ tels que

$$\begin{aligned} \lambda^2 &= \frac{gH}{h^2}\left(Q^H_4(\omega^{k_1})^2 + Q^H_4(\omega^{k_2})^2\right) - f^2 \\ &= -\frac{gH}{h^2}\left(\left(\frac{\sin\left(\frac{2\pi k_1}{N}\right)}{\frac{2}{3} + \frac{1}{3}\cos\left(\frac{2\pi k_1}{N}\right)}\right)^2 + \left(\frac{\sin\left(\frac{2\pi k_2}{N}\right)}{\frac{2}{3} + \frac{1}{3}\cos\left(\frac{2\pi k_2}{N}\right)}\right)^2\right) - f^2 \\ &< 0. \end{aligned}$$

Donc λ est imaginaire pure et la condition (2.164) est vérifiée si

$$\Delta t|\lambda| \leq \Delta t \max\left(\sqrt{\frac{2gH}{h^2}\max_{\mathbb{R}}|b(x)|^2 + f^2}, f\right) \leq 2\sqrt{2}, \tag{2.165}$$

avec, pour $x \in \mathbb{R}$,

$$b(x) = \frac{\sin(x)}{\frac{2}{3} + \frac{1}{3}\cos(x)}. \tag{2.166}$$

Or $b(x)$ est borné et vérifie $b(x) \in [-\sqrt{3}, \sqrt{3}]$ pour tout $x \in \mathbb{R}$. La condition (2.165) devient

$$\begin{aligned}\Delta t &\leq \min\left(\frac{2\sqrt{2}}{\sqrt{\frac{2gH}{h^2}\max_{\mathbb{R}}|b(x)|^2 + f^2}}, \frac{2\sqrt{2}}{f}\right)\\ &\leq \min\left(\frac{2\sqrt{2}}{\sqrt{\frac{6gH}{h^2} + f^2}}, \frac{2\sqrt{2}}{f}\right)\\ &\leq \frac{2\sqrt{2}}{\sqrt{\frac{6gH}{h^2} + f^2}},\end{aligned}$$

ce qui démontre le résultat de stabilité attendu. □

D'après les propositions 1.22 et 1.24, le spectre de l'opérateur de filtrage est inclus dans $[-1, 1]$:

$$\mathrm{Sp}\left(\mathcal{F}_{2J}\right) \subset [-1, 1]. \tag{2.167}$$

Même en présence d'un opérateur de filtrage, le schéma est stable sous la condition (2.163). La présence d'un tel opérateur de filtrage atténue les ondes hautes fréquences en dissipant un minimum les ondes basses fréquences. Ainsi, la stabilité est accrue et les oscillations parasites sont atténuées.

De plus, le schéma donné dans l'algorithme 5 conserve la quantité de matière.

Proposition 2.13. *L'algorithme 5 (avec ou sans filtrage) conserve la masse au sens où, pour tout $n \in \mathbb{N}$, on a*

$$(\eta^{n+1}, \mathtt{1})_{h,pér} = (\eta^n, \mathtt{1})_{h,pér} \tag{2.168}$$

où (η^n) est issu de l'algorithme 5.

Démonstration. Pour tout $\mathfrak{b} \in L^2_{h,\text{pér}}$, on a

$$(\delta^H_{4,x}\mathfrak{b}, \mathtt{1})_{h,\text{pér}} = (\delta^H_{4,y}\mathfrak{b}, \mathtt{1})_{h,\text{pér}} = 0, \tag{2.169}$$

car les opérateurs $\delta^H_{4,x}$ et $\delta^H_{4,y}$ sont anti-symétriques. De là, il découle que pour $i \in \{1, 2, 3, 4\}$, on a

$$(K^{(i)}_\eta, \mathtt{1})_{h,\text{pér}} = 0. \tag{2.170}$$

Ainsi, on a

$$(\eta^{n+1}, \mathtt{1})_{h,\text{pér}} = \left(\mathcal{F}_{2J}\left(\eta^n + \frac{\Delta t}{6}\left(K^{(1)}_\eta + 2K^{(2)}_\eta + 2K^{(3)}_\eta + K^{(4)}_\eta\right)\right), \mathtt{1}\right)_{h,\text{pér}}. \tag{2.171}$$

En utilisant la symétrie de $\mathcal{F}_{2J}$ et $\mathcal{F}_{2J}(\mathtt{1}) = \mathtt{1}$, on obtient

$$(\eta^{n+1}, \mathtt{1})_{h,\text{pér}} = \left(\eta^n + \frac{\Delta t}{6}\left(K^{(1)}_\eta + 2K^{(2)}_\eta + 2K^{(3)}_\eta + K^{(4)}_\eta\right), \mathtt{1}\right)_{h,\text{pér}}. \tag{2.172}$$

En utilisant (2.170), on obtient

$$(\eta^{n+1}, \mathtt{1})_{h,\text{pér}} = (\eta^n, \mathtt{1})_{h,\text{pér}}. \tag{2.173}$$

Donc $(\eta^n, \mathtt{1})_{h,\text{pér}}$ est conservé d'une itération à l'autre. □

2.4.2 Résultats numériques

On considère dans cette section un test numérique effectué sur le système périodique (2.122). Les constantes physiques sont

- $g = 9.80616\mathrm{m\,s}^{-2}$,
- $H = 100\mathrm{m}$,
- $f = 2\Omega_c \sin\theta_0$ avec $\Omega_c = 7.292 \times 10^{-5}\mathrm{s}^{-1}$ et $\theta_0 = \pi/2$.

Les données initiales sont $u_0 \equiv 0$ et $v_0 \equiv 0$. La hauteur de fluide initiale est donnée par

$$\eta_0(x,y) = \exp\left(-\frac{r(x,y)^2}{0.01}\right), \tag{2.174}$$

avec $r(x,y) = (x-0.5)^2 + (y-0.5)^2$ et $(x,y) \in [0,1]^2$.

Sur la Figure 2.5, on représente la solution approchée aux temps $t = 0$, $t = 0.5$ et $t = 1$. La Figure 2.6 permet d'une part de confirmer la conservation de la masse et d'autre part d'observer l'erreur relative sur l'énergie au cours du temps.

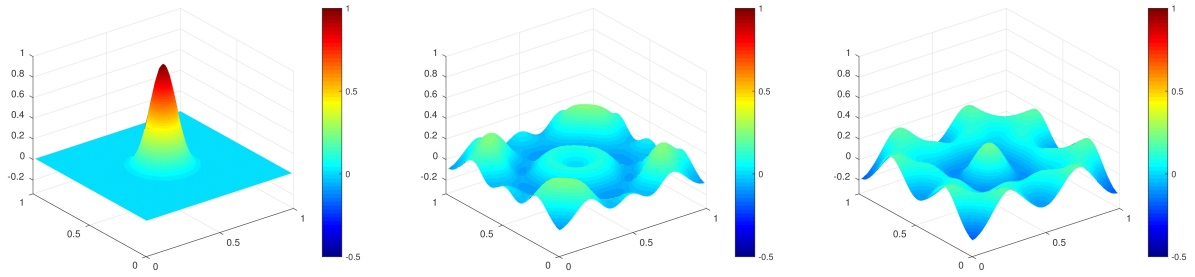


FIGURE 2.5 – Solutions numériques pour l'équation des ondes avec paramètre de Coriolis (2.122) aux temps $t = 0$, $t = 0.5$ et $t = 1$, $N = 64$, $\Delta t_\infty \approx 5.7616 \times 10^{-4}$. La solution est obtenue par l'algorithme 5 avec un filtrage d'ordre 10 pour la condition initiale (2.174).

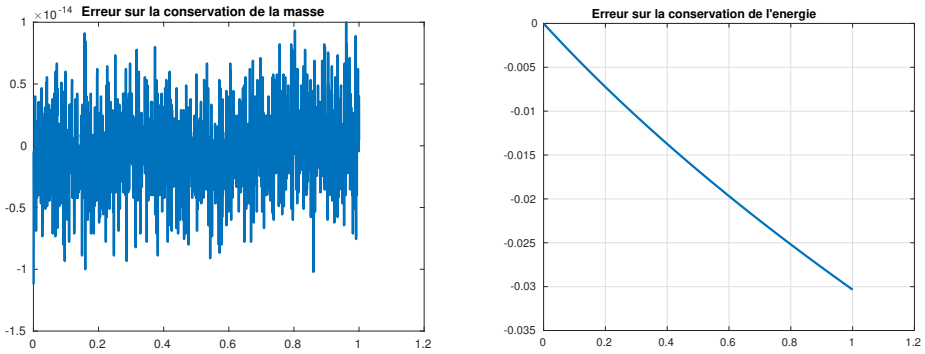


FIGURE 2.6 – Erreur relative lors de la résolution de (2.122) sur la conservation de la masse et de l'énergie obtenue pour le test (2.174) en utilisant l'algorithme 5 avec un filtrage d'ordre 10, $N = 64$, $\Delta t_\infty \approx 5.7616 \times 10^{-4}$.

La Table 2.5 montre la recherche du pas de temps maximal pour lequel l'algorithme 5 est stable pour la donnée initiale (2.174). Le temps final est $t = 2$ et le paramètre de la grille est $N = 32$. On

constate que plus l'ordre du filtre est bas, plus il est possible de choisir un pas de temps grand. Dans la Table 2.6, on observe que choisir un filtrage d'ordre trop bas a des répercussions sur la conservation de l'énergie. Un filtrage d'ordre 2 ne permet pas de conserver l'énergie et un filtrage d'ordre 4 donne une dégradation de l'ordre de convergence sur la conservation. En revanche les filtrages d'ordre 8 et 10 donnent des résultats semblables. On souhaite utiliser un schéma stable pour lequel l'énergie est conservée à un ordre proche de 4. Le filtrage d'ordre 10 est un bon compromis, il atténue les hautes fréquences, ce qui permet une bonne stabilité. De plus, suffisamment de fréquences sont conservées pour permettre une conservation de l'énergie satisfaisante.

Ordre du filtre $\mathcal{F}$: 2J	**Instable pour $\Delta t =$**	**Stable pour $\Delta t =$**
Pas de filtrage	$1.16(-3)$	$\Delta t_\infty \approx 1.15(-3)$
10	$1.23(-3)$	$1.22(-3)$
8	$1.26(-3)$	$1.25(-3)$
6	$1.32(-3)$	$1.31(-3)$
4	$1.41(-3)$	$1.4(-3)$
2	$1.66(-3)$	$1.65(-3)$

TABLE 2.5 – Recherche du pas de temps maximum pour l'algorithme 5 de résolution de l'équation (2.122). On cherche la valeur de Δt maximale pour que la méthode soit stable jusqu'à $t = 2$ pour la donnée initiale (2.174) avec $N = 32$. On constate que plus le filtrage est d'ordre bas, plus le pas de temps peut être choisi grand.

N et Δt_∞	**Pas de filtre**	**Ordre 10**	**Ordre 8**	**Ordre 6**	**Ordre 4**	**Ordre 2**
32 et $1.1523(-3)$	$2.5340(-1)$	$2.5367(-1)$	$2.5709(-1)$	$3.0173(-1)$	$6.2132(-1)$	$9.3717(-1)$
64 et $5.7616(-4)$	$3.0328(-2)$	$3.0347(-2)$	$3.0802(-2)$	$4.3763(-2)$	$3.0419(-1)$	$9.3712(-1)$
128 et $2.8808(-4)$	$1.2226(-3)$	$1.2227(-3)$	$1.2301(-3)$	$1.9323(-3)$	$7.7084(-2)$	$9.3398(-1)$
Ordre :	3.85	3.85	3.85	3.64	1.51	$2.46(-3)$

TABLE 2.6 – Equation (2.122). Convergence de la conservation de l'énergie pour l'algorithme 5 et la donnée initiale (2.174). On représente l'erreur relative maximale pour $t < 1$. Un filtre d'ordre 2 ne permet pas de conserver l'énergie. Avec un filtre d'ordre 4 l'énergie est conservée à un ordre bas. Pour un filtre d'ordre plus élevé, la convergence se fait à un ordre plus élevé.

2.5 Equation de Burgers

Dans cette section, on considère l'équation de Burgers en contexte périodique [6, 14, 86] :

$$\left\{ \begin{array}{rcl} \dfrac{\partial u}{\partial t} + 2\pi \dfrac{\partial}{\partial x}\left(\dfrac{u^2}{2}\right) & = & 0 \\ u(t=0,x) & = & u_0(x) \end{array} \right. \quad \text{pour } x \in [0, 2\pi] \text{ et } t \geq 0. \tag{2.175}$$

Une donnée initiale u_0 périodique non constante, conduit nécessairement à l'apparition d'un choc en temps fini. Le but de cette étude est double :

1. Nous utilisons le schéma hermitien centré $\delta^H_{4,x}$ donné par la définition 1.5 avec le filtre linéaire $\mathcal{F}_{2J,x}$ (1.163) pour l'équation (2.175). Il est bien connu qu'un tel schéma est insuffisant pour obtenir une solution numérique sans oscillations. Le but est ici d'observer précisément la manière dont les filtres linéaires $\mathcal{F}_{2J,x}$ agissent à l'approche du choc. Nous renvoyons à [22, 87] pour des études plus avancées utilisant des opérateurs de type filtre non linéaire, ou des opérateurs d'hyperdiffusion.

2. L'équation de Burgers périodique a été utilisée par [10] pour construire une famille particulière d'équations aux dérivées partielles sur la sphère conduisant à des chocs. Une telle solution sera analysée numériquement dans le chapitre 5.

Nous supposerons que la donnée initiale u_0 est périodique et est telle que $u_0 \in \mathcal{C}^1([0,2\pi])$ et $u_0' \in L^\infty([0,2\pi])$.

Théorème 2.3. *Soit $u_0 \in \mathcal{C}^1$, 2π-périodique. Alors le problème (2.175) admet une unique solution 2π-périodique $u \in \mathcal{C}^1\left(\left[0, -\frac{1}{\inf_{x\in[0,2\pi]}(2\pi u_0'(x))}\right] \times [0,2\pi]\right)$ qui satisfait*

$$u(t, 2\pi u_0(x)t + x) = u_0(x). \tag{2.176}$$

Démonstration. On étudie l'équation (2.175) par la méthode des caractéristiques. Dans un premier temps, on suppose que $u \in \mathcal{C}^1$ solution de (2.175) existe. Soit $X : \mathbb{R}^+ \mapsto \mathbb{R}$ une courbe telle que

$$\left\{ \begin{array}{rcl} X'(t) & = & 2\pi u(t, X(t)) \\ X(0) & = & x \end{array} \right. . \tag{2.177}$$

D'après le théorème de Cauchy-Lipschitz, comme $X \mapsto 2\pi u(t,X)$ est $\mathcal{C}^1$, il existe une solution maximale à (2.177).

Dès lors, on a que $g : t \mapsto u(t, X(t))$ est constante. En effet

$$\begin{aligned} g'(t) &= \frac{\partial u}{\partial t}(t, X(t)) + X'(t)\frac{\partial u}{\partial x}(t, X(t)) \\ &= \frac{\partial u}{\partial t}(t, X(t)) + 2\pi u(t, X(t))\frac{\partial u}{\partial x}(t, X(t)) \\ &= \frac{\partial u}{\partial t}(t, X(t)) + 2\pi \frac{\partial}{\partial x}\left(\frac{u(t, X(t))^2}{2}\right) \\ &= 0. \end{aligned}$$

Comme $t \mapsto u(t, X(t))$ est constante, on a en particulier $2\pi u(t, X(t)) = 2\pi u_0(x)$ donc X est solution de

$$\left\{ \begin{array}{rcl} X'(t) & = & 2\pi u_0(x) \\ X(0) & = & x \end{array} \right. . \tag{2.178}$$

Donc X est donné par

$$X(t) = 2\pi u_0(x)t + x. \tag{2.179}$$

u est constante le long de X donc u vérifie

$$u(t, 2\pi u_0(x)t + x) = u_0(x). \tag{2.180}$$

Cependant, ce résultat n'est vrai que si (2.175) admet une solution de classe $\mathcal{C}^1$. Posons $X_t(x) = 2\pi u_0(x)t + x$, donc

$$X_t'(x) = 2\pi u_0'(x)t + 1. \tag{2.181}$$

Comme u_0 est régulière sur un compact, il existe $m \in \mathbb{R}$ tel que

$$m = \inf_{x\in[0,2\pi]} \left(2\pi u_0'(x)\right). \tag{2.182}$$

Trois cas de figures se présentent alors :

- Le cas $m > 0$ est impossible. En effet, si $m > 0$, alors pour tout $x \in [0, 2\pi]$, on a $u_0' > 0$, donc

$$\begin{aligned} u_0(2\pi) &= u_0(0) + \int_0^{2\pi} u_0'(\tau)d\tau \\ &\geq u_0(0) + m \\ &> u_0(0) \end{aligned}$$

  et u_0 n'est pas périodique (car $u_0(0) > u_0(2\pi)$), ce qui est absurde.

- Si $m = 0$ alors pour tout $x \in [0, 2\pi]$ et $t > 0$, $u_0'(x) \geq 0$ et u_0 périodique. Donc u_0 est constante et on a
$$u(x,t) = u_0(x) = C^{\text{ste}}. \tag{2.183}$$
La solution u existe et est définie pour $t \in [0, +\infty[$.

- Si $m < 0$ alors en posant $T = -1/m$ (si $m = 0$, on a $T = +\infty$), on a pour tout $t \in [0, T[$
$$X_t'(x) \geq m(t - T) > 0 \tag{2.184}$$
alors X_t est i.e bijection de $[0, 2\pi]$ dans $\mathbb{R}$.

Si $u \in \mathcal{C}^1$ est solution (2.175) telle que X_t est i.e bijection alors
$$u(t,x) = u_0(X_t^{-1}(x)) \tag{2.185}$$
et X_t est une bijection si $m \geq 0$ ou si $m < 0$ et $t \in [0, T[$.

Vérifions que $u : (t,x) \mapsto u(t,x) = u_0(X_t^{-1}(x))$ est bien solution et de classe $\mathcal{C}^1$ (pour t tel que X_t^{-1} est bien défini). X_t est de classe $\mathcal{C}^1$, si $X_t'(x) > 0$ pour tout $x \in [0, 2\pi]$ alors X_t est inversible, X_t^{-1} est $\mathcal{C}^1$. Pour tout $x \in [0, 2\pi]$, on a
$$(X_t^{-1})'(x) = \frac{1}{X_t'(X_t^{-1}(x))}, \tag{2.186}$$
donc en dérivant u par rapport à x, on obtient
$$\begin{aligned}\frac{\partial u}{\partial x}(t,x) &= \frac{u_0'(X_t^{-1}(x))}{X_t'(X_t^{-1}(x))} \\ &= \frac{u_0'(X_t^{-1}(x))}{1 + 2\pi u_0'(X_t^{-1}(x))},\end{aligned}$$
d'où la formule suivante
$$\frac{\partial u}{\partial x}(t,x) = \frac{u_0'(X_t^{-1}(x))}{1 + 2\pi u_0'(X_t^{-1}(x))}. \tag{2.187}$$
Or, on sait que
$$\begin{aligned}X_t(x) &= x + 2\pi t u_0(x) \\ &= x + 2\pi t u(t, 2\pi u_0(x)t + x) \\ &= x + 2\pi t u(t, X_t(x)).\end{aligned}$$
Donc $x = X_t(x) - 2\pi t u(t, X_t(x))$, d'où
$$X_t^{-1}(x) = x - 2\pi t u(t,x). \tag{2.188}$$
De là, on peut déduire la dérivée de X_t^{-1} en temps :
$$\frac{\partial X_t^{-1}}{\partial t}(x) = -2\pi u(t,x) - 2\pi t \frac{\partial u}{\partial t}(t,x). \tag{2.189}$$
On en déduit :
$$\begin{aligned}\frac{\partial u}{\partial t}(tx,) &= \frac{\partial}{\partial t} u_0(X_t^{-1}(x)) \\ &= \frac{\partial X_t^{-1}}{\partial t}(x) u_0'(X_t^{-1}(x)) \\ &= (-2\pi u(t,x) - 2\pi t \frac{\partial u}{\partial t}(t,x)) u_0'(X_t^{-1}(x)).\end{aligned}$$

Il découle :

$$\frac{\partial u}{\partial t}(t,x)(1+2\pi t u_0'(X_t^{-1}(x))) = -2\pi u(t,x)u_0'(X_t^{-1}(x)). \tag{2.190}$$

Donc, en divisant, on a

$$\begin{aligned}\frac{\partial u}{\partial t}(t,x) &= -\frac{2\pi u(t,x)u_0'(X_t^{-1}(x))}{1+2\pi t u_0'(X_t^{-1}(x))}\\ &= -2\pi u(t,x)\frac{u_0'(X_t^{-1}(x))}{1+2\pi u_0'(X_t^{-1}(x))}\\ &= -2\pi u(t,x)\frac{\partial u}{\partial x}(t,x) \text{ d'après (2.187).}\\ &= -2\pi\frac{\partial}{\partial x}\left(\frac{u(t,x)^2}{2}\right).\end{aligned}$$

Il est à vérifier que u est 2π-périodique pour la variable x. X_t est bijective donc pour tout x, il existe un unique y tel que $y = X_t^{-1}(x)$. Alors, par définition de X_t, on a

$$X_t(y) = 2\pi u_0(y)t + y, \tag{2.191}$$

d'où

$$x = 2\pi u_0(X_t^{-1}(x))t + X_t^{-1}(x). \tag{2.192}$$

Cette dernière égalité entraîne

$$\begin{aligned}x+2\pi &= 2\pi u_0(X_t^{-1}(x))t + X_t^{-1}(x) + 2\pi\\ &= 2\pi u_0(X_t^{-1}(x)+2\pi)t + X_t^{-1}(x) + 2\pi \text{ par périodicité de } u_0,\\ &= X_t(X_t^{-1}(x)+2\pi) \text{ par définition de } X_t.\end{aligned}$$

En tenant compte de ce résultat, on a

$$\begin{aligned}u(t,x+2\pi) &= u_0(X_t^{-1}(x+2\pi))\\ &= u_0(X_t^{-1}(x)+2\pi)\\ &= u_0(X_t^{-1}(x))\\ &= u(t,x).\end{aligned}$$

Ainsi u est périodique. Donc u est bien solution de (2.175) et par composition c'est une fonction $\mathcal{C}^1$ sur les intervalles souhaités. □

Exemple : Si u_0 est définie pour $x \in [0, 2\pi]$ par

$$u_0(x) = \sin(x) \tag{2.193}$$

l'équation (2.175) admet une unique solution de classe $\mathcal{C}^1$ pour $t \in [0, 1/(2\pi)]$. Au delà de cet intervalle, des chocs apparaissent et la solution de (2.175) n'est plus régulière.

Nous observons les performances du schéma numérique donné par l'algorithme 6 pour l'équation (2.175) où $\mathbf{u}^n$ est une approximation de $u(t^n, \cdot)^*$.

Algorithme 6 : Schémas en temps RK4 avec étape de filtrage pour l'équation périodique (2.175)

$\mathtt{u}^0 = u_0^*$ connu,
for $n = 0, 1, \ldots$ **do**
  $K^{(1)} = -\pi \delta_{4,x}^H \left((\mathtt{u}^n)^2 \right)$,
  $K^{(2)} = -\pi \delta_{4,x}^H \left(\left(\mathtt{u}^n + \frac{\Delta t}{2} K^{(1)} \right)^2 \right)$,
  $K^{(3)} = -\pi \delta_{4,x}^H \left(\left(\mathtt{u}^n + \frac{\Delta t}{2} K^{(2)} \right)^2 \right)$,
  $K^{(4)} = -\pi \delta_{4,x}^H \left(\left(\mathtt{u}^n + \Delta t K^{(3)} \right)^2 \right)$,
  $\mathtt{u}^{n+1} = \mathcal{F}_{2J,x} \left(\mathtt{u}^n + \frac{\Delta t}{6} \left(K^{(1)} + 2K^{(2)} + 2K^{(3)} + K^{(4)} \right) \right)$.
end for

On peut montrer la propriété de conservation suivante :

Proposition 2.14. *Pour tout $n \in \mathbb{N}$, si $(\mathtt{u}^n)$ est calculée par l'algorithme 6 alors*

$$(\mathtt{u}^{n+1}, \mathtt{1})_{h,pér} = (\mathtt{u}^n, \mathtt{1})_{h,pér}. \tag{2.194}$$

où $\mathtt{1}$ est la fonction de grille constante égale à 1.

Les résultats sont donnés pour la condition initiale u_0 (2.193) et pour l'algorithme 6 sans filtrage sur la Figure 2.7. Des oscillations apparaissent et dégradent le résultat. Au temps $t = 0.5$, les oscillations deviennent beaucoup trop importantes et l'ordinateur ne donne plus de résultats.

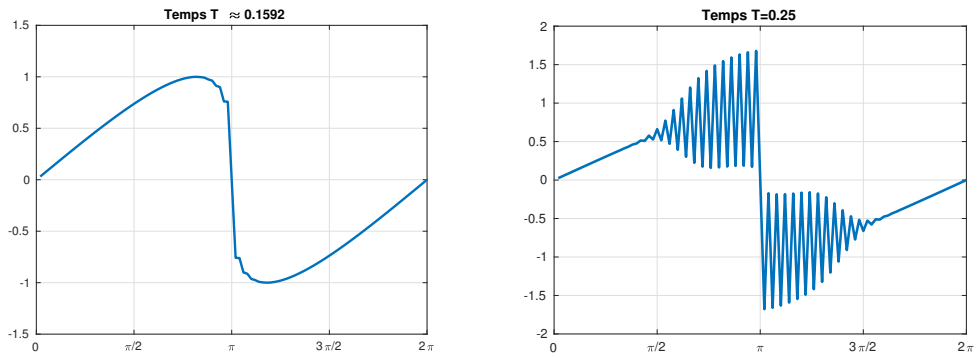


FIGURE 2.7 – Résultats pour l'équation (2.175) résolue par l'algorithme 6 sans opérateur de filtrage (c'est à dire $\mathcal{F}_{2J,x} = \mathrm{Id}$) à différents temps pour la résolution de l'équation (2.175). On choisit ici $N = 100$ et $\Delta t = 10^{-3}$. Le temps d'apparition du choc est $t = 1/(2\pi) \approx 0.1592$.

Sur la Figure 2.8, on compare la solution obtenue à l'aide de l'algorithme 6 en utilisant différents ordres pour l'opérateur de filtrage.

Le filtrage d'ordre 2 est trop dissipatif. Les filtrages d'ordres plus élevés représentent mieux le choc et permettent au schéma de rester stable jusqu'à $t = 10/(2\pi)$ au minimum. Le filtre d'ordre 10 est un bon compromis entre la précision souhaitée, la stabilité, la conservation et la bonne représentation des chocs. En effet, il conserve suffisamment d'ondes pour représenter le choc sans que des ondes parasites rendent le calcul impossible. Comme nous utilisons un schéma centré, des oscillations restent cependant présente mais un tel schéma permet de vérifier exactement la conservation de la masse.

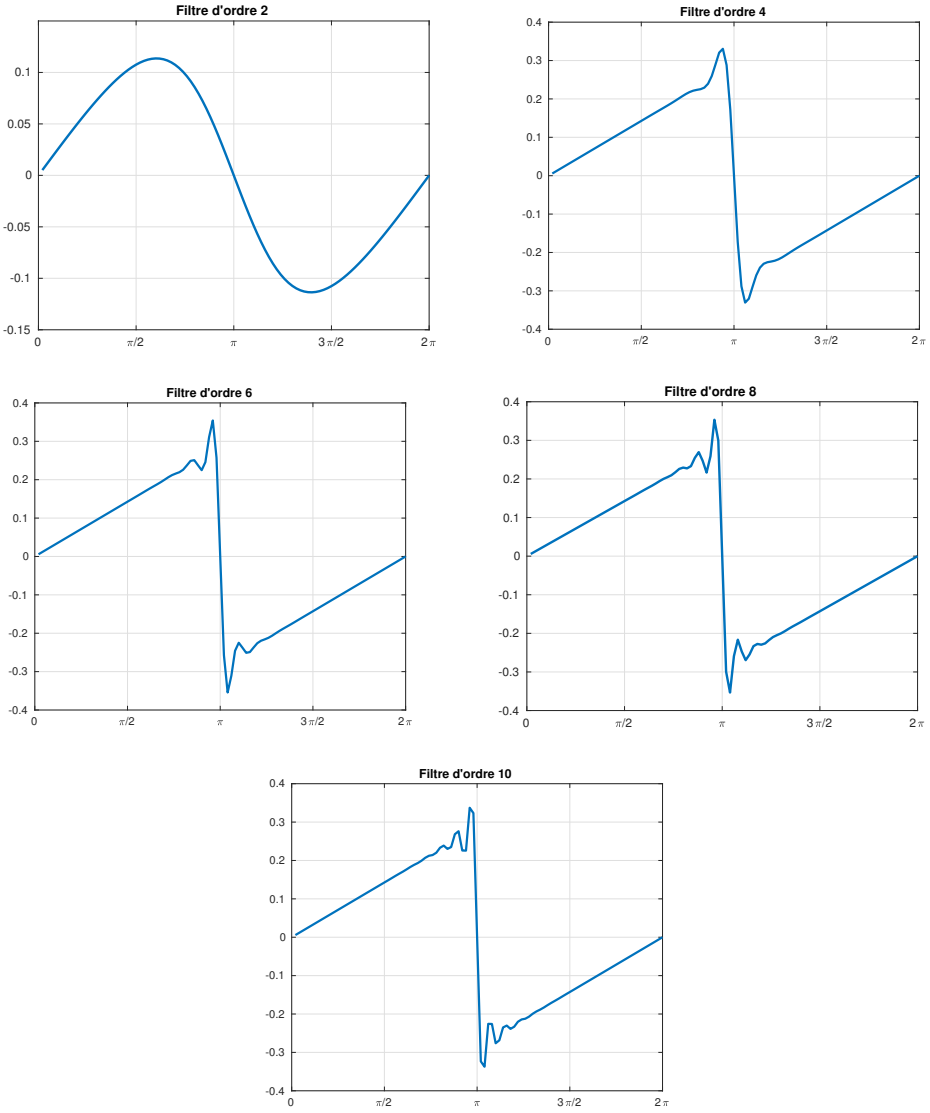


FIGURE 2.8 – Résultats pour l'équation (2.175) résolue par l'algorithme 6 avec différents opérateurs de filtrage pour la résolution de l'équation (2.175). On choisit ici $N = 100$ et $\Delta t = 10^{-3}$. Le temps final est $T = 10/(2\pi) \approx 1.5915$. On présente les résultats pour les filtres d'ordres 2, 4, 6, 8 et 10.

Chapitre 3

Grille Cubed-Sphere

3.1 Définition géométrique de la Cubed-Sphere

3.1.1 La sphère $\mathbb{S}^2_a$

L'objectif de ce chapitre est de construire le maillage qui sera utilisée pour résoudre des équations aux dérivées partielles sur la Sphère. Pour cela, nous avons besoin de définitions utiles dans la suite. On note $\mathbb{S}^2_a$ *la sphère* de centre $\mathbf{O}(0,0,0) \in \mathbb{R}^3$ et de rayon $a > 0$:

$$\mathbb{S}^2_a = \left\{ \mathbf{x}(x,y,z) \in \mathbb{R}^3 \text{ tels que } x^2 + y^2 + z^2 = a^2 \right\}. \tag{3.1}$$

Un *grand cercle* est un cercle de centre $\mathbf{O}$ et de rayon a tracé sur la sphère $\mathbb{S}^2_a$. Soit C un grand cercle et $\mathbf{x}_0 \in C$ un point fixé. On choisit l'un des deux sens de parcours le long de C à partir de $\mathbf{x}_0$ et on définit *l'abscisse curviligne* de $\mathbf{x} \in C$ par la distance séparant $\mathbf{x}$ de $\mathbf{x}_0$ le long de C. La relation suivante est vérifiée :

$$\text{longarc}(\mathbf{x}_0\mathbf{x}) = a\alpha \tag{3.2}$$

où $\alpha \in [0, 2\pi[$ désigne l'angle $\widehat{\mathbf{x_0}\mathbf{x}} = (\mathbf{O}\mathbf{x_0}, \mathbf{O}\mathbf{x})$ dans le sens choisi (Fig. 3.1). L'angle α est *l'angle géodésique* entre $\mathbf{x}_0$ et $\mathbf{x}$.

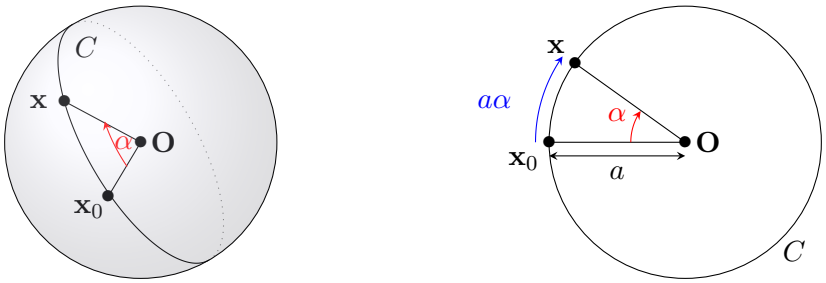


FIGURE 3.1 – Un grand cercle sur la sphère $\mathbb{S}^2_a$. L'angle α est tel que $\widehat{\mathbf{x_0}\mathbf{x}} = \alpha$ et l'abscisse curviligne de $\mathbf{x}$ comptée à partir de $\mathbf{x}_0$ est $a\alpha$.

Soit $\overline{\mathbf{x}} \in \mathbb{S}^2_a$ fixé. Soient C_1 et C_2 deux grands cercles distincts arbitraires tels que $\overline{\mathbf{x}} \in C_1 \cap C_2$. Soient α et β les abscisses curvilignes le long de C_1 et C_2, définies à partir de $\mathbf{x}_{1,0} \in C_1$ et $\mathbf{x}_{2,0} \in C_2$ arbitrairement choisis. On définit les vecteurs $\mathbf{e}_\alpha(\overline{\mathbf{x}})$ et $\mathbf{e}_\beta(\overline{\mathbf{x}})$ (Fig. 3.2) par

$$\mathbf{e}_\alpha(\overline{\mathbf{x}}) = \frac{d\mathbf{x}}{d\alpha}(\overline{\mathbf{x}}) \text{ et } \mathbf{e}_\beta(\overline{\mathbf{x}}) = \frac{d\mathbf{x}}{d\beta}(\overline{\mathbf{x}}). \tag{3.3}$$

Les vecteurs $\mathbf{e}_\alpha(\overline{\mathbf{x}})$ et $\mathbf{e}_\beta(\overline{\mathbf{x}})$ sont tangents aux cercles C_1 et C_2. On note $\alpha \mapsto \mathbf{x}(\alpha)$ (resp. $\beta \mapsto \mathbf{x}(\beta)$) le paramétrage du cercle C_1 (resp. C_2) par l'angle α (resp. β).

En tout point $\mathbf{x} \in \mathbb{S}^2_a$, le *plan tangent* $\mathbb{T}_\mathbf{x}\mathbb{S}^2_a$ est défini par :

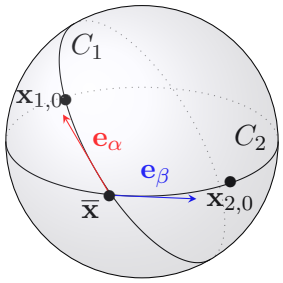


FIGURE 3.2 – Vecteurs $\mathbf{e}_\alpha$ et $\mathbf{e}_\beta$ associés au cercles C_1 et C_2 de $\mathbb{S}^2_a$.

Définition 3.1. *On appelle plan tangent à la sphère $\mathbb{S}^2_a$ au point $\overline{\mathbf{x}}$ le plan :*

$$\mathbb{T}_{\overline{\mathbf{x}}}\mathbb{S}^2_a = \left\{ \mathbf{m} \in \mathbb{R}^3 \ \overrightarrow{\overline{\mathbf{x}}\mathbf{m}} \perp \overrightarrow{O\overline{\mathbf{x}}} \right\}. \tag{3.4}$$

Proposition 3.1. *Soit $\overline{\mathbf{x}} \in \mathbb{S}^2_a$, alors :*

$$\mathbf{e}_\alpha(\overline{\mathbf{x}}),\ \mathbf{e}_\beta(\overline{\mathbf{x}}) \in \mathbb{T}_{\overline{\mathbf{x}}}\mathbb{S}^2_a. \tag{3.5}$$

Démonstration. Soit $\mathbf{x}$ est un point quelconque de $\mathbb{S}^2_a$. Nous adoptons la notation $\|\mathbf{x}\| = \|\mathbf{Ox}\|$. On a $\|\mathbf{x}\|^2 = \mathbf{x}\cdot\mathbf{x} = a^2$ constant. Alors en dérivant par rapport à α, on a :

$$\begin{aligned} 0 &= \frac{d}{d\alpha}(a^2) \\ &= \frac{d}{d\alpha}(\mathbf{x}\cdot\mathbf{x}) \\ &= 2\mathbf{x}\cdot\frac{d\mathbf{x}}{d\alpha}. \end{aligned}$$

En particulier, en $\overline{\mathbf{x}}$, on a :

$$\overline{\mathbf{x}}\cdot\mathbf{e}_\alpha(\overline{\mathbf{x}}) = 0 \tag{3.6}$$

donc $\mathbf{e}_\alpha(\overline{\mathbf{x}}) \in \mathbb{T}_{\overline{\mathbf{x}}}\mathbb{S}^2_a$. En dérivant par rapport à β, on a $\mathbf{e}_\beta(\overline{\mathbf{x}}) \in \mathbb{T}_{\overline{\mathbf{x}}}\mathbb{S}^2_a$. □

Les vecteurs $\mathbf{e}_\alpha(\overline{\mathbf{x}})$ et $\mathbf{e}_\beta(\overline{\mathbf{x}})$ sont dans le plan $\mathbb{T}_{\overline{\mathbf{x}}}\mathbb{S}^2_a$. De plus, $C_1 \neq C_2$ donc $\mathbf{e}_\alpha(\overline{\mathbf{x}})$ et $\mathbf{e}_\beta(\overline{\mathbf{x}})$ ne sont pas colinéaires. On en déduit que $\mathbf{e}_\alpha(\overline{\mathbf{x}})$ et $\mathbf{e}_\beta(\overline{\mathbf{x}})$ engendrent $\mathbb{T}_{\overline{\mathbf{x}}}\mathbb{S}^2_a$. En général, $\mathbf{e}_\alpha(\overline{\mathbf{x}})$ et $\mathbf{e}_\beta(\overline{\mathbf{x}})$ ne sont pas orthogonaux.

Définition 3.2. *On définit $(\mathbf{e}^\alpha(\overline{\mathbf{x}}), \mathbf{e}^\beta(\overline{\mathbf{x}}))$ la base duale de $(\mathbf{e}_\alpha(\overline{\mathbf{x}}), \mathbf{e}_\beta(\overline{\mathbf{x}}))$. Les vecteurs $\mathbf{e}^\alpha(\overline{\mathbf{x}})$ et $\mathbf{e}^\beta(\overline{\mathbf{x}})$ sont les vecteurs de $\mathbb{T}_{\overline{\mathbf{x}}}\mathbb{S}^2_a$ vérifiant :*

$$\left\{ \begin{array}{ccccc} \mathbf{e}_\alpha(\overline{\mathbf{x}})\cdot\mathbf{e}^\alpha(\overline{\mathbf{x}}) & = & 1 & = & \mathbf{e}_\beta(\overline{\mathbf{x}})\cdot\mathbf{e}^\beta(\overline{\mathbf{x}}) \\ \mathbf{e}_\alpha(\overline{\mathbf{x}})\cdot\mathbf{e}^\beta(\overline{\mathbf{x}}) & = & 0 & = & \mathbf{e}_\beta(\overline{\mathbf{x}})\cdot\mathbf{e}^\alpha(\overline{\mathbf{x}}). \end{array} \right. \tag{3.7}$$

Proposition 3.2. *La base $(\mathbf{e}^\alpha(\overline{\mathbf{x}}), \mathbf{e}^\beta(\overline{\mathbf{x}}))$, duale de $(\mathbf{e}_\alpha(\overline{\mathbf{x}}), \mathbf{e}_\beta(\overline{\mathbf{x}}))$ est définie de façon unique.*

Démonstration. Il suffit de remarquer que

$$\left[\mathbf{e}_\alpha(\overline{\mathbf{x}}), \mathbf{e}_\beta(\overline{\mathbf{x}})\right]^T \cdot \left[\mathbf{e}^\alpha(\overline{\mathbf{x}}), \mathbf{e}^\beta(\overline{\mathbf{x}})\right] = \mathrm{Id}. \tag{3.8}$$

Alors

$$\left[\mathbf{e}^\alpha(\overline{\mathbf{x}}), \mathbf{e}^\beta(\overline{\mathbf{x}})\right] = \left[\mathbf{e}_\alpha(\overline{\mathbf{x}}), \mathbf{e}_\beta(\overline{\mathbf{x}})\right]^{-T}. \tag{3.9}$$

□

A partir de ces différentes bases, le *gradient* est donné par la définition suivante.

Définition 3.3. *Soit $h : \mathbb{S}^2_a \to \mathbb{R}$ une fonction régulière. Le gradient de h, $\nabla_T h(\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}^2_a$, est défini par*

$$\nabla_T h(\overline{\mathbf{x}}) = \frac{\partial}{\partial \alpha}\left(h(\overline{\mathbf{x}})\right)_{|\mathbf{x}\in C_1} \mathbf{e}^{\alpha}(\overline{\mathbf{x}}) + \frac{\partial}{\partial \beta}\left(h(\overline{\mathbf{x}})\right)_{|\mathbf{x}\in C_2} \mathbf{e}^{\beta}(\overline{\mathbf{x}}). \tag{3.10}$$

Pour alléger les notations, nous noterons abusivement les quantités h, $\mathbf{e}_\alpha$, $\mathbf{e}^\beta$ etc. au lieu de $h(\bar{\mathbf{x}})$, $\mathbf{e}_\alpha(\bar{\mathbf{x}})$, $\mathbf{e}^\beta(\bar{\mathbf{x}})$, etc.

Proposition 3.3. *Soit $u : \mathbb{R}^3 \mapsto \mathbb{R}$ et $\mathbf{x} \in \mathbb{R}^3$. On pose $\hat{u} := u_{|\mathbb{S}^2_a}$ la restriction de u à la sphère $\mathbb{S}^2_a$. Alors $\nabla_T \hat{u}(\mathbf{x})$ est la projection orthogonale de $\nabla_{\mathbb{R}^3} u(\mathbf{x})$ sur le plan tangent $\mathbb{T}_{\mathbf{x}}\mathbb{S}^2_a$:*

$$\nabla_T \hat{u} = \nabla_{\mathbb{R}^3} u - \mathbf{n}\,(\mathbf{n} \cdot \nabla_{\mathbb{R}^3} u) \tag{3.11}$$

avec $\mathbf{n}$ la normale unitaire extérieure.

Démonstration. On montre que les deux termes sont égaux dans trois directions distinctes. On pose $\mathbf{n}$ la normale unitaire extérieure à la sphère $\mathbb{S}^2_a$.

- Direction $\mathbf{e}_\alpha$: d'une part on a :

$$\nabla_T \hat{u} \cdot \mathbf{e}_\alpha = \frac{\partial \hat{u}}{\partial \alpha}_{|\mathbf{x}\in C_1} = \frac{\partial u}{\partial \alpha}_{|\mathbf{x}\in C_1}. \tag{3.12}$$

  D'autre part, on a :

$$(\nabla_{\mathbb{R}^3} u - \mathbf{n}\,(\mathbf{n} \cdot \nabla_{\mathbb{R}^3} u)) \cdot \mathbf{e}_\alpha = \nabla_{\mathbb{R}^3} u \cdot \mathbf{e}_\alpha \tag{3.13}$$

  car $\mathbf{n}$ est normal à la sphère donc $\mathbf{n}$ est normal à $\mathbf{e}_\alpha$. Or :

$$\nabla_{\mathbb{R}^3} u \cdot \mathbf{e}_\alpha = \frac{\partial u}{\partial \alpha}_{|\mathbf{x}\in C_1} = \nabla_T \hat{u} \cdot \mathbf{e}_\alpha. \tag{3.14}$$

- Dans la direction $\mathbf{e}_\beta$, on a de la même manière :

$$\nabla_{\mathbb{R}^3} u \cdot \mathbf{e}_\beta = \nabla_T \hat{u} \cdot \mathbf{e}_\beta. \tag{3.15}$$

- Dans la direction $\mathbf{n}$ on a d'une part :

$$\nabla_T \hat{u} \cdot \mathbf{n} = 0 \tag{3.16}$$

  car $\nabla_T \hat{u}$ est tangent à la sphère. D'autre part :

$$(\nabla_{\mathbb{R}^3} u - \mathbf{n}\,(\mathbf{n} \cdot \nabla_{\mathbb{R}^3} u)) \cdot \mathbf{n} = \nabla_{\mathbb{R}^3} u \cdot \mathbf{n} - \mathbf{n} \cdot \nabla_{\mathbb{R}^3} u = 0. \tag{3.17}$$

Or, $(\mathbf{e}_\alpha, \mathbf{e}_\beta, \mathbf{n})$ est une base de $\mathbb{R}^3$, d'où l'égalité souhaitée. □

3.1.2 Définition de la Cubed-Sphere

Dans cette partie, on définit la Grille Cubed-Sphere associée à la base orthonormée $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ de $\mathbb{R}^3$. On définit les points suivants (Voir Figure 3.3) sur la sphère $\mathbb{S}^2_a$:

- **N** le point de coordonnées dans $\mathbb{R}^3$: $(0, 0, a)$,
- **S** le point de coordonnées $(0, 0, -a)$,
- **E** le point de coordonnées $(0, a, 0)$,
- **W** le point de coordonnées $(0, -a, 0)$,
- **F** le point de coordonnées $(a, 0, 0)$,

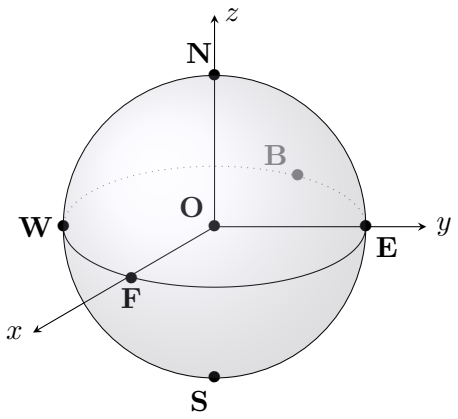
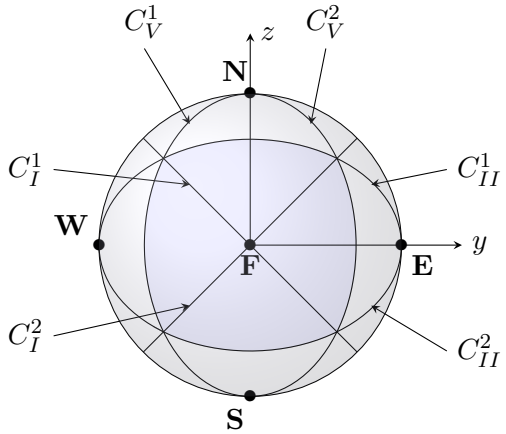


FIGURE 3.3 – Sphère $\mathbb{S}_a^2$ avec les 6 points **N** (Nord), **S** (Sud), **E** (Est), **W** (Ouest), **F** (Avant) et **B** (Arrière).

- **B** le point de coordonnées $(-a, 0, 0)$.

Un grand cercle est l'intersection de la sphère $\mathbb{S}_a^2$ avec un plan contenant le point **O**. On définit les grands cercles suivants (Fig. 3.4) :

- $C_V^1 = \text{Vect}(\mathbf{i} - \mathbf{j}, \mathbf{k}) \cap \mathbb{S}_a^2$,
- $C_V^2 = \text{Vect}(\mathbf{i} + \mathbf{j}, \mathbf{k}) \cap \mathbb{S}_a^2$, il s'agit d'une rotation de C_V^1 d'un angle de $\pi/2$ autour de (Oz),
- $C_{II}^1 = \text{Vect}(\mathbf{i} + \mathbf{k}, \mathbf{j}) \cap \mathbb{S}_a^2$,
- $C_{II}^2 = \text{Vect}(\mathbf{i} - \mathbf{k}, \mathbf{j}) \cap \mathbb{S}_a^2$, le cercle C_{II}^2 est une rotation de C_{II}^1 d'un angle de $\pi/2$ autour de (Oy),
- $C_I^1 = \text{Vect}(\mathbf{j} - \mathbf{k}, \mathbf{i}) \cap \mathbb{S}_a^2$,
- $C_I^2 = \text{Vect}(\mathbf{j} + \mathbf{k}, \mathbf{i}) \cap \mathbb{S}_a^2$, C_I^2 est une rotation de C_I^1 autour de (Ox) d'un angle de $\pi/2$.



Panel (I)

FIGURE 3.4 – Les 6 grands cercles C_I^1, C_I^2, C_{II}^1, C_{II}^2, C_V^1 et C_V^1 vus depuis le panel (I).

La construction de la Cubed-Sphere fait intervenir six zones recouvrant la sphère $\mathbb{S}_a^2$ appelées *panels*. Chaque panel est délimité par quatre grands cercles (Fig. 3.5).

Définition 3.4.
- ***Panels* (I) *et* (III) *:*** *Les panels (I) et (III) sont définis par les points de $\mathbb{S}^2_a$ délimités par les grands cercles C^1_V, C^2_V, C^1_{II} et C^2_{II}. Le panel (III) est le symétrique du panel (I) par la symétrie ponctuelle de centre* **O**. *Le panel (I) ne contient que des points (x, y, z) tels que $x > 0$, le panel (III) ne contient que des points tels que $x < 0$.*
- ***Panels* (II) *et* (IV) *:*** *Les panels (II) et (IV) sont définis par les points de $\mathbb{S}^2_a$ délimités par les grands cercles C^1_V, C^2_V, C^1_I et C^2_I. Le panel (IV) est le symétrique du panel (II) par la symétrie ponctuelle de centre* **O**. *Le panel (II) ne contient que des points (x, y, z) tels que $y > 0$, le panel (IV) ne contient que des points tels que $y < 0$.*
- ***Panels* (V) *et* (VI) *:*** *Les panels (V) et (VI) sont définis par les points de $\mathbb{S}^2_a$ délimités par les grands cercles C^1_I, C^2_I, C^1_{II} et C^2_{II}. Le panel (VI) est le symétrique du panel (V) par la symétrie ponctuelle de centre* **O**. *Le panel (V) ne contient que des points (x, y, z) tels que $z > 0$, le panel (VI) ne contient que des points tels que $z < 0$.*

La grille Cubed-Sphere est constituée de l'intersection d'un ensemble de grands cercles sur chaque panel. On commence par introduire le système de coordonnées (ξ, η) associés aux cercles centraux de chaque panel. Les paramètres ξ et η sont définis comme suit :

Définition 3.5. *Les lignes de coordonnées $\xi = 0$ et $\eta = 0$ sont les grands cercles équatoriaux, notés respectivement $C^{(1)}_0$ et $C^{(2)}_0$ (en pointillés gras sur la Figure sur 3.6 et sur 3.7). Les deux cercles passent par les points centraux de chaque panel.*

1. *$C^{(1)}_0$ est le grand cercle passant par les points* **N**, **F** *et* **S**.
2. *$C^{(2)}_0$ est le grand cercle passant par les points* **E**, **F** *et* **W**.

Les cercles $C^{(1)}_0$ et $C^{(2)}_0$ se coupent orthogonalement en **F** *et* **B**. *La donnée ξ est l'angle géodésique mesuré sur $C^{(2)}_0$ et η l'angle géodésique mesuré sur $C^{(1)}_0$. La valeur $\xi = 0$ correspond à l'équateur ordinaire et $\eta = 0$ correspond à "l'équateur vertical".*

Sur chaque panel, tout $\mathbf{x}$ du panel est localisé par les angles géodésiques ξ et η (voir Fig. 3.6). Un panel est donné par le domaine

$$-\frac{\pi}{4} \leq \xi, \eta \leq \frac{\pi}{4}, \tag{3.18}$$

Soit $\mathbf{x}_{i,j}$ un point du panel $(k) \in \{(I), (II), (III), (IV), (V), (VI)\}$. C'est un point du maillage si ses coordonnées (ξ_i, η_j) sont données par :

$$\xi_i = i\Delta\xi \text{ et } \eta_j = j\Delta\eta \text{ avec } -\frac{N}{2} \leq i, j \leq \frac{N}{2}, \tag{3.19}$$

Dans ce qui suit, nous supposerons N pair. $\Delta\xi$ et $\Delta\eta$ représentent le pas angulaire séparant régulièrement les grands cercles dans le système de coordonnées (ξ, η). On a

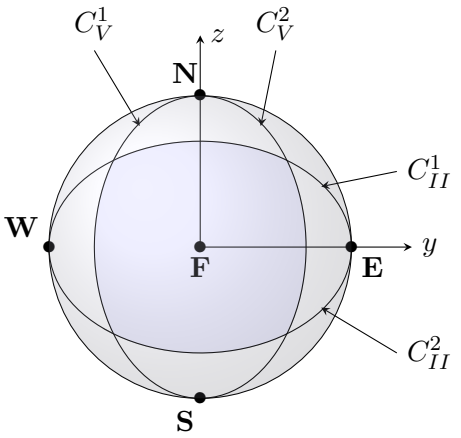
$$\Delta\xi = \Delta\eta = \frac{\pi}{2N}. \tag{3.20}$$

Le cercle $C^{(2)}_0$ fait le tour de la sphère en formant un angle longitudinal de 2π, on peut donc insérer les 4 panels notés (I), (II), (III) et (IV). De même, le long de $C^{(1)}_0$, sont présents les panels (I), (V), (III) et (VI).

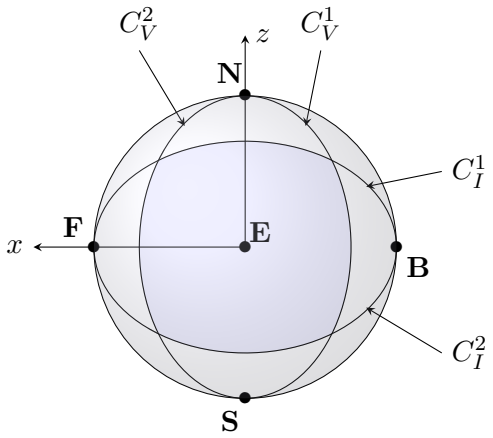
On note $C^{(1)}_i$ le grand cercle obtenu par rotation de $C^{(1)}_0$ d'un angle géodésique $i\Delta\xi$ autour de l'axe (Oz) et $C^{(2)}_j$ le cercle obtenu par rotation de $C^{(2)}_0$ d'angle $j\Delta\eta$ autour de (Oy).

Le maillage associé au panel (I) est constitué des points d'intersections des $N+1$ cercles $(C^{(1)}_i)_{-N/2 \leq i \leq N/2}$ et des $N+1$ cercles $(C^{(2)}_j)_{-N/2 \leq j \leq N/2}$ (voir Figure 3.7) sur le panel (I). Le même procédé est reproduit sur chaque panel, il y a donc $(N+1)^2$ points d'intersections sur un panel.

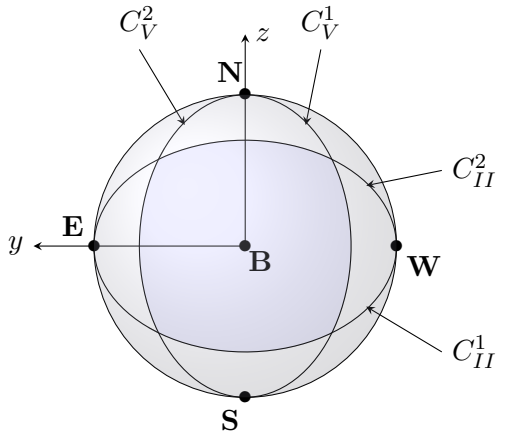
En reproduisant le procédé pour chaque panel, on constitue la Cubed-Sphere associée à la base $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ et de paramètre $N \in \mathbb{N}^\star$.



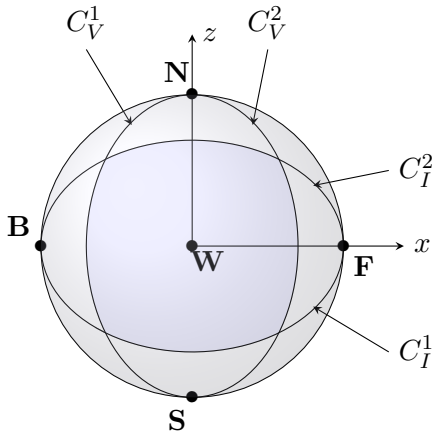
Panel (I)



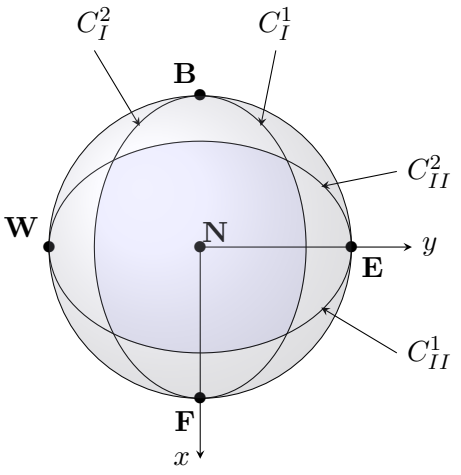
Panel (II)



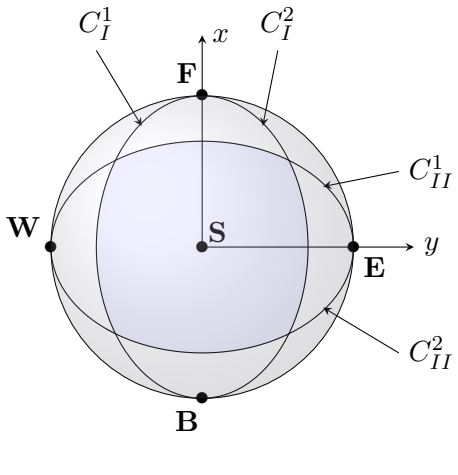
Panel (III)



Panel (IV)



Panel (V)



Panel (VI)

FIGURE 3.5 – Délimitations des panels (I) à (VI) à l'aide des grands cercles

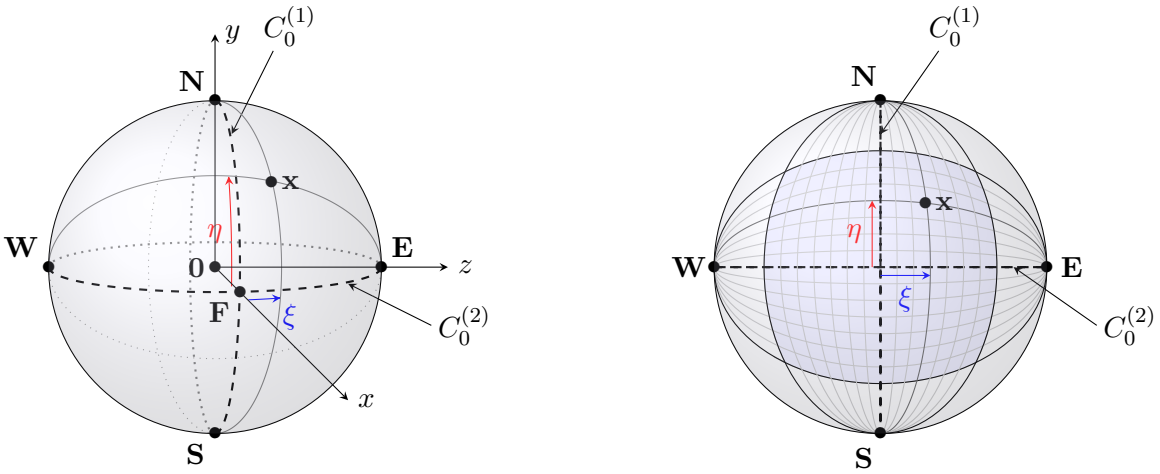


FIGURE 3.6 – Sur un panel, un point **x** est localisé par ξ et η.

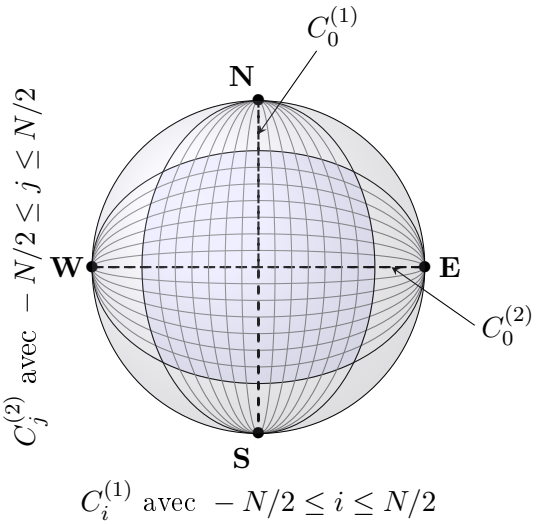


FIGURE 3.7 – Le panel (I) est constitué des points d'intersections d'un ensemble de grands cercles.

Définition 3.6. *La Cubed-Sphere est une grille de la sphère $\mathbb{S}_a^2$. La sphère est couverte par 6 panels identiques notés panel (I) (Front), (II) (East), (III) (Bottom), (IV) (West), (V) (North) et (VI) (South). Chaque panel est doté d'un système de coordonnées :*

$$\left(\xi^{(k)}, \eta^{(k)}\right), \ -\frac{\pi}{4} \leq \xi^{(k)}, \eta^{(k)} \leq \frac{\pi}{4}, \ (I) \leq (k) \leq (VI) \tag{3.21}$$

défini précédemment. Les points de la Cubed-Sphere sont notés $\mathbf{x}_{i,j}^{(k)}$. Ils sont définis par leurs coordonnées $\left(\xi_i^{(k)}, \eta_j^{(k)}\right)$ avec :

$$\xi_i^{(k)} = i\Delta\xi, \ \eta_j^{(k)} = j\Delta\eta, \ -N/2 \leq i, j \leq N/2 \text{ et } (I) \leq (k) \leq (VI), \tag{3.22}$$

où le pas de discrétisation est :

$$\Delta\xi = \Delta\eta = \frac{\pi}{2N}. \tag{3.23}$$

Proposition 3.4. *La Cubed-Sphere est composée de $6N^2 + 2$ points.*

Démonstration. Il y a 6 intérieurs de panels de $(N-1)^2$ points, 12 arrêtes de $N-1$ points et 8 sommets. Ainsi le nombre de points sur la Cubed-Sphere est :

$$6(N-1)^2 + 12(N-1) + 8 = 6N^2 + 2. \tag{3.24}$$

□

Les points $\mathbf{x}_{i,j}^{(k)}$ de chaque panel se répartissent en trois catégories :

- Les points intérieurs si :
$$-\frac{N}{2} + 1 \leq i, j \leq \frac{N}{2} - 1 \tag{3.25}$$
Ils sont au nombre de $(N-1)^2$ par panel.

- Les $4(N-1)$ points de bords de chaque panel, si :
$$\left[j = \pm\frac{N}{2} \text{ et } -\frac{N}{2} + 1 \leq i \leq \frac{N}{2} - 1\right] \text{ ou } \left[i = \pm\frac{N}{2} \text{ et } -\frac{N}{2} + 1 \leq j \leq \frac{N}{2} - 1\right] \tag{3.26}$$

- Les 4 points de coins si :
$$i, j = \pm\frac{N}{2}. \tag{3.27}$$

3.2 Coordonnées Gnomoniques

On considère un cube inscrit dans la sphère $\mathbb{S}_a^2$. Le demi côté de ce cube mesure $R = \frac{\sqrt{3}}{3}a$. Chaque face du cube est donnée par :

- la face centrée sur $F' = (R, 0, 0)$:
$$\left\{\mathbf{x}' = (R, y', z') \in \mathbb{R}^3 \text{ tels que } -R \leq y', z' \leq R\right\} \tag{3.28}$$

- la face centrée sur $B' = (-R, 0, 0)$:
$$\left\{\mathbf{x}' = (-R, y', z') \in \mathbb{R}^3 \text{ tels que } -R \leq y', z' \leq R\right\}, \tag{3.29}$$

- la face centrée sur $E' = (0, R, 0)$:
$$\left\{\mathbf{x}' = (x', R, z') \in \mathbb{R}^3 \text{ tels que } -R \leq x', z' \leq R\right\}, \tag{3.30}$$

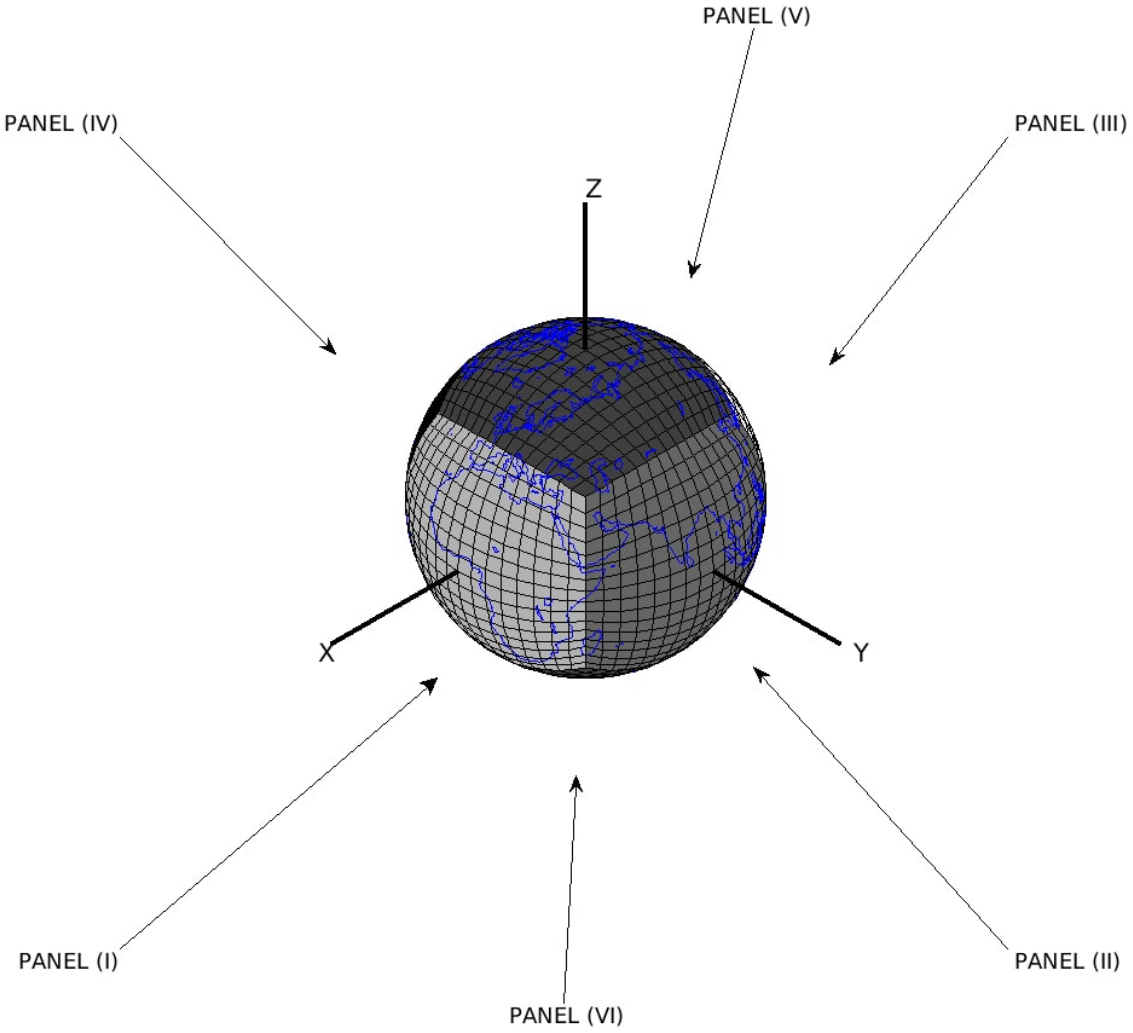


FIGURE 3.8 – Cubed-Sphere avec $N = 16$.

- la face centrée sur $W' = (0, -R, 0)$:

$$\left\{\mathbf{x}' = (x', -R, z') \in \mathbb{R}^3 \text{ tels que } -R \leq x', z' \leq R\right\}, \tag{3.31}$$

- la face centrée sur $N' = (0, 0, R)$:

$$\left\{\mathbf{x}' = (x', y', R) \in \mathbb{R}^3 \text{ tels que } -R \leq x', y' \leq R\right\}, \tag{3.32}$$

- la face centrée sur $S' = (0, 0, -R)$:

$$\left\{\mathbf{x}' = (x', y', -R) \in \mathbb{R}^3 \text{ tels que } -R \leq x', y' \leq R\right\}. \tag{3.33}$$

Le point $\mathbf{F}$ est la projection gnomonique du point F' sur la sphère $\mathbb{S}^2_a$, $\mathbf{E}$ celle de E', etc. Si l'on considère par exemple le panel (I), un point $\mathbf{x}' = (x', y', z')$ de la face centrée sur F' est projeté en $\mathbf{x} = (x, y, z)$ un point du panel (I) (voir Fig. 3.9). Chaque panel est la projection de l'une des faces du cube.

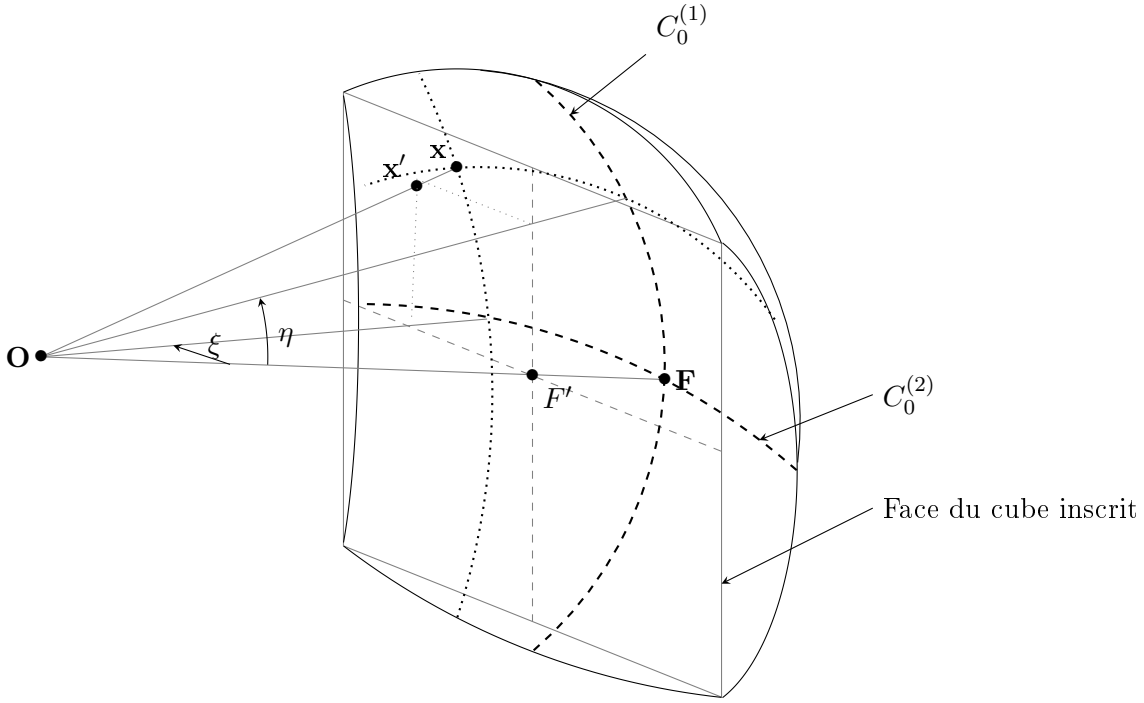


FIGURE 3.9 – Projection gnomonique.

On a les relations suivantes :

$$\tan\xi = \frac{y'}{x'} = \frac{y}{x} \text{ et } \tan\eta = \frac{z'}{x'} = \frac{z}{x} \tag{3.34}$$

Or $\mathbf{x}'(x', y', z')$ est un point de la face du cube centrée en F', donc $x' = R = \frac{\sqrt{3}}{3}a$:

$$\tan\xi = \frac{y'}{R} = \frac{y}{x} \text{ et } \tan\eta = \frac{z'}{R} = \frac{z}{x} \tag{3.35}$$

Sur chaque panel, on définit les coordonnées gnomoniques (X, Y) par :

Définition 3.7. *Les coordonnées gnomoniques* $(X, Y) \in [-1, 1]^2$ *sont définies par :*

$$X = \tan\xi \text{ et } Y = \tan\eta. \tag{3.36}$$

Un point de la sphère est localisé de manière unique par sa face et ses coordonnées gnomoniques. Si on se donne (X, Y) un couple de coordonnées gnomoniques du panel (I), on a :

$$\begin{cases} x^2 + y^2 + z^2 = a^2 \\ X = \dfrac{y}{x} \\ Y = \dfrac{z}{x} \end{cases} \tag{3.37}$$

ainsi $x^2\left(1 + X^2 + Y^2\right) = a^2$, d'où :

$$\left\{ \begin{array}{rcl} x & = & \pm\dfrac{a}{\sqrt{1+X^2+Y^2}} \quad = \quad \dfrac{a}{\sqrt{1+X^2+Y^2}} \\ y & = & xX \\ z & = & xY. \end{array} \right. \tag{3.38}$$

Le signe de x est prescrit car $\mathbf{x}$ est un point du panel (I) qui ne contient que des points d'abscisse positive. De plus, la fonction tan est une bijection de $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$ dans $[-1, 1]$, donc

Théorème 3.1. *Pour chaque panel, les coordonnées gnomoniques* $(X, Y) \in [-1, 1]^2$ *et* $(\xi, \eta) \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^2$ *forment deux systèmes de coordonnées admissibles.*

On peut dériver les coordonnées d'un point $\mathbf{x}(x, y, z) \in \mathbb{R}^3$ en fonction de ξ et η :

$$\frac{\partial y}{\partial \xi} = \frac{\partial x}{\partial \xi} X + x\frac{\partial X}{\partial \xi} = \frac{\partial x}{\partial \xi} X + x(1 + X^2) \tag{3.39}$$

$$\frac{\partial z}{\partial \xi} = \frac{\partial x}{\partial \xi} Y + x\frac{\partial Y}{\partial \xi} = \frac{\partial x}{\partial \xi} Y. \tag{3.40}$$

Le calcul de ces dérivées dépend de $\dfrac{\partial x}{\partial \xi}$:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \xi}(x^2 + y^2 + z^2) \\ &= 2x\frac{\partial x}{\partial \xi} + 2y\frac{\partial y}{\partial \xi} + 2z\frac{\partial z}{\partial \xi} \\ &= x\frac{\partial x}{\partial \xi}(1 + X^2 + Y^2) + x^2 X(1 + X^2) \\ &= x\frac{\partial x}{\partial \xi}\delta^2 + xy(1 + X^2), \end{aligned}$$

en posant $\delta = \sqrt{1 + X^2 + Y^2}$. Ainsi, chaque dérivée est connue et :

$$\left\{ \begin{array}{rcl} \dfrac{\partial x}{\partial \xi} & = & -\dfrac{y(1+X^2)}{\delta^2} \\ \dfrac{\partial y}{\partial \xi} & = & x\dfrac{1+X^2}{\delta^2}(1+Y^2) \\ \dfrac{\partial z}{\partial \xi} & = & -\dfrac{yY(1+X^2)}{\delta^2}. \end{array} \right. \tag{3.41}$$

De la même manière, en dérivant par rapport à η :

$$\left\{ \begin{array}{rcl} \dfrac{\partial x}{\partial \eta} & = & -z\dfrac{1+Y^2}{\delta^2} \\ \dfrac{\partial y}{\partial \eta} & = & -zX\dfrac{1+Y^2}{\delta^2} \\ \dfrac{\partial z}{\partial \eta} & = & -x(1+X^2)\dfrac{1+Y^2}{\delta^2}. \end{array} \right. \tag{3.42}$$

On en déduit la base sur le panel (I) : $(\mathbf{g}_\xi, \mathbf{g}_\eta)$, donnée par :

$$\mathbf{g}_\xi = \frac{\partial \mathbf{x}}{\partial \xi} = \frac{1+X^2}{\delta^2} \begin{bmatrix} -y \\ x(1+Y^2) \\ -yY \end{bmatrix} \text{ et } \mathbf{g}_\eta = \frac{\partial \mathbf{x}}{\partial \xi} = \frac{1+Y^2}{\delta^2} \begin{bmatrix} -z \\ -zX \\ x(1+X^2) \end{bmatrix}. \tag{3.43}$$

Des calculs similaires peuvent être effectuées sur les autres panels. Les résultats sont donnés dans la Table 3.1.

Le couple de vecteurs $(\mathbf{g}^\xi, \mathbf{g}^\eta)$ est la base duale de $(\mathbf{g}_\xi, \mathbf{g}_\eta)$. Cette base doit vérifier les relations suivantes :

$$\left\{ \begin{array}{rcccl} \mathbf{g}^\xi \cdot \mathbf{g}_\xi & = & 1 & = & \mathbf{g}^\eta \cdot \mathbf{g}_\eta \\ \mathbf{g}^\xi \cdot \mathbf{g}_\eta & = & 0 & = & \mathbf{g}^\eta \cdot \mathbf{g}_\xi. \end{array} \right. \tag{3.44}$$

Les vecteurs $\mathbf{g}^\xi$ et $\mathbf{g}^\eta$ sont des vecteurs de $\mathbb{T}_{\mathbf{x}}\mathbb{S}^2_a$. Il existe A, B, C et D tels que :

$$\left\{ \begin{array}{rcl} \mathbf{g}^\xi & = & A\mathbf{g}_\xi + B\mathbf{g}_\eta \\ \mathbf{g}^\eta & = & C\mathbf{g}_\xi + D\mathbf{g}_\eta \end{array} \right. \tag{3.45}$$

En effectuant des produit scalaires de (3.45) par $\mathbf{g}_\xi$ et $\mathbf{g}_\eta$, on obtient :

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \times \begin{bmatrix} \mathbf{g}_\xi \cdot \mathbf{g}_\xi & \mathbf{g}_\xi \cdot \mathbf{g}_\eta \\ \mathbf{g}_\eta \cdot \mathbf{g}_\xi & \mathbf{g}_\eta \cdot \mathbf{g}_\eta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.46}$$

On en déduit :

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} \mathbf{g}_\xi \cdot \mathbf{g}_\xi & \mathbf{g}_\xi \cdot \mathbf{g}_\eta \\ \mathbf{g}_\eta \cdot \mathbf{g}_\xi & \mathbf{g}_\eta \cdot \mathbf{g}_\eta \end{bmatrix}^{-1}. \tag{3.47}$$

Définition 3.8. *La matrice* $\mathbf{G}$ *est la métrique associée à* $(\mathbf{g}_\xi, \mathbf{g}_\eta)$ *en* $\mathbf{x}$ *:*

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}_\xi \cdot \mathbf{g}_\xi & \mathbf{g}_\xi \cdot \mathbf{g}_\eta \\ \mathbf{g}_\eta \cdot \mathbf{g}_\xi & \mathbf{g}_\eta \cdot \mathbf{g}_\eta \end{bmatrix}. \tag{3.48}$$

Proposition 3.5. *La métrique* $\mathbf{G}$ *est invariante par changement de panel.*

Démonstration. Soit $\mathbf{x}$ un point d'un panel (k) de la Cubed-Sphere de coordonnées gnomoniques (X, Y) et $\mathbf{x}'$ un point d'un autre panel (k') de la Cubed-Sphere ayant les mêmes coordonnées gnomoniques (X, Y). Il existe une rotation R_f qui permet de transformer tout point du panel (k) en un point du panel (k') de mêmes coordonnées gnomoniques. De là, il découle :

$$\mathbf{x}' = R_f \mathbf{x}. \tag{3.49}$$

Cette rotation est indépendante de ξ et de η. Donc :

$$\mathbf{g'}_\xi = \frac{\partial}{\partial \xi}(R_f \mathbf{x}) = R_f \frac{\partial \mathbf{x}}{\partial \xi} = R_f \mathbf{g}_\xi. \tag{3.50}$$

De même, on a $\mathbf{g'}_\eta = R_f \mathbf{g}_\eta$. Ainsi, si $\mathbf{G}$ est la métrique en $\mathbf{x}$ et $\mathbf{G}'$ la métrique en $\mathbf{x}'$, on a :

$$\begin{array}{rcl} \mathbf{G}' & = & \begin{bmatrix} \mathbf{g'}_\xi \cdot \mathbf{g'}_\xi & \mathbf{g'}_\xi \cdot \mathbf{g'}_\eta \\ \mathbf{g'}_\eta \cdot \mathbf{g'}_\xi & \mathbf{g'}_\eta \cdot \mathbf{g'}_\eta \end{bmatrix} \\ & = & \begin{bmatrix} R_f\mathbf{g}_\xi \cdot R_f\mathbf{g}_\xi & R_f\mathbf{g}_\xi \cdot R_f\mathbf{g}_\eta \\ R_f\mathbf{g}_\eta \cdot R_f\mathbf{g}_\xi & R_f\mathbf{g}_\eta \cdot R_f\mathbf{g}_\eta \end{bmatrix} \\ & = & \begin{bmatrix} \mathbf{g}_\xi \cdot R_f^T R_f \mathbf{g}_\xi & \mathbf{g}_\xi \cdot R_f^T R_f \mathbf{g}_\eta \\ \mathbf{g}_\eta \cdot R_f^T R_f \mathbf{g}_\xi & \mathbf{g}_\eta \cdot R_f^T R_f \mathbf{g}_\eta \end{bmatrix} \\ & = & \begin{bmatrix} \mathbf{g}_\xi \cdot \mathbf{g}_\xi & \mathbf{g}_\xi \cdot \mathbf{g}_\eta \\ \mathbf{g}_\eta \cdot \mathbf{g}_\xi & \mathbf{g}_\eta \cdot \mathbf{g}_\eta \end{bmatrix} \\ & = & \mathbf{G}. \end{array}$$

Donc la métrique $\mathbf{G}$ est invariante par changement de panel. □

L'expression de la métrique $\mathbf{G}$ en fonction de (X, Y) est

$$\mathbf{G} = \begin{bmatrix} G_{1,1} & G_{1,2} \\ G_{2,1} & G_{2,2} \end{bmatrix} = a^2 \frac{(1+X^2)(1+Y^2)}{\delta^4} \begin{bmatrix} 1+X^2 & -XY \\ -XY & 1+Y^2 \end{bmatrix}. \tag{3.51}$$

De même, on a

$$\mathbf{G}^{-1} = \begin{bmatrix} G^{1,1} & G^{1,2} \\ G^{2,1} & G^{2,2} \end{bmatrix} = \frac{\delta^2}{a^2(1+X^2)(1+Y^2)} \begin{bmatrix} 1+Y^2 & XY \\ XY & 1+X^2 \end{bmatrix}. \tag{3.52}$$

La base duale $(\mathbf{g}^\xi, \mathbf{g}^\eta)$ sur le panel (I) est donnée par :

$$\left\{ \begin{array}{rcl} \mathbf{g}^\xi &=& G^{1,1}\mathbf{g}_\xi + G^{1,2}\mathbf{g}_\eta \\ \mathbf{g}^\eta &=& G^{2,1}\mathbf{g}_\eta + G^{2,2}\mathbf{g}_\eta \end{array} \right. \tag{3.53}$$

d'où :

$$\mathbf{g}^\xi = \frac{1}{x(1+X^2)} \begin{bmatrix} -X \\ 1 \\ 0 \end{bmatrix} \text{ et } \mathbf{g}^\eta = \frac{1}{x(1+Y^2)} \begin{bmatrix} -Y \\ 0 \\ 1 \end{bmatrix}. \tag{3.54}$$

Les champs de vecteur $(\mathbf{g}^\xi, \mathbf{g}^\eta)$ et $(\mathbf{g}_\xi, \mathbf{g}_\eta)$ sont tangents à la sphère et sont fonctions de ξ et η. On définit les *symboles de Christoffel* par :

Définition 3.9. *Les symboles de Christoffel $\Gamma^\tau_{\kappa,\nu}$ (avec κ, ν et τ dans $\{\xi, \eta\}$), sont définis par :*

$$\left\{ \begin{array}{rcl} \dfrac{\partial \mathbf{g}_\xi}{\partial \xi} &=& \Gamma^\xi_{\xi,\xi}\mathbf{g}_\xi + \Gamma^\eta_{\xi,\xi}\mathbf{g}_\eta + \Gamma^r_{\xi,\xi}\mathbf{n} \\ \dfrac{\partial \mathbf{g}_\xi}{\partial \eta} &=& \Gamma^\xi_{\eta,\xi}\mathbf{g}_\xi + \Gamma^\eta_{\eta,\xi}\mathbf{g}_\eta + \Gamma^r_{\eta,\xi}\mathbf{n} \\ \dfrac{\partial \mathbf{g}_\eta}{\partial \xi} &=& \Gamma^\xi_{\xi,\eta}\mathbf{g}_\xi + \Gamma^\eta_{\xi,\eta}\mathbf{g}_\eta + \Gamma^r_{\xi,\eta}\mathbf{n} \\ \dfrac{\partial \mathbf{g}_\eta}{\partial \eta} &=& \Gamma^\xi_{\eta,\eta}\mathbf{g}_\xi + \Gamma^\eta_{\eta,\eta}\mathbf{g}_\eta + \Gamma^r_{\eta,\eta}\mathbf{n} \end{array} \right. \tag{3.55}$$

où $\mathbf{n}$ est le vecteur unitaire normal extérieur à la sphère $\mathbb{S}^2_a$:

$$\mathbf{n} = \frac{\mathbf{x}}{a}. \tag{3.56}$$

Les symboles de Christoffel s'écrivent en fonction de $\mathbf{g}_\xi$, $\mathbf{g}_\eta$ ainsi que de $\mathbf{g}^\xi$, $\mathbf{g}^\eta$ et de la normale extérieure $\mathbf{n}$ grâce à la proposition suivante :

Panel	**Coord. Gnomoniques** (X, Y)	**Bases** $(\mathbf{g}_\xi, \mathbf{g}_\eta)$ et $\left(\mathbf{g}^\xi, \mathbf{g}^\eta\right)$
(I)	$X = \dfrac{y}{x}, Y = \dfrac{z}{x}$	$\mathbf{g}_\xi = \dfrac{1+X^2}{\delta^2}\begin{bmatrix}-y\\ x(1+Y^2)\\ -yY\end{bmatrix}$,$\mathbf{g}_\eta = \dfrac{1+Y^2}{\delta^2}\begin{bmatrix}-z\\ -zX\\ x(1+X^2)\end{bmatrix}$
		$\mathbf{g}^\xi = \dfrac{1}{x(1+X^2)}\begin{bmatrix}-X\\ 1\\ 0\end{bmatrix}$ et $\mathbf{g}^\eta = \dfrac{1}{x(1+Y^2)}\begin{bmatrix}-Y\\ 0\\ 1\end{bmatrix}$
(II)	$X = -\dfrac{x}{y}, Y = \dfrac{z}{y}$	$\mathbf{g}_\xi = \dfrac{1+X^2}{\delta^2}\begin{bmatrix}-y(1+Y^2)\\ x\\ xY\end{bmatrix}$, $\mathbf{g}_\eta = \dfrac{1+Y^2}{\delta^2}\begin{bmatrix}zX\\ -z\\ y(1+X^2)\end{bmatrix}$
		$\mathbf{g}^\xi = \dfrac{1}{y(1+X^2)}\begin{bmatrix}-1\\ -X\\ 0\end{bmatrix}$, $\mathbf{g}^\eta = \dfrac{1}{y(1+Y^2)}\begin{bmatrix}0\\ -Y\\ 1\end{bmatrix}$
(III)	$X = -\dfrac{y}{x}, Y = -\dfrac{z}{x}$	$\mathbf{g}_\xi = \dfrac{1+X^2}{\delta^2}\begin{bmatrix}-y\\ x(1+Y^2)\\ yY\end{bmatrix}$, $\mathbf{g}_\eta = \dfrac{1+Y^2}{\delta^2}\begin{bmatrix}-z\\ -zX\\ x(1+X^2)\end{bmatrix}$
		$\mathbf{g}^\xi = \dfrac{1}{x(1+X^2)}\begin{bmatrix}-X\\ 1\\ 0\end{bmatrix}$, $\mathbf{g}^\eta = \dfrac{1}{x(1+Y^2)}\begin{bmatrix}-Y\\ 0\\ -1\end{bmatrix}$
(IV)	$X = \dfrac{x}{y}, Y = -\dfrac{z}{y}$	$\mathbf{g}_\xi = \dfrac{1+X^2}{\delta^2}\begin{bmatrix}-y(1+Y^2)\\ x\\ -xY\end{bmatrix}$, $\mathbf{g}_\eta = \dfrac{1+Y^2}{\delta^2}\begin{bmatrix}-zX\\ z\\ -y(1+X^2)\end{bmatrix}$
		$\mathbf{g}^\xi = \dfrac{1}{y(1+X^2)}\begin{bmatrix}-1\\ -X\\ 0\end{bmatrix}$, $\mathbf{g}^\eta = \dfrac{1}{y(1+Y^2)}\begin{bmatrix}0\\ -Y\\ -1\end{bmatrix}$
(V)	$X = \dfrac{y}{z}, Y = \dfrac{x}{z}$	$\mathbf{g}_\xi = \dfrac{1+X^2}{\delta^2}\begin{bmatrix}-yY\\ z(1+Y^2)\\ -y\end{bmatrix}$, $\mathbf{g}_\eta = \dfrac{1+Y^2}{\delta^2}\begin{bmatrix}-z(1+X^2)\\ xX\\ x\end{bmatrix}$
		$\mathbf{g}^\xi = \dfrac{1}{z(1+X^2)}\begin{bmatrix}0\\ 1\\ -X\end{bmatrix}$, $\mathbf{g}^\eta = \dfrac{1}{z(1+Y^2)}\begin{bmatrix}-1\\ 0\\ -Y\end{bmatrix}$
(VI)	$X = -\dfrac{y}{z}, Y = -\dfrac{x}{z}$	$\mathbf{g}_\xi = \dfrac{1+X^2}{\delta^2}\begin{bmatrix}-yY\\ -z(1+Y^2)\\ y\end{bmatrix}$, $\mathbf{g}_\eta = \dfrac{1+Y^2}{\delta^2}\begin{bmatrix}-z(1+X^2)\\ -xX\\ x\end{bmatrix}$
		$\mathbf{g}^\xi = \dfrac{1}{z(1+X^2)}\begin{bmatrix}0\\ -1\\ -X\end{bmatrix}$, $\mathbf{g}^\eta = \dfrac{1}{z(1+Y^2)}\begin{bmatrix}-1\\ 0\\ -Y\end{bmatrix}$

TABLE 3.1 – Coordonnées gnomoniques en fonction de x, y et z et bases $(\mathbf{g}_\xi, \mathbf{g}_\eta)$ et $\left(\mathbf{g}^\xi, \mathbf{g}^\eta\right)$ en fonction de x, y et z sur chaque panel, avec $\delta = \sqrt{1+X^2+Y^2}$.

Proposition 3.6. *Les relations suivantes sont vérifiées :*

$$\begin{cases} \Gamma^{\xi}_{\xi,\xi} = \left[\frac{\partial \mathbf{g}_\xi}{\partial \xi}\right]\cdot \mathbf{g}^\xi = -\left[\frac{\partial \mathbf{g}^\xi}{\partial \xi}\right]\cdot \mathbf{g}_\xi \\ \Gamma^{\eta}_{\xi,\xi} = \left[\frac{\partial \mathbf{g}_\xi}{\partial \xi}\right]\cdot \mathbf{g}^\eta = -\left[\frac{\partial \mathbf{g}^\eta}{\partial \xi}\right]\cdot \mathbf{g}_\xi \\ \Gamma^{\xi}_{\xi,\eta} = \left[\frac{\partial \mathbf{g}_\eta}{\partial \xi}\right]\cdot \mathbf{g}^\xi = -\left[\frac{\partial \mathbf{g}^\xi}{\partial \xi}\right]\cdot \mathbf{g}_\eta \\ \Gamma^{\eta}_{\xi,\eta} = \left[\frac{\partial \mathbf{g}_\eta}{\partial \xi}\right]\cdot \mathbf{g}^\eta = -\left[\frac{\partial \mathbf{g}^\eta}{\partial \xi}\right]\cdot \mathbf{g}_\eta \\ \Gamma^{\eta}_{\eta,\eta} = \left[\frac{\partial \mathbf{g}_\eta}{\partial \eta}\right]\cdot \mathbf{g}^\eta = -\left[\frac{\partial \mathbf{g}^\eta}{\partial \eta}\right]\cdot \mathbf{g}_\eta \\ \Gamma^{\xi}_{\eta,\eta} = \left[\frac{\partial \mathbf{g}_\eta}{\partial \eta}\right]\cdot \mathbf{g}^\xi = -\left[\frac{\partial \mathbf{g}^\xi}{\partial \eta}\right]\cdot \mathbf{g}_\eta \\ \Gamma^{r}_{\eta,\eta} = \left[\frac{\partial \mathbf{g}_\eta}{\partial \eta}\right]\cdot \mathbf{n} = -\frac{1}{a}|\mathbf{g}_\eta|^2 \\ \Gamma^{r}_{\xi,\xi} = \left[\frac{\partial \mathbf{g}_\xi}{\partial \xi}\right]\cdot \mathbf{n} = -\frac{1}{a}|\mathbf{g}_\xi|^2 \\ \Gamma^{r}_{\xi,\eta} = \left[\frac{\partial \mathbf{g}_\eta}{\partial \xi}\right]\cdot \mathbf{n} = -\frac{1}{a}\mathbf{g}_\eta\cdot\mathbf{g}_\xi. \end{cases} \tag{3.57}$$

Démonstration. Nous ne démontrons que la première égalité, les autres s'obtiennent de la même manière.

$$\left[\frac{\partial \mathbf{g}_\xi}{\partial \xi}\right]\cdot \mathbf{g}^\xi = \left[\Gamma^{\xi}_{\xi,\xi}\mathbf{g}_\xi + \Gamma^{\eta}_{\xi,\xi}\mathbf{g}_\eta + \Gamma^{r}_{\xi,\xi}\mathbf{n}\right]\cdot \mathbf{g}^\xi. \tag{3.58}$$

Or $\mathbf{g}_\xi\cdot\mathbf{g}^\xi = 1$ et $\mathbf{g}_\eta\cdot\mathbf{g}^\xi = 0$, d'où la première partie :

$$\Gamma^{\xi}_{\xi,\xi} = \left[\frac{\partial \mathbf{g}_\xi}{\partial \xi}\right]\cdot \mathbf{g}^\xi. \tag{3.59}$$

D'autre part, on a :

$$\Gamma^{\xi}_{\xi,\xi} = \left[\frac{\partial \mathbf{g}_\xi}{\partial \xi}\right]\cdot \mathbf{g}^\xi = \frac{\partial}{\partial \xi}\underbrace{\left(\mathbf{g}_\xi\cdot\mathbf{g}^\xi\right)}_{=1} - \left[\frac{\partial \mathbf{g}^\xi}{\partial \xi}\right]\cdot \mathbf{g}_\xi = -\left[\frac{\partial \mathbf{g}^\xi}{\partial \xi}\right]\cdot \mathbf{g}_\xi \tag{3.60}$$

et la relation est démontrée. □

Remarque 3.1. *On note que* $\Gamma^{\xi}_{\eta,\xi} = \Gamma^{\xi}_{\xi,\eta}$ *et* $\Gamma^{\eta}_{\eta,\xi} = \Gamma^{\eta}_{\xi,\eta}$.

En effet :

$$\Gamma^{\eta}_{\xi,\eta} = \left(\frac{\partial \mathbf{g}_\eta}{\partial \xi}\right)\cdot\mathbf{g}^\eta = \left(\frac{\partial}{\partial \xi}\frac{\partial \mathbf{x}}{\partial \eta}\right)\cdot\mathbf{g}^\eta = \left(\frac{\partial}{\partial \eta}\frac{\partial \mathbf{x}}{\partial \xi}\right)\cdot\mathbf{g}^\eta = \left(\frac{\partial \mathbf{g}_\xi}{\partial \eta}\right)\cdot\mathbf{g}^\eta = \Gamma^{\eta}_{\eta,\xi}$$

de même pour $\Gamma^{\xi}_{\eta,\xi} = \Gamma^{\xi}_{\xi,\eta}$ *et* $\Gamma^{r}_{\eta,\xi} = \Gamma^{r}_{\xi,\eta}$.

Les symboles de Christoffel s'obtiennent en fonction des coordonnées gnomoniques (X, Y) de la façon qui suit. On peut calculer les dérivées suivantes :

$$\frac{\partial \mathbf{g}^\xi}{\partial \xi} = \frac{1}{\delta^2 x(1+X^2)}\begin{bmatrix} X^2Y^2-\delta^2 \\ -X(Y^2+\delta^2) \\ 0 \end{bmatrix} \text{ et } \frac{\partial \mathbf{g}^\xi}{\partial \eta} = \frac{1+Y^2}{\delta^2 x^2(1+X^2)}\begin{bmatrix} -X \\ 1 \\ 0 \end{bmatrix} \tag{3.61}$$

de même :

$$\frac{\partial \mathbf{g}^\eta}{\partial \xi} = \frac{X(1+X^2)}{x\delta^2(1+Y^2)}\begin{bmatrix} -Y \\ 0 \\ 1 \end{bmatrix} \text{ et } \frac{\partial \mathbf{g}^\eta}{\partial \eta} = \frac{1}{x\delta^2(1+Y^2)}\begin{bmatrix} X^2Y^2-\delta^2 \\ 0 \\ -Y(X^2-\delta^2) \end{bmatrix} \tag{3.62}$$

d'où les symboles de Christoffel :

$$\left\{\begin{array}{rcl}
\Gamma^{\xi}_{\xi,\eta} & = & -\dfrac{Y(1+Y^2)}{\delta^2}, \\
\Gamma^{\eta}_{\xi,\eta} & = & -\dfrac{X(1+X^2)}{\delta^2}, \\
\Gamma^{\xi}_{\eta,\eta} & = & 0, \\
\Gamma^{\eta}_{\xi,\xi} & = & 0, \\
\Gamma^{\eta}_{\eta,\eta} & = & \dfrac{2X^2Y}{\delta^2}, \\
\Gamma^{\xi}_{\xi,\xi} & = & \dfrac{2XY^2}{\delta^2}, \\
\Gamma^{r}_{\xi,\xi} & = & -\dfrac{a(1+Y^2)(1+X^2)^2}{\delta^4}, \\
\Gamma^{r}_{\eta,\eta} & = & -\dfrac{a(1+Y^2)^2(1+X^2)}{\delta^4}, \\
\Gamma^{r}_{\xi,\eta} & = & aXY\dfrac{(1+X^2)(1+Y^2)}{\delta^4}.
\end{array}\right. \tag{3.63}$$

De plus, la proposition suivante permet de donner une expression des symboles de Christoffel sur chaque panel.

Proposition 3.7. *Les symboles de Christoffel sont invariants par changement de panel.*

Démonstration. Soit $\mathbf{x}$ un point d'un panel (k) de la Cubed-Sphere de coordonnées gnomoniques (X, Y) et $\mathbf{x}'$ un point d'un autre panel (k') de la Cubed-Sphere ayant les mêmes coordonnées gnomoniques (X, Y). Il existe une rotation R_f qui permet de transformer tout point du panel (k) en un point du panel (k') de mêmes coordonnées gnomoniques :

$$\mathbf{x}' = R_f\mathbf{x}. \tag{3.64}$$

Cette rotation est indépendante de ξ et de η.

Soient τ, $\upsilon \in \{\xi, \eta\}$. On a :

$$\begin{aligned}
\left(\frac{\partial}{\partial\tau}R_f\mathbf{g}_\tau\right)\cdot(R_f\mathbf{g}_\upsilon) &= \left(R_f\frac{\partial}{\partial\tau}\mathbf{g}_\tau\right)\cdot(R_f\mathbf{g}_\upsilon) \\
&= \left(\frac{\partial}{\partial\tau}\mathbf{g}_\tau\right)\cdot\left(R_f^{-1}R_f\mathbf{g}_\upsilon\right) \\
&= \left(\frac{\partial}{\partial\tau}\mathbf{g}_\tau\right)\cdot(\mathbf{g}_\upsilon)
\end{aligned}$$

donc :

$$\Gamma^{\tau}_{\tau,\mu}(\mathbf{x}) = \Gamma^{\tau}_{\tau,\mu}(\mathbf{x}'). \tag{3.65}$$

D'où l'invariance par changement de panel. □

Proposition 3.8. *Les égalités suivantes sont vérifiées :*

$$\left\{\begin{array}{rcl}
\dfrac{\partial\mathbf{g}^{\xi}}{\partial\xi} & = & -\Gamma^{\xi}_{\xi,\xi}\mathbf{g}^{\xi} - \Gamma^{\xi}_{\xi,\eta}\mathbf{g}^{\eta} - \dfrac{1}{a}\mathbf{n} \\
\dfrac{\partial\mathbf{g}^{\xi}}{\partial\eta} & = & -\Gamma^{\xi}_{\xi,\eta}\mathbf{g}^{\xi} - \Gamma^{\xi}_{\eta,\eta}\mathbf{g}^{\eta} \\
\dfrac{\partial\mathbf{g}^{\eta}}{\partial\eta} & = & -\Gamma^{\eta}_{\xi,\eta}\mathbf{g}^{\xi} - \Gamma^{\eta}_{\eta,\eta}\mathbf{g}^{\eta} - \dfrac{1}{a}\mathbf{n} \\
\dfrac{\partial\mathbf{g}^{\eta}}{\partial\xi} & = & -\Gamma^{\eta}_{\xi,\xi}\mathbf{g}^{\xi} - \Gamma^{\eta}_{\xi,\eta}\mathbf{g}^{\eta}.
\end{array}\right. \tag{3.66}$$

Démonstration. $(\mathbf{g}^\xi, \mathbf{g}^\eta, \mathbf{n})$ forme une base de $\mathbb{R}^3$, donc il existe A_ξ, A_η et A_r tels que

$$\frac{\partial \mathbf{g}^\xi}{\partial \xi} = A_\xi \mathbf{g}^\xi + A_\eta \mathbf{g}^\eta + A_r \mathbf{n}. \tag{3.67}$$

Par produit scalaire, on a

$$\frac{\partial \mathbf{g}^\xi}{\partial \xi} \cdot \mathbf{g}_\xi = A_\xi = -\Gamma^\xi_{\xi,\xi}. \tag{3.68}$$

De la même manière, on a

$$\frac{\partial \mathbf{g}^\xi}{\partial \xi} \cdot \mathbf{g}_\eta = A_\eta = -\Gamma^\xi_{\xi,\eta}, \tag{3.69}$$

ainsi que

$$\frac{\partial \mathbf{g}^\xi}{\partial \xi} \cdot \mathbf{n} = A_r = -\mathbf{g}^\xi \cdot \frac{\partial \mathbf{n}}{\partial \xi} = -\frac{1}{a}\mathbf{g}^\xi \cdot \mathbf{g}_\xi = -\frac{1}{a}. \tag{3.70}$$

De la même manière, il existe B_ξ, B_η et B_r tels que

$$\frac{\partial \mathbf{g}^\xi}{\partial \eta} = B_\xi \mathbf{g}^\xi + B_\eta \mathbf{g}^\eta + B_r \mathbf{n} \tag{3.71}$$

et on a, par produit scalaire

$$B_\xi = -\Gamma^\xi_{\xi,\eta} \text{ et } B_\eta = -\Gamma^\xi_{\eta,\eta}. \tag{3.72}$$

De plus

$$B_r = \frac{\partial \mathbf{g}^\xi}{\partial \eta} \cdot \mathbf{n} = -\frac{1}{a}\mathbf{g}^\xi \cdot \mathbf{g}_\eta = 0. \tag{3.73}$$

Les résultats pour $\dfrac{\partial \mathbf{g}^\eta}{\partial \eta}$ et $\dfrac{\partial \mathbf{g}^\eta}{\partial \xi}$ se démontrent de la même manière. □

$\mathbf{G}$ et $\mathbf{G}^{-1}$	Symboles de Christoffels
$\mathbf{G} = a^2 \dfrac{(1+X^2)(1+Y^2)}{\delta^4} \begin{bmatrix} 1+X^2 & -XY \\ -XY & 1+Y^2 \end{bmatrix}$ $\mathbf{G}^{-1} = \dfrac{\delta^2}{a^2(1+X^2)(1+Y^2)} \begin{bmatrix} 1+Y^2 & XY \\ XY & 1+X^2 \end{bmatrix}$	$\Gamma^\xi_{\xi,\eta} = -\dfrac{Y(1+Y^2)}{\delta^2}$ $\Gamma^\eta_{\xi,\eta} = -\dfrac{X(1+X^2)}{\delta^2}$ $\Gamma^\xi_{\eta,\eta} = 0$ $\Gamma^\eta_{\xi,\xi} = 0$ $\Gamma^\eta_{\eta,\eta} = \dfrac{2X^2Y}{\delta^2}$ $\Gamma^\xi_{\xi,\xi} = \dfrac{2XY^2}{\delta^2}$ $\Gamma^r_{\xi,\xi} = -\dfrac{a(1+Y^2)(1+X^2)^2}{\delta^4}$ $\Gamma^r_{\eta,\eta} = -\dfrac{a(1+Y^2)^2(1+X^2)}{\delta^4}$ $\Gamma^r_{\xi,\eta} = aXY\dfrac{(1+Y^2)(1+X^2)}{\delta^4}$

TABLE 3.2 – Quelques invariants par changement de panels

Proposition 3.9. ***(Contraction des symboles de Christoffel)*** *Les égalités suivantes sont vérifiées :*

$$\begin{cases} \Gamma^{\eta}_{\eta,\xi} + \Gamma^{\xi}_{\xi,\xi} &= \dfrac{1}{\sqrt{\det(\mathbf{G})}} \dfrac{\partial}{\partial \xi} \left(\sqrt{\det(\mathbf{G})} \right) \\ \Gamma^{\xi}_{\xi,\eta} + \Gamma^{\eta}_{\eta,\eta} &= \dfrac{1}{\sqrt{\det(\mathbf{G})}} \dfrac{\partial}{\partial \eta} \left(\sqrt{\det(\mathbf{G})} \right). \end{cases} \tag{3.74}$$

Démonstration. Soient $i, j, k \in \{\xi, \eta\}$. En dérivant $g_{i,j} = \mathbf{g}_i \cdot \mathbf{g}_j$ par rapport à k, on trouve :

$$\begin{aligned} \partial_k(g_{i,j}) &= (\partial_k \mathbf{g}_i) \cdot \mathbf{g}_j + (\partial_k \mathbf{g}_j) \cdot \mathbf{g}_i \\ &= \left(\Gamma^{\xi}_{k,i} \mathbf{g}_\xi + \Gamma^{\eta}_{k,i} \mathbf{g}_\eta + \Gamma^{r}_{k,i} \mathbf{n} \right) \cdot \mathbf{g}_j + \left(\Gamma^{\xi}_{k,j} \mathbf{g}_\xi + \Gamma^{\eta}_{k,j} \mathbf{g}_\eta + \Gamma^{r}_{k,j} \mathbf{n} \right) \cdot \mathbf{g}_i \\ &= \Gamma^{\xi}_{k,i} g_{\xi,j} + \Gamma^{\eta}_{k,i} g_{\eta,j} + \Gamma^{\xi}_{k,j} g_{\xi,i} + \Gamma^{\eta}_{k,j} g_{\eta,i}. \end{aligned}$$

De la même manière, on a

$$\partial_i(g_{j,k}) = \Gamma^{\xi}_{i,j} g_{\xi,k} + \Gamma^{\eta}_{i,j} g_{\eta,k} + \Gamma^{\xi}_{i,k} g_{\xi,j} + \Gamma^{\eta}_{i,k} g_{\eta,j}, \tag{3.75}$$

ainsi que

$$\partial_k(g_{k,i}) = \Gamma^{\xi}_{j,k} g_{\xi,i} + \Gamma^{\eta}_{j,k} g_{\eta,i} + \Gamma^{\xi}_{j,i} g_{\xi,k} + \Gamma^{\eta}_{j,i} g_{\eta,k}. \tag{3.76}$$

En combinant ces trois relations, on obtient

$$\partial_k(g_{i,j}) + \partial_i(g_{j,k}) - \partial_j(g_{k,i}) = 2\Gamma^{\xi}_{k,i} g_{\xi,j} + 2\Gamma^{\eta}_{k,i} g_{\eta,j}. \tag{3.77}$$

Ainsi en considérant les cas $j = \xi$ et $j = \eta$, on trouve le système :

$$\begin{cases} \frac{1}{2} \left(\partial_k(g_{i,\xi}) + \partial_i(g_{\xi,k}) - \partial_\xi(g_{k,i}) \right) &= \Gamma^{\xi}_{k,i} g_{\xi,\xi} + \Gamma^{\eta}_{k,i} g_{\eta,\xi} \\ \frac{1}{2} \left(\partial_k(g_{i,\eta}) + \partial_i(g_{\eta,k}) - \partial_\eta(g_{k,i}) \right) &= \Gamma^{\xi}_{k,i} g_{\xi,\eta} + \Gamma^{\eta}_{k,i} g_{\eta,\eta}. \end{cases} \tag{3.78}$$

En remarquant que

$$\begin{bmatrix} g_{\xi,\xi} & g_{\xi,\eta} \\ g_{\eta,\xi} & g_{\eta,\eta} \end{bmatrix}^{-1} = \begin{bmatrix} g^{\xi,\xi} & g^{\xi,\eta} \\ g^{\eta,\xi} & g^{\eta,\eta} \end{bmatrix}, \tag{3.79}$$

on trouve

$$\begin{cases} \dfrac{1}{2} g^{\xi,\xi} (\partial_k g_{i,\xi} + \partial_i g_{\xi,k} - \partial_\xi g_{k,i}) + \dfrac{1}{2} g^{\xi,\eta} (\partial_k g_{i,\eta} + \partial_i g_{\eta,k} - \partial_\eta g_{k,i}) = \Gamma^{\xi}_{k,i} \\ \dfrac{1}{2} g^{\xi,\eta} (\partial_k g_{i,\xi} + \partial_i g_{\xi,k} - \partial_\xi g_{k,i}) + \dfrac{1}{2} g^{\eta,\eta} (\partial_k g_{i,\eta} + \partial_i g_{\eta,k} - \partial_\eta g_{k,i}) = \Gamma^{\eta}_{k,i}. \end{cases} \tag{3.80}$$

En particulier, en prenant $k = i = \xi$ dans la première équation ainsi que $k = \eta$ et $i = \xi$ dans la seconde on trouve

$$\begin{cases} \dfrac{1}{2} g^{\xi,\xi} \partial_\xi g_{\xi,\xi} + \dfrac{1}{2} g^{\xi,\eta} \left(2\partial_\xi g_{\xi,\eta} - \partial_\eta g_{\xi,\xi} \right) = \Gamma^{\xi}_{\xi,\xi} \\ \dfrac{1}{2} g^{\xi,\eta} \partial_\eta g_{\xi,\xi} + \dfrac{1}{2} g^{\eta,\eta} \partial_\xi g_{\eta,\eta} = \Gamma^{\eta}_{\eta,\xi}. \end{cases} \tag{3.81}$$

On somme ces deux équations pour obtenir l'équation suivante

$$\Gamma^{\eta}_{\eta,\xi} + \Gamma^{\xi}_{\xi,\xi} = \frac{1}{2} \left(g^{\xi,\xi} \partial_\xi g_{\xi,\xi} + 2 g^{\xi,\eta} \partial_\xi g_{\xi,\eta} + g^{\eta,\eta} \partial_\xi g_{\eta,\eta} \right). \tag{3.82}$$

Or, le calcul de $g^{\xi,\xi}$, $g^{\eta,\xi}$ et $g^{\eta,\eta}$ peut se faire grâce aux relations suivantes :

$$\begin{aligned} \mathbf{G}^{-1} &= \begin{bmatrix} g^{\xi,\xi} & g^{\xi,\eta} \\ g^{\xi,\eta} & g^{\eta,\eta} \end{bmatrix} \\ &= \frac{1}{\det(\mathbf{G})} \mathrm{comat}(\mathbf{G})^T \\ &= \frac{1}{\det(\mathbf{G})} \begin{bmatrix} g_{\eta,\eta} & -g_{\xi,\eta} \\ -g_{\xi,\eta} & g_{\xi,\xi} \end{bmatrix}. \end{aligned}$$

Donc par identification on a

$$\begin{aligned}\Gamma^{\eta}_{\eta,\xi} + \Gamma^{\xi}_{\xi,\xi} &= \frac{1}{2\det(\mathbf{G})}\left[g_{\eta,\eta}\partial_\xi g_{\xi,\xi} - 2g_{\xi,\eta}\partial_\xi g_{\xi,\eta} + g_{\xi,\xi}\partial_\xi g_{\eta,\eta}\right] \\ &= \frac{1}{2\det(\mathbf{G})}\partial_\xi\left(g_{\xi,\xi}g_{\eta,\eta} - g^2_{\xi,\eta}\right) \\ &= \frac{1}{2\det(\mathbf{G})}\partial_\xi \det(\mathbf{G}) \\ &= \frac{1}{\sqrt{\det(\mathbf{G})}}\partial_\xi(\sqrt{\det(\mathbf{G})}).\end{aligned}$$

La seconde égalité se montre de la même manière. □

3.3 Calcul intrinsèque sur la Cubed-Sphere

On a vu dans la section 3.2 que les coordonnées gnomoniques (X, Y) ainsi que les coordonnées (ξ, η) forment des systèmes de coordonnées admissibles sur chaque panel. Pour tout point $\mathbf{x}^{(k)}_{i,j} \in \mathbb{S}^2_a$ du panel (k) de la Cubed-Sphere, il existe deux cercles $C^{(1)}_i$ et $C^{(2)}_j$ tels que :

$$\mathbf{x}^{(k)}_{i,j} \in C^{(1)}_i \cap C^{(2)}_j. \tag{3.83}$$

L'angle α est l'angle géodésique entre $\mathbf{x}^{(k)}_{i,j}$ et $\mathbf{x}^{(k)}_{0,j}$ le long de $C^{(2)}_j$. L'angle β est l'angle géodésique entre $\mathbf{x}^{(k)}_{i,j}$ et $\mathbf{x}^{(k)}_{i,0}$ mesuré le long de $C^{(1)}_i$ (voir Fig. 3.10). Des angles géodésiques α et β peuvent être construits sur chaque panel, l'orientation de ces angles est donnée en Figure 3.11. Ainsi, chaque panel est construit à partir de sections de grands cercles.

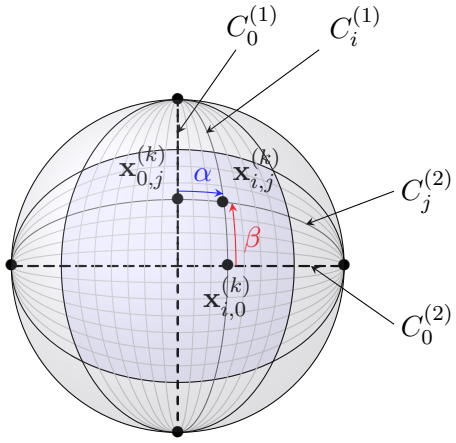


FIGURE 3.10 – Angles géodésiques α et β pour le point $x^{(k)}_{i,j}$.

Soit $\mathbf{x}^{(I)}_{i,j}$ un point du panel (I) de coordonnées (x, y, z) dans $\mathbb{R}^3$. Alors les relations suivantes sont vérifiées :

$$\left\{\begin{array}{ccccc} x & = & a\cos\alpha\cos\eta & = & a\cos\beta\cos\xi \\ y & = & a\sin\alpha & = & a\cos\beta\sin\xi \\ z & = & a\cos\alpha\sin\eta & = & a\sin\beta. \end{array}\right. \tag{3.84}$$

Des relations similaires existent sur tous les panels (voir Table 3.3).

Pour chaque panel, on peut déduire de (3.84) les expressions de α et β en fonction de ξ et η.

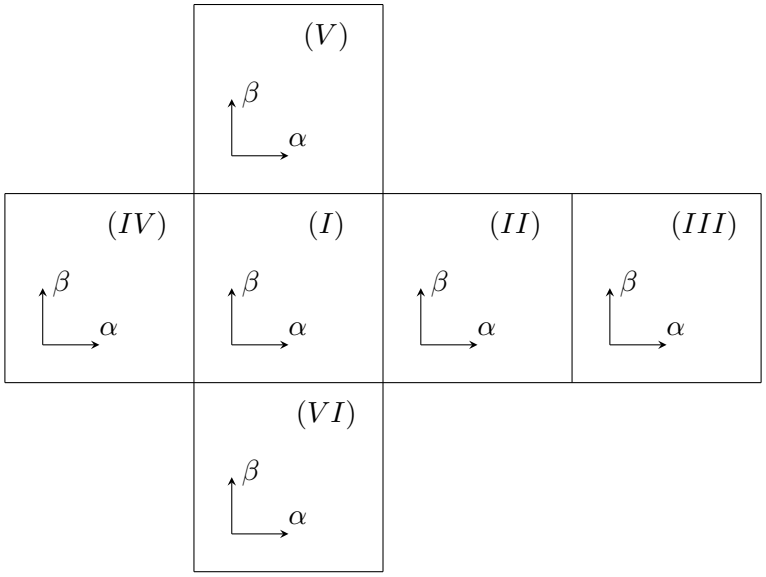


FIGURE 3.11 – Patron de la Cubed-Sphere avec orientation des directions α et β par panel.

Panel	(x, y, z) **fonction de** (α, η) **et de** (ξ, β)
(I)	$x = a\cos\alpha\cos\eta = a\cos\beta\cos\xi$ $y = a\sin\alpha = a\cos\beta\sin\xi$ $z = a\cos\alpha\sin\eta = a\sin\beta$
(II)	$x = -a\sin\alpha = -a\cos\beta\sin\xi$ $y = a\cos\alpha\cos\eta = a\cos\beta\cos\xi$ $z = a\cos\alpha\sin\eta = a\sin\beta$
(III)	$x = -a\cos\alpha\cos\eta = a\cos\beta\cos\xi$ $y = -a\sin\alpha = -a\cos\beta\sin\xi$ $z = a\cos\alpha\sin\eta = a\sin\beta$
(IV)	$x = a\sin\alpha = a\cos\beta\sin\xi$ $y = -a\cos\alpha\cos\eta = -a\cos\beta\cos\xi$ $z = a\cos\alpha\sin\eta = a\sin\beta$
(V)	$x = -a\cos\alpha\sin\eta = -a\sin\beta$ $y = a\sin\alpha = a\cos\beta\sin\xi$ $z = a\cos\alpha\cos\eta = a\cos\beta\cos\xi$
(VI)	$x = a\cos\alpha\sin\eta = a\sin\beta$ $y = a\sin\alpha = a\cos\beta\sin\xi$ $z = -a\cos\alpha\cos\eta = -a\cos\beta\cos\xi$

TABLE 3.3 – Coordonnées cartésiennes (x, y, z) en fonction des angles (α, η) et de (ξ, β) sur chaque panel.

Théorème 3.2. *Les systèmes (α, η) et (ξ, β) sont des systèmes de coordonnées admissibles sur chaque panel.*

Démonstration. Sur le panel (I) (d'après l'équation (3.84) et le tableau 3.3), on a :

$$\begin{cases} x &= a\cos\alpha\cos\eta \\ y &= a\sin\alpha \\ z &= a\cos\alpha\sin\eta \end{cases} \tag{3.85}$$

Or $x \neq 0$ sur le panel (I) donc

$$\begin{cases} \tan\eta = \dfrac{z}{x} \\ \sin\alpha = \dfrac{y}{a} \end{cases} \tag{3.86}$$

et par construction de la Cubed-Sphere, $\eta \in [-\pi/4, \pi/4]$ et $\alpha \in I \subset [-\pi/4, \pi/4]$. L'application

$$\varphi : (\alpha, \eta) \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^2 \mapsto (x, y, z) \in \mathbb{R}^3 \tag{3.87}$$

donnée par (3.85) est injective. La réciproque $\varphi^{-1} : (x, y, z) \in \mathrm{Im}(\varphi) \mapsto (\alpha, \eta) \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^2$ est, d'après (3.86), donnée par $\alpha = \arcsin(y/a)$ et $\eta = \arctan(z/x)$ car sin et tan sont des bijections sur $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$. Comme (x, y, z) est un système de coordonnées admissible sur le panel (I) alors (α, η) en est un aussi. De plus, on a le même type de relations sur les autres panels.

La démonstration est la même pour montrer que (ξ, β) est un système de coordonnées admissible par panel. □

Proposition 3.10. *Les angles α et β s'expriment en fonction des angles ξ et η par*

$$\begin{cases} \alpha(\xi, \eta) = \arctan\left[\dfrac{\tan\xi}{\sqrt{1+\tan^2\eta}}\right] \\ \beta(\xi, \eta) = \arctan\left[\dfrac{\tan\eta}{\sqrt{1+\tan^2\xi}}\right]. \end{cases} \tag{3.88}$$

Démonstration. D'après l'équation (3.85), on a

$$x^2 + y^2 = a^2\cos^2\beta, \tag{3.89}$$

ainsi que

$$z^2 = a^2\sin^2\beta, \tag{3.90}$$

d'où on déduit facilement :

$$\begin{aligned} \tan^2\beta &= \frac{z^2}{x^2+y^2} \\ &= \frac{Y^2}{1+X^2}. \end{aligned}$$

L'expression de β suivante se déduit

$$\beta(\xi, \eta) = \arctan\left[\frac{\tan\eta}{\sqrt{1+\tan^2\xi}}\right]. \tag{3.91}$$

De la même manière on obtient :

$$\alpha(\xi, \eta) = \arctan\left[\frac{\tan\xi}{\sqrt{1+\tan^2\eta}}\right]. \tag{3.92}$$

De plus, ces équations se retrouvent sur chaque panel. □

Théorème 3.3. *Les angles (α, β) forment un système de coordonnées admissible par panel.*

Démonstration. On définit l'application φ permettant de passer des coordonnées gnomoniques $(X, Y) = (\tan\xi, \tan\eta)$ à (α, β). Cette application est donnée par

$$\varphi : \begin{cases} [-1,1]^2 & \to & \mathrm{Im}(\varphi) \\ (X,Y) & \mapsto & (\alpha(X,Y), \beta(X,Y)) \end{cases} \tag{3.93}$$

avec $\alpha(X,Y)$ et $\beta(X,Y)$ vérifiant :

$$\alpha(X,Y) = \arctan\left[\frac{X}{\sqrt{1+Y^2}}\right], \text{ et } \beta(X,Y) = \arctan\left[\frac{Y}{\sqrt{1+X^2}}\right]. \tag{3.94}$$

L'application φ est continue, $[-1,1]^2$ est connexe donc $\mathrm{Im}(\varphi)$ est connexe. De plus, l'inclusion suivante est vérifiée :

$$\mathrm{Im}(\varphi) \subset \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^2. \tag{3.95}$$

L'application φ est surjective par construction. Montrons qu'elle est injective. Supposons qu'il existe (X_1, Y_1) et (X_2, Y_2) dans $[-1,1]^2$ tels que

$$\varphi(X_1, Y_1) = \varphi(X_2, Y_2). \tag{3.96}$$

Comme arctan est bijective de $[-1,1]$ dans $[-\pi/4, \pi/4]$, cette relation est équivalente à

$$\begin{cases} \dfrac{X_1}{\sqrt{1+Y_1^2}} = \dfrac{X_2}{\sqrt{1+Y_2^2}} \\ \dfrac{Y_1}{\sqrt{1+X_1^2}} = \dfrac{Y_2}{\sqrt{1+X_2^2}}. \end{cases} \tag{3.97}$$

On pose $a = \dfrac{X_1}{\sqrt{1+Y_1^2}}$ et $b = \dfrac{Y_1}{\sqrt{1+X_1^2}}$. Le système (3.97) implique :

$$\begin{cases} X_2^2 - a^2 Y_2^2 = a^2 \\ -b^2 X_2^2 + Y_2^2 = b^2. \end{cases} \tag{3.98}$$

Le système (3.98) est un système linéaire en (X_2^2, Y_2^2). De plus, ce système est inversible, en effet pour que le système ne soit pas inversible, il faudrait que

$$\begin{aligned} 0 &= \det \begin{bmatrix} 1 & -a^2 \\ -b^2 & 1 \end{bmatrix} \\ &= 1 - a^2 b^2. \end{aligned}$$

En remplaçant a et b par leur valeur, cette dernière relation implique

$$0 = 1 + X_1^2 + Y_1^2, \tag{3.99}$$

ce qui est impossible. Donc le système (3.98) admet une unique solution et cette dernière est donnée par

$$\begin{cases} X_2^2 = \dfrac{a^2(b^2+1)}{1-a^2b^2} = X_1^2 \\ Y_2^2 = \dfrac{b^2(a^2+1)}{1-a^2b^2} = Y_1^2. \end{cases} \tag{3.100}$$

Donc on obtient $X_1 = \pm X_2$ et $Y_1 = \pm Y_2$. Or si $X_1 = -X_2$, on a

$$\frac{X_1}{\sqrt{1+Y_1^2}} = -\frac{X_1}{\sqrt{1+Y_2^2}}, \tag{3.101}$$

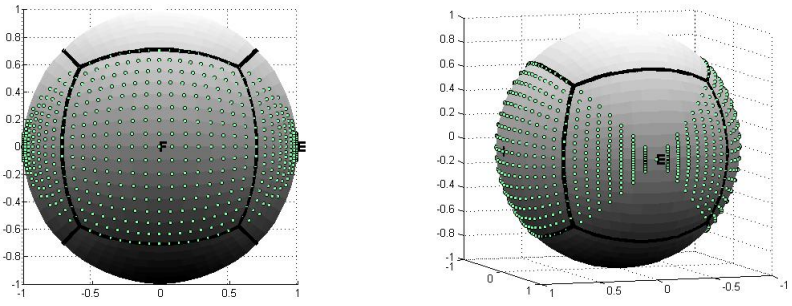


FIGURE 3.12 – L'ensemble (I_α) de grands cercles correspond aux isolignes η constant du panel (I) et du panel (III).

d'où on déduit directement

$$\sqrt{1+Y_1^2} = -\sqrt{1+Y_2^2}, \tag{3.102}$$

ce qui est absurde. Donc on ne peut pas avoir $X_1 = -X_2$. De la même manière, on ne peut pas avoir $Y_1 = -Y_2$. On en déduit que $(X_1, Y_1) = (X_2, Y_2)$ et l'application φ est bijective.

Enfin on a vu que (X, Y) est un système de coordonnées admissible par panel donc (α, β) est un système de coordonnées admissible par panel. □

Chacune des sections de grand cercle peut être complétée en un grand cercle. Le grand cercle $C_i^{(1)}$ correspond à une isoligne ξ constante et le grand cercle $C_j^{(2)}$ à une isoligne η constante. Considérons l'isoligne η constante sur le panel (I). Sur ce panel, l'angle α est tel que :

$$-\alpha_0(\eta) \leq \alpha \leq \alpha_0(\eta) \text{ avec } \alpha_0(\eta) = \arctan\left(\frac{\sqrt{2}}{2}\tan\eta\right). \tag{3.103}$$

Le grand cercle $C_j^{(2)}$ traverse sur les panels (II), (III) et (IV) dans cet ordre. On peut prolonger α comme étant l'angle curviligne le long du grand cercle complet. Sur le panel (II), le grand cercle $C_j^{(2)}$ coupe d'autres grands cercles (ceux permettant de construire le panel (II)) en $\mathbf{M}_k$ $(-N/2 \leq k \leq N/2)$ de coordonnées $(\xi_k^E = k\Delta\xi, \eta_k^E)$. Il coupe aussi le panel (III) mais correspond aux points de maillage par symétrie, soit le grand cercle du panel (III) correspondant à l'isoligne $\eta_k^B = \frac{\pi}{2} - \eta$. Enfin, le grand cercle $C_j^{(2)}$ traverse le panel (IV) aux points de coordonnées $(\xi_k^W = k\Delta\xi, \eta_k^W)$.

De la même manière, le grand cercle $C_i^{(1)}$ traverse les panels (V), (III) et (VI) dans cet ordre et est paramétré par l'angle β.

Les autres panels (II), (III), (IV), (V), (VI) sont traités de la même manière. Compte tenu des symétries de la Cubed-Sphere, six familles de grands cercles sont à considérer. On définit les familles de grands cercles suivants :

- (I_α) et (I_β) sont les grands cercles passant par le panel (I). Ils sont définis comme les isolignes en η et en ξ (du panel (I)) respectivement,
- (II_α) et (II_β) sont les grands cercles passant par le panel (II). Ils sont définis comme les isolignes en η et en ξ (du panel (II)) respectivement,
- (V_α) et (V_β) sont les grands cercles passant par le panel (V). Ils sont définis comme les isolignes en η et en ξ (du panel (V)) respectivement.

La structure de la Cubed-Sphere décrite par des grands cercles est à la base du procédé de calcul des dérivées hermitiennes que nous utilisons dans ce travail. Le calcul du gradient suit la procédure suivante :

1. Construction de données le long de grands cercles complets,

2. Calcul des dérivées hermitiennes sur les grands cercles,

3. Assemblage pour obtenir le gradient.

Pour calculer $\dfrac{\partial h}{\partial \alpha}_{|C_1}$ et $\dfrac{\partial h}{\partial \beta}_{|C_2}$ aux points de maillage, il est utile de connaître les coordonnées des points d'intersections de grands cercles entre les panels.

Soit $h_{i,j}^{(k)}$ donné avec $-N/2 \leq i,j \leq N/2$ et $(k) = (I),(II),(III),(IV),(V),(VI)$. Considérons le grand cercle $C_1 \in (I_\alpha)$. Il correspond à l'isoligne $\eta = \eta_0^F = j_0 \Delta\eta$. Les valeurs h_{i,j_0}^F sont localisée sur un grand cercle avec $-N/2 \leq i \leq N/2$. Le cercle C_1 traverse le panel (II) sur lequel il faut interpoler les données pour compléter les valeurs le long de C_1 sur le panel (II). Le cercle C_1 coupe chaque isoligne $\xi = \xi_{i_0}^E = i_0 \Delta\xi$ au point de coordonnées $(\xi_{i_0}^E, \beta_{i_0,j_0})$ dans le système de coordonnées (ξ^E, β) du panel (II). Un point $\mathbf{x}(x,y,z) \in \mathbb{R}^3$ d'un cercle de la famille (II_β) satisfait les relations :

$$\left\{ \begin{array}{rcl} x &=& -a\cos\beta\sin\xi^E \\ y &=& a\cos\beta\cos\xi^E \\ z &=& a\sin\beta. \end{array} \right. \tag{3.104}$$

Le point $\mathbf{x}(x,y,z)$ se situe à l'intersection de C_1 (isocline $\eta_{j_0}^F$ du panel (I)) avec le grand cercle correspondant à l'isocline $\xi = \xi_{i_0}^E = i_0\Delta\xi$. Alors x, y et z sont solutions du système suivant :

$$\left\{ \begin{array}{rcccl} x &=& a\cos\alpha_{i_0,j_0}\cos\eta_{j_0}^F &=& a\cos\beta_{i_0,j_0}\cos\xi_{i_0}^E \\ y &=& a\sin\alpha_{i_0,j_0} &=& a\cos\beta_{i_0,j_0}\cos\xi_{i_0}^E \\ z &=& a\cos\alpha_{i_0,j_0}\sin\eta_{j_0}^F &=& a\sin\beta_{i_0,j_0}. \end{array} \right. \tag{3.105}$$

De là, il découle :

$$\frac{z}{x} = \tan\eta_{j_0}^F = -\frac{\sin\beta_{i_0,j_0}}{\cos\beta_{i_0,j_0}\sin\xi_{i_0}^E} \tag{3.106}$$

donc :

$$\beta_{i_0,j_0} = \arctan\left[-\tan\eta_{j_0}^F \sin\xi_{i_0}^E\right]. \tag{3.107}$$

Ainsi, le point du panel (II) de coordonnées (ξ,β), représentant l'intersection de l'isocline $\eta_{j_0}^F$ du panel (I) avec l'isocline $\xi_{i_0}^E$ du panel (II), est donné par $(\xi_{i_0}^E, \beta_{i_0,j_0})$. En général, il ne s'agit pas d'un point de la Cubed-Sphere. Nous utilisons une méthode de spline cubique pour obtenir des valeurs $u_{i_0,j}^{(II)}$ interpolées sur la totalité de C. En continuant ce procédé, nous obtenons des valeurs sur le cercle C.

On calcule à présent les dérivées partielles $\dfrac{\partial h}{\partial \alpha}_{|C_1}$ et $\dfrac{\partial h}{\partial \beta}_{|C_2}$ en fonction des paramètres ξ et η. Par composition, les relations suivantes sont vérifiées :

$$\begin{array}{rcl} \dfrac{\partial h}{\partial \alpha}_{|C_1} &=& \dfrac{1}{\frac{\partial \alpha}{\partial \xi}_{|C_1}} \dfrac{\partial h}{\partial \xi}_{|C_1} \\ \dfrac{\partial h}{\partial \beta}_{|C_2} &=& \dfrac{1}{\frac{\partial \beta}{\partial \eta}_{|C_2}} \dfrac{\partial h}{\partial \eta}_{|C_2} \end{array} \tag{3.108}$$

car $\eta \mapsto \alpha$ est constant le long de $C_2 \in (I_\alpha)$ et $\xi \mapsto \beta$ est constant le long de $C_1 \in (I_\beta)$.

On calcule les vecteurs $\mathbf{g}_\alpha$, $\mathbf{g}_\beta$, $\mathbf{g}^\alpha$ et $\mathbf{g}^\beta$ en fonction de $\mathbf{g}_\xi$, $\mathbf{g}_\eta$, $\mathbf{g}^\xi$ et $\mathbf{g}^\eta$:

Proposition 3.11. *Les expressions suivantes sont vérifiées :*

- $\mathbf{e}_\alpha = \dfrac{1}{\frac{\partial \alpha}{\partial \xi}_{|C_1}} \mathbf{g}_\xi$ *le long de* C_1,

- $\mathbf{e}_\beta = \dfrac{1}{\frac{\partial\beta}{\partial\eta}_{|C_2}}\mathbf{g}_\eta$ *le long de* C_2,
- $\mathbf{e}^\alpha = \dfrac{\partial\alpha}{\partial\xi}_{|C_1}\mathbf{g}^\xi$ *le long de* C_1,
- $\mathbf{e}^\beta = \dfrac{\partial\beta}{\partial\eta}_{|C_2}\mathbf{g}^\eta$ *le long de* C_2.

Démonstration. • Par définition et composition, le long de C_1, on a :

$$\begin{aligned}\mathbf{g}_\xi &= \frac{d\mathbf{x}}{d\xi}_{|\eta} \\ &= \frac{\partial\alpha}{\partial\xi}_{|\eta}\frac{\partial\mathbf{x}}{\partial\alpha}_{|\eta} + \frac{\partial\beta}{\partial\xi}_{|\eta}\frac{\partial\mathbf{x}}{\partial\beta}_{|\eta} \\ &= \frac{\partial\alpha}{\partial\xi}_{|\eta}\frac{\partial\mathbf{x}}{\partial\alpha}_{|\eta}.\end{aligned}$$

De cette dernière relation, il découle la première égalité :

$$\mathbf{e}_\alpha = \frac{1}{\frac{\partial\alpha}{\partial\xi}_{|\eta}}\mathbf{g}_\xi. \tag{3.109}$$

De la même manière :

$$\mathbf{e}_\beta = \frac{1}{\frac{\partial\beta}{\partial\eta}_{|\xi}}\mathbf{g}_\eta. \tag{3.110}$$

- On pose $\mathbf{u} = \dfrac{\partial\alpha}{\partial\xi}_{|\eta}\mathbf{g}^\xi$ et $\mathbf{v} = \dfrac{\partial\beta}{\partial\eta}_{|\xi}\mathbf{g}^\eta$. Donc le long de C_1, on a :

$$\mathbf{e}_\alpha \cdot \mathbf{u} = \frac{\frac{\partial\alpha}{\partial\xi}_{|\eta}}{\frac{\partial\alpha}{\partial\xi}_{|\eta}}\mathbf{g}_\xi \cdot \mathbf{g}^\xi = 1. \tag{3.111}$$

De plus :

$$\mathbf{e}_\beta \cdot \mathbf{u} = \frac{\frac{\partial\alpha}{\partial\xi}_{|\eta}}{\frac{\partial\beta}{\partial\eta}_{|\xi}}\mathbf{g}_\eta \cdot \mathbf{g}^\xi = 0 \tag{3.112}$$

Donc $\mathbf{e}^\alpha = \mathbf{u}$. De la même manière, on montre que $\mathbf{e}^\beta = \mathbf{v}$ le long de C_2.

□

Théorème 3.4. *Soit* $h : \mathbb{S}^2_a \to \mathbb{R}$ *une fonction régulière. Alors :*

$$\nabla_T h = \frac{\partial h}{\partial\xi}_{|\eta=\bar\eta}\mathbf{g}^\xi + \frac{\partial h}{\partial\eta}_{|\xi=\bar\xi}\mathbf{g}^\eta. \tag{3.113}$$

Démonstration. Les égalités suivantes sont vérifiées grâce à la proposition 3.11 et aux équations (3.108) :

$$\frac{\partial h}{\partial\alpha}_{|C_1}\mathbf{e}^\alpha = \frac{1}{\alpha'(\xi)}\frac{\partial h}{\partial\xi}_{|\eta=\bar\eta}\alpha'(\xi)\mathbf{g}^\xi = \frac{\partial h}{\partial\xi}_{|\eta=\bar\eta}\mathbf{g}^\xi \tag{3.114}$$

De même on a :

$$\frac{\partial h}{\partial\beta}_{|C_2}\mathbf{e}^\beta = \frac{\partial h}{\partial\eta}_{|\xi=\bar\xi}\mathbf{g}^\eta \tag{3.115}$$

d'où le résultat à l'aide de la formule (3.10). □

3.4 Harmoniques Sphériques sur la Cubed-Sphere

Pour résoudre des problèmes sur la sphère, les harmoniques sphériques jouent un rôle particulièrement important [5, 37]. On les utilise pour décomposer des fonctions de carré intégrable. On note $\mathbf{Y}_m^l$ les harmoniques sphériques avec $l \in \mathbb{N}$ et $|m| \leq l$, $m \in \mathbb{Z}$. Les harmoniques sphériques sur $\mathbb{S}_a^2$ s'expriment en fonction de (λ, θ), les coordonnées longitude-latitude, par

$$\mathbf{Y}_m^l(\mathbf{x}) = \mathbf{Y}_m^l(\lambda, \theta) = \frac{(-1)^m}{a}\sqrt{\frac{(2l+1)}{4\pi}\frac{(l-|m|)!}{(l+|m|)!}}P_l^{|m|}(\sin(\theta))\exp(im\lambda), \tag{3.116}$$

où P_l^m désignent les polynômes de Legendre associés [5, 57]. Ils s'expriment grâce à la formule

$$P_l^m(x) = \frac{1}{2^l l!}(1-x^2)^{m/2}\frac{d^l}{dx^l}\left((x^2-1)^l\right). \tag{3.117}$$

En particulier, on a

$$\mathbf{Y}_0^0(\mathbf{x}) = \frac{1}{a\sqrt{4\pi}}. \tag{3.118}$$

Toute fonction $f : \mathbf{x} \in \mathbb{S}_a^2 \mapsto f(\mathbf{x})$ de carré intégrable s'écrit comme combinaison linéaire des harmoniques sphériques. Autrement dit, il existe une famille $(f_{m,l})_{l\in\mathbb{N},|m|\leq l}$ de $\mathbb{C}$ telle que

$$f = \sum_{l\in\mathbb{N}}\sum_{m=-l}^{l} f_{m,l}\mathbf{Y}_m^l. \tag{3.119}$$

Les harmoniques sphériques $\mathbf{Y}_m^l$ forment une famille orthonormée de $L^2(\mathbb{S}_a^2)$ [5], c'est à dire

$$\int_{\mathbb{S}_a^2} \mathbf{Y}_m^l(\mathbf{x})\bar{\mathbf{Y}}_{m'}^{l'}(\mathbf{x})d\sigma(\mathbf{x}) = \delta_{m,m'}\delta_{l,l'} \tag{3.120}$$

où $\delta_{p,q}$ désigne le symbole de Kronecker de p et q.

Dans cette partie, on s'intéresse à la version discrète sur la Cubed-Sphere de ces résultats. Pour cela nous définissons un produit scalaire pondéré sur le maillage et analysons la version discrète de l'équation (3.120). La conception et l'étude d'un produit scalaire sur la Cubed-Sphere sont liées à la conception d'une méthode de quadrature. Une étude a déjà été réalisée dans [70] dans le cadre de méthodes permettant d'approcher l'intégrale.

3.4.1 Produit scalaire discret sur la Cubed-Sphere

Sur un panel $(k) = (I), \cdots, (VI)$ donné, on note $I^{(k)}(f)$ l'intégrale d'une fonction $f : \mathbf{x} \in \mathbb{S}_a^2 \mapsto f(\mathbf{x}) \in \mathbb{C}$ (supposée intégrable) :

$$I^{(k)}(f) = \int_{(k)} f(\mathbf{x})d\sigma(\mathbf{x}). \tag{3.121}$$

Proposition 3.12. *Soit $f : \mathbf{x} \in \mathbb{S}_a^2 \mapsto f(\mathbf{x}) \in \mathbb{C}$ une fonction intégrable sur la sphère $\mathbb{S}_a^2$. Alors pour tout panel $(k) = (I), \cdots, (VI)$, nous avons*

$$I^{(k)} = \int_{[-\pi/4,\pi/4]^2} f(\xi,\eta)\sqrt{\det(\mathbf{G})}d\xi d\eta. \tag{3.122}$$

où (ξ, η) sont les coordonnées d'un point du panel définies dans la section 3.2.

Démonstration. On définit $\psi_{(k)}$ l'application permettant de passer des coordonnées (ξ, η) associées au panel (k) aux coordonnées cartésiennes . L'application $\psi_{(k)}$ est bien définie car (ξ, η) est un système de coordonnées admissibles sur le panel (k). On a

$$\psi_{(k)} : (\xi,\eta) \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^2 \mapsto \mathbf{x}(\xi,\eta) \in (k). \tag{3.123}$$

Par exemple, sur le panel (I) on a :

$$\psi_{(I)} : (\xi, \eta) \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^2 \mapsto \mathbf{x}(\xi, \eta) = (x, y, z) \in (I) \subset \mathbb{R}^3, \tag{3.124}$$

or les relations suivantes sont vérifiées :

$$\begin{cases} X = \tan\xi = \dfrac{y}{x} \\ Y = \tan\eta = \dfrac{z}{x} \\ a^2 = x^2 + y^2 + z^2. \end{cases} \tag{3.125}$$

Donc $\mathbf{x}(x, y, z)$ est donné par

$$\begin{cases} x &= \dfrac{a}{\sqrt{1 + \tan^2(\xi) + \tan^2(\eta)}} \\ y &= \dfrac{a\tan(\xi)}{\sqrt{1 + \tan^2(\xi) + \tan^2(\eta)}} \\ z &= \dfrac{a\tan(\eta)}{\sqrt{1 + \tan^2(\xi) + \tan^2(\eta)}}. \end{cases} \tag{3.126}$$

Des relations similaires peuvent être déduites sur tous les panels.

$$\begin{aligned} \int_{(k)} f(\mathbf{x}) d\sigma(\mathbf{x}) &= \int_{[-\pi/4, \pi/4]^2} f_k(\psi_{(k)}(\xi, \eta)) |\det J_{\psi_{(k)}}(\xi, \eta)| d\xi d\eta \\ &= \int_{[-\pi/4, \pi/4]^2} f(\psi_{(k)}(\xi, \eta)) \sqrt{\det(\mathbf{G})} d\xi d\eta \end{aligned}$$

avec $J_{\psi_{(k)}}$ la matrice Jacobienne de $\psi_{(k)}$. En notant $f_k = f \circ \psi_{(k)}$, on a

$$I^{(k)} = \int_{[-\pi/4, \pi/4]^2} f_k(\xi, \eta) \sqrt{\det(\mathbf{G})} d\xi d\eta. \tag{3.127}$$

□

Les panels $(I), \cdots, (VI)$ couvrent la totalité de la Cubed-Sphere et sont disjoints. Donc en notant

$$I(f) = \int_{\mathbb{S}_a^2} f(\mathbf{x}) d\sigma(\mathbf{x}), \tag{3.128}$$

on obtient la relation

$$I(f) = \sum_{(k)=(I)}^{(VI)} I^{(k)}(f). \tag{3.129}$$

Par analogie avec l'intégrale, nous nous intéressons à des produits scalaires de la forme

$$< \mathfrak{u}, \mathfrak{v} >_{\mathrm{CS}} = \sum_{(k)=(I)}^{(VI)} \left(\Delta\xi \Delta\eta \sum_{i=-N/2}^{N/2} \sum_{j=-N/2}^{N/2} \omega_{i,j} \mathfrak{u}_{i,j}^{(k)} \bar{\mathfrak{v}}_{i,j}^{(k)} \sqrt{\bar{\mathbf{G}}_{i,j}} \right) \tag{3.130}$$

où $\mathfrak{u}$ et $\mathfrak{v}$ sont des fonctions de grille sur la Cubed-Sphere. De plus, nous notons $\bar{\mathbf{G}}_{i,j} = \det(\mathbf{G}(\xi_i, \eta_j))$. Pour tout $-N/2 \leq i, j \leq N/2$, on note $\omega_{i,j} > 0$ un poids donné.

Proposition 3.13. *$< \cdot, \cdot >_{CS}$ définit un produit scalaire hermitien sur l'espace des fonctions de grille sur la Cubed-Sphere.*

3.4.2 Harmoniques sphériques sur la Cubed-Sphere

Pour les produits scalaires de la forme $< \cdot, \cdot >_{\mathrm{CS}}$, on analyse l'orthogonalité des harmoniques sphériques restreintes au maillages. Pour cela, on commence par observer les propriétés de symétrie suivantes :

Lemme 3.1. *Soit $l \in \mathbb{N}$ et m tel que $|m| \leq l$. Alors à tout harmonique sphérique $\mathbf{Y}_m^l$, si $(\lambda, \theta) \in]0, 2\pi] \times]-\pi/2, \pi/2[$ représente les coordonnées longitude-latitude, on a les relations de symétrie suivantes :*

- $\mathbf{Y}_m^l(\lambda, \theta) = (-1)^{m+l} \mathbf{Y}_m^l(\lambda, -\theta)$,
- $\mathbf{Y}_m^l(\lambda + \alpha, \theta) = e^{im\alpha} \mathbf{Y}_m^l(\lambda, \theta)$,
- $\mathbf{Y}_m^l(\lambda, \theta) = \bar{\mathbf{Y}}_m^l(-\lambda, \theta)$.

En tenant compte de ces symétries sur les harmoniques sphériques, il suffit d'ajouter des symétries semblables aux poids $\omega_{i,j}$ dans la définition du produit scalaire $< \cdot, \cdot >_{\mathrm{CS}}$ pour que des compensations apparaissent entre les panels de la Cubed-Sphere. On a le résultat suivant :

Théorème 3.5. *Soient $\mathbf{Y}_m^l$ et $\mathbf{Y}_{m'}^{l'}$ deux harmoniques sphériques avec $l, l' \in \mathbb{N}$, $|m| \leq l$ et $|m'| \leq l'$. On suppose que pour tout $-N/2 \leq i, j \leq N/2$, les poids $(\omega_{i,j})_{-N/2 \leq i,j \leq N/2}$ satisfont*

$$\omega_{i,j} = \omega_{-i,j} = \omega_{-i,-j} = \omega_{i,-j} > 0 \tag{3.131}$$

alors les fonctions de grilles restreintes à la Cubed-Sphere $\mathbf{Y}_m^{l,*}$ et $\mathbf{Y}_{m'}^{l',*}$ satisfont

$$< \mathbf{Y}_m^{l,*}, \mathbf{Y}_{m'}^{l',*} >_{CS} = 0 \tag{3.132}$$

si $m + m'$ ou $l + l'$ est impair. En particulier on a

$$< \mathbf{Y}_m^{l,*}, \mathbf{Y}_0^{0,*} >_{CS} = 0 \tag{3.133}$$

lorsque

- *m impair ou bien,*
- *l impair ou bien,*
- *l pair et $m \equiv 2[4]$.*

Démonstration. Soit $(k) = (I), \cdots, (VI)$. Nous adoptons la notation

$$S_{(k)} = \Delta\xi \Delta\eta \sum_{i=-N/2}^{N/2} \sum_{j=-N/2}^{N/2} \omega_{i,j} (\mathbf{Y}_m^{l,*})_{i,j}^{(k)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{i,j}^{(k)} \sqrt{\bar{\mathbf{G}}_{i,j}}. \tag{3.134}$$

Donc la relation suivante est vérifiée :

$$S = < \mathbf{Y}_m^{l,*}, \mathbf{Y}_{m'}^{l',*} >_{\mathrm{CS}} = \sum_{(k)=(I)}^{(VI)} S_{(k)}. \tag{3.135}$$

D'après les relations de symétrie sur $\omega_{i,j}$ (3.131) et sur les harmoniques sphériques (lemme 3.1), on a

$$\begin{aligned} S_{(I)} &= \Delta\xi \Delta\eta \sum_{i=-N/2}^{N/2} \sum_{j=-N/2}^{N/2} \omega_{i,j} (\mathbf{Y}_m^{l,*})_{i,j}^{(I)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{i,j}^{(I)} \sqrt{\bar{\mathbf{G}}_{i,j}} \\ &= (-1)^{l+l'} \Delta\xi \Delta\eta \sum_{i=-N/2}^{N/2} \sum_{j=-N/2}^{N/2} \omega_{i,j} (\mathbf{Y}_m^{l,*})_{i,j}^{(III)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{i,j}^{(III)} \sqrt{\bar{\mathbf{G}}_{i,j}} \\ &= (-1)^{l+l'} S_{(III)}. \end{aligned}$$

De la même manière, on a

$$S_{(II)} = (-1)^{l+l'} S_{(IV)}. \tag{3.136}$$

Les panels (V) et (VI) étant symétriques par rapport à l'équateur, en utilisant le premier point du lemme 3.1, on a

$$S_{(V)} = (-1)^{m+m'+l+l'} S_{(VI)}. \tag{3.137}$$

En combinant (3.136) et (3.137), on montre que

$$S = (1+(-1)^{l+l'})(S_{(I)} + S_{(II)}) + (1+(-1)^{l+l'+m+m'})S_{(V)}. \tag{3.138}$$

Les points du panel (I) se déduisent par rotation d'angle $\pi/2$ à partir de ceux du panel (II). De plus, d'après le lemme 3.1, on a

$$\mathbf{Y}_m^l(\lambda+\pi/2,\theta) = \exp\left(-im\frac{\pi}{2}\right)\mathbf{Y}_m^l(\lambda,\theta). \tag{3.139}$$

Partant de ces relations entre les panels (I) et (II), on trouve

$$S = (1+(-1)^{l+l'})\left(\exp\left(-i(m+m')\frac{\pi}{2}\right)+1\right)S_{(I)} + (1+(-1)^{l+l'+m+m'})S_{(V)}. \tag{3.140}$$

Le problème est à présent de savoir si on peut avoir $S_{(I)} = 0$ ou $S_{(V)} = 0$.

Considérons d'abord le panel (I). On a

$$\begin{aligned} S_{(I)} &= \Delta\xi\Delta\eta \sum_{i=-N/2}^{N/2}\sum_{j=-N/2}^{N/2} \omega_{i,j}(\mathbf{Y}_m^{l,*})_{i,j}^{(I)}(\bar{\mathbf{Y}}_{m'}^{l',*})_{i,j}^{(I)}\sqrt{\bar{\mathbf{G}}_{i,j}} \\ &= S_1 + S_2 + S_3 + S_4, \end{aligned}$$

avec S_1 donné par

$$\begin{aligned} S_1 =& \Delta\xi\Delta\eta \sum_{i=-N/2}^{1}\sum_{j=1}^{N/2}\omega_{i,j}(\mathbf{Y}_m^{l,*})_{i,j}^{(I)}(\bar{\mathbf{Y}}_{m'}^{l',*})_{i,j}^{(I)}\sqrt{\bar{\mathbf{G}}_{i,j}} \\ &+ \frac{\Delta\xi\Delta\eta}{2}\sum_{i=-N/2}^{1}\omega_{i,0}(\mathbf{Y}_m^{l,*})_{i,0}^{(I)}(\bar{\mathbf{Y}}_{m'}^{l',*})_{i,0}^{(I)}\sqrt{\bar{\mathbf{G}}_{i,0}} \\ &+ \frac{\Delta\xi\Delta\eta}{2}\sum_{j=1}^{N/2}\omega_{0,j}(\mathbf{Y}_m^{l,*})_{0,j}^{(I)}(\bar{\mathbf{Y}}_{m'}^{l',*})_{0,j}^{(I)}\sqrt{\bar{\mathbf{G}}_{0,j}} \\ &+ \frac{\Delta\xi\Delta\eta}{4}(\mathbf{Y}_m^{l,*})_{0,0}^{(I)}(\bar{\mathbf{Y}}_{m'}^{l',*})_{0,0}^{(I)}\sqrt{\bar{\mathbf{G}}_{0,0}}, \end{aligned}$$

S_2 donné par

$$\begin{aligned} S_2 =& \Delta\xi\Delta\eta \sum_{i=-N/2}^{1}\sum_{j=-N/2}^{1}\omega_{i,j}(\mathbf{Y}_m^{l,*})_{i,j}^{(I)}(\bar{\mathbf{Y}}_{m'}^{l',*})_{i,j}^{(I)}\sqrt{\bar{\mathbf{G}}_{i,j}} \\ &+ \frac{\Delta\xi\Delta\eta}{2}\sum_{i=-N/2}^{1}\omega_{i,0}(\mathbf{Y}_m^{l,*})_{i,0}^{(I)}(\bar{\mathbf{Y}}_{m'}^{l',*})_{i,0}^{(I)}\sqrt{\bar{\mathbf{G}}_{i,0}} \\ &+ \frac{\Delta\xi\Delta\eta}{2}\sum_{j=-N/2}^{1}\omega_{0,j}(\mathbf{Y}_m^{l,*})_{0,j}^{(I)}(\bar{\mathbf{Y}}_{m'}^{l',*})_{0,j}^{(I)}\sqrt{\bar{\mathbf{G}}_{0,j}} \\ &+ \frac{\Delta\xi\Delta\eta}{4}(\mathbf{Y}_m^{l,*})_{0,0}^{(I)}(\bar{\mathbf{Y}}_{m'}^{l',*})_{0,0}^{(I)}\sqrt{\bar{\mathbf{G}}_{0,0}}, \end{aligned}$$

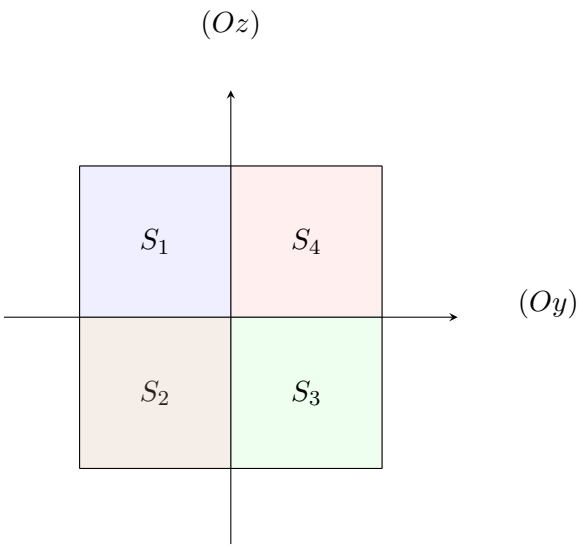


FIGURE 3.13 – Représentation schématique des zones S_1 à S_4 sur le panel (I)

S_3 donné par

$$\begin{aligned} S_3 =& \Delta\xi\Delta\eta \sum_{i=1}^{N/2} \sum_{j=-N/2}^{1} \omega_{i,j} (\mathbf{Y}_m^{l,*})_{i,j}^{(I)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{i,j}^{(I)} \sqrt{\bar{\mathbf{G}}_{i,j}} \\ &+ \frac{\Delta\xi\Delta\eta}{2} \sum_{i=1}^{N/2} \omega_{i,0} (\mathbf{Y}_m^{l,*})_{i,0}^{(I)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{i,0}^{(I)} \sqrt{\bar{\mathbf{G}}_{i,0}} \\ &+ \frac{\Delta\xi\Delta\eta}{2} \sum_{j=-N/2}^{1} \omega_{0,j} (\mathbf{Y}_m^{l,*})_{0,j}^{(I)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{0,j}^{(I)} \sqrt{\bar{\mathbf{G}}_{0,j}} \\ &+ \frac{\Delta\xi\Delta\eta}{4} (\mathbf{Y}_m^{l,*})_{0,0}^{(I)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{0,0}^{(I)} \sqrt{\bar{\mathbf{G}}_{0,0}}, \end{aligned}$$

S_4 donné par

$$\begin{aligned} S_4 =& \Delta\xi\Delta\eta \sum_{i=1}^{N/2} \sum_{j=1}^{N/2} \omega_{i,j} (\mathbf{Y}_m^{l,*})_{i,j}^{(I)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{i,j}^{(I)} \sqrt{\bar{\mathbf{G}}_{i,j}} \\ &+ \frac{\Delta\xi\Delta\eta}{2} \sum_{i=1}^{N/2} \omega_{i,0} (\mathbf{Y}_m^{l,*})_{i,0}^{(I)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{i,0}^{(I)} \sqrt{\bar{\mathbf{G}}_{i,0}} \\ &+ \frac{\Delta\xi\Delta\eta}{2} \sum_{j=1}^{N/2} \omega_{0,j} (\mathbf{Y}_m^{l,*})_{0,j}^{(I)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{0,j}^{(I)} \sqrt{\bar{\mathbf{G}}_{0,j}} \\ &+ \frac{\Delta\xi\Delta\eta}{4} (\mathbf{Y}_m^{l,*})_{0,0}^{(I)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{0,0}^{(I)} \sqrt{\bar{\mathbf{G}}_{0,0}}. \end{aligned}$$

Une représentation des zones attribuées à S_1, S_2, S_3 et S_4 sur le panel (I) est donnée en Figure 3.13.

Compte tenu des symétries des harmoniques sphériques ainsi que des symétries des poids $\omega_{i,j}$, on a

$$\left\{ \begin{array}{rcl} S_3 & = & \bar{S}_1 \\ S_4 & = & \bar{S}_2 \\ \bar{S}_1 & = & (-1)^{m+m'+l+l'} S_1 \end{array} \right. \tag{3.141}$$

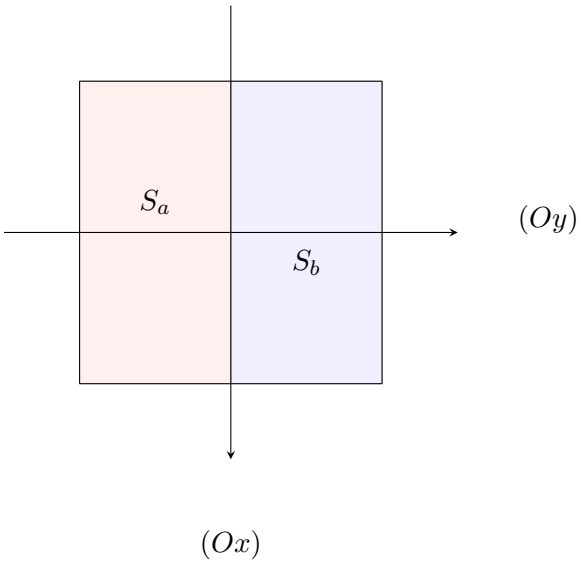


FIGURE 3.14 – Représentation schématique des zones S_a et S_b sur le panel V

donc il vient :

$$\begin{aligned} S_{(I)} &= S_1 + S_2 + S_3 + S_4 \\ &= S_1 + S_2 + \bar{S}_1 + \bar{S}_2 \\ &= (1 + (-1)^{l+l'+m+m'})(S_1 + S_2) \end{aligned}$$

De plus, on a le même type de résultat pour $S_{(II)}$.

Pour le panel (V), on nomme S_a et S_b les quantités suivantes :

$$\begin{aligned} S_a =& \Delta\xi\Delta\eta \sum_{i=-N/2}^{1} \sum_{j=-N/2}^{N/2} \omega_{i,j} (\mathbf{Y}_m^{l,*})_{i,j}^{(I)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{i,j}^{(I)} \sqrt{\bar{\mathbf{G}}_{i,j}} \\ &+ \frac{\Delta\xi\Delta\eta}{2} \sum_{j=-N/2}^{N/2} \omega_{0,j} (\mathbf{Y}_m^{l,*})_{0,j}^{(I)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{0,j}^{(I)} \sqrt{\bar{\mathbf{G}}_{0,j}}, \end{aligned}$$

S_b est donné par la relation suivante

$$\begin{aligned} S_b =& \Delta\xi\Delta\eta \sum_{i=1}^{N/2} \sum_{j=-N/2}^{N/2} \omega_{i,j} (\mathbf{Y}_m^{l,*})_{i,j}^{(I)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{i,j}^{(I)} \sqrt{\bar{\mathbf{G}}_{i,j}} \\ &+ \frac{\Delta\xi\Delta\eta}{2} \sum_{j=-N/2}^{N/2} \omega_{0,j} (\mathbf{Y}_m^{l,*})_{0,j}^{(I)} (\bar{\mathbf{Y}}_{m'}^{l',*})_{0,j}^{(I)} \sqrt{\bar{\mathbf{G}}_{0,j}}. \end{aligned}$$

Les zones attribuées à S_a et S_b sont représentées schématiquement dans la Figure 3.14.

On note que :

$$\begin{aligned} S_{(V)} &= S_a + S_b \\ &= S_a + (-1)^{l-l'} S_a \\ &= (1 + (-1)^{l+l'}) S_a. \end{aligned}$$

En faisant le bilan et en considérant les égalités démontrées, on obtient :

$$S = (1+(-1)^{l+l'})\left(\exp\left(-i(m+m')\frac{\pi}{2}\right)+1\right)(1+(-1)^{l+l'+m+m'})(S_1+S_2)+... \\ ...+(1+(-1)^{l+l'+m+m'})(1+(-1)^{l+l'})S_a. \quad (3.142)$$

Donc $S=0$ si $m+m'$ est impair ou si $l+l'$ est impair.

Dans le cas où $m=0$ et $l=0$, on a

$$< \mathbf{Y}_m^{l,*}, \mathbf{Y}_0^{0,*} >_{\mathrm{CS}} = (1+(-1)^l)\left(1+\exp\left(-im\frac{\pi}{2}\right)\right)(1+(-1)^{l+m})(S_1+S_2)+(1+(-1)^{l+m})(1+(-1)^l)S_a. \quad (3.143)$$

On a déjà vu que $< \mathbf{Y}_m^{l,*}, \mathbf{Y}_0^{0,*} >_{\mathrm{CS}} = 0$ si l est impair ou m impair. Dans le cas où $l \equiv 2[4]$, on a

$$< \mathbf{Y}_m^{l,*}, \mathbf{Y}_0^{0,*} >_{\mathrm{CS}} = 4S_a. \quad (3.144)$$

De plus $S_a = 0$ d'après le second point du lemme 3.1. Le résultat est alors prouvé. □

Dans le cas où $m+m'$ et $l+l'$ sont pairs, le produit scalaire $< \mathbf{Y}_m^l, \mathbf{Y}_{m'}^{l'} >_{\mathrm{CS}}$ n'est a priori pas nul. Un choix judicieux des coefficients $(\omega_{i,j})_{-N/2\leq i,j\leq N/2}$ permet d'assurer que cette quantité reste faible lorsque $m \neq m'$ et $l \neq l'$. C'est l'objet de la prochaine section.

3.4.3 Quadrature sur la sphère

Les équations que nous allons résoudre sur la sphère sont des relations de conservation. De manière à étudier les propriétés de conservation du schéma utilisé, il est utile de connaître une formule de quadrature adaptée. Dans [3, 34], la méthode de quadrature est conçue pour être adaptée aux harmoniques sphériques, voir aussi [62] ou les livres [5, 47]. Les méthodes de quadratures considérées visent à approcher l'intégrale

$$I(f) = \int_{\mathbb{S}_a^2} f(\mathbf{x}) d\sigma(\mathbf{x}). \quad (3.145)$$

On note que lorsque $f = \mathbf{Y}_m^l$ est une harmonique sphérique, il s'agit d'un cas particulier de la formule (3.120) avec $m' = l' = 0$. Ainsi, on a

$$I(\mathbf{Y}_m^l) = 0 \quad (3.146)$$

sauf si $m = l = 0$. Pour approcher $I(f)$, on considère des méthodes de quadrature de la forme

$$Q(f) = \sum_p \omega_p f(\mathbf{x}_p) \quad (3.147)$$

où $f : \mathbf{x} \in \mathbb{S}_a^2 \mapsto f(\mathbf{x}) \in \mathbb{C}$ est une fonction définie sur la sphère. Les points $(\mathbf{x}_p)_p$ représentent un nombre fini de points de $\mathbb{S}_a^2$, les valeurs $(\omega_p)_p$ sont des poids permettant d'approcher (3.145).

Compte tenu de la structure de la Cubed-Sphere, il est possible de construire des méthodes de quadrature par panel. On cherche des formules de quadrature proches du produit scalaire $< \cdot, \cdot >_{\mathrm{CS}}$ car si $f : \mathbb{S}_a^2 \mapsto \mathbb{C}$ est une fonction de $L^2(\mathbb{S}_a^2, \mathbb{C})$ alors

$$\int_{\mathbb{S}_a^2} f(\mathbf{x}) d\sigma(\mathbf{x}) = < f, 1 >_{L^2(\mathbb{S}_a^2,\mathbb{C})} . \quad (3.148)$$

On utilise donc des formules de quadrature de la forme

$$Q(\mathfrak{f}) = \sum_{(k)=(I)}^{(VI)} Q^{(k)}(\mathfrak{f}) \quad (3.149)$$

avec $Q^{(k)}(f^*)$ une formule de quadrature par panel de la forme

$$Q^{(k)}(\mathfrak{f}) = \Delta\xi\Delta\eta \sum_{i=-N/2}^{N/2} \sum_{j=-N/2}^{N/2} \omega_{i,j} \mathfrak{f}_{i,j}^{(k)} \sqrt{\bar{\mathbf{G}}_{i,j}} \approx \int_{(k)} f(\mathbf{x}) d\sigma(\mathbf{x}) \tag{3.150}$$

avec $\bar{\mathbf{G}}_{i,j} = \det(\mathbf{G}_{i,j})$. Ces formules de quadratures sont étudiées dans [70] et donnent de très bon résultats grâce aux symétries de la Cubed-Sphere. De plus, il est possible de les améliorer en perturbant les valeurs de $(\omega_{i,j})_{-N/2\leq i,j\leq N/2}$. Le lien avec le produit scalaire du chapitre 3 peut être fait directement en notant que

$$Q(\mathfrak{f}) = < \mathfrak{f}, \mathfrak{1} >_{\mathrm{CS}}, \tag{3.151}$$

où $\mathfrak{1}$ est la fonction de grille constante égale à 1. Or, $\mathbf{1} = a\sqrt{4\pi}\mathbf{Y}_0^{0,*}$, donc en utilisant le théorème 3.5, on obtient le corollaire suivant :

Corollaire 3.1. *Toute formule de quadrature Q vérifiant*

$$\omega_{i,j} = \omega_{-i,j} = \omega_{i,-j} = \omega_{-i,-j} \tag{3.152}$$

pour tout $-N/2 \leq i,j \leq N/2$ satisfait exactement la relation

$$Q\left(\mathbf{Y}_m^{l,*}\right) = 0, \tag{3.153}$$

avec $l \in \mathbb{N}$ et $|m| \leq l$, si

- *m est impair ou,*
- *l impair ou,*
- *l pair et $m \equiv 2[4]$.*

pour tout N paramètre de la Cubed-Sphere.

Dans la suite de ce chapitre, nous étudions différents choix de $(\omega_{i,j})_{-N/2\leq i,j\leq N/2}$ donnant des formules de quadratures (3.147) sur la sphère $\mathbb{S}_a^2$.

3.4.4 Formules de quadrature de type trapèze

Soient $\tilde{f} : (\xi,\eta) \in \mathbb{R}^2 \mapsto \tilde{f}(\xi,\eta) \in \mathbb{C}$ une fonction régulière et $(a,b) \in \mathbb{R}^2$ deux réels tels que $a < b$. Alors il existe un unique polynôme $P(\xi,\eta) \in \mathbb{C}[\xi,\eta]$ de degré 1 en ξ et 1 en η tel que

$$\left\{ \begin{array}{rcl} P(a,a) & = & \tilde{f}(a,a) \\ P(a,b) & = & \tilde{f}(a,b) \\ P(b,a) & = & \tilde{f}(b,a) \\ P(b,b) & = & \tilde{f}(b,b). \end{array} \right. \tag{3.154}$$

Ce polynôme est égal à, pour tous $(\xi,\eta) \in \mathbb{R}^2$,

$$P(\xi,\eta) = \frac{\tilde{f}(b,b)}{(a-b)^2}(\xi-a)(\eta-a) - \frac{\tilde{f}(a,b)}{(a-b)^2}(\xi-a)(\eta-b) - \frac{\tilde{f}(b,a)}{(a-b)^2}(\xi-b)(\eta-a) + \frac{\tilde{f}(a,a)}{(a-b)^2}(\xi-b)(\eta-b). \tag{3.155}$$

L'idée de la méthode des trapèzes est d'approcher l'intégrale de $\tilde{f}$ à l'aide de l'intégrale de P, c'est à dire

$$\int_{[a,b]^2} \tilde{f}(\xi,\eta) d\xi d\eta \approx (b-a)^2 \frac{\tilde{f}(a,a) + \tilde{f}(a,b) + \tilde{f}(b,a) + \tilde{f}(b,b)}{4}. \tag{3.156}$$

La méthode des trapèzes composite consiste à juxtaposer un ensemble de carrés pour calculer une intégrale. Plus le nombre de carrés est grand, plus l'intégrale approchée sera précise. On remarque que

$$\int_{[-\pi/4,\pi/4]^2} \tilde{f}(\xi,\eta)d\xi d\eta = \sum_{i=-N/2}^{N/2-1}\sum_{j=-N/2}^{N/2-1}\int_{\xi_i}^{\xi_{i+1}}\int_{\eta_i}^{\eta_{i+1}} \tilde{f}(\xi,\eta)d\xi d\eta, \tag{3.157}$$

avec $\xi_i = \dfrac{\pi}{4} + i\Delta\xi$ et $\eta_j = \dfrac{\pi}{4} + j\Delta\eta$. Les pas d'espace $\Delta\xi$ et $\Delta\eta$ sont

$$\Delta\xi = \Delta\eta = \frac{\pi}{2N}. \tag{3.158}$$

On souhaite approcher l'intégrale

$$\int_{[-\pi/4,\pi/4]^2} \tilde{f}(\xi,\eta)d\xi d\eta \tag{3.159}$$

en considérant la formule (3.156) sur chaque carré $[\xi_i,\xi_{i+1}]\times[\eta_i,\eta_{i+1}]$. La formule de quadrature prend alors la forme suivante :

$$\begin{aligned}
\int_{[-\pi/4,\pi/4]^2} \tilde{f}(\xi,\eta)d\xi d\eta &\approx \Delta\xi\Delta\eta \sum_{i=-N/2}^{N/2-1}\sum_{j=-N/2}^{N/2-1} \frac{\tilde{f}(\xi_i,\eta_j)+\tilde{f}(\xi_{i+1},\eta_j)+\tilde{f}(\xi_i,\eta_{j+1})+\tilde{f}(\xi_{i+1},\eta_{j+1})}{4} \\
&= \Delta\xi\Delta\eta \sum_{i=-N/2+1}^{N/2-1}\sum_{j=-N/2+1}^{N/2-1} \tilde{f}(\xi_i,\eta_j) + ... \\
&\quad ... + \frac{\Delta\xi\Delta\eta}{2}\left[\sum_{i=-N/2+1}^{N/2-1}\Big(\tilde{f}(\xi_i,\eta_{N/2})+\tilde{f}(\xi_i,\eta_{-N/2})\Big)\right] + ... \\
&\quad ... + \frac{\Delta\xi\Delta\eta}{2}\left[\sum_{j=-N/2+1}^{N/2-1}\Big(\tilde{f}(\xi_{N/2},\eta_j)+\tilde{f}(\xi_{-N/2},\eta_j)\Big)\right] + ... \\
&\quad ... + \frac{\Delta\xi\Delta\eta}{4}\Big(\tilde{f}(\xi_{N/2},\eta_{N/2})+\tilde{f}(\xi_{-N/2},\eta_{N/2})+\tilde{f}(\xi_{N/2},\eta_{-N/2})+\tilde{f}(\xi_{-N/2},\eta_{-N/2})\Big) \\
&= \Delta\xi\Delta\eta \sum_{i=-N/2}^{N/2}\sum_{j=-N/2}^{N/2} \omega_{i,j}\tilde{f}(\xi_i,\eta_j).
\end{aligned}$$

Les coefficients $(\omega_{i,j})_{-N/2\le i,j\le N/2}$ sont donnés par

- $\omega_{-\frac{N}{2},-\frac{N}{2}} = \omega_{\frac{N}{2},-\frac{N}{2}} = \omega_{-\frac{N}{2},\frac{N}{2}} = \omega_{\frac{N}{2},\frac{N}{2}} = \frac{1}{4}$,
- $\omega_{i,\frac{N}{2}} = \omega_{i,-\frac{N}{2}} = \frac{1}{2}$ pour $-\frac{N}{2}+1 \le i \le \frac{N}{2}-1$,
- $\omega_{\frac{N}{2},j} = \omega_{-\frac{N}{2},j} = \frac{1}{2}$ pour $-\frac{N}{2}+1 \le j \le \frac{N}{2}-1$,
- $\omega_{i,j} = 1$ dans tous les autres cas.

Soit $f : \mathbf{x} \in \mathbb{S}^2_a \mapsto f(\mathbf{x}) \in \mathbb{C}$ une fonction régulière. On peut appliquer la méthode des trapèzes composites à la fonction $\tilde{f}$ définie par

$$\tilde{f}(\xi,\eta) = f(\xi,\eta)\sqrt{\det(\mathbf{G}(\xi,\eta))} \text{ avec } -\pi/4 \le \xi,\eta \le \pi/4. \tag{3.160}$$

On obtient la formule de quadrature Q_{tpz} agissant sur les fonctions de grille sur la Cubed-Sphere et définie par

$$Q_{\mathrm{tpz}}(\mathfrak{f}) = \sum_{(k)=(I)}^{(VI)} Q^{(k)}_{\mathrm{tpz}}(\mathfrak{f}). \tag{3.161}$$

Pour tout $(k) = (I), ..., (VI)$, on a

$$Q^{(k)}_{\mathrm{tpz}}(\mathfrak{f}) = \Delta\xi\Delta\eta \sum_{i=-N/2}^{N/2} \sum_{j=-N/2}^{N/2} \omega_{i,j} \mathfrak{f}^{(k)}_{i,j} \sqrt{\bar{\mathbf{G}}_{i,j}}, \tag{3.162}$$

où $\bar{\mathbf{G}}_{i,j} = \det(\mathbf{G}(\xi_i, \eta_j))$ pour tout $-N/2 \leq i, j \leq N/2$.

Théorème 3.6. *La formule Q_{tpz} est consistante à l'ordre 2. De plus, pour toute fonction $f : \mathbf{x} \in \mathbb{S}^2_a \mapsto f(\mathbf{x}) \in \mathbb{C}$ régulière, on a*

$$\begin{aligned}
\int_{(k)} f(\mathbf{x}) d\sigma(\mathbf{x}) = Q^{(k)}_{tpz}(f^*) + ... \\
... + \frac{\Delta\xi^3}{12} \left(\sum_{i=-N/2+1}^{N/2-1} \left(\partial_\eta \tilde{f}_{i,N/2} - \partial_\eta \tilde{f}_{i,-N/2} \right) + \sum_{j=-N/2+1}^{N/2-1} \left(\partial_\eta \tilde{f}_{N/2,j} - \partial_\eta \tilde{f}_{-N/2,j} \right) \right) - ... \\
... - \frac{\Delta\xi^3}{24} \left(\partial_\eta \tilde{f}_{-N/2,N/2} - \partial_\eta \tilde{f}_{-N/2,-N/2} + \partial_\eta \tilde{f}_{N/2,N/2} - \partial_\eta \tilde{f}_{N/2,-N/2} \right) - ... \\
... - \frac{\Delta\xi^3}{24} \left(\partial_\xi \tilde{f}_{N/2,N/2} + \partial_\xi \tilde{f}_{N/2,-N/2} - \partial_\xi \tilde{f}_{-N/2,N/2} - \partial_\xi \tilde{f}_{-N/2,-N/2} \right) + ... \\
... + \mathcal{O}\left(\Delta\xi^4\right).
\end{aligned}$$

en notant que $\Delta\xi = \Delta\eta$ ainsi que $\partial_\xi f_{i,j} = \partial_\xi f(\xi_i, \eta_j)$ et $\partial_\eta f_{i,j} = \partial_\eta f(\xi_i, \eta_j)$ pour tous $-N/2 \leq i, j \leq N/2$, et

$$\tilde{f}(\xi, \eta) = f(\xi, \eta) \sqrt{\det(\mathbf{G}(\xi, \eta))} \text{ avec } -\pi/4 \leq \xi, \eta \leq \pi/4. \tag{3.163}$$

Démonstration. La fonction $\tilde{f}$ est régulière sur $[-\pi/4, \pi/4]^2$ comme produit de fonctions régulières. La formule d'Euler MacLaurin [27, 45, 63] permet de donner une estimation de l'intégrale sur $[a, b] \subset \mathbb{R}$:

$$\frac{1}{b-a} \int_a^b \tilde{f}(x) dx = \frac{\tilde{f}(a) + \tilde{f}(b)}{2} - \sum_{j=1}^{n} (b-a)^{2j-1} \frac{b_{2j}}{(2j)!} \left(\tilde{f}^{(2j-1)}(b) - \tilde{f}^{(2j-1)}(a) \right) + \mathcal{O}\left((b-a)^{2n+2}\right) \tag{3.164}$$

avec (b_{2j}) les nombres de Bernoulli [21]. En particulier on note que $b_2 = 1/6$, donc

$$\frac{1}{b-a} \int_a^b \tilde{f}(x) dx = \frac{\tilde{f}(a) + \tilde{f}(b)}{2} - \frac{(b-a)^2}{6 \cdot 2!} \left(\tilde{f}'(a) - \tilde{f}'(b) \right) - \frac{(b-a)^3}{30 \cdot 4!} \left(\tilde{f}^{(3)}(a) - \tilde{f}^{(3)}(b) \right) + \mathcal{O}\left((b-a)^4\right). \tag{3.165}$$

Ainsi, la formule suivante est vérifiée :

$$\begin{aligned}
\int_{-\pi/4}^{\pi/4} \tilde{f}(\xi, \eta) d\xi &= \sum_{i=-N/2}^{N/2-1} \int_{\xi_i}^{\xi_{i+1}} \tilde{f}(\xi, \eta) d\xi \\
&= \frac{\Delta\xi}{2} \tilde{f}(\xi_{-N/2}, \eta) + \Delta\xi \sum_{-N/2+1}^{N/2-1} \tilde{f}(\xi_i, \eta) + \frac{\Delta\xi}{2} \tilde{f}(\xi_{N/2}, \eta) - ... \\
&\quad ... - \frac{\Delta\xi^3}{24} \left(\sum_{i=-N/2+1}^{N/2-1} \left(\partial_\eta \tilde{f}_{i,N/2} - \partial_\eta \tilde{f}_{i,-N/2} \right) + \sum_{j=-N/2+1}^{N/2-1} \left(\partial_\eta \tilde{f}_{N/2,j} - \partial_\eta \tilde{f}_{-N/2,j} \right) \right) - ... \\
&\quad ... - \frac{\Delta\xi^2}{12} \left(\partial_\xi \tilde{f}(\xi_{N/2}, \eta) - \partial_\xi \tilde{f}(\xi_{-N/2}, \eta) \right) + ... \\
&\quad ... - \frac{\Delta\xi^3}{720} \left(\partial^{(3)}_\xi \tilde{f}(\xi_{N/2}, \eta) - \partial^{(3)}_\xi \tilde{f}(\xi_{-N/2}, \eta) \right) + \mathcal{O}\left(\Delta\xi^4\right).
\end{aligned}$$

On intègre cette dernière relation par rapport à $\eta \in [-\pi/4, \pi/4]$ et en utilisant à nouveau la formule d'Euler-MacLaurin, on obtient

$$\begin{aligned}\int_{(k)} f(\mathbf{x})d\sigma(\mathbf{x}) &= Q^{(k)}_{\mathrm{tpz}}(f^*) + ... \\ &... + \frac{\Delta\xi^3}{12}\left(\sum_{i=-N/2+1}^{N/2-1}\left(\partial_\eta \tilde{f}_{i,N/2} - \partial_\eta \tilde{f}_{i,-N/2}\right) + \sum_{j=-N/2+1}^{N/2-1}\left(\partial_\eta \tilde{f}_{N/2,j} - \partial_\eta \tilde{f}_{-N/2,j}\right)\right) - ... \\ &... - \frac{\Delta\xi^3}{24}\left(\partial_\eta \tilde{f}_{-N/2,N/2} - \partial_\eta \tilde{f}_{-N/2,-N/2} + \partial_\eta \tilde{f}_{N/2,N/2} - \partial_\eta \tilde{f}_{N/2,-N/2}\right) - ... \\ &... - \frac{\Delta\xi^3}{24}\left(\partial_\xi \tilde{f}_{N/2,N/2} + \partial_\xi \tilde{f}_{N/2,-N/2} - \partial_\xi \tilde{f}_{-N/2,N/2} - \partial_\xi \tilde{f}_{-N/2,-N/2}\right) + ... \\ &... + \mathcal{O}\left(\Delta\xi^4\right).\end{aligned}$$

Cette dernière relation permet d'obtenir

$$Q_{\mathrm{tpz}}(f^*) - \int_{(k)} f(\mathbf{x})d\sigma(\mathbf{x}) = \mathcal{O}(\Delta\xi^2). \tag{3.166}$$

Alors par somme sur (k), on montre que la formule Q_{tpz} est consistante avec l'intégrale à l'ordre 2. □

Remarque 3.2. *On note que si $-N/2 \leq i, j \leq N/2$ alors on a*

$$\omega_{i,j} = \omega_{-i,j} = \omega_{i,-j} = \omega_{-i,-j} > 0 \tag{3.167}$$

donc le corollaire 3.1 est vérifié pour la formule de quadrature Q_{tpz}.

3.4.5 Formule de quadrature de type Simpson

Dans cette section, on considère N pair. La formule de quadrature de Simpson est basée sur une approximation polynomiale de $\tilde{f}$ en trois points. L'approximation de Simpson s'exprime sous la forme :

$$\int_a^b \tilde{f}(\xi,\eta)d\xi \approx \frac{b-a}{6}\left(\tilde{f}(a,\eta) + 4\tilde{f}\left(\frac{a+b}{2},\eta\right) + \tilde{f}(b,\eta)\right) \tag{3.168}$$

Il s'agit d'une approximation plus précise que la méthode des trapèzes. En effet si $g : x \in [a,b] \mapsto g(x) \in \mathbb{C}$ alors

$$\int_a^b g(x)dx - \frac{b-a}{6}\left(g(a) + 4g\left(\frac{a+b}{2}\right) + g(b)\right) = \mathcal{O}((b-a)^5). \tag{3.169}$$

Il s'agit d'une méthode classique d'approximation d'intégrale [44].

Ainsi, la formule de Simpson composite s'obtient en juxtaposant des intervalles de la forme $[\xi_{i-1}, \xi_{i+1}]$ pour calculer l'intégrale sur $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$:

$$\int_{-\pi/4}^{\pi/4} \tilde{f}(\xi,\eta)d\xi = \sum_{i=-N/2,\mathrm{pair}}^{N/2} \int_{\xi_{i-1}}^{\xi_{i+1}} \tilde{f}(\xi,\eta)d\xi. \tag{3.170}$$

En considérant l'approximation de Simpson sur chaque intervalle $[\xi_{i-1}, \xi_{i+1}]$, on obtient

$$\int_{-\pi/4}^{\pi/4} \tilde{f}(\xi,\eta)d\xi = \frac{\Delta\xi}{3}\left[\tilde{f}(\xi_{-N/2},\eta) + 2\sum_{i=-N/4-1}^{N/4} \tilde{f}(\xi_{2i},\eta) + 4\sum_{i=-N/4}^{N/4} \tilde{f}(\xi_{2i-1},\eta) + \tilde{f}(\xi_{N/2},\eta)\right] + \mathcal{O}\left(\Delta\xi^4\right). \tag{3.171}$$

On intègre cette équation par rapport à η et on utilise la même approximation dans la direction η :

$$\int_{-\frac{\pi}{4}}^{\frac{\pi}{4}}\int_{-\frac{\pi}{4}}^{\frac{\pi}{4}} \tilde{f}(\xi,\eta)d\xi d\eta = \Delta\xi\Delta\eta \sum_{i=-N/2}^{N/2}\sum_{j=-N/2}^{N/2} \omega_{i,j}\tilde{f}(\xi_i,\eta_j) + \mathcal{O}\left(\Delta\xi^4\right) \tag{3.172}$$

où les coefficients $(\omega_{i,j})_{-N/2\leq i,j\leq N/2}$ sont donnés par

- $\omega_{\frac{N}{2},\frac{N}{2}} = \omega_{\frac{N}{2},-\frac{N}{2}} = \omega_{-\frac{N}{2},\frac{N}{2}} = \omega_{-\frac{N}{2},-\frac{N}{2}} = 1/9$,
- $\omega_{\frac{N}{2},i} = \omega_{-\frac{N}{2},i} = \omega_{i,\frac{N}{2}} = \omega_{i,-\frac{N}{2}} = 4/9$ si i est pair,
- $\omega_{\frac{N}{2},i} = \omega_{-\frac{N}{2},i} = \omega_{i,\frac{N}{2}} = \omega_{i,-\frac{N}{2}} = 2/9$ si i est impair,
- $\omega_{i,j} = 16/9$ si i et j sont pairs,
- $\omega_{i,j} = 4/9$ si i et j sont impairs,
- $\omega_{i,j} = 8/9$ dans les autres cas.

On déduit la formule de quadrature sur la sphère. Soit $\tilde{f}$ la fonction défini par

$$\tilde{f}(\xi,\eta) = f(\xi,\eta)\sqrt{\det(\mathbf{G}(\xi,\eta)}. \tag{3.173}$$

On trouve la formule de quadrature de type Simpson, pour tout $(k) = (I), ..., (VI)$, on a

$$Q_{\text{sps}}^{(k)}(\mathfrak{f}) = \Delta\xi\Delta\eta \sum_{i=-N/2}^{N/2}\sum_{j=-N/2}^{N/2} \omega_{i,j}\mathfrak{f}_{i,j}^{(k)}\sqrt{\bar{\mathbf{G}}_{i,j}}, \tag{3.174}$$

où $\bar{\mathbf{G}}_{i,j} = \det(\mathbf{G}(\xi_i,\eta_j))$ pour tout $-N/2 \leq i,j \leq N/2$. On note aussi la formule de quadrature

$$Q_{\text{sps}}(\mathfrak{f}) = \sum_{(k)=(I)}^{(VI)} Q_{\text{sps}}^{(k)}(\mathfrak{f}). \tag{3.175}$$

La formule Q_{sps} est consistante à l'ordre 4 au sens du théorème suivant :

Théorème 3.7. *Soit $f : \mathbf{x} \in \mathbb{S}_a^2 \mapsto f(\mathbf{x}) \in \mathbb{C}$ une fonction régulière, alors*

$$\int_{\mathbb{S}_a^2} f(\mathbf{x})d\sigma(\mathbf{x}) - Q_{sps}(f^*) = \mathcal{O}\left(\Delta\xi^4\right), \tag{3.176}$$

ainsi que pour $(k) = (I), ..., (VI)$

$$\int_{(k)} f(\mathbf{x})d\sigma(\mathbf{x}) - Q_{sps}^{(k)}(f^*) = \mathcal{O}\left(\Delta\xi^4\right), \tag{3.177}$$

lorsque N est pair.

Remarque 3.3. *Comme pour la formule de quadrature Q_{tpz}, si $-N/2 \leq i,j \leq N/2$ alors on a*

$$\omega_{i,j} = \omega_{-i,j} = \omega_{i,-j} = \omega_{-i,-j} > 0 \tag{3.178}$$

donc le corollaire 3.1 est vérifié pour la formule de quadrature Q_{sps}.

3.4.6 Formule de quadrature de type Q_α

La formule de quadrature Q_{tpz} est précise à l'ordre 2. De plus, elle ne dépend pas de la parité de N le paramètre de la Cubed-Sphere. Dans cette partie, nous considérons la formule de quadrature Q_α basée sur une perturbation de Q_{tpz}.

Définition 3.10. *On définit Q_α la formule de quadrature agissant sur les fonctions de grilles sur la Cubed-Sphere $\mathfrak{f}$ par*

$$Q_\alpha(\mathfrak{f}) = \sum_{(k)=(I)}^{(VI)} Q_\alpha^{(k)}(\mathfrak{f}) \tag{3.179}$$

où $Q_\alpha^{(k)}$ est la règle de quadrature par panel donnée par

$$Q_\alpha^{(k)}(\mathfrak{f}) = \sum_{i=-\frac{N}{2}}^{\frac{N}{2}} \sum_{j=-\frac{N}{2}}^{\frac{N}{2}} \Delta\xi\Delta\eta\omega_{i,j}\mathfrak{f}_{i,j}^{(k)}\sqrt{\bar{\mathbf{G}}^{(k)}(\xi_i,\eta_j)}. \tag{3.180}$$

Les coefficients $(\omega_{i,j})_{-N/2\leq i,j\leq N/2}$ vérifient :

- $\omega_{-\frac{N}{2},-\frac{N}{2}} = \omega_{\frac{N}{2},-\frac{N}{2}} = \omega_{-\frac{N}{2},\frac{N}{2}} = \omega_{\frac{N}{2},\frac{N}{2}} = \alpha$,
- $\omega_{i,\frac{N}{2}} = \omega_{i,-\frac{N}{2}} = \frac{1}{2}$ *pour* $-\frac{N}{2}+1 \leq i \leq \frac{N}{2}-1$,
- $\omega_{\frac{N}{2},j} = \omega_{-\frac{N}{2},j} = \frac{1}{2}$ *pour* $-\frac{N}{2}+1 \leq j \leq \frac{N}{2}-1$,
- $\omega_{i,j} = 1$ *dans tous les autres cas.*

Le paramètre α est un paramètre qui permettra d'optimiser cette formule de quadrature. En particulier, on a

$$Q_{\text{tpz}} = Q_{1/4}. \tag{3.181}$$

La formule ainsi perturbée permet de retrouver des résultats de précision semblables à ceux que Q_{tpz}. En effet, on a la proposition suivante :

Proposition 3.14. *Soit $f : \mathbf{x} \in \mathbb{S}_a^2 \to f(\mathbf{x}) \in \mathbb{C}$ une fonction régulière alors :*

$$I^{(k)}(f) - Q_\alpha^{(k)}(f^*) = \mathcal{O}\left(\Delta\xi^2\right) \tag{3.182}$$

pour tout $(k) \in \{(I), ..., (VI)\}$. De plus, on a

$$I(f) - Q_\alpha(f^*) = \mathcal{O}(\Delta\xi^2). \tag{3.183}$$

Démonstration. La fonction f est régulière, donc on a

$$\begin{aligned} I^{(k)}(f) - Q_\alpha^{(k)}(f^*) &= I^{(k)}(f) - Q_{\text{tpz}}^{(k)}(f^*) + Q_{\text{tpz}}^{(k)}(f^*) - Q_\alpha^{(k)}(f^*) \\ &= Q_{\text{tpz}}^{(k)}(f^*) - Q_\alpha^{(k)}(f^*) + \mathcal{O}(\Delta\xi^2) \\ &= 3\Delta\xi\Delta\eta\left(\frac{1}{4}-\alpha\right)\sum_{(\xi,\eta)\in C} f(\xi,\eta)\sqrt{\det(\mathbf{G}(\xi,\eta))} + \mathcal{O}(\Delta\xi^2) \end{aligned}$$

où C désigne l'ensemble des coins de la Cubed-Sphere.

De plus $(\xi,\eta) \in [-\pi/4,\pi/4]^2 \mapsto f(\xi,\eta)\sqrt{\det(\mathbf{G}(\xi,\eta))}$ est continue sur un compact donc bornée. Ainsi

$$I^{(k)}(f) - Q_\alpha^{(k)}(f^*) = \mathcal{O}(\Delta\xi^2). \tag{3.184}$$

En sommant cette formule sur chaque panel, on montre la consistance de Q_α avec l'intégrale sur la sphère $\mathbb{S}_a^2$. □

Proposition 3.15. *La méthode Q_α est d'ordre 2.*

Démonstration. La preuve de ce résultat consiste à montrer que la formule Q_α est d'ordre 2 mais pas d'ordre supérieur. On pose

$$\bar{\mathbf{G}}(\mathbf{x}) = \det(\mathbf{G}(\mathbf{x})). \tag{3.185}$$

On choisit f tel que pour tout $\mathbf{x} \in \mathbb{S}_a^2$, on ait $f(\mathbf{x}) = \dfrac{1}{\sqrt{\overline{\mathbf{G}}(\mathbf{x})}}$. Alors on sait que l'égalité suivante est exactement vérifiée pour tout N paramètre de maillage de la Sphère.

$$Q_{\text{tpz}}(f^*) - I(f) = 0 \tag{3.186}$$

De plus pour tout $(\xi, \eta) \in C = \{(\pi/4, \pi/4), (-\pi/4, \pi/4), (\pi/4, -\pi/4), (-\pi/4, -\pi/4)\}$, on a :

$$f(\xi, \eta)\sqrt{\overline{\mathbf{G}}(\xi, \eta} = 1 \tag{3.187}$$

d'où :

$$\begin{aligned} Q_\alpha(f^*) - Q_{\text{tpz}}(f) &= 3\Delta\xi\Delta\eta\left(\alpha - \frac{1}{4}\right)\sum_{(\xi,\eta)\in C} f(\xi, \eta)\sqrt{\overline{\mathbf{G}}(\xi, \eta)} \\ &= 24\Delta\xi\Delta\eta\left(\alpha - \frac{1}{4}\right), \end{aligned}$$

donc :

$$\begin{aligned} Q_\alpha(f^*) - I(f) &= Q_\alpha(f^*) - Q_{\text{tpz}}(f^*) + Q_{\text{tpz}}(f^*) - I(f) \\ &= 24\Delta\xi\Delta\eta\left(\alpha - \frac{1}{4}\right) + \mathcal{O}\left(\Delta\xi^2\right) \end{aligned}$$

La méthode de quadrature Q_α ne peut donc pas être d'un ordre supérieur à 2. □

Proposition 3.16. *Soit $(k) = (I), \cdots, (VI)$ un panel fixé. Le coefficient $\alpha = 1/3$ optimise la quantité*

$$|Q_\alpha^{(k)}(\mathbf{1}) - I^{(k)}(\mathbf{1})| \tag{3.188}$$

au sens où

$$|Q_{1/3}^{(k)}(\mathbf{1}) - I^{(k)}(\mathbf{1})| = \mathcal{O}\left(\Delta\xi^4\right). \tag{3.189}$$

Démonstration. On pose $f = \sqrt{\det(\overline{\mathbf{G}})}$, alors par comparaison avec la méthode des trapèzes, on a

$$\begin{aligned} Q_\alpha^{(k)}(\mathbf{1}) - I^{(k)}(\mathbf{1}) &= Q_\alpha(\mathbf{1}) - Q_{\text{tpz}}(\mathbf{1}) + Q_{\text{tpz}}(\mathbf{1}) - I^{(k)}(\mathbf{1}) \\ &= \Delta\xi^2\left(\alpha - \frac{1}{4}\right)\left(f_{N/2,N/2} + f_{-N/2,N/2} + f_{N/2,-N/2} + f_{-N/2,-N/2}\right) + ... \\ &\quad ... + \frac{\Delta\xi^3}{12}\left(\sum_{i=-N/2+1}^{N/2-1}\left(\partial_\eta\tilde{f}_{i,N/2} - \partial_\eta\tilde{f}_{i,-N/2}\right) + \sum_{j=-N/2+1}^{N/2-1}\left(\partial_\eta\tilde{f}_{N/2,j} - \partial_\eta\tilde{f}_{-N/2,j}\right)\right) - ... \\ &\quad ... - \frac{\Delta\xi^3}{24}\left(\partial_\eta\tilde{f}_{-N/2,N/2} - \partial_\eta\tilde{f}_{-N/2,-N/2} + \partial_\eta\tilde{f}_{N/2,N/2} - \partial_\eta\tilde{f}_{N/2,-N/2}\right) - ... \\ &\quad ... - \frac{\Delta\xi^3}{24}\left(\partial_\xi\tilde{f}_{N/2,N/2} + \partial_\xi\tilde{f}_{N/2,-N/2} - \partial_\xi\tilde{f}_{-N/2,N/2} - \partial_\xi\tilde{f}_{-N/2,-N/2}\right) + ... \\ &\quad ... + \mathcal{O}\left(\Delta\xi^4.\right). \end{aligned}$$

On note en particulier que

$$f_{N/2,N/2} = f_{-N/2,N/2} = f_{N/2,-N/2} = f_{-N/2,-N/2} = \sqrt{\bar{\mathbf{G}}(\pi/4, \pi/4)} = \frac{4a^2}{3\sqrt{3}} \tag{3.190}$$

en posant $\bar{\mathbf{G}} = \det(\mathbf{G})$.

De plus, en dérivant $\sqrt{\bar{\mathbf{G}}} = a^2 \dfrac{(1+X^2)(1+Y^2)}{(1+X^2+Y^2)^{3/2}}$ avec $X = \tan(\xi)$ et $Y = \tan(\eta)$ on obtient les relations suivantes :

$$\partial_\xi f = X\left(\frac{3Y^2}{1+X^2+Y^2} - 1\right)\sqrt{\bar{\mathbf{G}}} \tag{3.191}$$

$$\partial_\eta f = Y\left(\frac{3X^2}{1+X^2+Y^2} - 1\right)\sqrt{\bar{\mathbf{G}}}. \tag{3.192}$$

Ainsi :

- $\partial_\xi f_{N/2,j} = -\partial_\xi f_{-N/2,j} = 4\dfrac{Y^4-1}{(Y^2+2)^{5/2}}a^2$,
- $\partial_\eta f_{i,N/2} = -\partial_\eta f_{i,-N/2} = 4\dfrac{X^4-1}{(X^2+2)^{5/2}}a^2$.

d'où le terme d'ordre 2 est

$$\Delta\xi^2 a^2\left(\alpha - \frac{1}{4}\right)\frac{16}{3\sqrt{3}} + \frac{4}{3}S \tag{3.193}$$

avec $S = \Delta\xi \displaystyle\sum_{i=-N/2+1}^{N/2-1} g(i\Delta\xi)$ avec g la fonction $g : x \mapsto g(x) = \dfrac{\tan(x)^4-1}{(\tan(x)^2+2)^{5/2}}$.

Comme $g\left(-\frac{\pi}{4}\right) = g\left(\frac{\pi}{4}\right) = 0$ et par propriété de la formule des trapèzes, on note que :

$$S = \frac{g\left(-\frac{\pi}{4}\right) + g\left(\frac{\pi}{4}\right)}{2} + \int_{-\pi/4}^{\pi/4} g(x)dx + \mathcal{O}\left(\Delta\xi^2\right) = -\frac{1}{3\sqrt{3}} + \mathcal{O}\left(\Delta\xi^2\right). \tag{3.194}$$

Donc le terme d'ordre 2 s'annule à condition que :

$$\left(\alpha - \frac{1}{4}\right)\frac{16}{3\sqrt{3}} - \frac{4}{9\sqrt{3}} = 0. \tag{3.195}$$

Cette condition est vérifiée pour $\alpha = 1/3$.

Le terme d'ordre 3 s'écrit

$$-\frac{\Delta\xi^3}{24}\left[\partial_\eta f_{-\frac{N}{2},\frac{N}{2}} - \partial_\eta f_{-\frac{N}{2},-\frac{N}{2}} + \partial_\eta f_{\frac{N}{2},\frac{N}{2}} - \partial_\eta f_{\frac{N}{2},-\frac{N}{2}}\right] - \frac{\Delta\xi^3}{24}\left[\partial_\xi f_{\frac{N}{2},\frac{N}{2}} + \partial_\xi f_{\frac{N}{2},-\frac{N}{2}} - \partial_\xi f_{-\frac{N}{2},\frac{N}{2}} - \partial_\xi f_{-\frac{N}{2},-\frac{N}{2}}\right] \tag{3.196}$$

on note que :

$$\partial_\eta f_{\frac{N}{2},\frac{N}{2}} = \partial_\eta f_{-\frac{N}{2},\frac{N}{2}} = \partial_\eta f_{\frac{N}{2},-\frac{N}{2}} = \partial_\eta f_{-\frac{N}{2},-\frac{N}{2}} = 0. \tag{3.197}$$

De la même manière, en ξ :

$$\partial_\xi f_{\frac{N}{2},\frac{N}{2}} = \partial_\xi f_{-\frac{N}{2},\frac{N}{2}} = \partial_\xi f_{\frac{N}{2},-\frac{N}{2}} = \partial_\xi f_{-\frac{N}{2},-\frac{N}{2}} = 0 \tag{3.198}$$

d'où le résultat :

$$|Q_{1/3}^{(k)}(\mathbf{1}) - I^{(k)}(\mathbf{1})| = \mathcal{O}\left(\Delta\xi^4\right) \tag{3.199}$$

□

Remarque 3.4. *Ce résultat d'optimalité s'écrit également sur la sphère $\mathbb{S}_a^2$:*

$$|Q_{1/3}(\mathbf{1}) - I(\mathbf{1})| = \mathcal{O}\left(\Delta\xi^4\right). \tag{3.200}$$

Remarque 3.5. *La formule de quadrature Q_α vérifie le corollaire 3.1.*

3.4.7 Résultats numériques pour les formules de quadratures

Pour tester numériquement les performances des formules de quadrature introduites dans les sections précédentes, on s'intéresse à un ensemble de fonctions sur la sphère dont l'intégrale est connue. Ces fonctions sont utilisées pour tester la précision des formules de quadrature sphériques [34]. Elles sont données pour $(x, y, z) \in \mathbb{S}_a^2$ par

- $f_0(x, y, z) = 1$,
$$\int_{\mathbb{S}_a^2} f_0(\mathbf{x}) d\sigma(\mathbf{x}) = 4\pi a^2, \tag{3.201}$$

- $f_1(x, y, z) = 1 + x + y^2 + y \cdot x^2 + x^4 + y^5 + x^2 \cdot y^2 \cdot z^2$,
$$\int_{\mathbb{S}_a^2} f_1(\mathbf{x}) d\sigma(\mathbf{x}) = \frac{216\pi}{35} a^2, \tag{3.202}$$

- la fonction f_2 vérifie
$$\begin{aligned} f_2(x, y, z) \quad = \quad & \frac{3}{4} \exp\left[-\frac{(9x-2)^2}{4} - \frac{(9y-2)^2}{4} - \frac{(9z-2)^2}{4}\right] + ... \\ & ... + \frac{3}{4} \exp\left[-\frac{(9x+1)^2}{49} - \frac{9y+1}{10} - \frac{9z+1}{10}\right] + ... \\ & ... + \frac{1}{2} \exp\left[-\frac{(9x-7)^2}{4} - \frac{(9y-3)^3}{4} - \frac{(9z-5)^2}{4}\right] + ... \\ & ... + \frac{1}{5} \exp\left[-(9x-4)^2 - (9y-7)^2 - (9z-5)^2\right] \end{aligned} \tag{3.203}$$

  et s'intègre :
$$\int_{\mathbb{S}_a^2} f_2(\mathbf{x}) d\sigma(\mathbf{x}) = a^2 \cdot 6.6961822200736179523... \tag{3.204}$$

- $f_3(x, y, z) = \frac{1+\tanh(-9x-9y+9z)}{9}$,
$$\int_{\mathbb{S}_a^2} f_3(\mathbf{x}) d\sigma(\mathbf{x}) = \frac{4\pi}{9} a^2, \tag{3.205}$$

- $f_4(x, y, z) = \frac{1+\text{sign}(-9x-9y+9z)}{9}$,
$$\int_{\mathbb{S}_a^2} f_4(\mathbf{x}) d\sigma(\mathbf{x}) = \frac{4\pi}{9} a^2, \tag{3.206}$$

- $f_5(x, y, z) = \frac{1-\text{sign}(\pi x+y)}{\nu}$,
$$\int_{\mathbb{S}_a^2} f_5(\mathbf{x}) d\sigma(\mathbf{x}) = \frac{4\pi}{\nu} a^2. \tag{3.207}$$

  Dans les tests effectués, nous choisissons $\nu = 7$.

On note en particulier que la fonction f_1 est polynomiale donc régulière et elle ne fait intervenir qu'un nombre fini d'harmoniques sphériques dans sa décomposition. Les fonctions f_2 et f_3 font intervenir un plus grand nombre d'harmoniques sphériques et sont régulières. Les fonctions f_4 et f_5 sont discontinues.

Pour chacune de ces fonctions, on mesure l'erreur relative donnée par

$$e_r(f) = \frac{|I(f) - Q(f^*)|}{|I(f)|} \tag{3.208}$$

où $I(f)$ représente l'intégrale exacte sur la sphère $\mathbb{S}_a^2$, $Q(f^*)$ est la valeur approchée obtenue par une formule de quadrature. Nous retenons l'erreur maximale obtenue après 1000 rotations aléatoires de f.

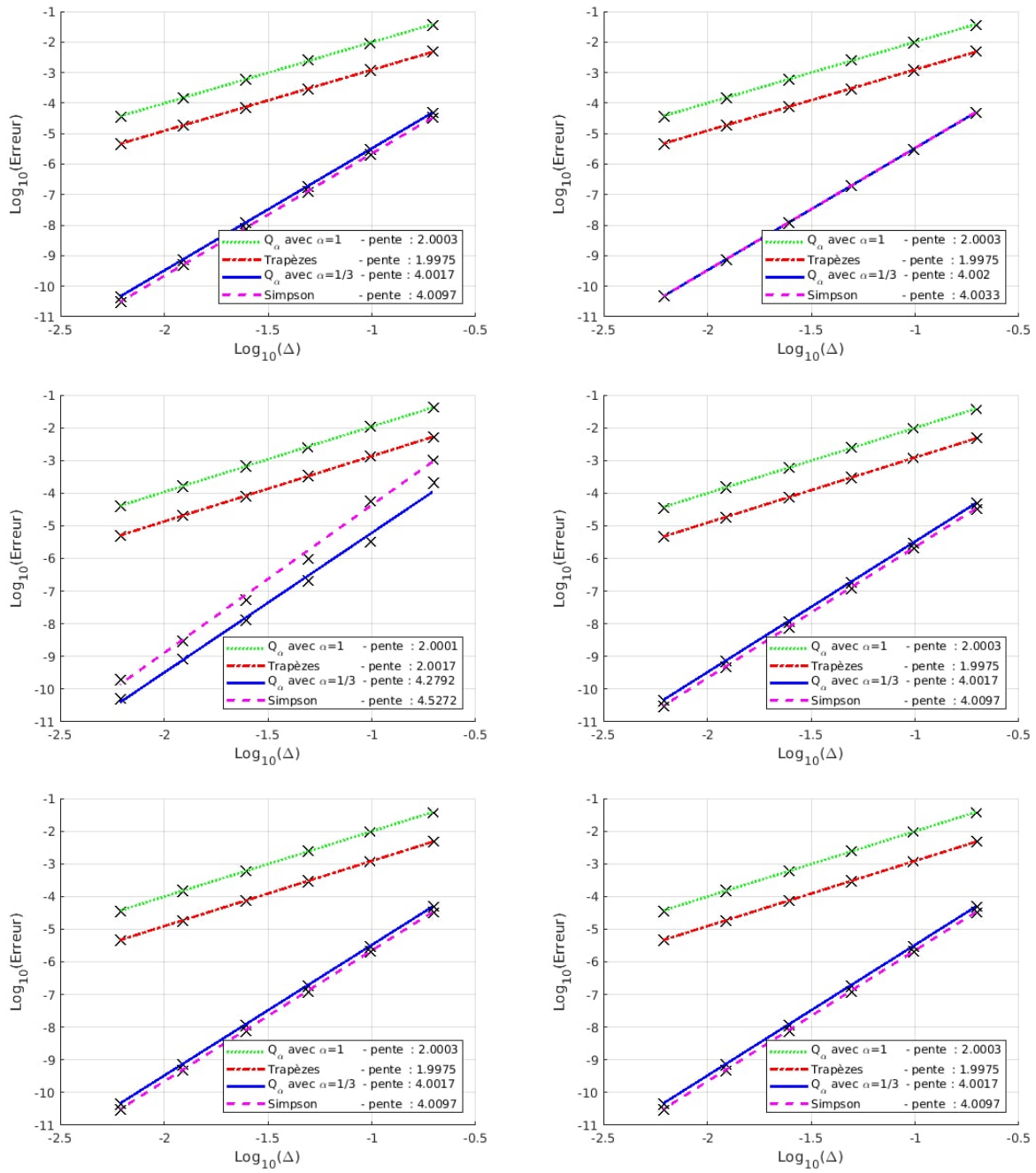


FIGURE 3.15 – Taux de convergence pour différentes méthodes de quadratures pour les fonctions tests $(f_p)_{0\leq p\leq 5}$. Nous retenons l'erreur maximale après 1000 rotations aléatoires pour chaque grille Cubed-Sphere de paramètre $N = 8$, $N = 16$, $N = 32$, $N = 64$, $N = 128$ et $N = 256$. De haut en bas et de gauche à droite, les fonctions sont f_0, f_1, f_2, f_3, f_4 et f_5. Le taux de convergence est proche de 4 pour Q_{sps} et $Q_{1/3}$, il est proche de 2 pour Q_{tpz} et Q_1.

Cette erreur est mesurée pour différentes grilles Cubed-Sphere de taille $N \times N \times 6$ avec $N = 8$, $N = 16$, $N = 32$, $N = 64$, $N = 128$ et $N = 256$. On peut ainsi mesurer l'ordre de convergence. Les résultats sont donnés dans la figure 3.15.

Les taux de convergence observées sont ceux attendus. Pour Q_{tpz} et Q_α avec $\alpha = 1$, le taux de convergence est proche de 2. Pour la formule de quadrature Q_{sps}, l'ordre de convergence est proche de 4. La formule de quadrature Q_α avec $\alpha = 1/3$ présente de meilleurs résultats que ceux attendus. On observe que la méthode est d'ordre 4 pour tous les tests effectués.

Nous avons étudié ici différents type de formules de quadrature sur la grille Cubed-Sphere qui sera utilisée pour la résolution des équations. Toutes les formules étudiées définissent un produit scalaire possédant de bonnes propriétés d'orthogonalité des harmoniques sphériques.

La formule Q_{tpz} est basée sur la formule des trapèzes composites, cette formule est consistante avec l'intégrale à l'ordre 2. La formule Q_{sps} est conçue à partir de la méthode de Simpson, elle est consistante à l'ordre 4 lorsque N est pair. Cette propriété de précision est fausse lorsque N est impair. Les formules de quadrature Q_α sont des perturbations de la formule Q_{tpz}. Le paramètre $\alpha > 0$ permet de pondérer les coins de la Cubed-Sphere. Quel que soit le choix de α, la formule Q_α est consistante avec l'intégrale au moins à l'ordre 2. Lorsque $\alpha = 1/3$, on prouve que la méthode est plus précise pour intégrer les fonctions constantes.

Lors des tests numériques, nous avons pu vérifier les ordres de convergence théoriques. De plus, la formule $Q_{1/3}$ semble particulièrement efficace puisqu'une convergence à l'ordre 4 est observée.

Dans la suite de ce travail, nous retenons la formule de quadrature Q_α avec $\alpha = 1/3$ pour étudier les propriétés de conservation des schémas numériques utilisés.

Chapitre 4

Approximation des opérateurs différentiels sur la Cubed-Sphere

4.1 Opérateurs différentiels sur la Cubed-sphere

4.1.1 Définition des opérateurs

Opérateur Gradient

Soit $\mathbf{x}_{i,j}^{(k)}$ un point de la Cubed-Sphere avec $-N/2 \leq i,j \leq N/2$ et $(k) = (I)\cdots(VI)$. Il existe deux grands cercles $C_i^{(1)}$ et $C_j^{(2)}$ tels que $\mathbf{x}_{i,j}^{(k)} \in C_i^{(1)} \cap C_j^{(2)}$. Les angles α et β paramètrent respectivement $C_i^{(1)}$ et $C_j^{(2)}$.

Le gradient en $\mathbf{x}_{i,j}^{(k)}$ de la fonction régulière $h : \mathbf{x} \in \mathbb{S}_a^2 \mapsto h(\mathbf{x})$ s'exprime par

$$\nabla_T h = \frac{\partial h}{\partial \alpha}_{|C_j^{(2)}} \mathbf{g}^\alpha + \frac{\partial h}{\partial \beta}_{|C_i^{(1)}} \mathbf{g}^\beta. \tag{4.1}$$

Le cercle $C_i^{(1)}$ (resp. $C_j^{(2)}$) est l'isoligne $\xi = \xi_i$ (resp $\eta = \eta_j$). D'après le théorème 3.4, le gradient s'exprime également par

$$\nabla_T h = \frac{\partial h}{\partial \xi}_{|\eta_j} \mathbf{g}^\xi + \frac{\partial h}{\partial \eta}_{|\xi_i} \mathbf{g}^\eta. \tag{4.2}$$

Donc si on sait calculer des approximations des dérivées partielles $\partial_\xi h$ et $\partial_\eta h$ le long des grands cercles, alors on dispose d'une approximation du gradient.

Opérateurs divergence et rotationnel

Soit $\mathbf{v} : \mathbf{x} \in \mathbb{S}_a^2 \mapsto \mathbf{v}(\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2$ un champ de vecteur tangent à la sphère. On définit la *divergence* et le *rotationnel* de $\mathbf{v}$, notés $\nabla_T \cdot \mathbf{v}$ et $\nabla_T \wedge \mathbf{v}$ par

Définition 4.1. *Soit $\mathbf{v} : \mathbf{x} \in \mathbb{S}_a^2 \mapsto \mathbf{v}(\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2$ un champ de vecteur régulier sur la sphère. Alors la divergence de $\mathbf{v}$ en $\mathbf{x} \in C_i^{(1)} \cap C_j^{(2)}$ est donnée par :*

$$\nabla_T \cdot \mathbf{v} = \frac{\partial \mathbf{v}}{\partial \alpha}_{|C_j^{(2)}} \cdot \mathbf{g}^\alpha + \frac{\partial \mathbf{v}}{\partial \beta}_{|C_i^{(1)}} \cdot \mathbf{g}^\beta. \tag{4.3}$$

La notation $\cdot$ désigne le produit scalaire usuel dans $\mathbb{R}^3$ dans le terme de droite.

Définition 4.2. *Soit $\mathbf{v} : \mathbf{x} \in \mathbb{S}_a^2 \mapsto \mathbf{v}(\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2$ un champ de vecteur régulier sur la sphère. Alors le rotationnel de $\mathbf{v}$ en $\mathbf{x} \in C_i^{(1)} \cap C_j^{(2)}$ est donné par*

$$\nabla_T \wedge \mathbf{v} = \mathbf{g}^\alpha \wedge \frac{\partial \mathbf{v}}{\partial \alpha}_{|C_j^{(2)}} + \mathbf{g}^\beta \wedge \frac{\partial \mathbf{v}}{\partial \beta}_{|C_i^{(1)}} \tag{4.4}$$

où $\wedge$ désigne le produit vectoriel dans le terme de droite.

Les opérateurs gradient, divergence, rotationnel donnés ici sont écrits en coordonnées (α, β) [75]. Elles sont intrinsèques à la sphère. Dans les équations de la climatologie, l'opérateur vorticité est utilisé.

Définition 4.3. *La vorticité du champ de vecteurs $\mathbf{v}$ est la composante normale du rotationnel :*

$$vort(\mathbf{v}) = (\nabla_T \wedge \mathbf{v}) \cdot \mathbf{n} \tag{4.5}$$

avec $\mathbf{n}$ le vecteur unitaire extérieur à la sphère en $\mathbf{x} \in \mathbb{S}_a^2$. Le vecteur $\mathbf{n}$ est

$$\mathbf{n}(\mathbf{x}) = \frac{1}{a}\mathbf{x}. \tag{4.6}$$

On déduit de la proposition 3.11 les expressions suivantes :

Théorème 4.1. *Soit $h : \mathbf{x} \in \mathbb{S}_a^2 \mapsto h(\mathbf{x})$ une fonction régulière et $\mathbf{v} : \mathbf{x} \in \mathbb{S}_a^2 \mapsto \mathbf{v}(\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2$ un champ de vecteurs régulier. Alors en $\mathbf{x}_{i,j}^k$, un point de la Cubed-Sphere, les égalités suivantes sont satisfaites :*

- ***Gradient** :*

$$\nabla_T h = \frac{\partial h}{\partial \xi}_{|\eta_j} \mathbf{g}^\xi + \frac{\partial h}{\partial \eta}_{|\xi_i} \mathbf{g}^\eta, \tag{4.7}$$

- ***Divergence** :*

$$\nabla_T \cdot \mathbf{v} = \frac{\partial \mathbf{v}}{\partial \xi}_{|\eta_j} \cdot \mathbf{g}^\xi + \frac{\partial \mathbf{v}}{\partial \eta}_{|\xi_i} \cdot \mathbf{g}^\eta, \tag{4.8}$$

- ***Vorticité** :*

$$vort(\mathbf{v}) = (\nabla_T \cdot \mathbf{v}) \cdot \mathbf{n} = \left(\mathbf{g}^\xi \wedge \frac{\partial \mathbf{v}}{\partial \xi}_{|\eta_j} + \mathbf{g}^\eta \wedge \frac{\partial \mathbf{v}}{\partial \eta}_{|\xi_i} \right) \cdot \mathbf{n}. \tag{4.9}$$

Pour calculer une valeur approchée des opérateurs gradient, divergence et vorticité aux points du maillage de la Cubed-Sphere, il faut calculer une valeur approchée de la dérivée d'une fonction le long d'un grand cercle. C'est à dire, il faut évaluer $\tilde{\delta}_\xi^H f_{i,j}^{(k)}$ et $\tilde{\delta}_\eta^H f_{i,j}^{(k)}$ tels que

$$\begin{cases} \tilde{\delta}_\xi^H f_{i,j}^{(k)} \to \partial_\xi f(\mathbf{x}_{i,j}^k) & \text{lorsque } \Delta\xi \to 0 \\ \tilde{\delta}_\eta^H f_{i,j}^{(k)} \to \partial_\eta f(\mathbf{x}_{i,j}^k) & \text{lorsque } \Delta\eta \to 0 \end{cases} \tag{4.10}$$

Dans la section suivante, on détaille une procédure numérique pour calculer ces dérivées partielles approchées. De plus, on détermine l'erreur de troncature associée.

4.1.2 Approximation de dérivées sur les grands cercles

Soit $f : \mathbf{x} \in \mathbb{S}_a^2 \mapsto f(\mathbf{x})$ la fonction que l'on souhaite dériver le long des grands cercles aux points du maillage de la Cubed-Sphere. Si $\mathbf{x}_{i,j}^{(k)}$ est un point de la Cubed-Sphere avec $(k) = (I) \cdots (VI)$, ainsi que $-N/2 \leq i, j \leq N/2$, on souhaite calculer une valeur approchée de $\partial_\xi f(\mathbf{x}_{i,j}^{(k)})$ et $\partial_\eta f(\mathbf{x}_{i,j}^{(k)})$. On suppose $(k) = (I)$, la méthode étant identique sur les autres panels. Il existe deux grands cercles de la Cubed-Sphere $C_i^{(1)} \in (I_\beta)$ et $C_j^{(2)} \in (I_\alpha)$ tels que

$$\mathbf{x}_{i,j}^{(k)} \in C_i^{(1)} \cap C_j^{(2)}. \tag{4.11}$$

$C_i^{(1)}$ est une isoligne en $\xi = \xi_i$ et $C_j^{(2)}$ est une isoligne en $\eta = \eta_j$. Pour calculer une valeur approchée de $\partial_\xi f(\mathbf{x}_{i,j}^{(I)})$, on souhaite connaître des données de f en un ensemble de points le long du cercle $C_j^{(2)}$. On note $\mathbf{m}_p$ avec $0 \leq p \leq 4N-1$ les points de $C_j^{(2)}$ construits de la manière suivante :

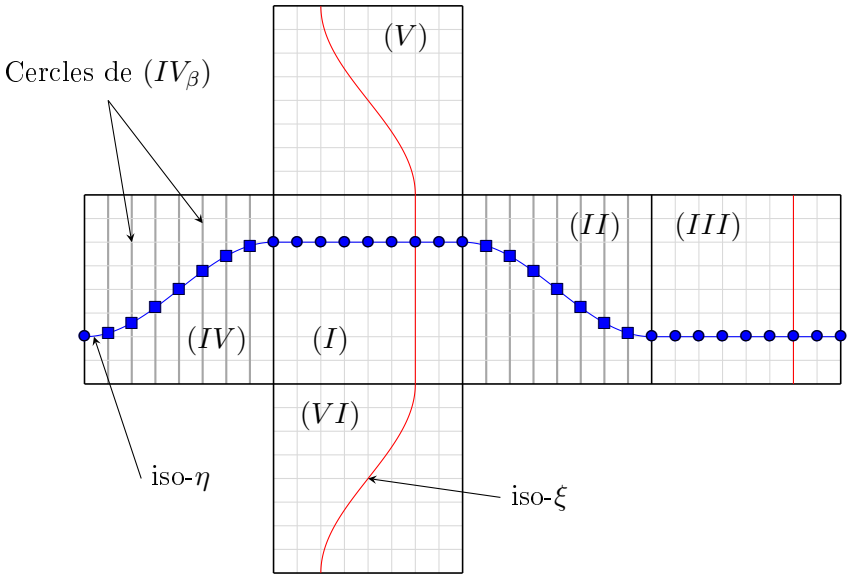


FIGURE 4.1 – Grands cercles $C_j^{(2)}$ (horizontaux) et $C_i^{(1)}$ verticaux. Ces grands cercles sont associés aux panels (I) et (III). Ces cercles ne passent pas par des points du la Cubed-Sphere sur les panels (II) et (IV).

- si $0 \leq p \leq N$ alors $\mathbf{m}_p = \mathbf{x}^{(I)}_{p-N/2,j}$, il s'agit des points du cercle $C_j^{(2)}$ associés au panel (I) sur le maillage. Ils sont représentés par des ronds bleus sur les Figures 4.1 et 4.2,

- si $N+1 \leq p \leq 2N-1$ alors les points ne font pas partie du maillage. Il s'agit des points d'intersections de $C_j^{(2)}$ avec les isolignes $\xi = \xi_i^{(II)}$ du panel (II), c'est à dire l'intersection de $C_j^{(2)}$ avec les cercles de (II_β). Ces points sont représentés par des carrés bleus dans les Figures 4.1 et 4.2. Les coordonnées sont calculées dans (3.107),

- si $2N \leq p \leq 3N$ alors $\mathbf{m}_p = \mathbf{x}^{(III)}_{p-3N/2}$. Il s'agit des points de la Cubed-Sphere du panel (III) représentés par des ronds bleus sur la Figure 4.1,

- si $3N+1 \leq p \leq 4N-1$ alors les points $\mathbf{m}_p$ sont les points d'intersection de $C_j^{(2)}$ avec les cercles de (IV_β). Il ne s'agit pas de points de la grille. Ces points sont représentés par des carrés bleus dans la Figure 4.1.

La périodicité sur les grands cercles permet d'assurer que $\mathbf{m}_p = \mathbf{m}_{p+4N}$ pour tout $p \in \mathbb{Z}$. De plus, la construction de la Cubed-Sphere permet d'assurer un paramétrage du grand cercle $C_j^{(2)}$. Chaque point est associé à des coordonnées (ξ_p, η_p). Les valeurs de ξ_p donnent un paramétrage des points $\mathbf{m}_p$ le long du grand cercle. On a de plus

$$\xi_{p+1} = \xi_p + \Delta\xi. \tag{4.12}$$

Supposant les valeurs $f_p = f(\mathbf{m}_p)$ connues pour tout p vérifiant $0 \leq p \leq 4N-1$ alors on peut évaluer la dérivée approchée $\tilde{\delta}^H_\xi f^{(k)}_{i,j}$ de $\partial_\xi f(\mathbf{x}^{(k)}_{i,j})$ à l'ordre 4 grâce à la formule de dérivation hermitienne (1.127)

$$\tilde{\delta}^H_\xi f^{(k)}_{i,j} = \delta^H_{4,\xi} f_p. \tag{4.13}$$

Cependant, les valeurs de f_p ne sont pas toutes connues car les points $\mathbf{m}_p$ ne sont pas des points du maillage lorsque $N+1 \leq p \leq 2N-1$ ou $3N+4 \leq p \leq 4N-1$. Il faut donc construire un procédé permettant d'obtenir des valeurs en ces points. Le procédé utilisé dans ce travail est basé sur une interpolation à l'aide de Splines Cubiques [2].

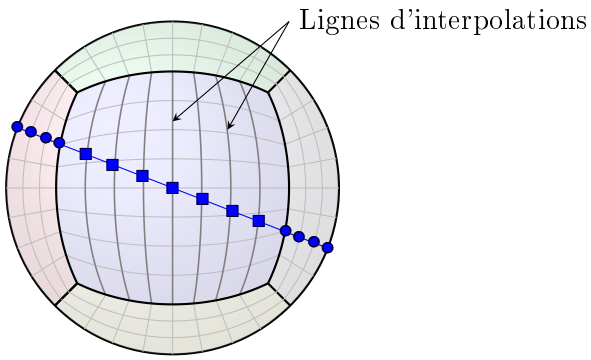


FIGURE 4.2 – Un grand cercle $C_j^{(2)}$. La ligne bleue représente une isoligne $\eta = \eta_j$ du panel (I) vue depuis le panel (II). Les cercles bleus représentent des points de la Cubed-Sphere et de l'isoligne $\eta = \eta_j$. Les carrés bleus ne sont pas sur la Cubed-Sphere. Il s'agit de l'isoligne $\eta = \eta_j$ et de portions de grands cercles du panel (II).

On souhaite évaluer une valeur f_p approchant $f(\mathbf{m}_p)$ avec $N+1 \leq p \leq 2N-1$ ou $3N+4 \leq p \leq 4N-1$. Bien que $\mathbf{m}_p$ ne soit pas un point de la Cubed-Sphere, on sait que

$$\mathbf{m}_p \in C_j^{(2)} \cap C \tag{4.14}$$

où $C_j^{(2)}$ est une isoligne pour $\eta = \eta_j$ constant pour le panel (I) et C est une isoligne ξ constant pour le panel (II) (si $N+1 \leq p \leq 2N-1$) ou pour le panel (IV) si $3N+4 \leq p \leq 4N-1$ (lignes en gras sur les Figures 4.1.et 4.2). La méthode consiste à utiliser les points de la Cubed-Sphere présents sur le cercle C pour construire une fonction d'interpolation de type Spline Cubique puis d'évaluer cette fonction au point du maillage $\mathbf{m}_p$. Si on note P_C la fonction d'interpolation en question, on a alors $f_p = P_C(\mathbf{m}_p)$. L'interpolation s'effectuant sur un grand cercle, il s'agit d'une fonction d'interpolation en dimension 1. On utilise l'interpolation par splines cubiques. On a alors [2] :

$$f_p = P_C(\mathbf{m}_p) = f(\mathbf{m}_p) + \mathcal{O}(\Delta\eta^4). \tag{4.15}$$

La fonction P_C ne dépend pas du point $\mathbf{m}_p$ mais uniquement des données aux points de la Cubed-Sphere sur le panel choisi et le long du cercle C (voir Fig. 4.3).

Le procédé est symétrique pour reconstruire les données sur un grand cercle $C_i^{(1)}$ ou le grand cercle d'un autre panel. Une fois les données $(f_p)_{0\leq p\leq 4N-1}$ construites, on peut calculer l'approximation de la dérivée grâce à $\partial_\xi f_p \approx \delta^H_{\Delta\xi} f_p$ que l'on restreint aux points du maillage. Pour tout $-N/2 \leq i,j \leq N/2$, on a

$$\left\{ \begin{array}{rcl} \tilde{\delta}^H_\xi f^{(I)}_{i,j} & = & \delta^H_{4,\xi} f_{i-N/2} \\ \tilde{\delta}^H_\eta f^{(III)}_{i,j} & = & \delta^H_{4,\xi} f_{i+3N/2} \end{array} \right. \quad \text{pour } C_j^{(2)} \text{ fixé.} \tag{4.16}$$

L'algorithme de calcul des dérivées hermitiennes en ξ sur un panel (k) est donné par l'algorithme 7.

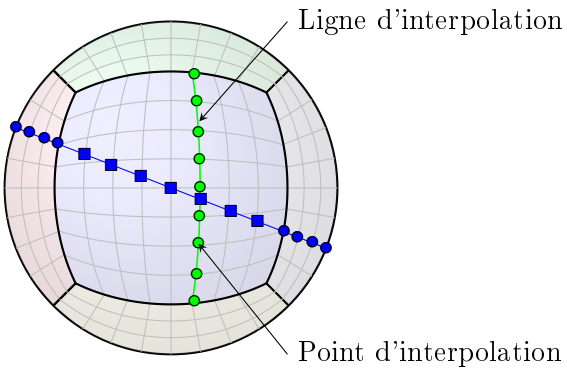


FIGURE 4.3 – Grand cercle $C_j^{(2)}$ et portions de grands cercles du panel (II). La ligne bleue représente une isoligne $\eta = \eta_j$ du panel (I) vue depuis le panel (II). Les cercles bleus représentent des points de la Cubed-Sphere contenus dans l'isoligne $\eta = \eta_j$. Les carrés bleus sont des points de l'isoligne $\eta = \eta_j$ qui ne sont pas sur la Cubed-Sphere. En vert, une section du grand cercle utilisée pour l'interpolation spline cubique.

Algorithme 7 : Calcul de $\tilde{\delta}_\xi^H f_{i,j}^{(I)}$ et $\tilde{\delta}_\xi^H f_{i,j}^{(III)}$

for $j = -N/2, \ldots, N/2,$ **do**
  pour un grand cercle $C_j^{(2)}$ fixé,
  for $p = 0, 1, \ldots 4N-1$ définir les points $\mathbf{m}_p \in C_j^{(2)}$ du grand cercle bleu **do**
    $\mathbf{m}_p = \mathbf{m}_{p-N/2,j}^{(I)}$ pour $0 \leq p \leq N$ donc $f_p = f(\mathbf{m}_p)$,
    $f_p = P_{C_p}(\mathbf{m}_p)$, où P_{C_p} est la fonction d'interpolation utilisant les points de l'isoligne $\xi = \xi_{p-N/2+1}^{(II)}$ pour $N+1 \leq p \leq 2N-1$,
    $\mathbf{m}_p = \mathbf{m}_{p-3N/2,j}^{(III)}$ pour $2N \leq p \leq 3N-1$ donc $f_p = f(\mathbf{m}_p)$,
    $f_p = P_{C_p}(\mathbf{m}_p)$, où P_{C_p} est la fonction d'interpolation utilisant les points de l'isoligne $\xi = \xi_{p-3N/2+1}^{(VI)}$ pour $3N+1 \leq p \leq 4N-1$.
  end for
  Calcul de $\delta_{4,\xi}^H(f_p)_p$,
  Affectation $\tilde{\delta}_\xi^H f_{i,j}^{(I)} = \delta_{4,\xi}^H f_{i+N/2}$ pour tout $-N/2 \leq i \leq N/2$,
  Affectation $\tilde{\delta}_\xi^H f_{i,j}^{(III)} = \delta_{4,\xi}^H f_{i+3N/2}$ pour tout $-N/2 \leq i \leq N/2$.
end for

De la même manière, l'algorithme de construction des dérivées approchées en η sur les panels (I) et (III) est donné par l'algorithme 8.

Algorithme 8 : Calcul de $\tilde{\delta}_\eta^H f_{i,j}^{(I)}$ et $\tilde{\delta}_\eta^H f_{i,j}^{(III)}$

for $i = -N/2, \ldots, N/2,$ **do**
  pour un grand cercle $C_i^{(1)}$ fixé,
  for $p = 0, 1, \ldots 4N-1$ définir les points $\mathbf{m}_p \in C_i^{(1)}$ **do**
    $\mathbf{m}_p = \mathbf{m}_{i,p-N/2}^{(I)}$ pour $0 \leq p \leq N$ donc $f_p = f(\mathbf{m}_p)$,
    $f_p = P_{C_p}(\mathbf{m}_p)$, où P_{C_p} est la fonction d'interpolation utilisant les points de l'isoligne $\eta = \eta_{p-N/2+1}^{(V)}$ pour $N+1 \leq p \leq 2N-1$,
    $\mathbf{m}_p = \mathbf{m}_{i,5N/2-p}^{(III)}$ pour $2N \leq p \leq 3N-1$ donc $f_p = f(\mathbf{m}_p)$ (Attention à l'orientation sur les panels),
    $f_p = P_{C_p}(\mathbf{m}_p)$, où P_{C_p} est la fonction d'interpolation utilisant les points de l'isoligne $\eta = \eta_{p-3N/2+1}^{(VI)}$ pour $3N+1 \leq p \leq 4N-1$.
  end for
  Calcul de $\delta_{4,\eta}^H f_p$,
  Affectation $\tilde{\delta}_\eta^H f_{i,j}^{(I)} = \delta_{4,\eta}^H f_{i+N/2}$ pour tout $-N/2 \leq j \leq N/2$,
  Affectation $\tilde{\delta}_\eta^H f_{i,j}^{(III)} = \delta_{4,\eta}^H f_{i+3N/2}$ pour tout $-N/2 \leq j \leq N/2$.
end for

Le processus est utilisé sur chaque panel $(k) = (I), \ldots, (VI)$. On obtient alors les approximations des dérivées en ξ et en η sur chaque point du maillage Cubed-Sphere $\mathbf{x}_{i,j}^{(k)}$ avec $-N/2 \leq i, j \leq N/2$.

Théorème 4.2. *Pour tous $-N/2 \leq i, j \leq N/2$ et $(k) = (I), \ldots, (VI)$ et pour $f : \mathbf{x} \in \mathbb{S}_a^2 \mapsto f(\mathbf{x}) \in \mathbb{R}$ une fonction régulière, on a*

$$\left\{ \begin{array}{rcl} \tilde{\delta}_\xi^H f_{i,j}^{(k)} & = & \partial_\xi f(\mathbf{x}_{i,j}^{(k)}) + \mathcal{O}(\Delta\eta^3) \\ \tilde{\delta}_\eta^H f_{i,j}^{(k)} & = & \partial_\eta f(\mathbf{x}_{i,j}^{(k)}) + \mathcal{O}(\Delta\xi^3) \end{array} \right. \tag{4.17}$$

Démonstration. La preuve est la même sur chaque panel, on se restreint ici au panel (I). De plus, par symétrie, nous ne montrons que le premier résultat :

$$\tilde{\delta}_\xi^H f_{i,j}^{(k)} = \partial_\xi f(\mathbf{x}_{i,j}^{(k)}) + \mathcal{O}(\Delta\eta^3). \tag{4.18}$$

Le procédé de construction des valeurs sur un grand cercle $C_j^{(2)}$, aux points $\mathbf{m}_p$ avec $0 \leq p \leq 4N-1$ donne

$$f_p = f(\mathbf{m}_p) + \mathcal{O}(\Delta\eta^4) \tag{4.19}$$

car nous utilisons une interpolation Spline cubique.

Donc dans le calcul de la dérivée Hermitienne, on a

$$\delta_\xi f_p = \frac{f_{p+1} - f_{p-1}}{2\Delta\xi} = \frac{f(\mathbf{m}_{p+1}) - f(\mathbf{m}_{p-1})}{2\Delta\xi} + \mathcal{O}\left(\frac{\Delta\eta^4}{\Delta\xi}\right). \tag{4.20}$$

On a $\Delta\xi = \Delta\eta = \pi/(2N)$. Alors en composant par σ_ξ^{-1}, on a

$$\begin{aligned} \delta_{4,\xi}^H f_p &= \sigma_\xi^{-1} \circ \delta_{2,\xi} f_p \\ &= \sigma_\xi^{-1} \circ \left(\frac{f(\mathbf{m}_{p+1}) - f(\mathbf{m}_{p-1})}{2\Delta\xi} + \mathcal{O}\left(\Delta\eta^3\right) \right) \\ &= \sigma_\xi^{-1} \circ \left(\frac{f(\mathbf{m}_{p+1}) - f(\mathbf{m}_{p-1})}{2\Delta\xi} \right) + \mathcal{O}\left(\Delta\eta^3\right) \\ &= \partial_\xi f(\mathbf{m}_p) + \mathcal{O}\left(\Delta\eta^3\right) + \mathcal{O}\left(\Delta\xi^4\right) \\ &= \partial_\xi f(\mathbf{m}_p) + \mathcal{O}\left(\Delta\xi^3\right). \end{aligned}$$

On assigne les dérivées aux points des panels à l'aide de $\mathbf{m}_p = \mathbf{x}^{(k)}_{p-N/2,j}$ pour $p = 0 \dots N$. Le résultat est alors :

$$\tilde{\delta}^H_\xi f^{(k)}_{i,j} = \delta^H_{4,\xi} f_{i+N/2} = \partial_\xi f(\mathbf{x}^{(k)}_{i+N/2,j}) + \mathcal{O}\left(\Delta\xi^3\right). \tag{4.21}$$

Le résultat concernant la dérivée en η est obtenu de la même manière. □

D'après le théorème 4.2, la méthode de calcul est d'ordre au moins 3. Cependant, on note que l'interpolation (cause d'une possible perte de précision) n'intervient qu'en dehors des panels où l'on souhaite calculer la dérivée. Lorsque l'on souhaite calculer une approximation de $\partial_\xi f(\mathbf{x}^{(I)}_{i,j})$, l'interpolation intervient sur les panels (II) et (IV) mais pas sur le panel (I). Dans la pratique, on s'attend à ce que la méthode soit d'ordre 4, en particulier loin des bords des panels et c'est bien ce qui est observé.

4.2 Opérateur gradient discret

4.2.1 Construction et consistance de l'opérateur gradient discret

Soit $h : \mathbf{x} \in \mathbb{S}^2_a \mapsto h(\mathbf{x}) \in \mathbb{R}$ une fonction régulière sur la sphère. On note $h^{(k)}_{i,j} = h(\mathbf{x}^{(k)}_{i,j})$, avec $-N/2 \leq i,j \leq N/2$ et $(k) = (I), \dots, (VI)$, la valeur de h au point $\mathbf{x}^{(k)}_{i,j}$ de la Cubed-Sphere. Nous notons $(\mathbf{g}^\xi)^{(k)}_{i,j} = \mathbf{g}^\xi(\mathbf{x}^{(k)}_{i,j})$ et $(\mathbf{g}^\eta)^{(k)}_{i,j} = \mathbf{g}^\eta(\mathbf{x}^{(k)}_{i,j})$.

Définition 4.4. *On définit l'opérateur gradient discret par*

$$\nabla_{T,\Delta} h^{(k)}_{i,j} = \tilde{\delta}^H_\xi h^{(k)}_{i,j} (\mathbf{g}^\xi)^{(k)}_{i,j} + \tilde{\delta}^H_\eta h^{(k)}_{i,j} (\mathbf{g}^\eta)^{(k)}_{i,j} \tag{4.22}$$

avec $-N/2 \leq i,j \leq N/2$ et $(k) = (I), \dots, (VI)$ ainsi que $\tilde{\delta}^H_\xi h^{(k)}_{i,j}$ et $\tilde{\delta}^H_\eta h^{(k)}_{i,j}$ obtenus grâce aux algorithmes 7 et 8.

L'opérateur gradient discret de h, $\nabla_{T,\Delta} h^{(k)}_{i,j}$ donne une approximation du gradient en $\mathbf{x}^{(k)}_{i,j}$. En effet, le résultat de consistance suivant est vérifié :

Proposition 4.1. *Soit $h : \mathbf{x} \in \mathbb{S}^2_a \mapsto h(\mathbf{x}) \in \mathbb{R}$ une fonction régulière sur la sphère. Pour tous $-N/2 \leq i,j \leq N/2$ et $(k) = (I), \dots, (VI)$, on a*

$$\nabla_{T,\Delta} h^{(k)}_{i,j} - (\nabla_T h)^{*,(k)}_{i,j} = \mathcal{O}\left(\Delta\xi^3\right) \tag{4.23}$$

où * désigne l'opérateur de restriction à la Cubed-Sphere et $\Delta\xi = \Delta\eta$.

Démonstration. Ce résultat est une conséquence immédiate de la linéarité du gradient discret et du théorème 4.2. □

De plus, on note que par construction

$$\nabla_{T,\Delta} h^{(k)}_{i,j} \in \mathbb{T}_{\mathbf{x}^{(k)}_{i,j}} \mathbb{S}^2_a. \tag{4.24}$$

Une valeur du gradient est associée à chaque point de chaque panel. Lorsque le point appartient à plusieurs panels, on adopte la moyenne des différentes valeurs de gradient. Si le point est à l'intersection d'exactement deux panels, il s'agit d'une demi-somme. Si le point est un coin de la Cubed-Sphere, il s'agit d'un tiers de somme.

4.2.2 Tests numériques

Nous testons numériquement l'algorithme de gradient présenté ici pour évaluer son comportement. On se donne une fonction $h : \mathbf{x} \in \mathbb{S}_a^2 \mapsto h(\mathbf{x})$ tel que le gradient $\nabla_T h$ est connu. Nous comparons le gradient approché et le gradient exact. Si h est une fonction sphérique de carré intégrable, alors h se décompose comme une somme d'harmoniques sphériques. De plus, les harmoniques sphériques sont des restrictions de polynômes sur la Sphère $\mathbb{S}_a^2$. Un test pertinent est donc de choisir h de type monomiale :

$$\left\{ \begin{array}{rcl} \hat{h}(x,y,z) & = & x^p y^q z^r, \\ h & = & \hat{h}_{|\mathbb{S}_a^2}. \end{array} \right. \quad \text{avec } p,q,r \in \mathbb{N}. \tag{4.25}$$

En se basant sur la proposition 3.3, le gradient de h est obtenu grâce à la relation

$$\nabla_T h = \nabla_{\mathbb{R}^3} \hat{h} - \mathbf{n} \left(\mathbf{n} \cdot \nabla_{\mathbb{R}^3} \hat{h} \right) \tag{4.26}$$

avec $\mathbf{n}(\mathbf{x}) = \mathbf{x}/a$ la normale extérieure en $\mathbf{x} \in \mathbb{S}_a^2$. Le gradient $\nabla_{\mathbb{R}^3} \hat{h}$ est donné dans la base $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ par

$$\begin{aligned} \nabla_{\mathbb{R}^3} \hat{h} &= \frac{\partial \hat{h}}{\partial x}\mathbf{i} + \frac{\partial \hat{h}}{\partial y}\mathbf{j} + \frac{\partial \hat{h}}{\partial z}\mathbf{k} \\ &= p x^{p-1} y^q z^r \mathbf{i} + q x^p y^{q-1} z^r \mathbf{j} + r x^p y^q z^{r-1} \mathbf{k}. \end{aligned}$$

Si $\mathbf{u}$ est une fonction vectorielle définie sur la Cubed-Sphere, alors

$$\mathbf{u}_{i,j}^{(k)} = u_{i,j}^{(k)} \mathbf{i} + v_{i,j}^{(k)} \mathbf{j} + w_{i,j}^{(k)} \mathbf{k} \tag{4.27}$$

On définit la norme $\mathcal{N}$ par

$$\mathcal{N}(\mathbf{u}) = \max_{-N/2 \leq i,j \leq N/2} \max_{(k)=(I)\ldots(VI)} \max(|u_{i,j}^{(k)}|, |v_{i,j}^{(k)}|, |w_{i,j}^{(k)}|). \tag{4.28}$$

On mesure l'erreur faite sur le calcul du gradient approché par

$$e_\Delta = \frac{\mathcal{N}\left(\nabla_{T,\Delta}(h^*) - (\nabla_T h)^*\right)}{\mathcal{N}\left((\nabla_T h)^*\right)}. \tag{4.29}$$

Sur la Figure 4.4, on trace le logarithme décimal de l'erreur en fonction du logarithme décimal de $\frac{\pi a}{2N} = a\Delta\xi = a\Delta\eta$. La pente de cette courbe donne l'ordre de convergence de la méthode sur le test effectué. On observe que les points sont parfaitement alignés. De plus les niveaux d'erreurs observés sont très bons, même sur les grilles les plus grossières avec $N = 32$ et $N = 64$.

Les ordres estimés sont meilleurs que l'ordre 3 montré dans la proposition 4.1. Lorsque $(p,q,r) = (1,2,3)$, l'ordre de convergence calculé est 3.8787, il est de 3.8868 lorsque $(p,q,r) = (2,2,2)$. Cela confirme l'ordre démontré et nous donne un ordre proche de 4 dans la pratique. Les erreurs effectives sont données en fonction de N dans la Table 4.1.

4.2.3 Une variante de l'opérateur de gradient discret

Gradient discret utilisant un schéma compact d'ordre 8

Une variante pour le calcul du gradient approché est d'utiliser un schéma aux différences finies en dimension 1 d'ordre plus élevé que $\delta_{4,x}^H$ (qui est d'ordre 4). L'opérateur centré hermitien $\delta_{8,x}^H$ (1.130) est un opérateur d'approximation de la dérivée première d'ordre 8. On remplace $\delta_{4,\xi}^H$ et $\delta_{4,\eta}^H$ dans les algorithmes 7 et 8 de calcul de dérivées sur les grands cercles par $\delta_{8,\xi}^H$ et $\delta_{8,\eta}^H$, le reste étant inchangé. Le nouveau gradient approché vérifie toujours la proposition 4.1, le facteur limitant la montée en ordre étant, comme précédemment, la fonction d'interpolation de type spline.

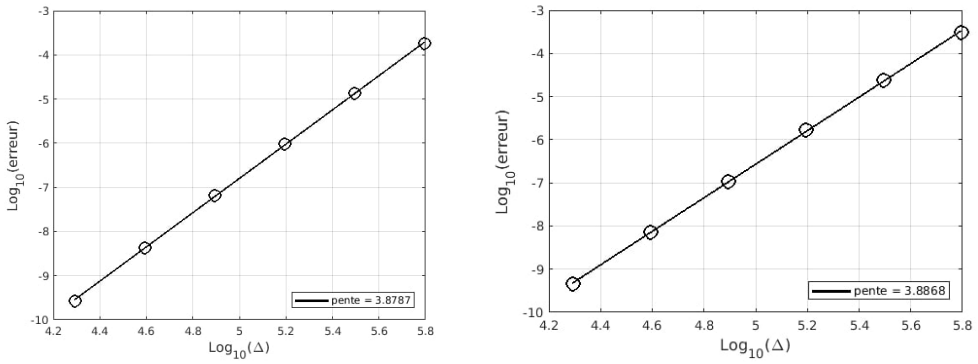


FIGURE 4.4 – Erreur sur le gradient en fonction de $\Delta = a\Delta\xi$. Convergence du gradient de (4.25) avec $(p, q, r) = (1, 2, 3)$ à gauche et avec $(p, q, r) = (2, 2, 2)$ à droite.

N	$(\mathbf{1}, \mathbf{2}, \mathbf{3})$	$(\mathbf{2}, \mathbf{2}, \mathbf{2})$
16	$1.8636(-4)$	$3.1368(-4)$
32	$1.3545(-5)$	$2.3910(-5)$
64	$9.7564(-7)$	$1.6716(-6)$
128	$6.5592(-8)$	$1.1088(-7)$
256	$4.2563(-9)$	$7.1437(-9)$
511	$2.7115(-10)$	$4.5361(-10)$
ordre estimé	3.88	3.89

TABLE 4.1 – Table des erreurs en fonction du paramètre de maillage N pour le calcul du gradient. On mesure l'erreur relative pour le calcul du gradient de (4.25) avec $(p, q, r) = (1, 2, 3)$ à gauche et avec $(p, q, r) = (2, 2, 2)$ à droite.

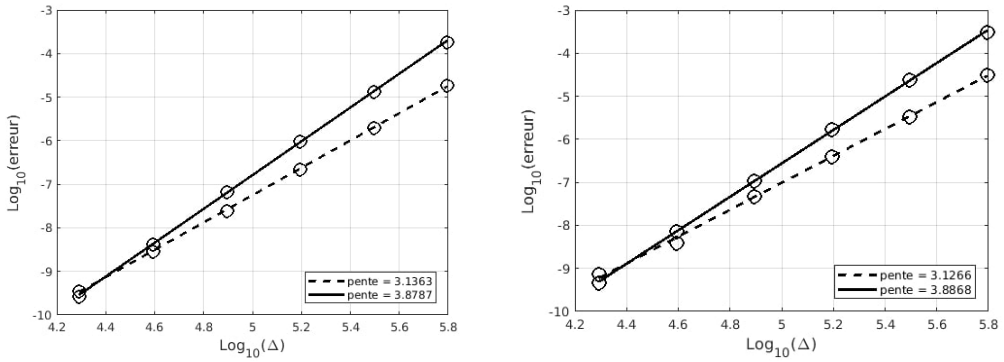


FIGURE 4.5 – Erreur sur le calcul du gradient en fonction de $\Delta = a\Delta\xi$. La convergence du gradient de (4.25) avec $(p, q, r) = (1, 2, 3)$ est à gauche et avec $(p, q, r) = (2, 2, 2)$ est à droite. Nous comparons l'utilisation de $\delta^H_{4,x}$ en trait plein avec l'utilisation de $\delta^H_{8,x}$ en pointillés. L'ordre de convergence du gradient utilisant $\delta^H_{8,x}$ est plus faible que celui utilisant $\delta^H_{4,x}$.

	$(\mathbf{p},\mathbf{q},\mathbf{r})=(\mathbf{1},\mathbf{2},\mathbf{3})$		$(\mathbf{p},\mathbf{q},\mathbf{r})=(\mathbf{2},\mathbf{2},\mathbf{2})$	
N	$\delta^H_{4,x}$	$\delta^H_{8,x}$	$\delta^H_{4,x}$	$\delta^H_{8,x}$
16	1.8636(−4)	1.8475(−5)	3.1368(−4)	3.1349(−5)
32	1.3545(−5)	2.0130(−6)	2.3910(−5)	3.4020(−6)
64	9.7564(−7)	2.1919(−7)	1.6716(−6)	3.9216(−7)
128	6.5592(−8)	2.4346(−8)	1.1088(−7)	4.7945(−8)
256	4.2563(−9)	2.9159(−9)	7.1437(−9)	3.8473(−9)
511	2.7115(−10)	3.5643(−10)	4.5361(−10)	7.2161(−10)
ordre estimé	3.88	3.14	3.89	3.13

TABLE 4.2 – Table de convergence du gradient approché de (4.25) utilisant $\delta^H_{8,x}$ ou $\delta^H_{4,x}$. L'ordre de convergence du gradient utilisant $\delta^H_{8,x}$ est plus faible que celui utilisant $\delta^H_{4,x}$. Pour un maillage plus grossier ($N=32$ ou $N=64$), l'erreur est plus faible pour le gradient utilisant $\delta^H_{8,x}$.

Sur la Figure 4.5, nous présentons les courbes de convergence comparant le schéma utilisant $\delta^H_{4,x}$ et $\delta^H_{8,x}$. On constate que le premier schéma est meilleur en ordre que le schéma utilisant $\delta^H_{8,x}$. En revanche, l'erreur pour des petites valeurs de N est plus faible pour le schéma utilisant $\delta^H_{8,x}$ comme on le voit sur la Table 4.2.

On souhaite connaître la localisation spatiale sur la sphère de l'erreur de manière à connaître les zones dans lesquelles le gradient approché est le moins performant. Pour cela, nous représentons sur la Figure 4.6, l'erreur spatiale

$$\mathbf{x}^{(k)}_{i,j} \mapsto \frac{\max_{\mathbf{u}=\mathbf{i},\mathbf{j},\mathbf{k}} \left| \left(\nabla_{T,\mathrm{app}} h^{(k)}_{i,j} - \nabla_T h^{(k)}_{i,j} \right) \cdot \mathbf{u} \right|}{\max_{\mathbf{u}=\mathbf{i},\mathbf{j},\mathbf{k}} \left(\left| \nabla_T h^{(k)}_{i,j} \cdot \mathbf{u} \right| \right)} \tag{4.30}$$

où $\nabla_{T,\mathrm{app}} h^{(k)}_{i,j}$ désigne le gradient approché calculé en utilisant $\delta^H_{8,x}$ la dérivée Hermitienne d'ordre 8 et $\mathbf{x}^{(k)}_{i,j}$ est un point de la Cubed-Sphere, $(k)=(I),\ldots,(VI)$ et $-N/2 \leq i,j \leq N/2$. L'erreur est

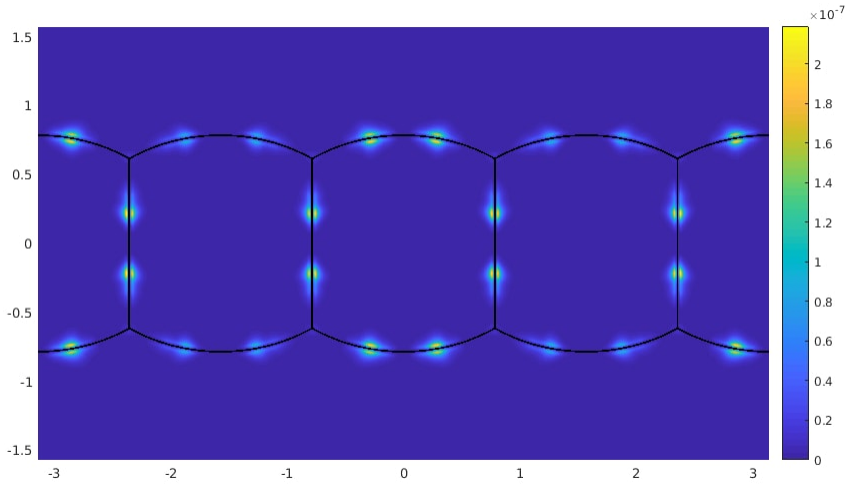


FIGURE 4.6 – Erreur relative pour le calcul du gradient $\nabla_{T,\mathrm{app}} h^{(k)}_{i,j}$ avec $N=63$ et $h(x,y,z)=xy^2z^3$.

plus importante là où l'interpolation est la plus mauvaise. Sur les grands cercles associés aux isolignes $\xi=\pm\frac{\pi}{4}$ ou $\eta=\pm\frac{\pi}{4}$ ou $\xi=0$ ou $\eta=0$, il n'y a pas d'interpolation car ces grands cercles passent

exactement à des points de grilles sur tous les panels. Il s'agit des grands cercles passant sur les bords des panels ainsi que les grands cercles équatoriaux. L'erreur introduite par l'interpolation est particulièrement visible loin de ces cercles. Elle reste cependant limitée.

Gradient sans interpolation

Nous avons vu que la méthode de calcul du gradient approché repose sur le calcul de dérivées hermitiennes périodiques le long de grands cercles. L'utilisation de splines cubiques empêche de passer à un ordre supérieur à 3 en changeant l'ordre de la dérivée Hermitienne. Il est légitime de se demander si un schéma décentré au bord des panels ne donne pas des résultats comparables de manière à n'utiliser que les points intérieurs du panel (k) pour calculer le gradient en $\mathbf{x}_{i,j}^{(k)}$ avec $(k) = (I), \ldots, (VI)$ et $-N/2 \leq i, j \leq N/2$. Un tel schéma est naturellement beaucoup plus simple que la méthode utilisant les splines cubiques présentée précédemment.

Soit $(\mathfrak{u}_p)_{-N/2 \leq p \leq N/2}$ une fonction de grille. On définit l'opérateur explicite $\delta_{\text{déc},x}$ par

$$\left\{\begin{array}{lcll} \delta_{\text{déc},x}\mathfrak{u}_p & = & \dfrac{1}{2h}(\mathfrak{u}_{p+1} - \mathfrak{u}_{p-1}) & \text{si } -N/2+1 \leq p \leq N/2-1 \\ \delta_{\text{déc},x}\mathfrak{u}_p & = & \dfrac{1}{h}\left(-\dfrac{103}{72}\mathfrak{u}_p + \dfrac{91}{36}\mathfrak{u}_{p+1} - \dfrac{7}{4}\mathfrak{u}_{p+2} + \dfrac{29}{36}\mathfrak{u}_{p+3} - \dfrac{11}{72}\mathfrak{u}_{p+4}\right) & \text{si } p = -N/2 \\ \delta_{\text{déc},x}\mathfrak{u}_p & = & \dfrac{1}{h}\left(\dfrac{103}{72}\mathfrak{u}_p - \dfrac{91}{36}\mathfrak{u}_{p+1} + \dfrac{7}{4}\mathfrak{u}_{p+2} - \dfrac{29}{36}\mathfrak{u}_{p+3} + \dfrac{11}{72}\mathfrak{u}_{p+4}\right) & \text{si } p = N/2 \end{array}\right. \tag{4.31}$$

ainsi que l'opérateur $\sigma_{\text{déc},x}$ par

$$\left\{\begin{array}{lcll} \sigma_{\text{déc},x}\mathfrak{u}_p & = & \dfrac{1}{6}\mathfrak{u}_{p+1} + \dfrac{4}{6}\mathfrak{u}_p + \dfrac{1}{6}\mathfrak{u}_{p-1} & \text{si } -N/2+1 \leq p \leq N/2-1 \\ \sigma_{\text{déc},x}\mathfrak{u}_p & = & \dfrac{1}{6}\mathfrak{u}_{p-1} + \dfrac{4}{6}\mathfrak{u}_p & \text{si } p = -N/2 \\ \sigma_{\text{déc},x}\mathfrak{u}_p & = & \dfrac{4}{6}\mathfrak{u}_p + \dfrac{1}{6}\mathfrak{u}_{p-1} & \text{si } p = N/2 \end{array}\right. \tag{4.32}$$

L'opérateur décentré au bord est construit de manière à être d'ordre 4 en conservant la structure tridiagonale de la partie implicite $\sigma_{\text{déc},x}$. Grâce à cette structure, le système à résoudre peut être résolu en utilisant l'algorithme de Thomas [20, 72].

L'opérateur hermitien $\delta^H_{\text{déc},x} = \sigma^{-1}_{\text{déc},x} \circ \delta_{\text{déc},x}$ est un opérateur d'approximation de la dérivée première d'ordre 4. Son décentrement au bord des panels permet d'obtenir une nouvelle procédure de calcul des dérivées approchées de $\partial_\xi h(x_{i,j}^{(k)})$ et $\partial_\eta h(x_{i,j}^{(k)})$ donnée par

$$\left\{\begin{array}{lcll} \partial_\xi h(x_{i,j}^{(k)}) & \approx & \delta^H_{\text{déc},\Delta\xi} h_{i,j}^{(k)} & j \text{ fixé}, \\ \partial_\xi h(x_{i,j}^{(k)}) & \approx & \delta^H_{\text{déc},\Delta\eta} h_{i,j}^{(k)} & i \text{ fixé}. \end{array}\right. \tag{4.33}$$

Puisque $\delta^H_{\text{déc},x}$ est un opérateur d'ordre 4, le gradient approché donné par

$$\nabla_{T,\text{déc}} h_{i,j}^{(k)} = \delta^H_{\text{déc},\Delta\xi} h_{i,j}^{(k)} (\mathbf{g}^\xi)_{i,j}^{(k)} + \delta^H_{\text{déc},\Delta\eta} h_{i,j}^{(k)} (\mathbf{g}^\eta)_{i,j}^{(k)} \tag{4.34}$$

est un gradient d'ordre 4. Nous observons sur la Figure 4.7 la convergence de ce gradient $\nabla_{T,\text{déc}} h_{i,j}^{(k)}$ en comparaison au gradient $\nabla_{T,\Delta} h_{i,j}^{(k)}$. Les ordres de convergence sont pratiquement les mêmes. Cependant, les niveaux d'erreurs à N fixé sont bien plus faibles pour $\nabla_{T,\Delta}$ qu'avec $\nabla_{T,\text{déc}}$ comme on peut le voir sur la Table 4.3.

Le gradient décentré au bord des panels $\nabla_{T,\text{déc}}$ possède des propriétés de convergence semblables à celles du gradient $\nabla_{T,\Delta}$. Cependant, les niveaux d'erreurs sont plus élevés. Le gradient utilisant les splines cubiques est plus précis sur un maillage fixé. Dans la suite de ce travail, nous retenons le gradient $\nabla_{T,\Delta}$ pour la résolution des équations aux dérivées partielles.

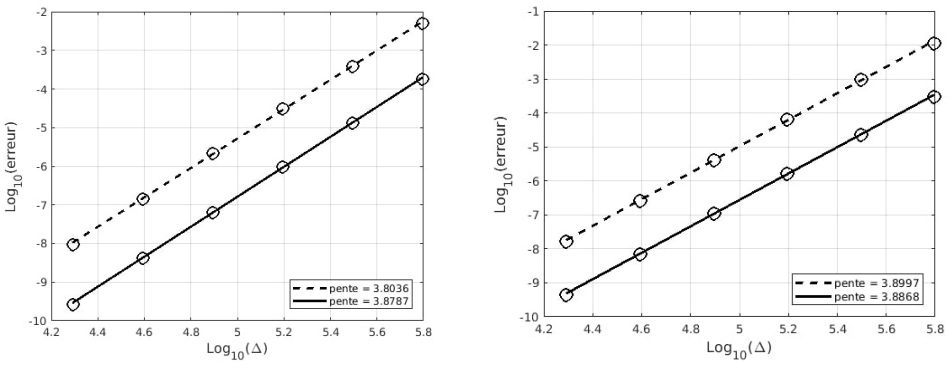


FIGURE 4.7 – Erreur sur le calcul du gradient en fonction de $\Delta = a\Delta\xi$. La convergence du gradient approché avec $(p, q, r) = (1, 2, 3)$ est à gauche et avec $(p, q, r) = (2, 2, 2)$ à droite. Nous comparons le gradient $\nabla_{T,\text{déc}}$ en pointillés et le gradient $\nabla_{T,\Delta}$ en trait plein.

	$(\mathbf{p}, \mathbf{q}, \mathbf{r}) = (\mathbf{1}, \mathbf{2}, \mathbf{3})$		$(\mathbf{p}, \mathbf{q}, \mathbf{r}) = (\mathbf{2}, \mathbf{2}, \mathbf{2})$	
N	$\nabla_{T,\Delta}$	$\nabla_{T,\text{déc}}$	$\nabla_{T,\Delta}$	$\nabla_{T,\text{déc}}$
16	$1.8636(-4)$	$5.0425(-3)$	$3.1368(-4)$	$1.1437(-2)$
32	$1.3545(-5)$	$3.9160(-4)$	$2.3910(-5)$	$9.7399(-4)$
64	$9.7564(-7)$	$3.1917(-5)$	$1.6716(-6)$	$6.6478(-5)$
128	$6.5592(-8)$	$2.2341(-6)$	$1.1088(-7)$	$4.2363(-6)$
256	$4.2563(-9)$	$1.4700(-7)$	$7.1437(-9)$	$2.6521(-7)$
511	$2.7115(-10)$	$9.4135(-9)$	$4.5361(-10)$	$1.6554(-8)$
ordre estimé	3.88	3.80	3.89	3.90

TABLE 4.3 – Erreur pour le calcul du gradient en fonction du paramètre de la Cubed-Sphere N. Table de convergence des gradients approchés $\nabla_{T,\Delta}$ et $\nabla_{T,\text{déc}}$.

4.3 Opérateur divergence discret

4.3.1 Construction et consistance de l'opérateur divergence discret

Soit $\mathbf{v} : \mathbf{x} \in \mathbb{S}_a^2 \mapsto \mathbf{v}(\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2$ un champ de vecteurs régulier sur la sphère. Il existe u, v et w des fonctions de coordonnées régulières de $\mathbb{S}_a^2$ dans $\mathbb{R}$ telles que

$$\mathbf{v}(\mathbf{x}) = u(\mathbf{x})\mathbf{i} + v(\mathbf{x})\mathbf{j} + w(\mathbf{x})\mathbf{k} \tag{4.35}$$

où $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ est la base canonique de $\mathbb{R}^3$ et $\mathbf{x} \in \mathbb{S}_a^2$. On définit $\tilde{\delta}_\xi^H \mathbf{v}_{i,j}^{(k)}$ et $\tilde{\delta}_\eta^H \mathbf{v}_{i,j}^{(k)}$ par

$$\left\{ \begin{array}{rcl} \tilde{\delta}_\xi^H \mathbf{v}_{i,j}^{(k)} & = & \tilde{\delta}_\xi^H u_{i,j}^{(k)}\mathbf{i} + \tilde{\delta}_\xi^H v_{i,j}^{(k)}\mathbf{j} + \tilde{\delta}_\xi^H w_{i,j}^{(k)}\mathbf{k} \\ \tilde{\delta}_\eta^H \mathbf{v}_{i,j}^{(k)} & = & \tilde{\delta}_\eta^H u_{i,j}^{(k)}\mathbf{i} + \tilde{\delta}_\eta^H v_{i,j}^{(k)}\mathbf{j} + \tilde{\delta}_\eta^H w_{i,j}^{(k)}\mathbf{k} \end{array} \right. \text{ pour } -N/2 \leq i,j \leq N/2,\ (k) = (I), \ldots, (VI). \tag{4.36}$$

où les fonctions de grille $(\tilde{\delta}_\xi^H u_{i,j}^{(k)}), (\tilde{\delta}_\eta^H u_{i,j}^{(k)}), \ldots$ sont construites grâce aux algorithmes 7 et 8.

Définition 4.5. *On définit l'opérateur divergence discrète par*

$$\nabla_{T,\Delta} \cdot \mathbf{v}_{i,j}^{(k)} = \tilde{\delta}_\xi^H \mathbf{v}_{i,j}^{(k)} \cdot (\mathbf{g}^\xi)_{i,j}^{(k)} + \tilde{\delta}_\eta^H \mathbf{v}_{i,j}^{(k)} \cdot (\mathbf{g}^\eta)_{i,j}^{(k)} \tag{4.37}$$

avec $-N/2 \leq i,j \leq N/2$ *et* $(k) = (I), \ldots, (VI)$.

Si un point $\mathbf{x}_{i,j}^{(k)}$ appartient à plusieurs panels, on y calcule plusieurs divergences approchées. Si le point est à l'intersection de deux panels, on conserve la moyenne des valeurs associées à chaque panel. Si le point est un coin de la Cubed-Sphere, on conserve la moyenne des trois valeurs. L'opérateur divergence discret de $(\mathbf{v}_{i,j}^{(k)})$ est une approximation de la divergence en $\mathbf{x}_{i,j}^{(k)}$. En effet :

Proposition 4.2. *Soit* $\mathbf{v} : \mathbf{x} \in \mathbb{S}_a^2 \mapsto \mathbf{v}(\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2$ *un champ de vecteur régulier sur la sphère. Alors pour tout* $-N/2 \leq i,j \leq N/2$ *et* $(k) = (I), \ldots, (VI)$, *on a*

$$\nabla_{T,\Delta} \cdot \mathbf{v}_{i,j}^{(k)} - (\nabla_T \cdot \mathbf{v})_{i,j}^{*,(k)} = \mathcal{O}\left(\Delta\xi^3\right) \tag{4.38}$$

où $*$ *désigne l'opérateur de restriction à la Cubed-Sphere et* $\Delta\xi = \Delta\eta$.

Démonstration. L'opérateur divergence étant linéaire, on utilise le théorème 4.2 pour démontrer le résultat. □

4.3.2 Tests numériques

Pour construire un test permettant d'évaluer les performances de la divergence approchée $\nabla_{T,\Delta}\cdot$, on utilise les propriétés suivantes :

Lemme 4.1. *Soit* $\mathbf{w} : \mathbf{x} \mapsto \mathbf{w}(\mathbf{x}) \in \mathbb{R}^3$ *un champ de vecteurs. On note* $\mathbf{n}(\mathbf{x}) = \mathbf{x}/a$ *le vecteur normal unitaire à* $\mathbb{S}_a^2$. *Si* $\mathbf{x} \in \mathbb{S}_a^2$, *on a*

$$\mathbf{F}(\mathbf{x}) = \mathbf{n}(\mathbf{x}) \wedge \mathbf{w}(\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2. \tag{4.39}$$

Soit $f : \mathbf{x} \in \mathbb{S}_a^2 \mapsto f(\mathbf{x}) \in \mathbb{C}$ une fonction régulière sur la sphère, alors la proposition suivante est vérifiée :

Proposition 4.3. *Pour tout* $\mathbf{x} \in \mathbb{S}_a^2$, *on pose* $\mathbf{F}(\mathbf{x}) = \mathbf{n}(\mathbf{x}) \wedge (f(\mathbf{x})\mathbf{u})$ *avec* $\mathbf{u} \in \mathbb{R}^3$ *un vecteur constant. Alors*

$$\nabla_T \cdot \mathbf{F} = \nabla_T f \cdot (\mathbf{n} \wedge \mathbf{u}) \tag{4.40}$$

où $\mathbf{n}$ *est le vecteur unitaire extérieur à* $\mathbb{S}_a^2$.

Démonstration.

- Soit $\mathbf{x} \in \mathbb{S}_a^2$, alors

$$\begin{aligned}\frac{\partial}{\partial \xi}\mathbf{n}(\mathbf{x}) &= \frac{1}{a}\frac{\partial \mathbf{x}}{\partial \xi} \\ &= \frac{1}{a}\mathbf{g}_\xi \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2\end{aligned}$$

De même,

$$\frac{\partial}{\partial \eta}\mathbf{n}(\mathbf{x}) = \frac{1}{a}\mathbf{g}_\eta \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2.$$

- On considère le terme en $\mathbf{g}^\xi$ de la divergence

$$\begin{aligned}\frac{\partial}{\partial \xi}\mathbf{F}\cdot\mathbf{g}^\xi &= \frac{\partial}{\partial \xi}\left(\mathbf{n}\wedge f\mathbf{w}\right)\cdot\mathbf{g}^\xi \\ &= \left(\frac{\partial \mathbf{n}}{\partial \xi}\wedge f\mathbf{w} + \mathbf{n}\wedge\frac{\partial}{\partial \xi}\left(f\mathbf{w}\right)\right)\cdot\mathbf{g}^\xi \\ &= \left(\mathbf{n}\wedge\mathbf{w}\right)\cdot\frac{\partial f}{\partial \xi}\mathbf{g}^\xi\end{aligned}$$

car $\mathbf{n} \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2$ et $\mathbf{w}$ est un vecteur constant. De la même manière, pour le terme en η, on a

$$\frac{\partial}{\partial \eta}\mathbf{F}\cdot\mathbf{g}^\eta = \left(\mathbf{n}\wedge\mathbf{w}\right)\cdot\frac{\partial f}{\partial \eta}\mathbf{g}^\eta.$$

- D'après (4.1), la divergence est :

$$\begin{aligned}\nabla_T\cdot\mathbf{F} &= \frac{\partial}{\partial \xi}\mathbf{F}\cdot\mathbf{g}^\xi + \frac{\partial}{\partial \eta}\mathbf{F}\cdot\mathbf{g}^\eta \\ &= \left(\mathbf{n}\wedge\mathbf{w}\right)\cdot\frac{\partial f}{\partial \xi}\mathbf{g}^\xi + \left(\mathbf{n}\wedge\mathbf{w}\right)\cdot\frac{\partial f}{\partial \eta}\mathbf{g}^\eta \\ &= \nabla_T f\cdot\left(\mathbf{n}\wedge\mathbf{u}\right).\end{aligned}$$

□

De plus, l'opérateur divergence est à moyenne nulle :

$$\int_{\mathbb{S}_a^2}\nabla_T\cdot\mathbf{u}(\mathbf{x})d\sigma(\mathbf{x}) = 0. \tag{4.41}$$

On a choisi de tester la divergence approchée $\nabla_{T,\Delta}\cdot$ de la définition 4.5 sur le champ de vecteurs

$$\mathbf{F}(\mathbf{x}) = x^p y^q z^r \mathbf{n}(\mathbf{x})\wedge\mathbf{u}, \tag{4.42}$$

avec $p, q, r \in \mathbb{N}$ et $\mathbf{u} = [1,1,1]^T$. La divergence d'un tel champ de vecteurs est donnée par la proposition 4.3 et la proposition 3.3. On mesure l'erreur relative :

$$\frac{\|\left(\nabla_{T,\Delta}\cdot\mathbf{F}^*\right) - \left(\nabla_T\cdot\mathbf{F}\right)^*\|_l}{\|\left(\nabla_T\cdot\mathbf{F}\right)^*\|_l} \tag{4.43}$$

avec $l \in \{1,2,\infty\}$. $*$ désigne la restriction à la Cubed-Sphere. On mesure aussi l'erreur sur la conservation, c'est à dire la valeur approchée de l'intégrale (4.41). Les résultats sont donnés dans les tables 4.4 et 4.5. Ils permettent de confirmer l'ordre 3 minimum obtenu théoriquement et donnent un ordre 4 numérique aussi bien sur l'erreur de la divergence que sur la conservation. La Table 4.5 présente une erreur de conservation proche de l'erreur machine quel que soit le maillage. Ce résultat est attribué aux symétries de h lorsque $(p,q,r) = (1,1,1)$.

Les courbes de convergence en Figure 4.8 sont pratiquement confondues et ont une pente proche de 4. De plus les points sont parfaitement alignés.

N	norme 1	norme 2	norme ∞	conservation
16	3.3707(−4)	3.1674(−4)	3.7839(−4)	1.6810(−6)
32	2.0751(−5)	1.9593(−5)	2.5503(−5)	1.1083(−7)
64	1.2964(−6)	1.2274(−6)	1.8238(−6)	7.0272(−9)
128	8.1231(−8)	7.6985(−8)	1.2133(−7)	4.4093(−10)
256	5.0903(−9)	4.8271(−9)	7.8110(−9)	2.7597(−11)
512	3.1888(−10)	3.0280(−10)	4.9564(−10)	1.7256(−12)
ordre estimé	4.00	4.00	3.90	3.98

TABLE 4.4 – Erreur pour le calcul de la divergence approchée $\nabla_{T,\Delta}\cdot$ en fonction de N le paramètre de la Cubed-Sphere. On calcule l'erreur lors du calcul de la divergence de (4.42) avec $(p,q,r) = (1,2,3)$ en normes 1, 2 et infinie. On vérifie aussi la conservation (4.41).

.

N	norme 1	norme 2	norme ∞	conservation
16	4.4656(−5)	5.2622(−5)	1.1018(−4)	2.2965(−18)
32	2.9404(−6)	3.4964(−6)	8.9432(−6)	9.5834(−18)
64	1.8859(−7)	2.2566(−7)	6.3993(−7)	6.9334(−18)
128	1.1972(−8)	1.4352(−8)	4.2842(−8)	7.5356(−18)
256	7.5548(−10)	9.0630(−10)	2.7721(−9)	1.3528(−17)
512	4.7499(−11)	5.7106(−11)	1.7640(−10)	1.7256(−18)
ordre estimé	3.97	3.97	3.86	–

TABLE 4.5 – Erreur pour le calcul de la divergence approchée $\nabla_{T,\Delta}\cdot$ en fonction de N le paramètre de la Cubed-Sphere. On calcule l'erreur lors du calcul de la divergence de (4.42) avec $(p,q,r) = (1,1,1)$ en normes 1, 2 et infinie. On vérifie aussi la conservation (4.41).

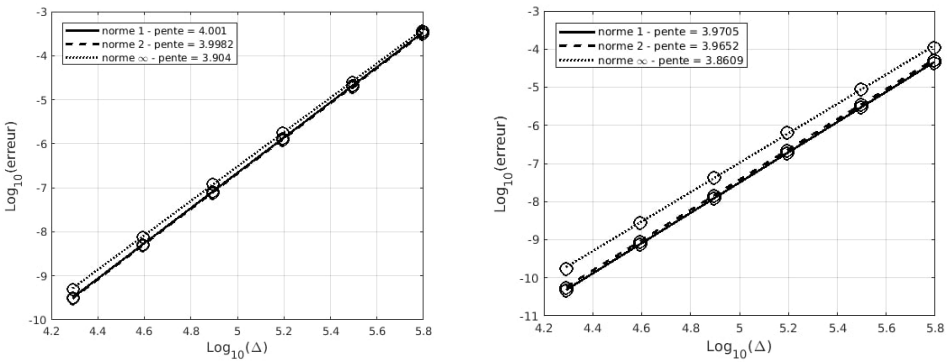


FIGURE 4.8 – Convergence de la divergence approchée $\nabla_{T,\Delta}$ avec $(p,q,r) = (1,2,3)$ à gauche et avec $(p,q,r) = (1,1,1)$ à droite avec différentes normes en fonction de $\Delta = a\Delta\xi$.

4.3.3 Variante de l'opérateur de divergence discret

La divergence utilisée dans la formule (4.5) est coûteuse en calcul. Il faut calculer six dérivées hermitiennes. Une méthode basée sur une divergence approchée moins coûteuse en calculs est donnée dans [25]. Cette méthode est basée sur l'égalité suivante :

Proposition 4.4. *Si* $\mathbf{F} : \mathbf{x} \in \mathbb{S}_a^2 \mapsto \mathbf{F}(\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2$ *est un champ de vecteurs réguliers sur la sphère, alors l'égalité suivante est vérifiée en* $\mathbf{x}_{i,j}^{(k)}$ *:*

$$\nabla_T \cdot \mathbf{F} = \frac{1}{\sqrt{\bar{\mathbf{G}}}} \left[\frac{\partial}{\partial \xi} \left(\sqrt{\bar{\mathbf{G}}} \mathbf{F} \cdot \mathbf{g}^{\xi} \right)_{\eta=\eta_j} + \frac{\partial}{\partial \eta} \left(\sqrt{\bar{\mathbf{G}}} \mathbf{F} \cdot \mathbf{g}^{\eta} \right)_{\xi=\xi_i} \right] \tag{4.44}$$

Démonstration. L'égalité suivante est vérifiée :

$$\begin{aligned} \frac{1}{\sqrt{\bar{\mathbf{G}}}} \frac{\partial}{\partial \xi} \left(\sqrt{\bar{\mathbf{G}}} \mathbf{F} \cdot g^{\xi} \right) &= \mathbf{g}^{\xi} \cdot \frac{\partial \mathbf{F}}{\partial \xi} + \mathbf{F} \cdot \mathbf{g}^{\xi} \frac{1}{\sqrt{\bar{\mathbf{G}}}} \frac{\partial \sqrt{\bar{\mathbf{G}}}}{\partial \xi} + \mathbf{F} \cdot \frac{\partial \mathbf{g}^{\xi}}{\partial \xi} \\ &= \mathbf{g}^{\xi} \cdot \frac{\partial \mathbf{F}}{\partial \xi} + \mathbf{F} \cdot \mathbf{g}^{\xi} \left(\Gamma_{\eta,\xi}^{\eta} + \Gamma_{\xi,\xi}^{\xi} \right) + \mathbf{F} \cdot \left(-\Gamma_{\xi,\xi}^{\xi} g^{\xi} - \Gamma_{\xi,\eta}^{\xi} \mathbf{g}^{\eta} - \frac{1}{a} \mathbf{n} \right) \\ &\qquad \text{d'après les propositions 3.9 et 3.8 ,} \\ &= \mathbf{g}^{\xi} \cdot \frac{\partial \mathbf{F}}{\partial \xi} + \mathbf{F} \cdot \mathbf{g}^{\xi} \Gamma_{\eta,\xi}^{\eta} - \mathbf{F} \cdot \mathbf{g}^{\eta} \Gamma_{\xi,\eta}^{\xi} \text{ car } \mathbf{F} \text{ est tangent à } \mathbb{S}_a^2. \end{aligned}$$

De la même manière, on a

$$\frac{1}{\sqrt{\bar{\mathbf{G}}}} \frac{\partial}{\partial \eta} \left(\sqrt{\bar{\mathbf{G}}} \mathbf{F} \cdot g^{\eta} \right) = \mathbf{g}^{\eta} \cdot \frac{\partial \mathbf{F}}{\partial \eta} + \mathbf{F} \cdot \mathbf{g}^{\eta} \Gamma_{\xi,\eta}^{\xi} - \mathbf{F} \cdot \mathbf{g}^{\xi} \Gamma_{\eta,\xi}^{\eta}. \tag{4.45}$$

En combinant ces deux égalités, on trouve

$$\nabla_T \cdot \mathbf{F} = \frac{1}{\sqrt{\bar{\mathbf{G}}}} \left[\frac{\partial}{\partial \xi} \left(\sqrt{\bar{\mathbf{G}}} \mathbf{F} \cdot \mathbf{g}^{\xi} \right)_{\eta=\eta_j} + \frac{\partial}{\partial \eta} \left(\sqrt{\bar{\mathbf{G}}} \mathbf{F} \cdot \mathbf{g}^{\eta} \right)_{\xi=\xi_i} \right], \tag{4.46}$$

ce qui conclut la preuve. □

Si l'on utilise la forme de la divergence (4.44) au lieu (4.8), il n'y a que deux dérivées hermitiennes à calculer au lieu de six. On se concentre sur le panel (I) pour détailler la procédure permettant de calculer une valeur approchée du terme

$$\frac{\partial}{\partial \xi} \left(\sqrt{\bar{\mathbf{G}}} \mathbf{F} \cdot \mathbf{g}^{\xi} \right)_{\eta=\eta_j}. \tag{4.47}$$

On utilise encore une fois la structure en grands cercles. La fonction

$$\xi \mapsto \sqrt{\bar{\mathbf{G}}} \mathbf{F} \cdot \mathbf{g}^{\xi} \tag{4.48}$$

est bien définie sur le panel (I) et peut s'étendre facilement au panel (III). En effet, si $\mathbf{x}(x, y, z) \in \mathbb{S}_a^2$ on a

$$\mathbf{g}^{\xi}(\mathbf{x}) = \frac{1}{1 + \frac{y^2}{x^2}} \begin{bmatrix} -\frac{y}{x^2} \\ \frac{1}{x} \\ 0 \end{bmatrix} \text{ ainsi que } \sqrt{\bar{\mathbf{G}}} = a^2 \frac{(1 + X^2)(1 + Y^2)}{(1 + X^2 + Y^2)^{3/2}} \tag{4.49}$$

avec $X = \frac{y}{x}$ et $Y = \frac{z}{|x|}$. Cependant, sur les panels (II) et (IV), l'abscisse x s'annule (en $(0, \pm a, 0)$), ce qui interdit d'utiliser l'expression (4.49). On remplace les fonctions $x \mapsto 1/x$ et $x \mapsto 1/|x|$ par des prolongements de classe $\mathcal{C}^2$ au voisinage de 0 :

$$\psi_1(x) = \begin{cases} \frac{1}{x} & \text{si } |x| \geq sa, \\ \frac{1}{s^6} x^5 - \frac{3}{s^4} x^3 + \frac{3}{s^2} x & \text{si } |x| \leq sa. \end{cases} \tag{4.50}$$

$$\psi_2(x) = \begin{cases} \frac{1}{|x|} & \text{si } |x| \geq sa, \\ \frac{3}{8s^5}|x|^4 - \frac{5}{4s^3}|x|^2 + \frac{15}{8s} & \text{si } |x| \leq sa, \end{cases} \tag{4.51}$$

où s est un paramètre de seuillage permettant de peu affecter la fonction initiale. Dans la pratique, nous utilisons $s = 0.05$ ce qui permet de peu perturber les fonctions initiales. On remplace dans (4.49) les termes en $1/x$ et $1/|x|$ par $\psi_1(x)$ et $\psi_2(x)$. Ainsi, on pose

$$\tilde{\mathbf{g}}^{\xi}(\mathbf{x}) = \frac{1}{1+(y\psi_2(x))^2}\begin{bmatrix} -y\psi_1(x^2) \\ \psi_1(x) \\ 0 \end{bmatrix}, \tilde{\mathbf{g}}^{\eta}(\mathbf{x}) = \frac{1}{1+(z\psi_2(x))^2}\begin{bmatrix} -z\psi_1(x^2) \\ 0 \\ \psi_1(x) \end{bmatrix} \text{ et } \sqrt{\tilde{\mathbf{G}}} = a^2\frac{(1+\tilde{X}^2)(1+\tilde{Y}^2)}{(1+\tilde{X}^2+\tilde{Y}^2)^{3/2}} \tag{4.52}$$

avec $\tilde{X} = y\psi_1(x)$ et $\tilde{Y} = z\psi_2(x)$. En remplaçant $\mathbf{g}^{\xi}$, $\mathbf{g}^{\eta}$ et $\sqrt{\mathbf{G}}$ par $\tilde{\mathbf{g}}^{\xi}$, $\tilde{\mathbf{g}}^{\eta}$ et $\sqrt{\tilde{\mathbf{G}}}$ données par (4.52), on définit l'opérateur de divergence approchée en $\mathbf{x}_{i,j}^{(k)}$

$$\nabla_{T,\Delta,2}\cdot\mathbf{F} = \frac{1}{\sqrt{\tilde{\mathbf{G}}}}\left[\tilde{\delta}_{\xi}^{H}\left(\sqrt{\tilde{\mathbf{G}}}\mathbf{F}\cdot\tilde{\mathbf{g}}^{\xi}\right) + \tilde{\delta}_{\eta}^{H}\left(\sqrt{\tilde{\mathbf{G}}}\mathbf{F}\cdot\tilde{\mathbf{g}}^{\eta}\right)\right], \tag{4.53}$$

où $\tilde{\delta}_{\xi}^{H} f$ et $\tilde{\delta}_{\eta}^{H} f$ désignent les dérivées hermitiennes en ξ et en η de f obtenues en utilisant les algorithmes 7 et 8.

De même que précédemment, on teste la précision de cette dernière sur le champ de vecteurs (4.42). Les résultats sont donnés dans les Tables 4.6 et 4.7.

N	**norme 1**	**norme 2**	**norme ∞**	**conservation**
16	4.4464(−4)	4.2619(−4)	5.7206(−4)	7.2201(−6)
32	2.7567(−5)	2.6414(−5)	4.3945(−5)	4.4825(−7)
64	1.7319(−6)	1.6606(−6)	3.0870(−6)	2.7976(−8)
128	1.0886(−8)	1.0438(−7)	2.0521(−7)	1.7475(−9)
256	6.8327(−9)	6.5488(−9)	1.3242(−8)	1.0921(−10)
512	4.2837(−10)	4.1045(−10)	8.4121(−10)	6.8258(−12)
ordre estimé	4.00	4.00	3.88	4.00

TABLE 4.6 – Table de convergence pour la divergence $\nabla_{T,\Delta,2}$ de (4.42) avec $(p,q,r) = (1,2,3)$ en normes 1, 2 et infinie. On vérifie aussi la conservation (4.41).

N	**norme 1**	**norme 2**	**norme ∞**	**conservation**
16	9.5570(−5)	1.1354(−4)	2.5127(−4)	9.9325(−18)
32	6.2727(−6)	7.4509(−6)	1.9429(−5)	9.0737(−18)
64	4.0293(−7)	4.7991(−7)	1.3703(−6)	1.4538(−17)
128	2.5561(−8)	3.0505(−8)	9.1295(−8)	1.0589(−17)
256	1.6119(−9)	1.9247(−9)	5.8964(−9)	2.2899(−18)
512	1.0140(−10)	1.2102(−10)	3.7487(−10)	1.0804(−17)
ordre estimé	3.97	3.97	3.88	–

TABLE 4.7 – Table de convergence pour la divergence $\nabla_{T,\Delta,2}$ de (4.42) avec $(p,q,r) = (1,1,1)$ en normes 1, 2 et infinie ainsi que l'erreur de conservation (4.41).

Les résultats numériques obtenus sur la divergence approchée $\nabla_{T,\Delta}\cdot$ (définition 4.5) et ceux obtenus sur $\nabla_{T,\Delta,2}\cdot$ (4.53) sont particulièrement proches l'un de l'autre. Le calcul de la divergence $\nabla_{T,\Delta}\cdot$ nécessite l'évaluation de 6 dérivées hermitiennes alors que l'évaluation de $\nabla_{T,\Delta}\cdot$ n'en nécessite que deux. Nous ne disposons pas de résultat théorique pour garantir la consistance de $\nabla_{T,\Delta,2}\cdot$ avec la divergence.

Les erreurs sont semblables pour chacune des divergences approchées et les ordres de convergences sont les mêmes. Cependant, nous disposons d'un résultat de consistance pour l'opérateur $\nabla_{T,\Delta}\cdot$. Dans la suite de ce travail, nous retiendrons l'opérateur divergence approchée $\nabla_{T,\Delta}\cdot$ pour la résolution des équations aux dérivées partielles.

4.4 Opérateur rotationnel discret

4.4.1 Construction et consistance de l'opérateur rotationnel discret

Soit $\mathbf{v} : \mathbf{x} \in \mathbb{S}_a^2 \mapsto \mathbf{v}(\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2$ un champ de vecteur régulier sur la sphère. En tout point de la Cubed-Sphere $\mathbf{x}_{i,j}^{(k)}$, avec $-N/2 \leq i,j \leq N/2$ et $(k) = (I), ..., (VI)$, les dérivées partielles approchées de $\mathbf{v}$ sont données par (4.36). On définit donc le rotationnel approché suivant

Définition 4.6. *On définit l'opérateur rotationnel discret par*

$$rot_\Delta(\mathbf{v}_{i,j}^{(k)}) = (\nabla_{T,\Delta} \wedge \mathbf{v}_{i,j}^{(k)}) \cdot \mathbf{n} = \left((\mathbf{g}^\xi)_{i,j}^{(k)} \wedge \tilde{\delta}_\xi^H \mathbf{v}_{i,j}^{(k)} + (\mathbf{g}^\eta)_{i,j}^{(k)} \wedge \tilde{\delta}_\eta^H \mathbf{v}_{i,j}^{(k)}\right) \cdot \mathbf{n}(\mathbf{x}_{i,j}^{(k)}) \in \mathbb{C} \tag{4.54}$$

avec $-N/2 \leq i,j \leq N/2$ *et* $(k) = (I), \ldots, (VI)$.

Le théorème 4.2 permet de montrer la consistance de cet opérateur d'approximation :

Proposition 4.5. *Soit* $\mathbf{v} : \mathbf{x} \in \mathbb{S}_a^2 \mapsto \mathbf{v}(\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2$ *un champ de vecteurs régulier sur* $\mathbb{S}_a^2$. *Pour tous* $-N/2 \leq i,j \leq N/2$ *et* $(k) = (I), ..., (VI)$, *on a*

$$rot_\Delta((\mathbf{v}^*)_{i,j}^{(k)}) - rot(\mathbf{v})_{i,j}^{*,(k)} = \mathcal{O}\left(\Delta\xi^3\right) \tag{4.55}$$

où * *désigne l'opérateur de restriction à la Cubed-Sphere et* $\Delta\xi = \Delta\eta$.

Si un point $\mathbf{x}_{i,j}^{(k)}$ appartient à plusieurs panels (c'est à dire i ou j vaut $\pm N/2$) alors plusieurs rotationnels approchés sont calculés en ce point. Si $\mathbf{x}_{i,j}^{(k)}$ est à l'intersection de deux panels, deux valeurs du rotationnel sont obtenues, nous retenons la moyenne de ces deux valeurs. Si $\mathbf{x}_{i,j}^{(k)}$ est à l'intersection de trois panels, nous conservons la moyenne des trois valeurs.

4.4.2 Tests numériques

Pour tester le rotationnel discret rot_Δ, on mesure l'erreur relative en norme l donnée par

$$e_l = \frac{\|\mathrm{rot}(\mathbf{v})^* - \mathrm{rot}_\Delta(\mathbf{v}^*)\|_l}{\|\mathrm{rot}(\mathbf{v})^*\|_l} \tag{4.56}$$

avec $l \in \{1, 2, \infty\}$.

Soit $\mathbf{v}$ un champ de vecteurs sur la sphère. Il existe v_λ et v_θ tels que

$$\mathbf{v} = v_\lambda \mathbf{e}_\lambda + v_\theta \mathbf{e}_\theta \tag{4.57}$$

où $(\lambda, \theta) \in]0, 2\pi] \times]-\pi/2, \pi/2[$ représentent les coordonnées longitude-latitude et $(\mathbf{e}_\lambda, \mathbf{e}_\theta)$ les vecteurs de base donnés (voir Annexe).

Supposons que $\mathbf{v}$ est un champ de vecteurs de la forme "champ zonal" :

$$\mathbf{v}(\lambda, \theta) = \cos^\alpha(\theta)\mathbf{e}_\lambda \tag{4.58}$$

alors le rotationnel de $\mathbf{v}$ est donné par

$$\mathrm{rot}(\mathbf{v}) = \frac{\alpha + 1}{a} \cos^{\alpha-1}\theta \sin\theta. \tag{4.59}$$

N	**norme 1**	**norme 2**	**norme ∞**
8	2.9158(−4)	3.3039(−4)	6.7103(−4)
16	1.7719(−5)	1.9906(−5)	4.0648(−5)
32	1.1025(−6)	1.2416(−6)	2.5207(−6)
64	6.9056(−8)	7.7821(−8)	1.6433(−7)
128	4.3244(−9)	4.8755(−9)	1.0822(−8)
256	2.7061(−10)	3.0522(−10)	6.9474(−10)
ordre estimé	4.01	4.01	3.97

TABLE 4.8 – Erreur pour le calcul du rotationnel approché rot_Δ en fonction de N le paramètre de la Cubed-Sphère. Table de convergence $\mathrm{rot}_\Delta(\mathbf{v})$ avec $\mathbf{v}(\lambda,\theta) = \cos^3(\theta)\,\mathbf{e}_\lambda$ en normes 1, 2 et ∞.

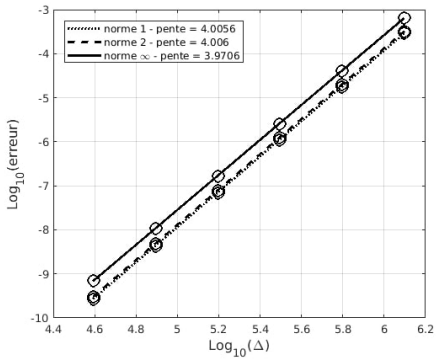


FIGURE 4.9 – Erreur pour le calcul du rotationnel approché rot_Δ en fonction de $\Delta = a\Delta\xi$. Convergence $\mathrm{rot}_\Delta(\mathbf{v})$ avec $\mathbf{v}(\lambda,\theta) = \cos^3(\theta)\,\mathbf{e}_\lambda$ en normes 1, 2 et ∞.

On choisit $\alpha = 3$ pour avoir un champ régulier aux pôles. Un tel champ représente un écoulement stationnaire zonale sur la sphère. La valeur maximale de ce dernier est atteinte à l'équateur et il s'atténue rapidement aux pôles. La Table 4.8 et la Figure 4.9 associée permettent de mettre en évidence une convergence à l'ordre 4.

Supposons un champ de vecteurs $\mathbf{v}$ défini dans $\mathbb{R}^3$ et tangent à la sphère :

$$\mathbf{v} : \mathbf{x} \in \mathbb{R}^3 \mapsto \mathbf{v}(\mathbf{x}) \in \mathbb{R}^3. \tag{4.60}$$

Alors le rotationnel sphérique rot et la vorticité de $\mathbb{R}^3$ sont liés par la proposition suivante :

Proposition 4.6. *Soit $\mathbf{v} : \mathbf{x} \in \mathbb{R}^3 \mapsto \mathbf{v}(\mathbf{x}) \in \mathbb{R}^3$ un champ de vecteurs. Soit $\hat{\mathbf{v}}$ la restriction de $\mathbf{v}$ à $\mathbb{S}^2_a$. On suppose que $\hat{\mathbf{v}}$ est un champ de vecteurs tangents à la sphère pour que $rot(\hat{\mathbf{v}})$ soit défini. Alors en tout point de la sphère $\mathbb{S}^2_a$ on a*

$$rot(\hat{\mathbf{v}}) = (\nabla_{\mathbb{R}^3} \wedge \mathbf{v}) \cdot \mathbf{n} \tag{4.61}$$

avec $\mathbf{n}$ la normale extérieure à la sphère $\mathbf{n} = \mathbf{x}/a$ et $\mathbf{x} \in \mathbb{S}^2_a$.

Démonstration. Le rotationnel d'un champ de vecteur de $\mathbf{v}$ s'exprime par

$$\nabla_{\mathbb{R}^3} \wedge \mathbf{v} = \mathbf{g}^\xi \wedge \frac{\partial \mathbf{v}}{\partial \xi} + \mathbf{g}^\eta \wedge \frac{\partial \mathbf{v}}{\partial \eta} + \mathbf{n} \wedge \frac{\partial \mathbf{v}}{\partial r} \tag{4.62}$$

avec r la coordonnée radiale. Or, la dernière composante est tangente à la sphère

$$\mathbf{n} \wedge \frac{\partial \mathbf{v}}{\partial r} \in \mathbb{TS}^2_a \tag{4.63}$$

donc

$$\left(\mathbf{n} \wedge \frac{\partial \mathbf{v}}{\partial r}\right) \cdot \mathbf{n} = 0. \tag{4.64}$$

Ainsi, on retrouve en tout point de la sphère $\mathbb{S}^2_a$:

$$(\nabla_{\mathbb{R}^3} \wedge \mathbf{v}) \cdot \mathbf{n} = \left(\mathbf{g}^\xi \wedge \frac{\partial \mathbf{v}}{\partial \xi} + \mathbf{g}^\eta \wedge \frac{\partial \mathbf{v}}{\partial \eta} + \mathbf{n} \wedge \frac{\partial \mathbf{v}}{\partial r}\right) \cdot \mathbf{n} \tag{4.65}$$

$$= \left(\mathbf{g}^\xi \wedge \frac{\partial \mathbf{v}}{\partial \xi} + \mathbf{g}^\eta \wedge \frac{\partial \mathbf{v}}{\partial \eta}\right) \cdot \mathbf{n} \tag{4.66}$$

$$= \mathrm{rot}(\mathbf{v}). \tag{4.67}$$

□

De plus, si $\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}$, alors

$$\nabla_{\mathbb{R}^3} \wedge \mathbf{F} = \begin{bmatrix} \partial_y F_z - \partial_z F_y \\ \partial_z F_x - \partial_x F_z \\ \partial_x F_y - \partial_y F_x \end{bmatrix}. \tag{4.68}$$

On choisit, pour le second test numérique, le champ de vecteurs tangents à la sphère suivant :

$$\mathbf{v}(x,y,z) = \mathbf{n}(x,y,z) \wedge \begin{bmatrix} \exp(y/a) \\ \exp(x/a) \\ \exp(z/a) \end{bmatrix} = \begin{bmatrix} (y/a)\exp(y/a) - (z/a)\exp(x/a) \\ (z/a)\exp(z/a) - (x/a)\exp(y/a) \\ (y/a)\exp(x/a) - (y/a)\exp(z/a) \end{bmatrix} \tag{4.69}$$

avec $(x,y,z) \in \mathbb{R}^3$. On a $\mathbf{v} \in \mathbb{TS}^2_a$. La vorticité de $\mathbf{v}$ est donnée par

$$\nabla_{\mathbb{R}^3} \wedge \mathbf{v} = \frac{1}{a} \begin{bmatrix} -(2 + z/a)\exp(z/a) \\ -(2 + x/a)\exp(x/a) \\ -(2 + y/a)\exp(y/a) \end{bmatrix}. \tag{4.70}$$

N	**norme 1**	**norme 2**	**norme ∞**
8	1.0377(−4)	1.2588(−4)	3.6670(−4)
16	6.3236(−6)	7.4682(−6)	2.2222(−5)
32	3.9444(−7)	4.6278(−7)	1.3713(−6)
64	2.4726(−8)	2.8931(−8)	8.5415(−8)
128	1.5500(−9)	1.8111(−9)	5.3339(−9)
256	9.7139(−11)	1.1342(−10)	3.3330(−10)
ordre estimé	4.00	4.01	4.01

TABLE 4.9 – Erreur pour le calcul du rotationnel discret rot_Δ en fonction du paramètre de la Cubed-Sphere N. La table de convergence de $\mathrm{rot}_\Delta(\mathbf{v})$ avec $\mathbf{v}$ donné par (4.69) en normes 1, 2 et ∞. On observe l'ordre 4 du rotationnel approché rot_Δ.

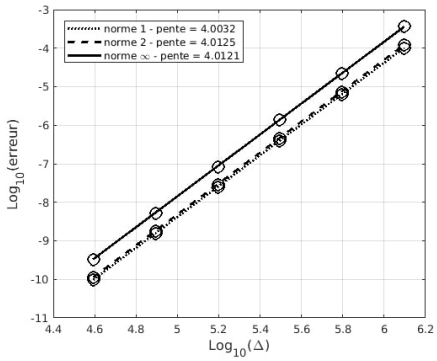


FIGURE 4.10 – Erreur pour le calcul du rotationnel discret rot_Δ en fonction de $\Delta = a\Delta\xi$. La convergence de $\mathrm{rot}_\Delta(\mathbf{v})$ avec $\mathbf{v}$ est donnée par (4.69) en normes 1, 2 et ∞.

On déduit le rotationnel sphérique de $\mathbf{v}$:

$$\mathrm{rot}(\mathbf{v}) = -\frac{1}{a}\left[\frac{x}{a}\left(2+\frac{z}{a}\right)\exp\left(\frac{z}{a}\right) + \frac{y}{a}\left(2+\frac{x}{a}\right)\exp\left(\frac{x}{a}\right) + \frac{z}{a}\left(2+\frac{y}{a}\right)\exp\left(\frac{y}{a}\right)\right], \tag{4.71}$$

avec $(x, y, z) \in \mathbb{S}_a^2$. On compare le rotationnel $\mathrm{rot}(\mathbf{v})$ avec le rotationnel approché $\mathrm{rot}_\Delta(\mathbf{v})$, la convergence de la méthode est donnée dans la Table 4.9.

On compare à présent dans la proposition suivante l'opérateur $\mathrm{rot}(\nabla_T \mathbf{v})$ avec le champ nul.

Proposition 4.7. *Soit $h : \mathbf{x} \in \mathbb{S}_a^2 \mapsto h(\mathbf{x})$ une fonction régulière. Alors*

$$\nabla_T \wedge \nabla_T h = \frac{1}{a}\mathbf{n} \wedge \nabla_T h. \tag{4.72}$$

En particulier, dans la direction de $\mathbf{n}$:

$$rot(\nabla_T h) = 0. \tag{4.73}$$

Démonstration. En coordonnées (ξ, η), on a

$$\nabla_T \wedge \nabla_T h = \mathbf{g}^\xi \wedge \frac{\partial}{\partial \xi}\left(\frac{\partial h}{\partial \xi}\mathbf{g}^\xi + \frac{\partial h}{\partial \eta}\mathbf{g}^\eta\right) + \mathbf{g}^\eta \wedge \frac{\partial}{\partial \eta}\left(\frac{\partial h}{\partial \xi}\mathbf{g}^\xi + \frac{\partial h}{\partial \eta}\mathbf{g}^\eta\right). \tag{4.74}$$

En utilisant la définition 3.9 des symboles de Christoffel, on a

$$\mathbf{g}^\xi \wedge \frac{\partial}{\partial \xi}\nabla_T h = -\Gamma^\xi_{\xi,\eta}\frac{\partial h}{\partial \xi} - \frac{1}{a}\frac{\partial h}{\partial \xi}\mathbf{g}^\xi \wedge \mathbf{n} + \frac{\partial^2 h}{\partial \xi \partial \eta}\mathbf{g}^\xi \wedge \mathbf{g}^\eta - \Gamma^\eta_{\xi,\eta}\frac{\partial h}{\partial \eta}\mathbf{g}^\xi \wedge \mathbf{g}^\eta. \tag{4.75}$$

De la même manière, on a

$$\mathbf{g}^\eta \wedge \frac{\partial}{\partial \eta}\nabla_T h = -\Gamma^\eta_{\eta,\xi}\frac{\partial h}{\partial \eta} - \frac{1}{a}\frac{\partial h}{\partial \eta}\mathbf{g}^\eta \wedge \mathbf{n} + \frac{\partial^2 h}{\partial \xi \partial \eta}\mathbf{g}^\eta \wedge \mathbf{g}^\xi - \Gamma^\xi_{\eta,\xi}\frac{\partial h}{\partial \xi}\mathbf{g}^\eta \wedge \mathbf{g}^\xi. \tag{4.76}$$

En sommant ces deux égalités et en utilisant

$$\mathbf{g}^\xi \wedge \mathbf{g}^\eta = -\mathbf{g}^\eta \wedge \mathbf{g}^\xi, \tag{4.77}$$

on obtient

$$\nabla_T \wedge \nabla_T h = \frac{1}{a}\mathbf{n} \wedge \left(\frac{\partial h}{\partial \xi}\mathbf{g}^\xi + \frac{\partial h}{\partial \eta}\mathbf{g}^\eta\right) = \frac{1}{a}\mathbf{n} \wedge \nabla_T h. \tag{4.78}$$

La seconde formule est obtenue en effectuant le produit scalaire par $\mathbf{n}$. □

Soit une fonction $h : \mathbb{S}_a^2 \to \mathbb{R}$ quelconque, on évalue numériquement la précision de la proposition 4.7. Dans la direction normale $\mathbf{n}$, on a :

$$\mathrm{rot}\left(\nabla_T h\right) = \left(\nabla_T \wedge \nabla_T h\right) \cdot \mathbf{n} = 0. \tag{4.79}$$

La convergence associée à ce résultat est donnée dans la Table 4.10 où l'on donne

$$e_s = \frac{1}{\sqrt[l]{4\pi a^2}}\|\mathrm{rot}_\Delta\left(\nabla_{T,\Delta} h^*\right) - \mathrm{rot}\left(\nabla_T h\right)^* \|_l \tag{4.80}$$

si $l = 1, 2$, et

$$e_\infty = \|\mathrm{rot}_\Delta\left(\nabla_{T,\Delta} h^*\right) - \mathrm{rot}\left(\nabla_T h\right)^* \|_\infty \tag{4.81}$$

en fonction de N, le paramètre du maillage. La fonction h choisie est

$$h(\lambda, \theta) = \cos^5(\theta)\sin(30\lambda). \tag{4.82}$$

avec (λ, θ) les coordonnées longitudes et latitudes. La présence de $\cos^5(\theta)$ permet d'éviter les problèmes de singularités aux pôles. Le terme $\sin(30\lambda)$ introduit des oscillations importantes dans la direction équatoriale ce qui rend la fonction difficile à représenter sur un maillage trop grossier.

La table 4.10 permet d'illustrer la convergence du gradient et de l'opérateur rotationnel. L'ordre de convergence est proche de 4 dès que le maillage est suffisamment fin. Si on considère la convergence avec un maillage plus grossier, la fonction h est mal représentée.

N	**norme 1**	**norme 2**	**norme ∞**
8	1.3263(−15)	2.2772(−15)	6.6711(−15)
16	1.0786(−13)	1.4654(−13)	4.9935(−13)
32	1.5870(−14)	2.0840(−14)	4.5428(−14)
64	8.5848(−16)	1.0911(−15)	2.6199(−15)
128	5.0055(−17)	6.3036(−17)	1.3325(−16)
256	3.0563(−18)	3.8445(−18)	9.4033(−18)
ordre estimé (1)	2.32	2.40	2.49
ordre estimé (2)	3.85	3.88	3.98

TABLE 4.10 – Table de convergence $\mathrm{rot}_\Delta\left(\nabla_{T,\Delta} h\right)$ avec $h(\lambda,\theta) = \cos^5(\theta)\sin(30\lambda)$ pour différentes normes. On donne en (1) l'ordre estimé prenant en compte tous les maillages et en (2) celui ne prenant pas en compte $N = 8$ car h est difficile à représenter sur un maillage grossier.

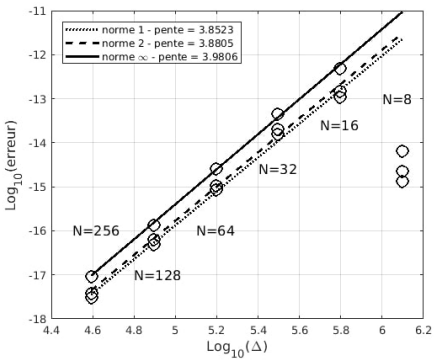


FIGURE 4.11 – Convergence $\mathrm{rot}_\Delta\left(\nabla_{T,\Delta} h\right)$ avec $h(\lambda,\theta) = \cos^5(\theta)\sin(30\lambda)$ et différentes normes. Pour $N = 8$, on constate que la fonction h est mal représentée. Dès que $N = 16$, la précision numérique de la relation $\mathrm{rot}(\nabla_T h) = 0$ est excellente sur cet exemple.

4.5 Opérateur de filtrage

4.5.1 Définition des opérateurs de filtrage

Les opérateurs de filtrages en ξ et en η, respectivement notés $\tilde{\mathcal{F}}_\xi$ et $\tilde{\mathcal{F}}_\eta$, sont construits de manière très similaire aux opérateurs de dérivations donnés par les algorithmes 7 et 8 en remplaçant l'opérateur hermitien $\delta^H_{4,\xi}$ (ou $\delta^H_{4,\eta}$) par un opérateur de filtrage $\mathcal{F}_{2J,\xi}$ (ou $\mathcal{F}_{2J,\eta}$) définis par (1.163).

Soit $h : \mathbf{x} \in \mathbb{S}^2_a \mapsto h(\mathbf{x}) \in \mathbb{R}$ une fonction régulière. L'opérateur de filtrage $\tilde{\mathcal{F}}_\xi$ agit sur les fonctions de grille grâce à un algorithme similaire à celui utilisé pour la dérivation (algorithme 7). On remplace, sur chaque grand cercle l'opérateur de dérivation hermitien $\delta^H_{4,\xi}$ par un opérateur de filtrage $\mathcal{F}_{2J,\xi}$. On définit $\tilde{\mathcal{F}}_\xi$ l'opérateur de filtrage dans la direction ξ par l'algorithme 9.

Algorithme 9 : Calcul de $\tilde{\mathcal{F}}_\xi(h)^{(I)}_{i,j}$ et $\tilde{\mathcal{F}}_\xi(h)^{(III)}_{i,j}$

for $j = -N/2, \dots, N/2,$ **do**
  pour un grand cercle $C^{(2)}_j$ fixé,
  for $p = 0, 1, \dots 4N-1$ définir les points $\mathbf{m}_p$, **do**
    $\mathbf{m}_p = \mathbf{m}^{(I)}_{p-N/2,j}$ pour $0 \leq p \leq N$ donc $h_p = h(\mathbf{m}_p)$,
    $h_p = P_{C_p}(\mathbf{m}_p)$, où P_{C_p} est la fonction d'interpolation utilisant les points de l'isoligne $\xi = \xi^{(II)}_{p-N/2+1}$ pour $N+1 \leq p \leq 2N-1$,
    $\mathbf{m}_p = \mathbf{m}^{(III)}_{p-3N/2,j}$ pour $2N \leq p \leq 3N-1$ donc $h_p = h(\mathbf{m}_p)$,
    $h_p = P_{C_p}(\mathbf{m}_p)$, où P_{C_p} est la fonction d'interpolation utilisant les points de l'isoligne $\xi = \xi^{(VI)}_{p-3N/2+1}$ pour $3N+1 \leq p \leq 4N-1$.
  end for
  Calcul de $\mathcal{F}_{2J,\xi}(h_p)$,
  Affectation $\tilde{\mathcal{F}}_\xi(h^{(I)}_{\xi,i,j}) = \mathcal{F}_{2J,\xi}(h)_{i+N/2}$ pour $-N/2 \leq i \leq N/2$,
  Affectation $\tilde{\mathcal{F}}_\xi(h^{(III)}_{\xi,i,j}) = \mathcal{F}_{2J,\xi}(h)_{i+3N/2}$ pour $-N/2 \leq i \leq N/2$.
end for

$\mathcal{F}_{2J,\xi}$ désigne un opérateur de filtrage 1D de la forme (1.163). Dans la direction de η, on définit l'opérateur de filtrage $\tilde{\mathcal{F}}_\eta$ agissant sur les fonctions de grille sur la Cubed-Sphere. L'algorithme 10 décrit l'opérateur $\tilde{\mathcal{F}}_\eta$:

Algorithme 10 : Calcul de $\tilde{\mathcal{F}}_\eta(h)^{(I)}_{i,j}$ et $\tilde{\mathcal{F}}_\xi(h)^{(III)}_{i,j}$

for $i = -N/2, \dots, N/2,$ **do**
  pour un grand cercle $C^{(1)}_i$ fixé,
  for $p = 0, 1, \dots 4N-1$ définir les points $\mathbf{m}_p$, **do**
    $\mathbf{m}_p = \mathbf{m}^{(I)}_{i,p-N/2}$ pour $0 \leq p \leq N$ donc $h_p = h(\mathbf{m}_p)$,
    $h_p = P_{C_p}(\mathbf{m}_p)$, où P_{C_p} est la fonction d'interpolation utilisant les points de l'isoligne $\eta = \eta^{(II)}_{p-N/2+1}$ pour $N+1 \leq p \leq 2N-1$,
    $\mathbf{m}_p = \mathbf{m}^{(III)}_{i,p-3N/2}$ pour $2N \leq p \leq 3N-1$ donc $h_p = h(\mathbf{m}_p)$,
    $h_p = P_{C_p}(\mathbf{m}_p)$, où P_{C_p} est la fonction d'interpolation utilisant les points de l'isoligne $\eta = \eta^{(VI)}_{p-3N/2+1}$ pour $3N+1 \leq p \leq 4N-1$.
  end for
  Calcul de $\mathcal{F}_{2J,\eta}(h_p)$,
  Affectation $\tilde{\mathcal{F}}_\eta(h^{(I)}_{i,j}) = \mathcal{F}_{2J,\eta}(h)_{j+N/2}$ pour $-N/2 \leq j \leq N/2$,
  Affectation $\tilde{\mathcal{F}}_\eta(h^{(III)}_{i,j}) = \mathcal{F}_{2J,\eta}(h)_{j+3N/2}$ pour $-N/2 \leq j \leq N/2$.
end for

Les opérateurs $\tilde{\mathcal{F}}_\xi$ et $\tilde{\mathcal{F}}_\eta$ définissent des perturbations de l'identité.

Proposition 4.8. *Pour toute fonction $h : \mathbf{x} \in \mathbb{S}_a^2 \mapsto h(\mathbf{x}) \in \mathbb{R}$ régulière, on a*

$$\tilde{\mathcal{F}}_\xi(h^*)_{i,j}^{(k)} = h(\mathbf{x}_{i,j}^{(k)}) + \mathcal{O}\left(\Delta\xi^{\min(4,2J)}\right), \tag{4.83}$$

et dans la direction η :

$$\tilde{\mathcal{F}}_\eta(h^*)_{i,j}^{(k)} = h(\mathbf{x}_{i,j}^{(k)}) + \mathcal{O}\left(\Delta\eta^{\min(4,2J)}\right), \tag{4.84}$$

pour tout $-N/2 \leq i,j \leq N/2$ et $(k) = (I), \cdots, (VI)$. L'ordre choisi pour l'opérateur de filtrage 1D est $2J$ et $\Delta\xi = \Delta\eta$ désigne le pas du maillage le long de l'équateur.

Démonstration. Par symétrie, on démontre seulement l'équation (4.83). Après la procédure d'interpolation, on a

$$\begin{cases} h_p &= h(\mathbf{m}_p) & \text{si } \mathbf{m}_p \text{ est un point de (I) ou (III),} \\ h_p &= h(\mathbf{m}_p) + \mathcal{O}(\Delta\eta^4) & \text{si } \mathbf{m}_p \text{ est un point de (II) ou (IV),} \end{cases} \tag{4.85}$$

sur chaque grand cercle. De plus, on a, après filtrage

$$\mathcal{F}_{2J,\xi}(h_p) = h_p + \mathcal{O}(\Delta\xi^{2J}), \tag{4.86}$$

donc

$$\begin{aligned} \mathcal{F}_{2J,\xi}(h_p) &= h(\mathbf{m}_p) + \mathcal{O}(\Delta\xi^{2J}) + \mathcal{O}(\Delta\eta^4) \\ &= h(\mathbf{m}_p) + \mathcal{O}\left(\Delta\xi^{\min(4,2J)}\right). \end{aligned}$$

En particulier, si $\mathbf{m}_p$ est un point du panel (I) ou (III), on obtient

$$\tilde{\mathcal{F}}_\xi(h^*)_{i,j}^{(k)} = h(\mathbf{x}_{i,j}^{(k)}) + \mathcal{O}\left(\Delta\xi^{\min(4,2J)}\right). \tag{4.87}$$

□

On souhaite filtrer dans les directions ξ et η sur chaque panel. Si h est une fonction sur la Cubed-Sphere, on calcule $\tilde{\mathcal{F}}_\xi(h^*)$ le filtrage de h^* dans la direction ξ. Enfin on calcule la version filtrée en η : $\tilde{\mathcal{F}}_\eta(\tilde{\mathcal{F}}_\xi(h^*))$, on compose les filtres $\tilde{\mathcal{F}}_\xi$ et $\tilde{\mathcal{F}}_\eta$. Par composition, en tout point de la Cubed-Sphere, on a

Proposition 4.9. *Pour toute fonction $h : \mathbf{x} \in \mathbb{S}_a^2 \mapsto h(\mathbf{x}) \in \mathbb{R}$ régulière, on a*

$$\tilde{\mathcal{F}}_\eta(\tilde{\mathcal{F}}_\xi(h^*)_{i,j}^{(k)}) = h(\mathbf{x}_{i,j}^{(k)}) + \mathcal{O}\left(\Delta\xi^{\min(4,2J)}\right), \tag{4.88}$$

ainsi que

$$\tilde{\mathcal{F}}_\xi(\tilde{\mathcal{F}}_\eta(h^*)_{i,j}^{(k)}) = h(\mathbf{x}_{i,j}^{(k)}) + \mathcal{O}\left(\Delta\xi^{\min(4,2J)}\right), \tag{4.89}$$

pour tout $-N/2 \leq i,j \leq N/2$ et $(k) = (I), \cdots, (VI)$, $\Delta\xi = \Delta\eta$.

Démonstration. D'après la proposition 4.8, on a

$$\begin{aligned} \tilde{\mathcal{F}}_\xi(\tilde{\mathcal{F}}_\eta(h^*))_{i,j}^{(k)} - h(\mathbf{x}_{i,j}^{(k)}) &= \tilde{\mathcal{F}}_\xi(\tilde{\mathcal{F}}_\eta(h^*))_{i,j}^{(k)} - \tilde{\mathcal{F}}_\eta(h^*)_{i,j}^{(k)} + \tilde{\mathcal{F}}_\eta(h^*)_{i,j}^{(k)} - h(\mathbf{x}_{i,j}^{(k)}) \\ &= \tilde{\mathcal{F}}_\xi(\tilde{\mathcal{F}}_\eta(h^*))_{i,j}^{(k)} - \tilde{\mathcal{F}}_\eta(h^*)_{i,j}^{(k)} + \mathcal{O}\left(\Delta\xi^{\min(4,2J)}\right) \text{ d'après (4.83)} \\ &= \mathcal{O}\left(\Delta\xi^{\min(4,2J)}\right) \text{ d'après (4.84),} \end{aligned}$$

d'où (4.88). De la même manière, on montre (4.89). □

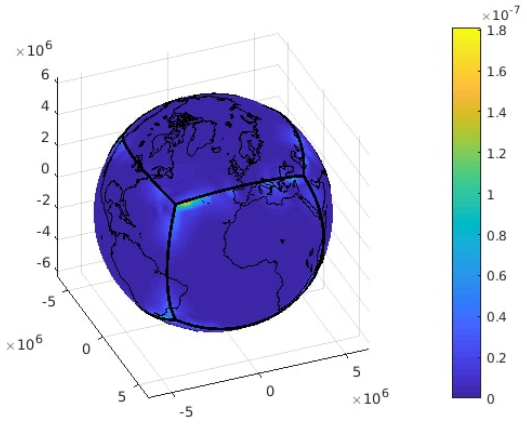


FIGURE 4.12 – Tracé de $\tilde{\mathcal{F}}_\xi(\tilde{\mathcal{F}}_\eta(h^*)) - \tilde{\mathcal{F}}_\eta(\tilde{\mathcal{F}}_\xi(h^*))$ avec h donné par (4.90) et $N = 16$.

Noter que les opérateurs $\tilde{\mathcal{F}}_\xi$ et $\tilde{\mathcal{F}}_\eta$ ne commutent pas. En effet si

$$h(x, y, z) = \exp\left(\frac{x}{a}\right) + \exp\left(\frac{y}{a}\right) + \exp\left(\frac{z}{a}\right) \tag{4.90}$$

pour tout $(x, y, z) \in \mathbb{S}^2_a$, on représente sur la Figure 4.12, la fonction de grille

$$\tilde{\mathcal{F}}_\xi(\tilde{\mathcal{F}}_\eta(h^*)) - \tilde{\mathcal{F}}_\eta(\tilde{\mathcal{F}}_\xi(h^*)). \tag{4.91}$$

On observe numériquement qu'elle est non nulle. On définit donc l'opérateur de filtrage $\mathcal{F}$ agissant sur les fonctions de grilles sur la Cubed-Sphere de façon symétrique par

$$\mathcal{F} = \frac{1}{2}\left(\tilde{\mathcal{F}}_\xi \circ \tilde{\mathcal{F}}_\eta + \tilde{\mathcal{F}}_\eta \circ \tilde{\mathcal{F}}_\xi\right). \tag{4.92}$$

L'opérateur $\mathcal{F}$ est une perturbation de l'identité au sens de la proposition suivante :

Proposition 4.10. *Pour toute fonction $h : \mathbf{x} \in \mathbb{S}^2_a \mapsto h(\mathbf{x}) \in \mathbb{R}$ régulière, on a*

$$\mathcal{F}(h^*)^{(k)}_{i,j} = h(\mathbf{x}^{(k)}_{i,j}) + \mathcal{O}\left(\Delta\xi^{\min(4,2J)}\right), \tag{4.93}$$

pour tout $-N/2 \leq i, j \leq N/2$ et $(k) = (I), \cdots, (VI)$, $\Delta\xi = \Delta\eta$.

Démonstration. D'après la proposition 4.9, chacun des opérateurs $\tilde{\mathcal{F}}_\xi \circ \tilde{\mathcal{F}}_\eta$ et $\tilde{\mathcal{F}}_\eta \circ \tilde{\mathcal{F}}_\xi$ est une perturbation de l'identité. Par moyenne, on déduit l'équation (4.93) des équations (4.88) et (4.89). □

L'opérateur de filtrage $\mathcal{F}$ filtre les données indépendamment dans la direction ξ et dans la direction η. Il dépend de l'ordre $2J$ choisi pour l'opérateur de filtrage de dimension 1 dans les algorithmes 9 et 10.

4.5.2 Résultats numériques pour l'opérateur de filtrage

Dans cette section, on évalue numériquement l'opérateur de filtrage $\mathcal{F}$ donné par (4.92). Pour cela, on mesure l'erreur

$$e_l = \frac{\|\mathcal{F}(h^*) - h^*\|_l}{\|h^*\|_l} \tag{4.94}$$

pour différentes fonctions h définies sur la sphère. On rappelle que $*$ désigne l'opérateur de restriction d'une fonction à la Cubed-Sphere. On considère les trois fonctions suivantes :

- **Fonction polynomiale :**

$$h(x,y,z) = \left(\frac{x}{a}\right)^2 \left(\frac{y}{a}\right)^3 \left(\frac{z}{a}\right)^4, \tag{4.95}$$

avec $(x,y,z) \in \mathbb{S}_a^2$,

- **Fonction de type exponentielle :**

$$h(x,y,z) = \exp\left(\frac{x}{a}\right) + \exp\left(\frac{y}{a}\right) + \exp\left(\frac{z}{a}\right) \tag{4.96}$$

avec $(x,y,z) \in \mathbb{S}_a^2$,

- **Fonction oscillante :**

$$h(\lambda,\theta) = \cos^5(\theta)\sin(30\lambda) \tag{4.97}$$

avec (λ,θ) les coordonnées d'un point dans le système longitude-latitude.

On compare, sur la Table 4.11, la valeur, pour différents maillages, de l'erreur relative en norme 1, en norme 2 et en norme infinie. On observe un ordre de convergence supérieur à 4. Les droites de régressions sont de pente supérieure à 4 comme on l'observe sur la Figure 4.13.

N	**norme 1**	**norme 2**	**norme ∞**
32	6.1250(−11)	2.0702(−10)	3.3071(−9)
64	1.7509(−12)	9.3707(−12)	2.1368(−10)
128	5.3780(−14)	4.2192(−13)	1.3738(−11)
256	1.7839(−15)	1.8822(−14)	8.6376(−13)
512	1.7228(−16)	8.5616(−16)	5.4099(−14)
ordre estimé	4.68	4.47	3.98

TABLE 4.11 – Table de convergence pour le filtre $\mathcal{F}$ utilisant le filtrage d'ordre 10 pour la fonction (4.96).

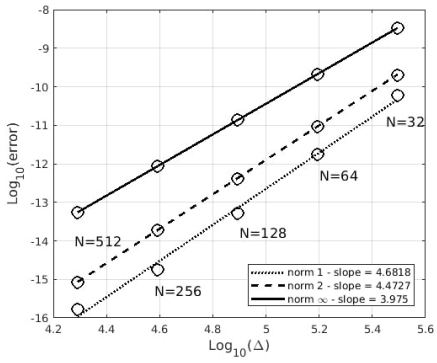


FIGURE 4.13 – Taux de convergence pour le filtre $\mathcal{F}$ utilisant le filtrage d'ordre 10 pour la fonction (4.96) en fonction de $\Delta = a\Delta\xi$.

Dans la Figure 4.14, on trace la fonction (4.96) ainsi que la fonction de grille

$$\frac{|\mathcal{F}(h^*) - h^*|}{\|h^*\|_\infty} \tag{4.98}$$

associée aux filtres d'ordre 2, 4, 6, 8 et 10. On constate que plus l'ordre du filtre est élevé moins la fonction est affectée par l'opérateur de filtrage $\mathcal{F}$. Pour les filtres d'ordre 8 et 10, on constate que

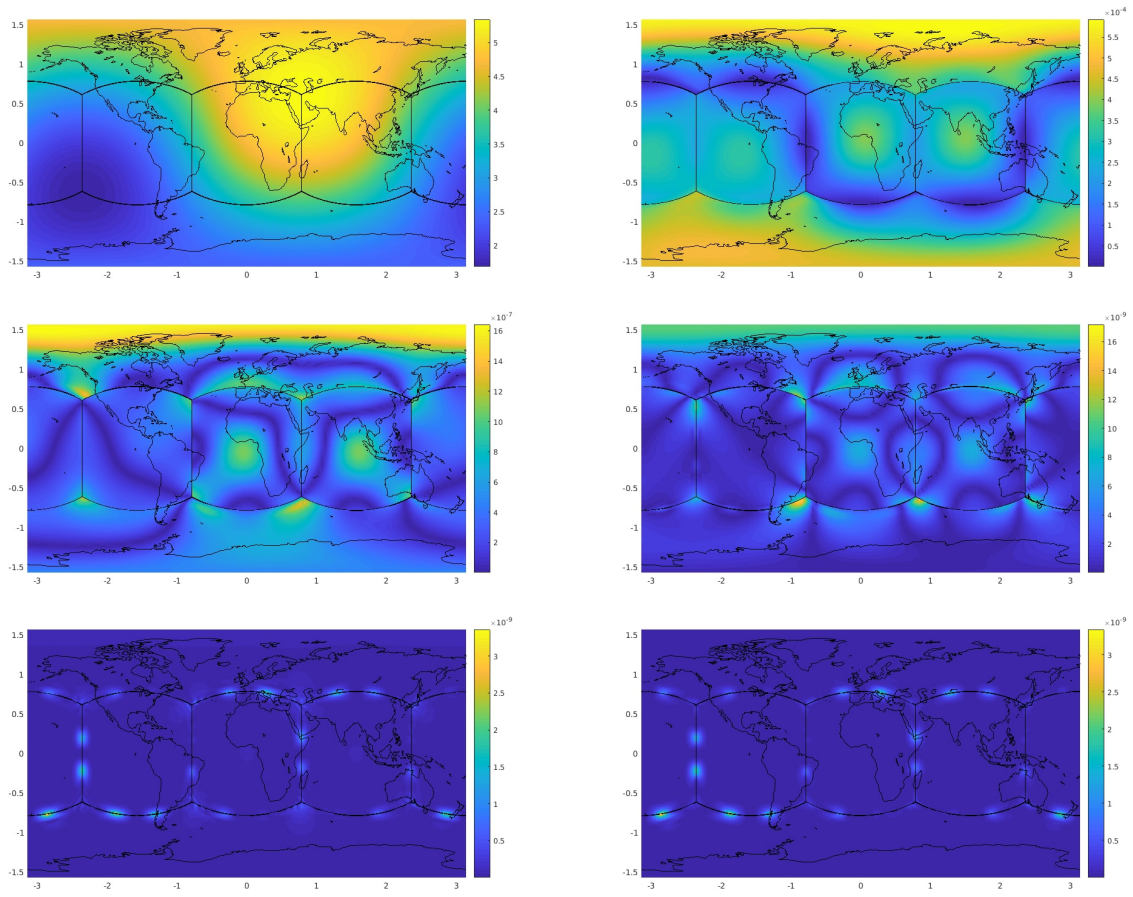


FIGURE 4.14 – De haut en bas et de gauche à droite, erreur (4.98) pour la fonction (4.96) associée à des filtrages 1D d'ordres 2, 4, 6, 8 et 10 pour $N = 32$. La première figure représente la fonction à filtrer.

Filtre : 2J	Fonction test	norme 1	norme 2	norme ∞
ordre 2	(4.95)	2.0767(−2)	2.0663(−2)	2.6936(−2)
	(4.96)	3.8365(−4)	3.8031(−4)	5.9219(−4)
	(4.97)	8.8026(−1)	8.3750(−1)	8.2422(−1)
ordre 4	(4.95)	4.4489(−4)	4.0346(−4)	5.5585(−4)
	(4.96)	5.8873(−7)	7.1369(−7)	1.6257(−6)
	(4.97)	2.7916(−1)	2.6217(−1)	2.5533(−1)
ordre 6	(4.95)	1.1255(−5)	9.7115(−6)	1.5661(−5)
	(4.96)	3.3056(−9)	4.3617(−9)	1.7010(−8)
	(4.97)	1.1461(−1)	1.0549(−1)	1.0099(−1)
ordre 8	(4.95)	4.2370(−7)	3.8887(−7)	1.3487(−6)
	(4.96)	8.4402(−11)	2.1848(−10)	3.3268(−9)
	(4.97)	5.2179(−2)	4.6915(−2)	4.3154(−2)
ordre 10	(4.95)	1.2671(−7)	2.5518(−7)	1.3316(−6)
	(4.96)	6.1250(−11)	2.0702(−10)	3.3071(−9)
	(4.97)	2.5576(−2)	2.2919(−2)	2.8680(−2)

TABLE 4.12 – Erreur relative pour différentes normes avec $N = 32$. La fonction (4.96) est la moins affectée par l'opérateur de filtrage. Les fonctions (4.95) et (4.97) sont beaucoup plus affectées, en particulier la fonction (4.97) oscille beaucoup et est filtrée de manière importante quel que soit l'ordre du filtre utilisé. Comme on s'y attendait, plus l'ordre du filtre est bas, plus la fonction est affectée par l'opérateur de filtrage.

l'erreur est particulièrement présente sur les bords de la Cubed-Sphere là où l'opérateur d'interpolation est le moins précis. C'est donc l'opérateur d'interpolation qui est ici la principale source d'erreur.

La Table 4.12 permet de confirmer que plus l'ordre du filtre est élevé, moins la fonction est affectée par l'opérateur de filtrage. La fonction (4.97) est fortement affectée par l'opérateur de filtrage, en effet la fonction oscille beaucoup et contient beaucoup de hautes fréquences.

En utilisant les mêmes grilles que sur la Table 4.11 et sur la Figure 4.13, on évalue l'ordre de convergence. Les résultats sont donnés pour la fonction (4.95) sur la Table 4.13, pour la fonction (4.96) sur la Table 4.14 et pour la fonction (4.97) sur la Table 4.15.

Ordre du filtre : $2J$	2	4	6	8	10
norme 1	2.00	4.00	5.91	4.81	4.68
norme 2	2.00	4.00	5.57	4.48	4.47
norme ∞	2.00	3.97	4.49	3.98	3.98

TABLE 4.13 – Ordre de convergence pour différents ordres des filtres $\mathcal{F}_{2J,\xi}$ et $\mathcal{F}_{2J,\eta}$ et pour la fonction (4.95). On utilise les paramètres de grille $N = 32$, $N = 64$, $N = 128$, $N = 256$ et $N = 512$.

Ordre du filtre : $2J$	**2**	**4**	**6**	**8**	**10**
norme 1	2.00	3.99	5.96	5.53	5.12
norme 2	2.00	4.00	5.82	4.69	4.57
norme ∞	2.00	4.00	4.92	4.05	4.04

TABLE 4.14 – Ordre de convergence pour différents ordres des filtres $\mathcal{F}_{2J,\xi}$ et $\mathcal{F}_{2J,\eta}$ et pour la fonction (4.96). On utilise les paramètres de grille $N = 32$, $N = 64$, $N = 128$, $N = 256$ et $N = 512$.

Ordre du filtre : $2J$	**2**	**4**	**6**	**8**	**10**
norme 1	2.12	3.92	5.67	6.66	6.80
norme 2	2.12	3.94	5.56	5.86	5.84
norme ∞	2.11	3.91	4.79	4.75	4.75

TABLE 4.15 – Ordre de convergence pour différents ordres des filtres $\mathcal{F}_{2J,\xi}$ et $\mathcal{F}_{2J,\eta}$ et pour la fonction (4.97). On utilise les paramètres de grille $N = 32$, $N = 64$, $N = 128$, $N = 256$ et $N = 512$.

Chapitre 5

Equations d'advection sphériques

5.1 Equations d'advection linéaires sur la sphère

Dans cette section, on s'intéresse à l'équation d'advection (5.1) :

$$\left\{ \begin{array}{rcl} \dfrac{\partial h}{\partial t} + \mathbf{c}(t,\mathbf{x}) \cdot \nabla_T h & = & 0 \\ h(0,\mathbf{x}) & = & h_0(\mathbf{x}) \end{array} \right. \quad \text{pour tout } \mathbf{x} \in \mathbb{S}^2_a \text{ et } t \geq 0, \tag{5.1}$$

sur la sphère $\mathbb{S}^2_a$. Le rayon terrestre est $a = 6371220$m. La fonction $\mathbf{c} : (t,\mathbf{x}) \in \mathbb{R}^+ \times \mathbb{S}^2_a \mapsto \mathbf{c}(t,\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}^2_a$ désigne un champ de vecteurs tangents à la sphère $\mathbb{S}^2_a$.

5.1.1 Résolution numérique

L'équation (5.1) est résolue par la méthode des lignes en utilisant l'opérateur gradient discret $\nabla_{T,\Delta}$ (définition 4.4). On pose J_Δ l'application agissant sur une fonction de grille $\mathfrak{h}$ donnée au temps t par

$$J_\Delta(t,\mathfrak{h}) = -\mathbf{c}(t,\cdot)^* \cdot \nabla_{T,\Delta}\mathfrak{h}. \tag{5.2}$$

En chaque point $\mathbf{x}^{(k)}_{i,j}$ de la Cubed-Sphere, on a

$$J_\Delta(t,\mathfrak{h})^{(k)}_{i,j} = -\mathbf{c}(t,\mathbf{x}^{(k)}_{i,j}) \cdot (\nabla_{T,\Delta}\mathfrak{h})^{(k)}_{i,j} \tag{5.3}$$

pour tous $-N/2 \leq i,j \leq N/2$ et $(k) = (I),...,(VI)$. La résolution en temps se fait par un algorithme de type RK4 couplé à un opérateur de filtrage $\mathcal{F}$ donné par (4.92). Cet algorithme est analogue à l'algorithme 4. Il est donné par :

Algorithme 11 : Schéma en temps RK4 avec étape de filtrage pour l'équation (5.1)

1: $h^0 = h_0^*$ connu,
2: **for** $n = 0, 1, \ldots$ **do**
3: $\quad K^{(1)} = J_\Delta(t^n, \mathfrak{h}^n)$,
4: $\quad K^{(2)} = J_\Delta\left(t^n + \frac{\Delta t}{2}, \mathfrak{h}^n + \frac{\Delta t}{2}K^{(1)}\right)$,
5: $\quad K^{(3)} = J_\Delta\left(t^n + \frac{\Delta t}{2}, \mathfrak{h}^n + \frac{\Delta t}{2}K^{(2)}\right)$,
6: $\quad K^{(4)} = J_\Delta\left(t^n + \Delta t\mathfrak{h}^n + \Delta t K^{(3)}\right)$,
7: $\quad \mathfrak{h}^{n+1} = \mathcal{F}\left(\mathfrak{h}^n + \dfrac{\Delta t}{6}\left(K^{(1)} + 2K^{(2)} + 2K^{(3)} + K^{(4)}\right)\right)$.
8: **end for**

La donnée $\mathfrak{h}^n$ désigne une approximation de $(h(t^n,\cdot))^*$ solution au temps $t^n = n\Delta t$ de (5.1). La solution de l'équation (5.1) étant connue dans les tests effectués, nous mesurons l'erreur relative au temps t^n par

$$e_l^n = \frac{\|\mathfrak{h}^n - h(t^n,\cdot)^*\|_l}{\|h(t^n,\cdot)^*\|_l} \tag{5.4}$$

où $l \in \{1,2,\infty\}$ et $\|\cdot\|_l$ désigne la norme 1, 2 ou ∞ calculée par

$$\|\mathfrak{q}\|_l = \Big(Q(|\mathfrak{q}|^l)\Big)^{1/l} \text{ avec } l = 1,2 \tag{5.5}$$

et Q un opérateur de quadrature numérique introduit dans [70], $Q = Q_{1/3}$. Pour la norme $\|\cdot\|_\infty$, on note

$$\|\mathfrak{q}\|_\infty = \max_{-N/2\le i,j\le N/2}\ \max_{(k)=(I)\ldots(VI)} |\mathfrak{q}_{i,j}^{(k)}|, \tag{5.6}$$

pour $\mathfrak{q}$ fonction de grille.

Dans les simulations numériques effectuées, le pas de temps Δt est proportionnel à $\Delta\xi$. La relation de proportionnalité est donnée par la condition

$$\mathrm{CFL} = \frac{u_0\Delta t}{a\Delta\xi} = C^{\mathrm{ste}}, \tag{5.7}$$

la valeur de u_0 est donnée par le contexte lors du test effectué.

5.1.2 Test de rotation solide

On considère d'abord le test numéro 1 de [85]. Il s'agit d'une rotation sans déformation de la condition initiale au cours du temps autour d'un axe incliné.

On considère (λ,θ) les coordonnées longitude-latitude associées au pôle Nord noté $\mathbf{N}$ et (λ',θ') les coordonnées longitude-latitude associées à un pôle Nord déplacé en $\mathbf{P}$ de coordonnées (λ_P,θ_P). La proposition suivante énonce le lien entre (λ,θ) et (λ',θ') :

Proposition 5.1. *Le changement de variables $(\lambda,\theta)\mapsto(\lambda',\theta')$ est donné par :*

$$\left\{\begin{array}{rcl} \theta' &=& \arcsin\left[\sin(\theta)\sin(\theta_P)+\cos(\theta)\cos(\theta_P)\cos(\lambda-\lambda_P)\right] \\ \lambda' &=& \arctan\left[\dfrac{\cos(\theta)\sin(\lambda-\lambda_P)}{\cos(\theta)\cos(\lambda-\lambda_P)\sin(\theta_P)-\sin(\theta)\cos(\theta_P)}\right]. \end{array}\right. \tag{5.8}$$

Le changement de variables inverse $(\lambda',\theta')\mapsto(\lambda,\theta)$ est donné par :

$$\left\{\begin{array}{rcl} \theta &=& \arcsin\left[\sin(\theta')\sin(\theta_P)-\cos(\theta')\cos(\theta_P)\cos(\lambda')\right] \\ \lambda &=& \lambda_P+\arctan\left[\dfrac{\cos(\theta')\sin(\lambda')}{\sin(\theta')\cos(\theta_P)+\cos(\theta')\cos(\lambda')\sin(\theta_P)}\right]. \end{array}\right. \tag{5.9}$$

Démonstration. Un point $\mathbf{x}\in\mathbb{S}_a^2$ a pour coordonnées (λ,θ) en longitude-latitude associées au pôle Nord et (λ',θ') en coordonnées latitude-longitude associées à un pôle déplacé en P. Le lien entre ces coordonnées se fait par rotations successives. En considérant les rotations liées au changement de pôle Nord, on a

$$\begin{bmatrix}\cos\theta'\cos\lambda'\\ \cos\theta'\sin\lambda'\\ \sin\theta'\end{bmatrix} = \begin{bmatrix}\cos(\theta_P-\pi/2) & 0 & \sin(\theta_p-\pi/2)\\ 0 & 1 & 0\\ -\sin(\theta_P-\pi/2) & 0 & \cos(\theta_P-\pi/2)\end{bmatrix}\begin{bmatrix}\cos\lambda_P & \sin\lambda_P & 0\\ -\sin\lambda_P & \cos\lambda_P & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\cos\theta\cos\lambda\\ \cos\theta\sin\lambda\\ \sin\theta\end{bmatrix} \tag{5.10}$$

$$= \begin{bmatrix}\sin\theta_P\cos\lambda_P & \sin\theta_P\sin\lambda_P & -\cos\theta_P\\ -\sin\lambda_P & \cos\lambda_P & 0\\ \cos\theta_P\cos\lambda_P & \cos\theta_P\sin\lambda_P & \sin\theta_P\end{bmatrix}\begin{bmatrix}\cos\theta\cos\lambda\\ \cos\theta\sin\lambda\\ \sin\theta\end{bmatrix}. \tag{5.11}$$

La seconde ligne de (5.11) donne

$$\begin{aligned}\cos\theta' \in \lambda' &= -\sin\lambda_P\cos\theta\cos\lambda + \cos\lambda_P\cos\theta\sin\lambda \\ &= \cos\theta(\cos\lambda_P\sin\lambda - \sin\lambda_P\cos\lambda) \\ &= \cos\theta\sin(\lambda-\lambda_P)\end{aligned}$$

ce qui donne l'équation :

$$\cos\theta'\sin\lambda' = \cos\theta\sin(\lambda-\lambda_P). \tag{5.12}$$

De la même manière, la troisième ligne de (5.11) permet d'obtenir

$$\sin\theta' = \cos\theta_P\cos\theta\cos(\lambda-\lambda_P) + \sin\theta_P\sin\theta. \tag{5.13}$$

D'après la première ligne de (5.11) :

$$\begin{aligned}\cos\theta'\cos\lambda' &= \sin\theta_P\cos\lambda_P\cos\theta\cos\lambda + \sin\theta_P\sin\lambda_P\cos\theta\sin\lambda - \cos\theta_P\sin\theta \\ &= \sin\theta_P\cos\theta\cos(\lambda-\lambda_P) - \cos\theta_P\sin\theta \\ &= \sin\theta_P\frac{\sin\theta'-\sin\theta_P\sin\theta}{\cos\theta_P} - \cos\theta_P\sin\theta \text{ d'après (5.13)} \\ &= \frac{\sin\theta_P\sin\theta'-\sin\theta}{\cos\theta_P}.\end{aligned}$$

D'où une troisième équation

$$\sin\theta = \sin\theta_P\sin\theta' - \cos\theta'\cos\lambda'\cos\theta_P. \tag{5.14}$$

Les équations démontrées sont les suivantes :

$$\left\{\begin{array}{rcl}\sin(\theta) &=& \sin(\theta_P)\sin(\theta') - \cos(\theta_P)\cos(\theta')\cos(\lambda') \\ \sin(\theta') &=& \sin(\theta)\sin(\theta_P) + \cos(\theta)\cos(\theta_P)\cos(\lambda-\lambda_P) \\ \cos(\theta)\sin(\lambda-\lambda_P) &=& \cos(\theta')\sin(\lambda').\end{array}\right. \tag{5.15}$$

En utilisant (5.15.b), la formule suivante est immédiate :

$$\theta' = \arcsin\left[\sin(\theta)\sin(\theta_P) + cos(\theta)cos(\theta_P)cos(\lambda-\lambda_P)\right]. \tag{5.16}$$

De plus, (5.15.a) et (5.15.c) donnent :

$$\left\{\begin{array}{rcl}\cos(\theta)\sin(\lambda-\lambda_P) &=& \cos(\theta')\sin(\lambda') \\ \cos(\theta)\cos(\lambda-\lambda_P) &=& \frac{\sin(\theta')\sin(\theta_P)-\sin(\theta)}{\cos(\theta_P)}.\end{array}\right. \tag{5.17}$$

Or :

$$\begin{aligned}\cos(\theta)\cos(\lambda-\lambda_P) &= \frac{\sin(\theta')\sin(\theta_P)-\sin(\theta)}{\cos(\theta_P)} \\ &= \frac{\sin(\theta)(\sin^2(\theta_P)-1)}{\cos(\theta_P)} + \cos(\theta)\cos(\lambda-\lambda_P)\sin(\theta_P) \\ &= \cos(\theta)\cos(\lambda-\lambda_P)\sin(\theta_P) - \sin(\theta)\cos(\theta_P).\end{aligned}$$

En utilisant

$$\tan(\lambda') = \frac{\cos(\theta')\sin(\lambda')}{\cos(\theta')\cos(\lambda')} \tag{5.18}$$

on obtient le changement de coordonnées (5.8) :

$$\left\{\begin{array}{rcl}\theta' &=& \arcsin\left[\sin(\theta)\sin(\theta_P) + \cos(\theta)\cos(\theta_P)\cos(\lambda-\lambda_P)\right] \\ \lambda' &=& \arctan\left[\dfrac{\cos(\theta)\sin(\lambda-\lambda_P)}{\cos(\theta)\cos(\lambda-\lambda_P)\sin(\theta_P) - \sin(\theta)\cos(\theta_P)}\right].\end{array}\right. \tag{5.19}$$

Inversement et par une démonstration similaire, on obtient

$$\left\{\begin{array}{rcl}\theta &=& \arcsin\left[\sin(\theta')\sin(\theta_P)-\cos(\theta')\cos(\theta_P)\cos(\lambda')\right]\\ \lambda &=& \lambda_P+\arctan\left[\dfrac{\cos(\theta')\sin(\lambda')}{\sin(\theta')\cos(\theta_P)+\cos(\theta')\cos(\lambda')\sin(\theta_P)}\right].\end{array}\right. \tag{5.20}$$

□

Proposition 5.2. *La solution de l'équation d'advection sur la sphère* (5.1) *avec*

$$\mathbf{c}(\mathbf{x})=\mathbf{c}(\lambda,\theta)=u_0\cos\theta\mathbf{e}_\lambda \tag{5.21}$$

est donnée pour $t\geq 0$ *par*

$$h(\mathbf{x},t)=h(\lambda,\theta,t)=h_0(\lambda-\omega_s t,\theta) \tag{5.22}$$

avec $u_0=a\omega_s$ *et* $\mathbf{x}$ *un point de la sphère* $\mathbb{S}^2_a$ *de coordonnées longitude-latitude* (λ,θ)*. On a également*

$$h(\mathbf{x},t)=h_0(R_{-t}\mathbf{x}) \tag{5.23}$$

où R_{-t} *est la matrice de rotation*

$$R_{-t}=\begin{pmatrix}\cos(-\omega_s t) & -\sin(-\omega_s t) & 0\\ \sin(-\omega_s t) & \cos(-\omega_s t) & 0\\ 0 & 0 & 1\end{pmatrix}. \tag{5.24}$$

Démonstration. On résout cette équation par la méthode des caractéristiques. Soit $X: t\in\mathbb{R}^+\mapsto X(t)=(\lambda(t),\theta(t))\in\mathbb{S}^2_a$ solution de

$$\left\{\begin{array}{rcl}\dfrac{dX}{dt} &=& \mathbf{c}(X(t))\\ X(0) &=& \mathbf{x}_0=(\lambda_0,\theta_0).\end{array}\right. \tag{5.25}$$

D'après le théorème de Cauchy-Lipschitz, il existe une telle courbe X solution maximale.

Si h est solution de (5.1), h est constante le long de X. En effet

$$\begin{aligned}\frac{dh}{dt}(X(t),t)&=\frac{\partial h}{\partial t}(X(t),t)+\frac{dX}{dt}(t)\cdot\nabla_T h(X(t),t)\\ &=\frac{\partial h}{\partial t}(X(t),t)+\mathbf{c}(X(t))\cdot\nabla_T h(X(t),t)\\ &=0.\end{aligned}$$

En exprimant X en coordonnée latitude-longitude, on obtient la formule

$$\frac{dX}{dt}=a\cos\theta\frac{d\lambda}{dt}\mathbf{e}_\lambda+a\frac{d\theta}{dt}\mathbf{e}_\theta. \tag{5.26}$$

En identifiant les termes dans le problème de Cauchy, on obtient

$$\left\{\begin{array}{rcl}\dfrac{d\lambda}{dt} &=& \omega_s\\ \dfrac{d\theta}{dt} &=& 0\end{array}\right. \tag{5.27}$$

d'où

$$\left\{\begin{array}{rcl}\lambda(t) &=& \omega_s t+\lambda_0\\ \theta(t) &=& \theta_0.\end{array}\right. \tag{5.28}$$

La fonction h est constante le long de la caractéristique $t\mapsto(\lambda(t),\theta(t))$, donc

$$h(\lambda(t),\theta(t),t)=h_0(\lambda_0,\theta_0)=h_0(\lambda-\omega_s t,\theta), \tag{5.29}$$

ce qui termine la preuve. □

Si $(\lambda_P, \theta_P) = (\pi, \pi/2 - \alpha)$, alors la matrice de rotation pour passer d'un système de coordonnées à l'autre est

$$P_\alpha = \begin{bmatrix} -\cos\alpha & 0 & -\sin\alpha \\ 0 & -1 & 0 \\ -\sin\alpha & 0 & \cos\alpha \end{bmatrix} \tag{5.30}$$

et on a le théorème suivant :

Théorème 5.1. *La solution de l'équation* (5.1) *avec*

$$\mathbf{c}_s(\mathbf{x}) = \mathbf{c}(\mathbf{x}) = \mathbf{c}(\lambda, \theta) = u_0 \left(\cos\theta\cos\alpha + \sin\theta\cos\lambda\sin\alpha\right)\mathbf{e}_\lambda - u_0 \sin\lambda\sin\alpha\mathbf{e}_\theta \tag{5.31}$$

est donnée pour $t \geq 0$ *par*

$$h(\mathbf{x}, t) = h_0(P_\alpha^{-1} R_{-t} P_\alpha \mathbf{x}) \tag{5.32}$$

où R_{-t} *est la matrice de rotation*

$$R_{-t} = \begin{bmatrix} \cos(-\omega_s t) & -\sin(-\omega_s t) & 0 \\ \sin(-\omega_s t) & \cos(-\omega_s t) & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.33}$$

, où $\omega_s = u_0/a$ *et* $\mathbf{x}$ *est un point de la sphère* $\mathbb{S}_a^2$.

Démonstration. La rotation P_α est inversible donc $\mathbf{x} \mapsto P_\alpha \mathbf{x}$ réalise une bijection de $\mathbb{S}_a^2$ dans $\mathbb{S}_a^2$.

Soit $g : (\mathbf{x}, t) \in \mathbb{S}_a^2 \times \mathbb{R}^+ \mapsto g(\mathbf{x}, t)$ la solution de

$$\left\{ \begin{array}{rcl} \dfrac{\partial g}{\partial t} + \mathbf{c}_g \cdot \nabla_T g & = & 0 \\ g(\mathbf{x}, 0) & = & h_0(P_\alpha^{-1}\mathbf{x}). \end{array} \right. \tag{5.34}$$

D'après la proposition 5.2, en tout point $\mathbf{x} \in \mathbb{S}_a^2$, on a

$$g(\mathbf{x}, t) = h_0(R_{-t} P_\alpha^{-1} \mathbf{x}). \tag{5.35}$$

En posant $h(\mathbf{x}, t) = g(P_\alpha \mathbf{x}, t)$, alors h est solution de (5.1) avec (5.31), en effet

$$\begin{aligned} \frac{\partial h}{\partial t}(\mathbf{x}, t) + \mathbf{c}(\mathbf{x}) \cdot \nabla_T h(\mathbf{x}, t) &= \frac{\partial g}{\partial t}(P_\alpha \mathbf{x}, t) + P_\alpha \mathbf{c}(\mathbf{x}) \cdot \nabla_T g(P_\alpha \mathbf{x}, t) \\ &= \frac{\partial g}{\partial t}(P_\alpha \mathbf{x}, t) + P\mathbf{c}_g(\mathbf{x}) \cdot \nabla_T g(P_\alpha \mathbf{x}, t) \\ &= \frac{\partial g}{\partial t}(P_\alpha \mathbf{x}, t) + u_0 \cos\theta \mathbf{e}_\lambda \cdot \nabla_T g(P_\alpha \mathbf{x}, t) \\ &= 0, \end{aligned}$$

car g est solution du problème (5.34). De plus, on a bien $h(\mathbf{x}, 0) = g(P_\alpha \mathbf{x}, 0) = h_0(\mathbf{x})$, donc en tout point $\mathbf{x} \in \mathbb{S}_a^2$, on a

$$\begin{aligned} h(\mathbf{x}, t) &= g(P_\alpha \mathbf{x}, t) \\ &= g_0(R_{-t} P_\alpha \mathbf{x}) \\ &= h_0(P_\alpha^{-1} R_{-t} P_\alpha \mathbf{x}) \end{aligned}$$

ce qui conclut la preuve. □

Le test numéro 1 présenté dans [85] consiste à comparer la solution numérique obtenue pour la résolution de (5.1) avec le champ de vitesse $\mathbf{c}$ donné par (5.31) et la donnée initiale donnée par la fonction localisée :

$$h_0(\lambda, \theta) = \left\{ \begin{array}{ccc} (h_0/2)(1 + \cos(\pi r/R)) & \text{si} & r < R \\ 0 & \text{si} & r \geq R \end{array} \right. \tag{5.36}$$

avec $h_0 = 1000\text{m}$ et $R = a/3$. La fonction r représente la distance sur la sphère entre le point de coordonnées (λ, θ) et le point de coordonnées (λ_C, θ_C). Elle est donnée par

$$r = a \arccos\left(\sin\theta_C \sin\theta + \cos\theta_C \cos\theta \cos(\lambda - \lambda_C)\right), \tag{5.37}$$

$(\lambda_C, \theta_C) = (3\pi/2, 0)$ est la position initiale du Bump. Dans ce test, on a $\omega_s = u_0/a = 2\pi/12$ jours^{-1}. Il s'agit d'une condition initiale de classe $\mathcal{C}^1$, elle n'est pas de classe $\mathcal{C}^2$. Dans les sections 5.1.3, les tests sont consacrés à des solutions plus régulières. Des tests existent avec des solutions initiales moins régulières (voir [65]), mais ici, nous nous limitons à (5.36).

Les Tables 5.1 et 5.2 donnent l'erreur obtenue sur 12 jours avec différentes tailles de grilles avec $\alpha = 0$ et $\alpha = \pi/4$. Le taux de convergence est compris entre 1.5 et 2.5. La convergence d'ordre 3 au moins était attendue en supposant la solution suffisamment régulière ce qui n'est pas le cas ici puisque la solution est seulement de classe $\mathcal{C}^1$.

Dans la Figure 5.2, on observe la localisation spatiale de l'erreur $h(t^n, \cdot)^* - \mathfrak{h}^n$ après une rotation complète de la solution initiale, au temps $t = 12$ jours ainsi que l'erreur relative au cours du temps pour $N = 40$. L'erreur est principalement localisée là où la fonction h est la moins régulière.

N	$\mathbf{e_1}$	$\mathbf{e_2}$	$\mathbf{e_\infty}$
40	4.3043(−2)	2.4784(−2)	2.0921(−2)
50	2.4403(−2)	1.4917(−2)	1.3748(−2)
60	1.5367(−2)	9.9131(−3)	1.0476(−3)
80	7.5508(−3)	5.3960(−3)	6.2646(−3)
100	4.3709(−3)	3.3958(−3)	4.4360(−3)
150	1.6538(−3)	1.4917(−3)	2.2885(−3)
Ordre estimé	2.47	2.12	1.67

TABLE 5.1 – Erreur et taux de convergence pour la rotation solide sur l'équation (5.1) en norme 1, 2 et ∞ pour $\alpha = 0$ et CFL $= 0.7$. Le filtre utilisé est le filtre d'ordre 10.

N	$\mathbf{e_1}$	$\mathbf{e_2}$	$\mathbf{e_\infty}$
40	3.7638(−2)	2.0633(−2)	1.4639(−2)
50	2.1323(−2)	1.2496(−2)	1.0056(−2)
60	1.3546(−2)	8.4518(−3)	7.3167(−3)
80	6.6905(−3)	4.6505(−3)	4.6060(−3)
100	3.9119(−3)	2.9448(−3)	3.1809(−3)
150	1.4922(−3)	1.3019(−3)	1.6341(−3)
Ordre estimé	2.44	2.08	1.66

TABLE 5.2 – Erreur et taux de convergence pour la rotation solide sur l'équation (5.1) en norme 1, 2 et ∞ pour $\alpha = \pi/4$ et CFL $= 0.7$. Le filtre utilisé est le filtre d'ordre 10.

Les valeurs obtenues par le schéma sont comparables à celles obtenues par [81, 80] à l'aide d'un schéma volumes finis d'ordre 4. La comparaison est donnée dans la Table 5.3. On constate que les valeurs des erreurs sont tout à fait comparables.

Pour analyser l'effet dissipatif de l'opérateur de filtrage, on compare la valeur du maximum de h_0 pour différents opérateurs de filtrage. Les résultats sont donnés sur la Table 5.4 et sur la Figure 5.3. On constate qu'un filtre d'ordre 2 est trop dissipatif et ne permet pas de conserver correctement la hauteur de h. Le filtre d'ordre 10 donne de bons résultats. On note sur la Figure 5.4 que sans filtrage, des oscillations parasites apparaissent et perturbent le calcul.

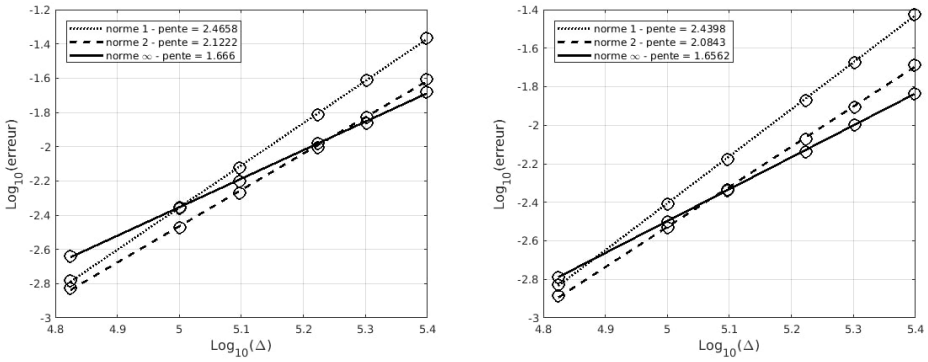


FIGURE 5.1 – Taux de convergence pour la rotation solide sur l'équation (5.1) en normes $\|\cdot\|_1$, $\|\cdot\|_2$ et $\|\cdot\|_\infty$ en fonction de $\Delta = a\Delta\xi$ pour $\alpha = 0$ (gauche) et $\alpha = \pi/4$ (droite) et CFL $= 0.7$. Le taux de convergence est quasiment identique pour $\alpha = \pi/4$ et $\alpha = 0$. Le filtre utilisé est le filtre d'ordre 10.

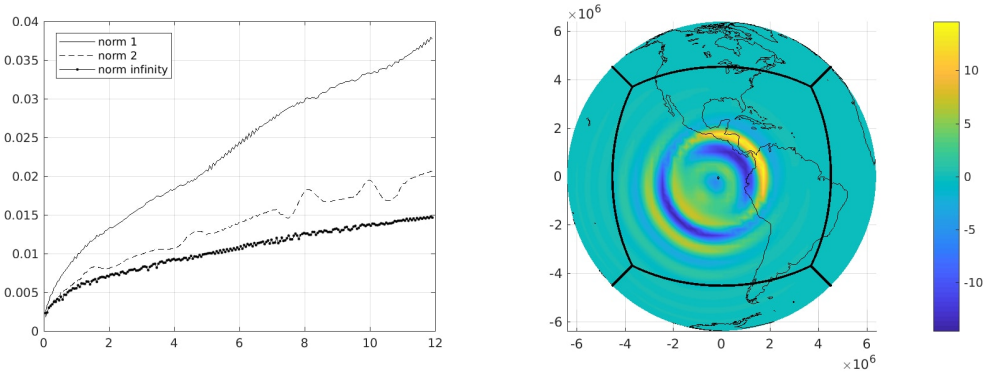


FIGURE 5.2 – Erreur relative pour l'équation (5.1) en norme 1, 2 et ∞ pour $\alpha = \pi/4$ (gauche) et localisation spatiale de l'erreur au temps $t = 12$ jours (droite) avec CFL $= 0.7$ et $N = 40$. Le filtre utilisé est le filtre d'ordre 10.

		$\mathbf{e}_1$		$\mathbf{e}_2$		$\mathbf{e}_\infty$	
CFL	α	[81]	Algo. 11	[81]	Algo. 11	[81]	Algo. 11
1.0	$\alpha = 0$	4.4262(−2)	5.4173(−2)	2.6982(−2)	3.2511(−2)	2.3012(−2)	2.6469(−2)
	$\alpha = \pi/4$	4.2173(−2)	5.1187(−2)	2.3674(−2)	2.9114(−2)	1.8696(−2)	2.2722(−2)
0.5	$\alpha = 0$	3.8326(−2)	4.0429(−2)	2.3194(−2)	2.2452(−2)	1.9969(−2)	1.8989(−2)
	$\alpha = \pi/4$	3.5096(−2)	3.4451(−2)	1.9601(−2)	1.8444(−2)	1.4171(−2)	1.4138(−2)

TABLE 5.3 – Erreur relative pour la rotation solide sur l'équation (5.1) en norme 1, 2 et ∞ pour $\alpha = \pi/4$ ainsi que $\alpha = 0$. Le pas de temps est issu de la condition CFL $= u_0\Delta t/a\Delta\xi$, le filtre est d'ordre 10. Les résultats obtenus sont pratiquement identiques à ceux obtenus par volumes finis d'ordre 4 dans [81]. Le paramètre de grille est $N = 40$.

N	20	40	80
Maxi. théorique	1000	1000	1000
Filtre d'ordre 10	968.87	990.68	997.45
Filtre d'ordre 8	915.22	995.34	996.77
Filtre d'ordre 6	767.00	996.30	996.86
Filtre d'ordre 4	436.37	795.56	969.92
Filtre d'ordre 2	47.52	91.45	170.70

TABLE 5.4 – Maximum au temps $t = 12$ jours de $\mathfrak{h}^n$ pour l'équation (5.1) discrétisée par l'algorithme 11 avec $\alpha = \pi/4$ et CFL $= 0.7$.

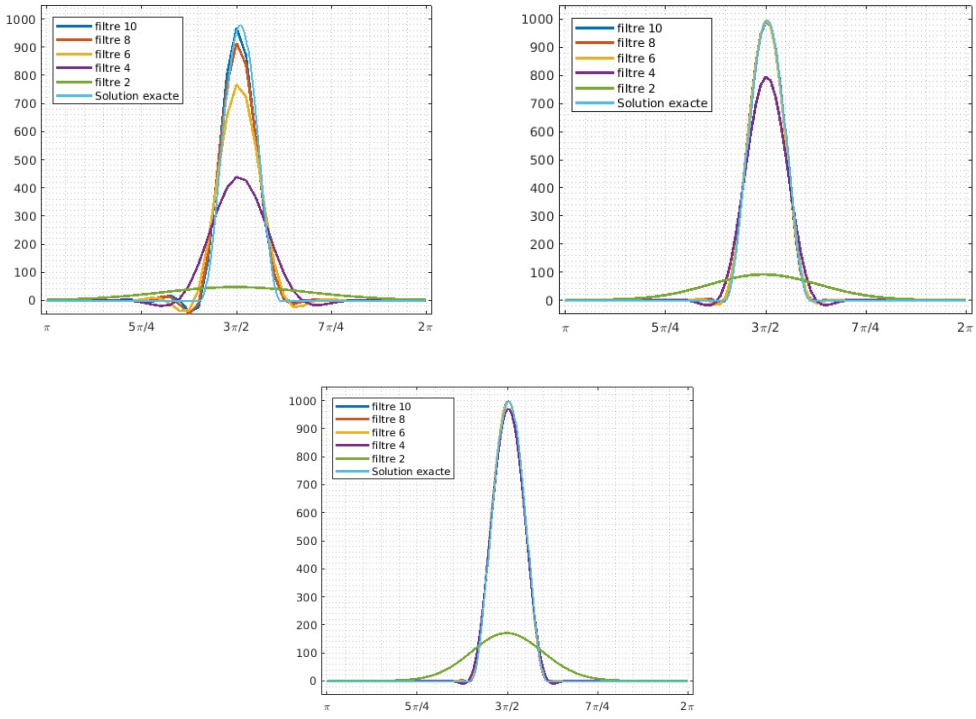


FIGURE 5.3 – Coupe au niveau de l'équateur du test 1 de [85] pour l'équation (5.1) au temps $t = 12$ jours avec $\alpha = \pi/4$ et CFL $= 0.7$. Les tailles de maillage sont $N = 20$ (haut, gauche), $N = 40$ (haut, droite) et $N = 80$ (bas). Plus l'ordre du filtre est bas, plus la solution est dissipée.

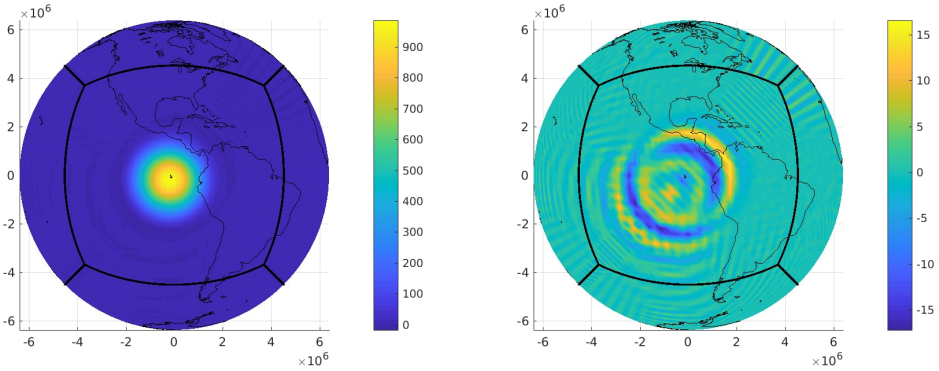


FIGURE 5.4 – Calcul de la solution au temps $t = 12$ jours sans opérateur de filtrage avec $N = 40$ et CFL $= 0.7$. Solution obtenue (gauche), erreur $\mathfrak{h}^n - h(t^n, \cdot)^*$ (droite) lorsque $\alpha = \pi/4$. On observe que sans filtrage, des oscillations sont présentes alors qu'elles ne le sont pas lorsque le filtrage est présent (voir Fig. 5.2).

5.1.3 Propagation d'un vortex

Vortex statique

Dorénavant, nous utilisons le filtrage d'ordre 10 lors de la résolution. Le test précédent est un test de déplacement sans déformation de la solution initiale. Dans [66], un autre test est construit pour que la condition initiale soit déformée au fil du temps. On considère $(\lambda_C, \theta_C) \in \mathbb{S}_a^2$ un point de la sphère. Le test consiste à suivre la propagation de deux vortex diamétralement opposés dont l'un est situé en (λ_C, θ_C). Les deux vortex s'enroulent autour de leurs centres respectifs au fil du temps, rendant la solution de plus en plus difficile à représenter sur un maillage fixe.

L'objectif est de résoudre l'équation (5.1) avec le champ $\mathbf{c}$ donné par l'équation :

$$\mathbf{c}_r(\mathbf{x}) = \mathbf{c}(\mathbf{x}) = u_r \mathbf{e}_\lambda + v_r \mathbf{e}_\theta \tag{5.38}$$

où $u_r, v_r : (\mathbf{x}) \in \mathbb{S}_a^2 \mapsto u_r(\mathbf{x}), v_r(\mathbf{x}) \in \mathbb{R}$ sont les fonctions définies par

$$\left\{ \begin{array}{rcl} u_r(\lambda, \theta) & = & a\omega_r(\theta') \left[\sin\theta_C \cos\theta - \cos\theta_C \cos(\lambda - \lambda_C) \sin\theta\right] \\ v_r(\lambda, \theta) & = & a\omega_r(\theta') \left[\cos\theta_C \sin(\lambda - \lambda_C)\right] \end{array} \right. \tag{5.39}$$

où (λ', θ') sont les coordonnées longitude-latitude associées au pôle Nord placé en (λ_C, θ_C). On déduit ces valeurs grâce aux équations (5.8). La vitesse de rotation du vortex est définie par $a\omega_r(\theta')$ avec la fonction ω_r définie par

$$\omega_r(\theta') = \left\{ \begin{array}{cl} V/a\rho & \text{si } \rho \neq 0 \\ 0 & \text{sinon.} \end{array} \right. \tag{5.40}$$

où $\rho = \rho_0 \cos(\theta')$ est une pseudo-distance au centre du vortex et $V = u_0 \dfrac{3\sqrt{3}}{2} \text{sech}^2(\rho) \tanh(\rho)$. Noter que $\dfrac{3\sqrt{3}}{2}$ est une constante de normalisation. On choisit $u_0 = 2\pi a/(12\text{jours})$ et $\rho_0 = 3$.

Une solution exacte de (5.1) est donnée par

$$h(t, \lambda, \theta) = 1 - \tanh\left[\frac{\rho}{\gamma} \sin(\lambda' - \omega_r(\theta')t)\right] \tag{5.41}$$

où γ est une constante influençant le gradient de la solution. Comme dans l'article [66], on choisit $\gamma = 5$.

N	$\mathbf{e_1}$	$\mathbf{e_2}$	$\mathbf{e_\infty}$
40	1.2170(−3)	5.2773(−3)	3.8615(−2)
50	4.4810(−4)	2.1100(−3)	1.6306(−2)
60	1.6313(−4)	8.2236(−4)	6.5687(−3)
80	2.8658(−5)	1.4710(−4)	1.4042(−3)
100	8.7526(−6)	4.1919(−5)	4.1512(−4)
150	1.1105(−6)	5.7646(−6)	6.2903(−5)
Ordre estimé	5.40	5.30	4.98

TABLE 5.5 – Erreur et taux de convergence pour le test du vortex stationnaire sur 12 jours pour l'équation (5.1) en normes $\|\cdot\|_1$, $\|\cdot\|_2$ et $\|\cdot\|_\infty$, CFL = 0.7 [66], le filtre est d'ordre 10. Le vortex est localisé en $(\lambda_C, \theta_C) = (\pi/4, \pi/4)$. La convergence se fait à un ordre supérieur ou égal à 5.

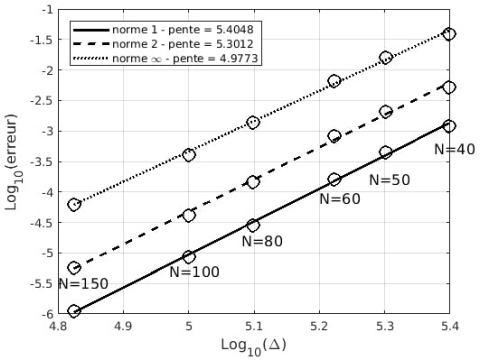


FIGURE 5.5 – Erreur et taux de convergence pour le test du vortex stationnaire sur l'équation (5.1) en normes $\|\cdot\|_1$, $\|\cdot\|_2$ et $\|\cdot\|_\infty$ et en fonction de $\Delta = a\Delta\xi$, avec CFL = 0.7 [66], le filtre est d'ordre 10. Le vortex est localisé en $(\lambda_C, \theta_C) = (\pi/4, \pi/4)$. L'ordre de convergence est d'environ 5 pour la norme $\|\cdot\|_\infty$ et supérieur pour les normes $\|\cdot\|_1$ et $\|\cdot\|_2$.

Sur la Table 5.5 et la Figure 5.5, on donne la convergence pour ce test avec $(\lambda_C, \theta_C) = (\pi/4, \pi/4)$. Les centres des vortex sont alors placés proches des coins de la Cubed-Sphere. La convergence se fait à un ordre supérieur ou égal à 5.

Sur une grille grossière ($N = 36$ correspondant à l'équateur à $\Delta\lambda = 2.5$deg.), on compare l'erreur au cours du temps pour deux valeurs différentes du pas de temps Δt. Les résultats sont donnés sur la Figure 5.6. Lorsque CFL $= u_0\Delta t/a\Delta\xi = 0.5$, il y a 288 pas de temps pour arriver au temps final. Lorsque CFL $= u_0\Delta t/a\Delta\xi = 0.05$, il y a 2880 pas de temps. Les erreurs obtenues sont tout a fait comparables à celles obtenues par la méthode de Galerkin Discontinu [64]. Avec 288 pas de temps, les erreurs spatiales et temporelles sont observées simultanément. L'erreur est sensiblement meilleure lorsque 2880 pas de temps sont utilisés.

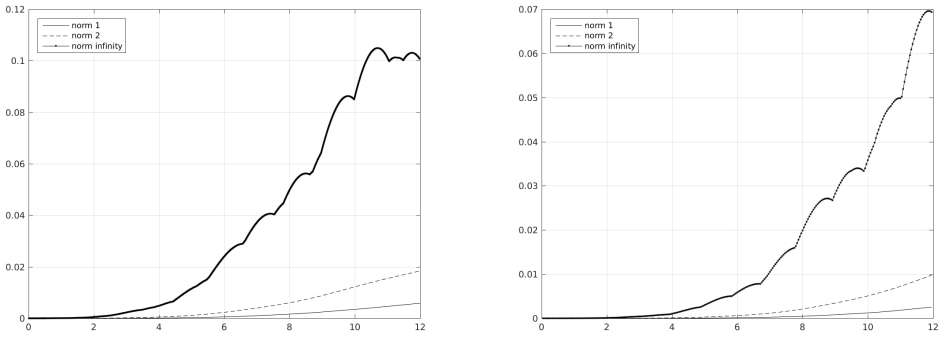


FIGURE 5.6 – Evolution de l'erreur sur $t = 12$ jours pour le cas test du vortex [66] avec $(\lambda_C, \theta_C) = (\pi/4, \pi/4)$. Les paramètres numériques sont $N = 40$, le filtrage utilisé est d'ordre 10. Le pas de temps est déduit des relations CFL $= 0.5$ (gauche), et CFL $= u_0\Delta t/\Delta\xi = 0.05$ (droite).

Sur la Figure 5.7, on représente la solution $\mathfrak{h}^n$ au temps $t = 12$ jours ainsi que l'erreur spatiale $\mathfrak{h}^n - h(t^n, \cdot)^*$ sur un maillage de paramètre $N = 40$. L'erreur est localisée au centre du vortex. Ce résultat était attendu, le vortex devient de plus en plus fin lorsque t grandit et la solution devient sous résolue. On vérifie cela sur la Figure 5.8 : on représente, au temps $t = 12$ jours, une coupe le long de l'équateur de la solution lorsque $(\lambda_C, \theta_C) = (3\pi/4, 0)$. On observe qu'avec une grille de paramètre $N = 25$, la solution est sous représentée en comparaison à une grille de paramètre $N = 50$.

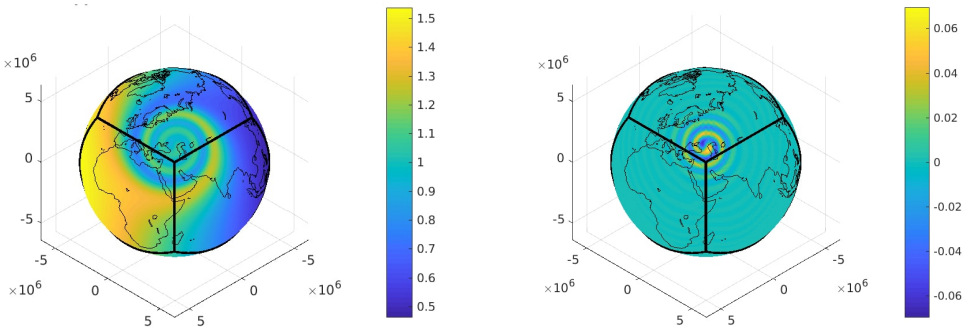


FIGURE 5.7 – Solution au temps $t = 12$ jours pour le vortex statique [66] avec $(\lambda_C, \theta_C) = (\pi/4, \pi/4)$. Les paramètres numériques sont $N = 40$ et CFL $= 0.7$, le filtrage utilisé est d'ordre 10. La solution $\mathfrak{h}^n$ (gauche), erreur spatiale $\mathfrak{h}^n - h(t^n, \cdot)^*$ (droite).

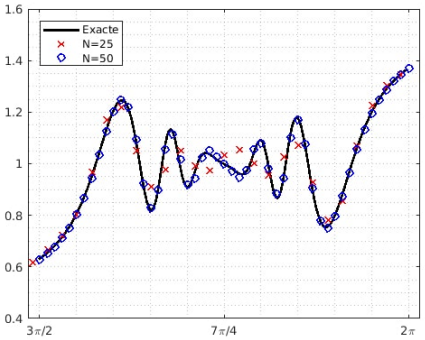


FIGURE 5.8 – Coupe le long de l'équateur de la solution au temps $t = 12$ jours pour le cas test du vortex [66] avec $(\lambda_C, \theta_C) = (3\pi/4, 0)$. Le pas de temps est issu de CFL $= 0.7$ et filtrage d'ordre 10. La solution sur grille grossière est moins bien représentée que celle sur grille fine.

Vortex avec rotation solide

Une variante du test [66] consiste à combiner la vitesse de rotation solide $\mathbf{c}_s$ (5.31) avec la vitesse de rotation du vortex $\mathbf{c}_r$ (5.38). Il s'agit du test présenté dans [64]. On considère l'équation d'advection (5.1) munie du champ de vitesse

$$\mathbf{c}(t, \mathbf{x}) = u\mathbf{e}_\lambda + v\mathbf{e}_\theta \tag{5.42}$$

où les fonctions u et v dépendent à présent du temps par

$$\left\{ \begin{array}{rcl} u(t,\lambda,\theta) & = & u_0 \left(\cos\theta\cos\alpha + \sin\theta\cos\lambda\sin\alpha\right) + a\omega_r \left(\sin\theta_C(t)\cos\theta - \cos\theta_C(t)\cos(\lambda - \lambda_C(t))\sin\theta\right) \\ v(t,\lambda,\theta) & = & -u_0 \sin\lambda\sin\alpha + a\omega_r \left(\cos\theta_C(t)\sin(\lambda - \lambda_C(t))\right), \end{array} \right. \tag{5.43}$$

La donnée $(\lambda_C(t), \theta_C(t)) \in \mathbb{S}^2_a$ correspond à la position du vortex au fil du temps. Cette dernière est donnée dans la base "tournée" d'un angle α par

$$\left\{ \begin{array}{rcl} \lambda'_C(t) & = & \lambda'_0 + \omega_s t \\ \theta'_C(t) & = & \theta'_0 \end{array} \right. \tag{5.44}$$

avec (λ'_0, θ'_0) la position initiale du vortex statique dans la base associée à $(\lambda_P, \theta_P) = (\pi, \pi/2 - \alpha)$. Dans le système de coordonnées longitude latitude associé au pôle Nord $\mathbf{N}$, on a $(\lambda_0, \theta_0) = (3\pi/2, 0)$. $\omega_s = u_0/a$ est la vitesse de rotation solide du vortex.

La solution exacte est alors donnée par (5.41) en déplaçant la position du vortex au fil du temps t. On note (λ, θ) les coordonnées longitude-latitude de $\mathbf{x} \in \mathbb{S}^2_a$. La solution exacte en $\mathbf{x}$ et au temps $t > 0$, notée $h(t, \mathbf{x})$, est calculée de la façon suivante :

1. Calculer (λ', θ') les coordonnées longitude-latitude associées au pôle de coordonnées $(\lambda_P, \theta_P) = (\pi, \pi/2 - \alpha)$. Pour cela, on utilise la formule (5.8).

2. Déplacer (λ', θ') sur la position du vortex par

$$\left\{ \begin{array}{rcl} \lambda'_s & = & \lambda' - \omega_s t \\ \theta'_s & = & \theta'. \end{array} \right. \tag{5.45}$$

   Il s'agit de l'action de la rotation solide.

3. Calcul de (λ_s, θ_s) en revenant dans le système de coordonnées longitude-latitude grâce à la formule (5.9) avec $(\lambda_P, \theta_P) = (\pi, \pi/2 - \alpha)$.

4. Calcul de $(\lambda_s'', \theta_s'')$ déduit de (λ_s, θ_s). Le point de coordonnées longitude latitude (λ_s, θ_s) a pour coordonnées longitude latitude $(\lambda_s'', \theta_s'')$ (pour le pôle de coordonnées (λ_C, θ_C) donné par (5.44)) grâce à la formule (5.8).

5. Calculer la solution exacte $h(t, \lambda, \theta)$ par

$$h(t, \lambda, \theta) = 1 - \tanh\left[\frac{\rho}{\gamma}\sin(\lambda_s'' - \omega_r(\theta_s'')t)\right], \tag{5.46}$$

avec ω_r donné par

$$\omega_r = \begin{cases} V/(a\rho) & \text{si } \rho \neq 0 \\ 0 & \text{sinon,} \end{cases} \tag{5.47}$$

et $\rho = \rho_0 \cos(\theta_s'')$ ainsi que $V = u_0 \dfrac{3\sqrt{3}}{2}\text{sech}^2(\rho)\tanh(\rho)$.

La solution exacte représente un vortex s'enroulant sur lui même. Les détails sont de plus en plus fins et à grille fixée, elle devient difficile à représenter. De plus, le centre des vortex se déplace sur un grand cercle de la sphère. En fonction de la valeur de α, les vortex passent plus ou moins loin des coins de la Cubed-Sphere.

Sur la Figure 5.9, on représente la solution aux temps $t = 3$, $t = 6$, $t = 9$ et $t = 12$ jours lorsque $\alpha = \pi/4$. On y observe le déplacement des vortex le long d'un grand cercle longeant les panels (V) et (VI).

Sur la Table 5.6 et la Figure 5.10, on représente le taux de convergence en utilisant différentes tailles de grilles et en conservant CFL $= 0.7$. On choisit $\alpha = \pi/4$ de manière à ce que les vortex longent les bords des panels comme c'est visible sur la Figure 5.9. Un tel choix vise à mettre en difficulté la méthode de résolution en faisant passer les détails fins de la solution sur les bords des panels. Les résultats permettent d'observer un ordre de convergence proche de 4 en norme $\|\cdot\|_1$ et $\|\cdot\|_2$. L'ordre de convergence est proche de 3.5 pour la norme $\|\cdot\|_\infty$.

N	$\mathbf{e_1}$	$\mathbf{e_2}$	$\mathbf{e_\infty}$
40	2.9241(−3)	1.0646(−2)	5.7267(−2)
50	1.3634(−3)	5.3187(−3)	3.3187(−2)
60	6.6453(−4)	2.7522(−3)	1.8792(−2)
80	2.0635(−4)	8.8170(−4)	6.3350(−3)
100	8.2353(−5)	3.5454(−4)	2.7479(−3)
150	1.6044(−5)	7.0918(−5)	5.9455(−4)
Ordre estimé	3.98	3.84	3.52

TABLE 5.6 – Erreur pour le test vortex avec rotation solide. On donne différentes erreurs et taux de convergence pour l'équation (5.1) avec le champ de vitesse (5.43) en norme 1, 2 et ∞, CFL $= 0.7$. On choisit $\alpha = \pi/4$ et le temps final $t = 12$ jours. Nous utilisons l'opérateur de filtrage d'ordre 10.

On a vu que le vortex statique est difficile à représenter lorsque t augmente. Le vortex en rotation représente la même fonction se déplaçant sur la sphère. La solution du vortex en rotation est aussi difficile à représenter lorsque t croît. Sur la Figure 5.11, on représente l'historique de l'erreur relative jusqu'à $t = 24$ jours. Le temps final pour ce test est usuellement $t = 12$ jours, mais on observe le comportement du schéma sur un temps plus long. Les résultats sont obtenus avec CFL $= 0.7$, l'obtention de résultats pour 24 jours sont obtenus après 457 pas de temps. Pour la grille de taille $40 \times 40 \times 6$, l'erreur finale est de 15.95% en norme $\|\cdot\|_\infty$, 3.67% pour la norme $\|\cdot\|_2$ et 1.36% en norme $\|\cdot\|_1$. Pour la grille $80 \times 80 \times 6$, on effectue 914 pas de temps pour arriver au temps $t = 24$ jours. L'erreur finale est de 9.63% en norme $\|\cdot\|_\infty$, 1.69% pour la norme $\|\cdot\|_2$ et 0.45% en norme $\|\cdot\|_1$.

Les erreurs au temps $t = 12$ jours sont comparables à celles obtenues par la méthode de Galerkin discontinue [64]. Nous comparons aussi notre schéma à des schémas de type volumes finis d'ordre élevé

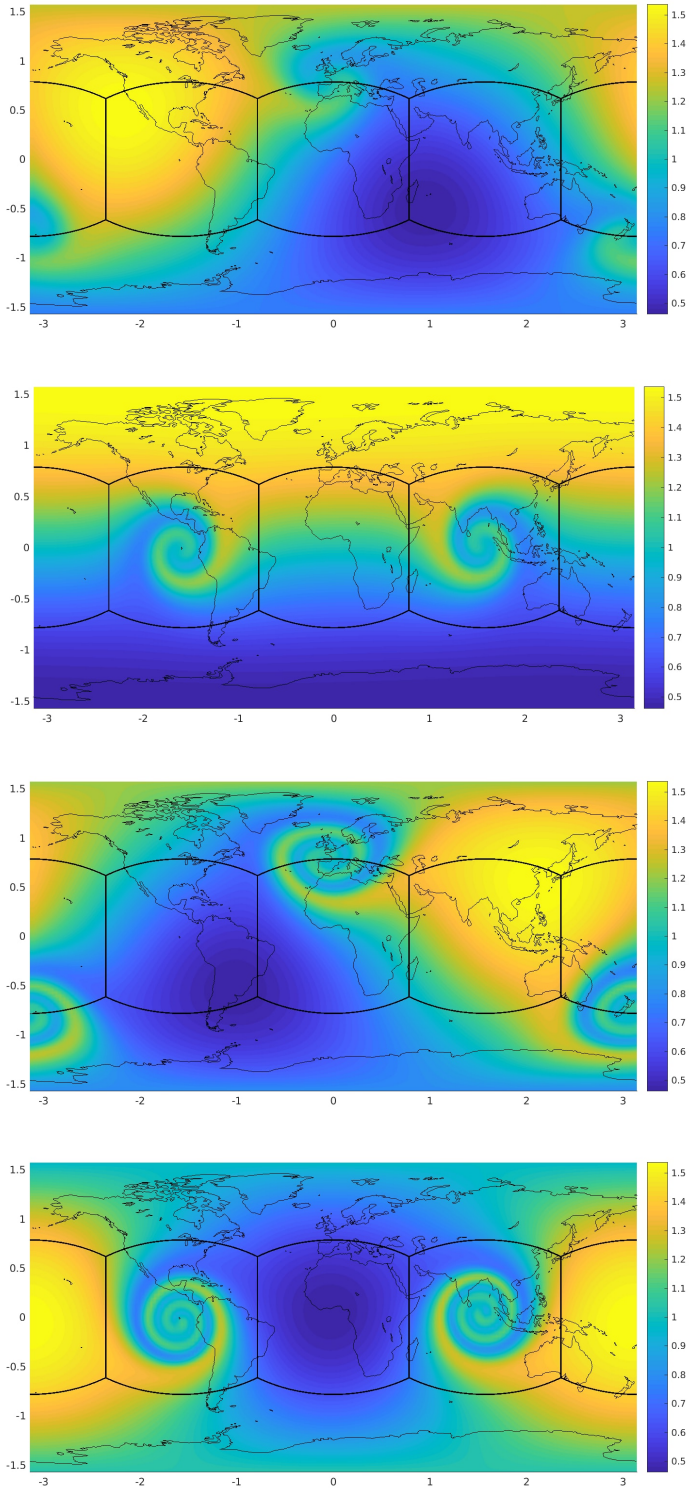


FIGURE 5.9 – Vortex avec rotation solide de [64]. On représente la solution (5.46) de l'équation de transport (5.1) avec le champ de vitesse (5.43) avec une grille de paramètre $N = 40$. On représente la solutions aux temps $t = 3$, $t = 6$, $t = 9$ et $t = 12$ jours (dans cet ordre, de haut en bas). En plus du déplacement des tourbillons, on observe que lors de la formation du vortex, la solution devient difficile à représenter.

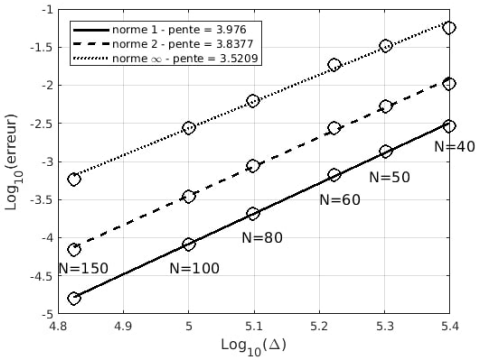


FIGURE 5.10 – Erreur pour le vortex avec rotation solide en fonction de $\Delta = a\Delta\xi$. On représente l'erreur et le taux de convergence pour l'équation (5.1) avec le champ de vitesse (5.43) en norme 1, 2 et ∞, CFL $= 0.7$, l'opérateur de filtrage est d'ordre 10. L'angle α est $\alpha = \pi/4$ et le temps final $t = 12$ jours. L'odre de convergence est proche de 4 pour les normes $\|\cdot\|_1$ et $\|\cdot\|_2$. Il est proche de 3.5 pour la norme $\|\cdot\|_\infty$.

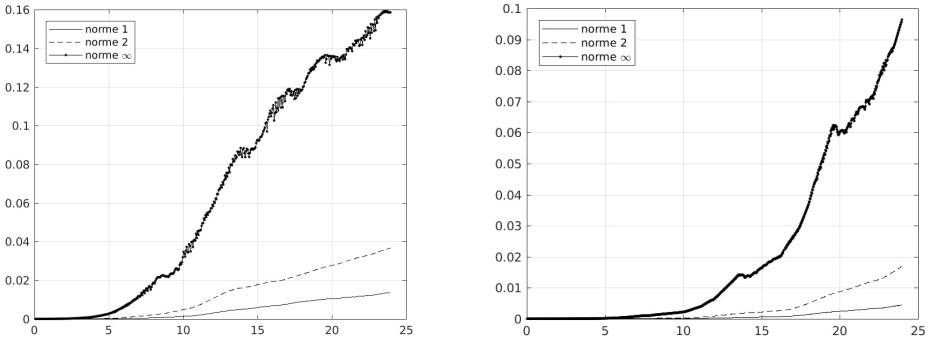


FIGURE 5.11 – Historique de l'erreur pour le test du vortex avec rotation solide. On représente l'historique de l'erreur pour l'équation (5.1) avec le champ de vitesse (5.43) en norme $\|\cdot\|_1$, $\|\cdot\|_2$ et $\|\cdot\|_\infty$ avec CFL $= 0.7$, le filtrage est d'ordre 10. On choisit $\alpha = \pi/4$. Le temps final est $t = 24$ jours. La grille est $40 \times 40 \times 6$, 457 pas de temps (gauche), la grille est $80 \times 80 \times 6$, l'algorithme effectue 914 pas de temps (droite). Sur la grille $80 \times 80 \times 6$, l'erreur est en dessous de 10% ce qui demeure acceptable.

[51]. Nous comparons les résultats au temps $t = 12$ jours. Les schémas volumes finis sont nommés WENO5 et KL4 dans [51]. Nous utilisons toujours $\alpha = \pi/4$. Sur la grille $80 \times 80 \times 6$ et après 750 pas de temps, le schéma WENO5 donne les erreurs relatives suivantes : $e_1 = 0.0021$, $e_2 = 0.0043$ et $e_\infty = 0.0191$. Le schéma KL4 obtient, dans le même contexte, les erreurs $e_1 = 0.0021$, $e_2 = 0.0043$ et $e_\infty = 0.0194$. Avec notre schéma, on obtient $e_1 = 1.67(-4)$, $e_2 = 7.23(-4)$ et $e_\infty = 5.75(-3)$. Les niveaux d'erreurs obtenus sont plus faibles que ceux obtenus par les méthodes de volumes finis WENO5 et KL4.

5.2 Equations de conservation non linéaire

L'équation d'advection (5.1) est un problème linéaire. Dans cette section, on s'intéresse à une équation non linéaire de type "Burgers" sphérique [10]. Les tests effectués concernent l'équation

$$\left\{ \begin{array}{rcll} \dfrac{\partial h}{\partial t} + \nabla_T \cdot F(h) & = & 0 & \\ h(0,\mathbf{x}) & = & h_0(\mathbf{x}) & \end{array} \right. \quad \text{avec } \mathbf{x} \in \mathbb{S}^2_a \text{ et } t \geq 0. \tag{5.48}$$

Nous choisissons $a = 1$. L'application $F : h \mapsto F(h) \in \mathbb{TS}^2$ transforme une fonction en un champ de vecteurs tangent à la sphère.

En particulier, on note que (5.48) est une loi de conservation. La relation suivante est vérifiée :

$$\frac{d}{dt} \int_{\mathbb{S}^2} h(t,\mathbf{x}) d\sigma(\mathbf{x}) = 0. \tag{5.49}$$

On a vu dans le lemme 4.1 que si $\mathbf{w} : \mathbf{x} \in \mathbb{S}^2 \mapsto \mathbf{w} \in \mathbb{R}^3$ est un champ de vecteurs de $\mathbb{R}^3$ et si $\mathbf{n}$ est la normale extérieure à la sphère, alors $\mathbf{F} = \mathbf{n} \wedge \mathbf{w}$ est un champ de vecteurs tangent à la sphère. On considère ici des champs de vecteurs de cette forme avec

$$\mathbf{w} = f_1\mathbf{i} + f_2\mathbf{j} + f_3\mathbf{k} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix}, \tag{5.50}$$

où $f_p : h \in L^2(\mathbb{S}^2_a, \mathbb{R}) \mapsto f_p(h)$ pour $1 \leq p \leq 3$.

Dans ce qui suit, on s'intéresse à deux tests pour cette équation, introduits dans [10]. Le premier permet d'analyser le comportement d'un schéma numérique lors de l'apparition d'un choc. Le second permet d'étudier la conservation d'une solution stationnaire.

5.2.1 Résolution numérique

Pour résoudre l'équation (5.48), on considère l'application J_Δ définie pour toute fonction de grille $\mathfrak{h}$ sur la sphère par

$$J_\Delta(t,\mathfrak{h}) = -\nabla_{T,\Delta} F(\mathfrak{h}). \tag{5.51}$$

Nous couplons cet opérateur d'approximation spatiale à un algorithme de résolution en temps. L'algorithme permettant la résolution de (5.48) est analogue à l'algorithme 11. Il s'écrit :

Algorithme 12 : Equation d'advection sphérique non linéaire (5.48)

$\mathfrak{h}^0 = h_0^*$ connu,
for $n = 0, 1, \ldots$ **do**
  $K^{(1)} = J_\Delta(t^n, \mathfrak{h}^n)$,
  $K^{(2)} = J_\Delta\left(t^n + \frac{\Delta t}{2}, \mathfrak{h}^n + \frac{\Delta t}{2} K^{(1)}\right)$,
  $K^{(3)} = J_\Delta\left(t^n + \frac{\Delta t}{2}, \mathfrak{h}^n + \frac{\Delta t}{2} K^{(2)}\right)$,
  $K^{(4)} = J_\Delta\left(t^n + \Delta t, \mathfrak{h}^n + \Delta t K^{(3)}\right)$,
  $\mathfrak{h}^{n+1} = \mathcal{F}\left(\mathfrak{h}^n + \dfrac{\Delta t}{6}\left(K^{(1)} + 2K^{(2)} + 2K^{(3)} + K^{(4)}\right)\right)$.
end for

L'opérateur $\mathcal{F}$ est de la forme (4.92) où $\tilde{\mathcal{F}}_\xi$ et $\tilde{\mathcal{F}}_\eta$ utilisent l'opérateur de filtrage en dimension 1 d'ordre 10 : $\mathcal{F}_{10,x}$ (1.163). Les opérateurs $\tilde{\mathcal{F}}_\xi$ et $\tilde{\mathcal{F}}_\eta$ sont calculés par les algorithmes 9 et 10.

5.2.2 Solution équatoriale périodique

Pour ce test, inspiré du premier test de [10], on considère les fonctions f_1 et f_2 nulles :

$$f_1 = f_2 \equiv 0. \tag{5.52}$$

Le champ de vecteurs F est alors donné pour toute fonction h par

$$F(h) = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \wedge \begin{bmatrix} f_1(h) \\ f_2(h) \\ f_3(h) \end{bmatrix} = \begin{bmatrix} y f_3(h) \\ -x f_3(h) \\ 0 \end{bmatrix} = -f_3(h)\cos(\theta)\mathbf{e}_\lambda. \tag{5.53}$$

On a en coordonnées longitude-latitude

$$\nabla_T \cdot F(h) = -\frac{\partial}{\partial \lambda} f_3(h). \tag{5.54}$$

L'équation (5.48) s'écrit alors en coordonnées longitude-latitude :

$$\frac{\partial h}{\partial t} - \frac{\partial}{\partial \lambda} f_3(h) = 0 \text{ en tout } \mathbf{x} \in \mathbb{S}^2_a \text{ et } t \geq 0. \tag{5.55}$$

Il découle la proposition suivante :

Proposition 5.3. *Soit $\tilde{h}$ la solution du problème périodique en dimension 1*

$$\begin{cases} \dfrac{\partial \tilde{h}}{\partial t} - \dfrac{\partial}{\partial \lambda} f_3(\tilde{h}) = 0 & \text{pour } \lambda \in [0, 2\pi[\text{ et } t > 0, \\ \tilde{h}(0, \lambda) = \tilde{h}_0(\lambda) \end{cases} \tag{5.56}$$

et soit $\hat{h} : \theta \in]-\pi/2, \pi/2[\mapsto \hat{h}(\theta) \in \mathbb{R}$ tel que

$$h_0(\mathbf{x}) = \tilde{h}_0(\lambda)\hat{h}(\theta). \tag{5.57}$$

Alors la solution du problème (5.48) *est donnée par*

$$h(t, \mathbf{x}) = h(t, \lambda, \theta) = \tilde{h}(t, \lambda)\hat{h}(\theta), \tag{5.58}$$

pour $t > 0$, $\mathbf{x} \in \mathbb{S}^2$ un point de la sphère de coordonnées longitude-latitude (λ, θ).

On pose $f_3(h) = -\pi h^2$. L'équation (5.56) est identique à l'équation (2.175). On compare une coupe le long de l'équateur de la solution calculée par l'algorithme 12 avec la solution calculée par l'algorithme 6 lorsque

$$\begin{cases} \tilde{h}_0(\lambda) &= \sin\lambda, \\ \hat{h}_0(\theta) &= \mathbf{1}_{[-\pi/12, \pi/12]}(\theta). \end{cases} \tag{5.59}$$

On rappelle que dans ce contexte, la solution de (5.56) est de classe C^1 pour $t \leq 1/(2\pi)$. Sur la Figure 5.12, on représente la solution aux temps $t = 1/(2\pi)$ et $t = 10/(2\pi)$ pour un paramètre de grille $N = 32$ pour la Cubed-Sphere (128 points de discrétisation sur l'équateur) et 128 points de discrétisation pour le problème 1D. Le pas de temps est $\Delta t = 0.005$. Les résultats des deux algorithmes sont très semblables et donnent des résultats satisfaisants même au delà du temps $1/(2\pi)$ au delà duquel la solution est moins régulière. De plus, le filtre d'ordre 10 est suffisant pour que les oscillations ne dégradent pas excessivement le résultat.

Sur la Figure 5.13, on représente l'historique de l'erreur de conservation au cours du temps :

$$|Q(\mathfrak{h}^n) - Q(h(t^n, \cdot)^*|. \tag{5.60}$$

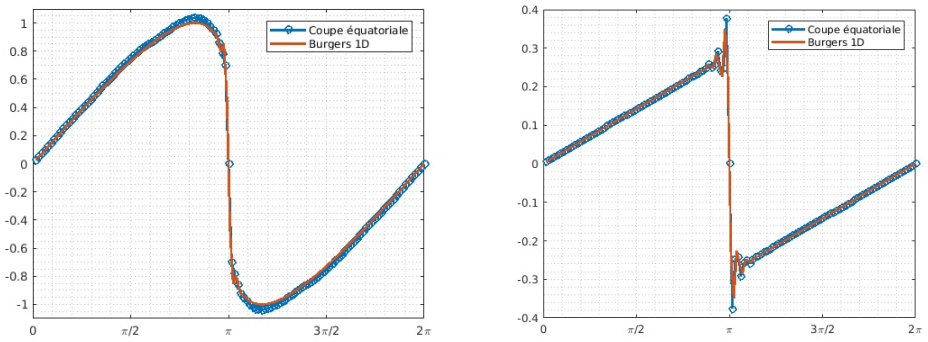


FIGURE 5.12 – Coupe de la solution équatoriale. On représente une coupe équatoriale de la solution de (5.48) et la solution de (5.56) pour le test périodique. On compare la solution au temps $t = 1/(2\pi)$ (gauche) et $t = 10/(2\pi)$ (droite). La grille Cubed-Sphere a pour paramètre $N = 32$ (128 points de discrétisation sur l'équateur). Le problème en dimension 1 est résolu avec 128 points de discrétisation. Le pas de temps est $\Delta t = 0.005$.

Nous n'utilisons pas l'erreur relative car pour les conditions initiales utilisées, on a

$$\int_{\mathbb{S}_a^2} h_0(\mathbf{x}) d\sigma(\mathbf{x}) = 0. \tag{5.61}$$

L'erreur de conservation est proche de $6.0(-5)$ lorsque $N = 32$ et proche de $1.5(-5)$ lorsque $N = 64$. L'erreur est cependant nettement plus importante au temps d'apparition du choc $t = 1/(2\pi) \approx 0.1592$.

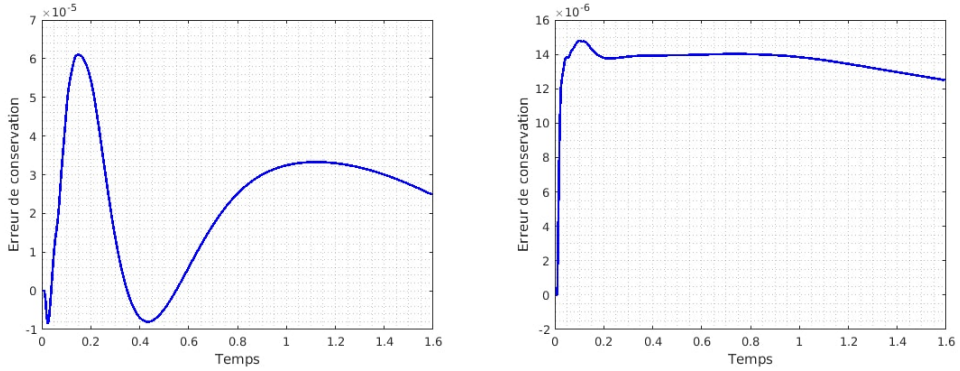


FIGURE 5.13 – Historique de l'erreur de conservation pour la solution équatoriale périodique. On représente l'erreur de conservation en fonction du temps t pour le test équatorial périodique (5.48). La grille Cubed-Sphere a pour paramètres $N = 32$ (gauche) et $N = 64$ (droite). Le pas de temps est le même dans les deux cas. On a $\Delta t = 0.005$, la simulation est faite en 318 pas de temps pour le temps final $t = 10/(2\pi)$. L'erreur n'est pas relative car la masse totale initiale est nulle. La masse totale mesurée est très bien conservée.

Ce test permet d'analyser le comportement du schéma numérique de l'algorithme 12 en présence d'un choc. Les résultats sont satisfaisants. Les oscillations qui apparaissent au delà du temps d'apparition du choc ne dégradent pas excessivement le calcul. Le filtrage d'ordre 10 symétrique est suffisant pour assurer un bon déroulement de la simulation. De plus, l'erreur sur la conservation de la masse reste faible au cours du temps.

Nous insistons sur le fait que le calcul effectué n'utilise aucun décentrement et aucun traitement de type reconstruction ou limitation de pente. Il n'est donc pas surprenant d'observer des oscillations non

linéaires (Fig. 5.12, droite). On constate que ces oscillations demeurent très limitées et très localisées. Une amélioration de ces résultats fera l'objet d'études ultérieures.

5.2.3 Solution stationnaire

Dans cette section, on construit une solution stationnaire de (5.48). Si la condition initiale h_0 est telle que

$$\nabla_T \cdot F(h_0) = 0, \tag{5.62}$$

alors h_0 est une solution stationnaire de (5.48).

On considère les fonctions f_1, f_2 et f_3 identiques, c'est à dire :

$$f_1(h) = f_2(h) = f_3(h) = f(h) \tag{5.63}$$

avec $h : \mathbb{S}_a^2 \to \mathbb{R}$ donnée. Alors, F est donnée par

$$F(h) = \mathbf{n} \wedge (f(h)(\mathbf{i} + \mathbf{j} + \mathbf{k})) . \tag{5.64}$$

Le calcul de $\nabla_T \cdot F(h)$ en fonction de h et f donne

$$\nabla_T \cdot F(h) = f'(h) \left((y-z)\frac{\partial h}{\partial x} + (z-x)\frac{\partial h}{\partial y} + (x-y)\frac{\partial h}{\partial z} \right). \tag{5.65}$$

On en déduit que indépendamment du choix de f, si $h_0(x, y, z) = \alpha(x + y + z)$, avec $\alpha \in \mathbb{R}$, on a

$$\nabla_T \cdot F(h_0) = 0, \tag{5.66}$$

et $h(t, \mathbf{x}) = h_0(\mathbf{x})$ est une solution stationnaire.

Dans [10], le test numéro 3 consiste à choisir

$$f_1(h) = f_2(h) = f_3(h) = \frac{h^2}{2}, \tag{5.67}$$

et la condition initiale

$$h_0(x, y, z) = \frac{x + y + z}{\sqrt{3}}. \tag{5.68}$$

Cette condition initiale est une solution stationnaire. On compare la solution calculée par l'algorithme 12 avec la solution initiale jusqu'au temps $t = 6$. On mesure l'erreur relative à chaque itération. Les résultats de convergence sont donnés sur la Table 5.7 et sur la Figure 5.14. L'ordre de convergence est proche de 4 en normes $\|\cdot\|_1$, $\|\cdot\|_2$ et $\|\cdot\|_\infty$. La conservation de la masse est vérifiée à un ordre supérieur à 5.

N	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_\infty$	**Conservation**
16	2.1446(−5)	1.5759(−5)	1.4251(−5)	3.6380(−7)
32	2.2000(−6)	1.1752(−6)	1.0776(−6)	8.3644(−9)
64	1.4092(−7)	7.9823(−8)	7.7308(−8)	9.6391(−11)
128	8.7856(−9)	5.0291(−9)	4.5510(−9)	8.4850(−12)
Ordre estimé	3.94	3.87	3.86	5.18

TABLE 5.7 – Erreur pour la solution stationnaire. Table de convergence pour le test stationnaire de l'équation (5.48). Le pas de temps est donné par $\Delta t = 0.96\Delta\xi/\pi$. Le temps final est $t = 6$. On mesure aussi l'erreur sur la conservation de la masse. Le taux de convergence est proche de 4. L'erreur de conservation mesurée n'est pas relative car la masse totale initiale est nulle. La conservation est excellente.

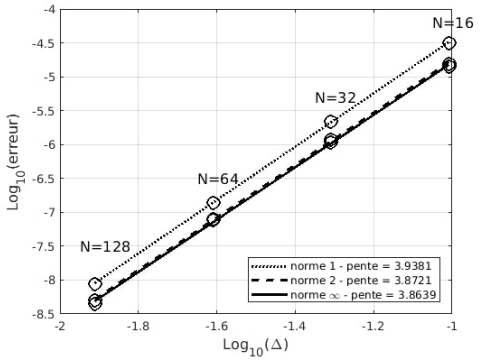


FIGURE 5.14 – Erreur et taux de convergence pour le test stationnaire de l'équation (5.48) en fonction de $\Delta = a\Delta\xi$. Le pas de temps est donné par $\Delta t = 0.96\Delta\xi/\pi$. Le temps final est $t = 6$.

Sur la Figure 5.15, on représente l'historique de l'erreur lorsque $N = 31$ et $\Delta t = 0.96\Delta\xi/\pi = 0.015$. On observe sur ces figures que l'erreur sur la fonction h est proche de 3×10^{-6}. L'erreur sur la conservation de la masse est proche de 10^{-8}. Les résultats sont très satisfaisants pour la conservation d'une solution stationnaire sur une équation non-linéaire. Sur la Figure 5.16, on représente la localisation spatiale de l'erreur ainsi que la solution calculée au temps $t = 6$. Les oscillations observées s'apparentent à des oscillations dispersives. Leur amplitude demeure très limitée.

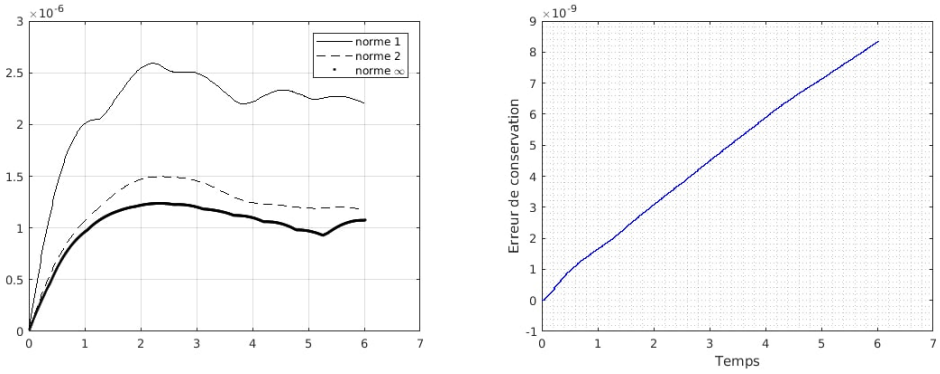


FIGURE 5.15 – Courbe d'erreur pour la solution stationnaire. L'erreur en norme et l'erreur de conservation est représentée pour le test stationnaire de l'équation (5.48). Le pas de temps est donné par $\Delta t = 0.96\Delta\xi/\pi$. Le temps final est $t = 6$. Le paramètre de la Cubed-Sphere est $N = 32$. L'erreur de conservation mesurée n'est pas relative car la masse totale initiale est nulle. L'erreur sur la conservation est très faible.

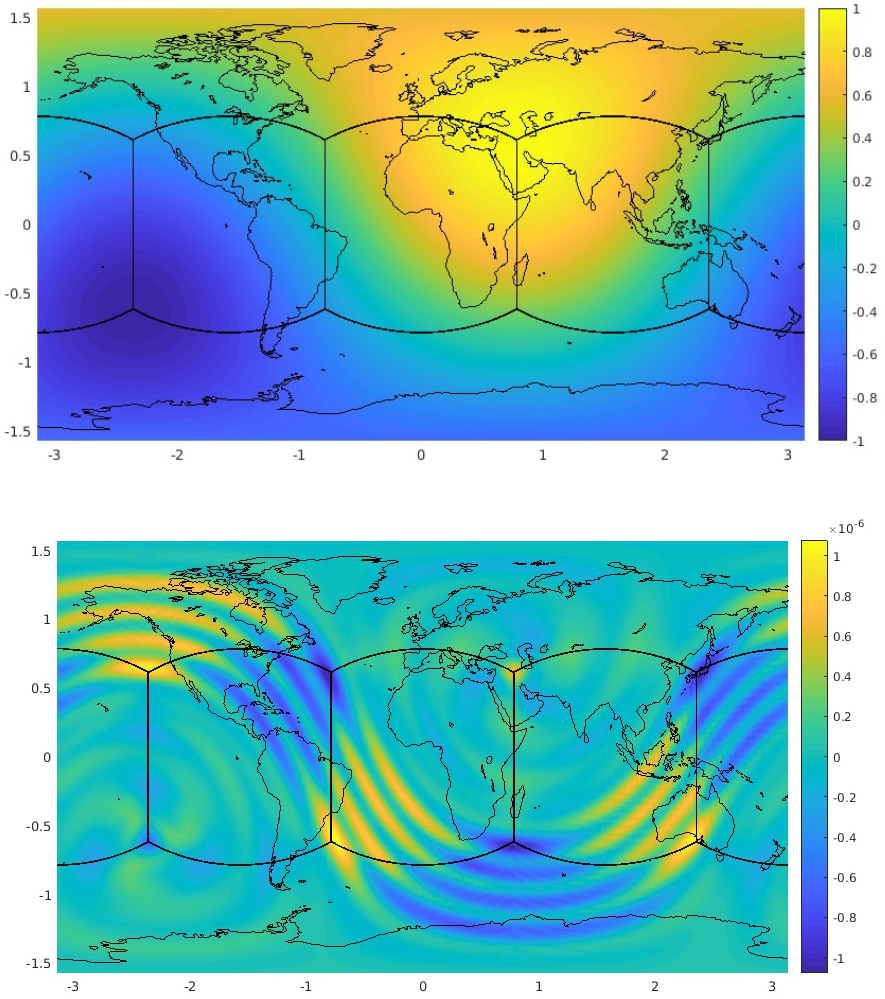


FIGURE 5.16 – Solution exacte (haut) et erreur (bas) pour le test stationnaire sur l'équation (5.48). Le paramètre de la Cubed-Sphere est $N = 32$. Le pas de temps est donné par $\Delta t = 0.96\Delta\xi/\pi = 0.015$. On représente les fonctions au temps $t = 6$.

Chapitre 6

Equations Shallow Water sphériques

6.1 Equation Shallow Water linéarisée

6.1.1 Propriétés de l'équation Shallow Water linéarisée

En l'absence de reliefs, l'équation Shallow Water s'écrit

$$\left\{\begin{array}{rcl} \dfrac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla_T)\,\mathbf{u} + f\mathbf{n}\wedge\mathbf{u} + g\nabla_T h &=& \mathbf{0} \\ \dfrac{\partial h}{\partial t} + \nabla_T\cdot(h\mathbf{u}) &=& 0, \end{array}\right. \tag{6.1}$$

où f désigne la force de Coriolis.

Pour linéariser le système d'équations (6.1) autour de la solution stationnaire $(H,\overline{\mathbf{u}}) = (H,\mathbf{0})$, nous considérons les solutions de la forme :

$$\left\{\begin{array}{rcl} h &=& H+\tilde{\eta} \\ \mathbf{u} &=& \tilde{\mathbf{u}} \end{array}\right. \tag{6.2}$$

où $(\tilde{\eta},\tilde{\mathbf{u}})$ représente une petite perturbation de l'état stationnaire $(H,\overline{\mathbf{u}}) = (H,\mathbf{0})$. En incorporant les solutions de cette forme à l'équation (6.1), on trouve :

$$\left\{\begin{array}{rcl} \dfrac{\partial \tilde{\mathbf{u}}}{\partial t} + (\tilde{\mathbf{u}}\cdot\nabla_T)\,\tilde{\mathbf{u}} + f\mathbf{n}\wedge\tilde{\mathbf{u}} + g\nabla_T\tilde{\eta} &=& \mathbf{0} \\ \dfrac{\partial \tilde{\eta}}{\partial t} + \nabla_T\cdot(\tilde{\eta}\tilde{\mathbf{u}}) + H\nabla_T\tilde{\mathbf{u}} &=& 0. \end{array}\right. \tag{6.3}$$

En négligeant les termes d'ordres 2 : $(\tilde{\mathbf{u}}\cdot\nabla_T)\tilde{\mathbf{u}}$ et $\nabla_T(\tilde{\eta}\tilde{\mathbf{u}})$, on obtient l'équation Shallow Water linéarisée :

$$\left\{\begin{array}{rcl} \dfrac{\partial \mathbf{u}}{\partial t} + f\mathbf{n}\wedge\mathbf{u} + g\nabla_T\eta &=& \mathbf{0} \\ \dfrac{\partial \eta}{\partial t} + H\nabla_T\cdot\mathbf{u} &=& 0. \end{array}\right. \tag{6.4}$$

Cette équation est munie d'une condition initiale. Pour simplifier les notations, nous notons $\mathbf{u}$ au lieu de $\tilde{\mathbf{u}}$ et η au lieu de $\tilde{\eta}$.

L'équation Shallow Water linéarisée (6.4) est une équation de conservation. La masse et l'énergie sont conservées au cours du temps.

Proposition 6.1. *Si* $(\mathbf{u},\eta)$ *est solution de* (6.4) *alors*

- *Conservation de la masse : la masse totale est conservée au cours du temps :*

$$\frac{d}{dt}\int_{\mathbb{S}^2_a} \eta(t,\mathbf{x})\,d\sigma(\mathbf{x}) = 0. \tag{6.5}$$

- *Conservation de l'énergie : l'énergie est conservée au cours du temps :*

$$\frac{d}{dt}\int_{\mathbf{S}_a^2} g\eta^2(t,\mathbf{x}) + H|\mathbf{u}(t,\mathbf{x})|^2 d\sigma(\mathbf{x}) = 0. \tag{6.6}$$

Démonstration. Conservation de la masse : nous intégrons sur $\mathbb{S}_a^2$ la seconde équation de (6.4) alors :

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{S}_a^2}\eta(t,\mathbf{x})d\sigma(\mathbf{x}) &= \int_{\mathbb{S}_a^2}\frac{\partial\eta}{\partial t}(t,\mathbf{x})d\sigma(\mathbf{x})\\ &= -H\int_{\mathbb{S}_a^2}\nabla_T\cdot\mathbf{u}(t,\mathbf{x})d\sigma(\mathbf{x})\\ &= 0.\end{aligned}$$

Conservation de l'énergie : en ce qui concerne la conservation de l'énergie, nous procédons par étapes.

- Premièrement, notons que $\mathbf{u}$ est orthogonal à $\mathbf{n}\wedge\mathbf{u}$ alors :

$$\begin{aligned}\int_{\mathbb{S}_a^2}\frac{\partial\mathbf{u}}{\partial t}\cdot\mathbf{u} &= \frac{1}{2}\frac{\partial}{\partial t}\int_{\mathbb{S}_a^2}|\mathbf{u}|^2\\ &= -g\int_{\mathbb{S}_a^2}\nabla_T\eta\cdot\mathbf{u},\end{aligned}$$

en d'autres termes :

$$\frac{1}{2}\frac{\partial}{\partial t}|||\mathbf{u}|||^2_{L^2(\mathbb{S}_a^2)} = -g\int_{\mathbb{S}_a^2}\nabla_T\eta\cdot\mathbf{u}. \tag{6.7}$$

- De la même manière, en multipliant par η, on a

$$\frac{\partial\eta}{\partial t}\cdot\eta = -H\eta\nabla_T\cdot\mathbf{u}. \tag{6.8}$$

En intégrant sur la sphère $\mathbb{S}_a^2$, on a

$$\frac{1}{2}\frac{\partial}{\partial t}\|\eta\|^2_{L^2(\mathbb{S}_a^2)} = -H\int_{\mathbb{S}_a^2}\eta\nabla\cdot\mathbf{u} \tag{6.9}$$

- De plus pour tout champ de vecteurs $\mathbf{A}\in\mathbb{TS}_a^2$ et pour toute fonction B, on a

$$\nabla_T\cdot(\mathbf{A}B) = (\mathbf{A}\cdot\nabla_T)\,B + (B\nabla_T\cdot\mathbf{A}) \tag{6.10}$$

On obtient alors

$$\begin{aligned}\frac{\partial}{\partial t}\int_{\mathbb{S}_a^2}\left(g\eta^2 + H|u|^2\right) &= -2gH\int_{\mathbb{S}_a^2}\eta\nabla_T\cdot\mathbf{u} + \mathbf{u}\cdot\nabla_T\eta d\sigma(\mathbf{x})\\ &= -2gH\int_{\mathbb{S}_a^2}\nabla_T\cdot(\eta\mathbf{u})d\sigma(\mathbf{x})\\ &= 0.\end{aligned}$$

Et l'énergie est bien conservée.

□

6.1.2 Résolution numérique

Dans cette section, nous nous intéressons à la résolution numérique de l'équation (6.4) à l'aide d'un schéma analogue à celui de l'algorithme 12. Les tests que nous effectuons sont de deux types. Il s'agit d'un test stationnaire et d'un test avec second membre. Ainsi, l'équation manipulée est de la forme

$$\left\{\begin{array}{rcl} \dfrac{\partial \mathbf{u}}{\partial t} + f\mathbf{n}\wedge\mathbf{u} + g\nabla_T\eta &=& S_{\mathbf{u}} \\ \dfrac{\partial \eta}{\partial t} + H\nabla_T\mathbf{u} &=& S_\eta, \end{array}\right. \tag{6.11}$$

où $S_{\mathbf{u}} : (t,\mathbf{x}) \in \mathbb{R}^+ \times \mathbb{S}_a^2 \mapsto S_{\mathbf{u}}(t,\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2$ et $S_\eta : (t,\mathbf{x}) \in \mathbb{R}^+ \times \mathbb{S}_a^2 \mapsto S_\eta(t,\mathbf{x}) \in \mathbb{R}$ sont des fonctions données.

Pour résoudre (6.11) en utilisant la méthode des lignes, nous définissons la fonction J_Δ agissant sur un couple de fonctions de grilles $\mathfrak{q} = (\mathfrak{u},\eta)$ par

$$J_\Delta(t,\mathfrak{q}) = J_\Delta(t,\mathfrak{u},\eta) = \left(-f^*\mathbf{n}^*\wedge\mathfrak{u}^* - g\nabla_{T,\Delta}\eta + S^*_{\mathbf{u}}, -H\nabla_{T,\Delta}\mathfrak{u} + S^*_\eta\right). \tag{6.12}$$

La résolution en temps est faite en utilisant l'algorithme de Runge-Kutta d'ordre 4 couplé à un opérateur de filtrage $\mathcal{F}$. L'algorithme 13 est l'algorithme de résolution numérique. Pour tout $n \in \mathbb{N}$, il permet de calculer $\mathfrak{q}^n = (\mathfrak{u}^n,\eta^n)$ où $\mathfrak{u}^n$ est une approximation de $\mathbf{u}(t^n,\cdot)^*$ et η^n est une approximation de $\eta(t^n,\cdot)^*$.

Algorithme 13 : Systèmes d'équations (6.11)

$\mathfrak{q}^0 = (\mathbf{u}_0^*,\eta_0^*)$ connu,
for $n = 0,1,\ldots$ **do**
  $K^{(1)} = J_\Delta(t^n,\mathfrak{q}^n)$,
  $K^{(2)} = J_\Delta\left(t^n + \frac{\Delta t}{2}, \mathfrak{q}^n + \frac{\Delta t}{2}K^{(1)}\right)$,
  $K^{(3)} = J_\Delta\left(t^n + \frac{\Delta t}{2}, \mathfrak{q}^n + \frac{\Delta t}{2}K^{(2)}\right)$,
  $K^{(4)} = J_\Delta\left(t^n + \Delta t, \mathfrak{q}^n + \Delta t K^{(3)}\right)$,
  $\mathfrak{q}^{n+1} = \mathcal{F}\left(\mathfrak{q}^n + \dfrac{\Delta t}{6}\left(K^{(1)} + 2K^{(2)} + 2K^{(3)} + K^{(4)}\right)\right)$.
end for

L'opérateur de filtrage $\mathcal{F}$ agit sur chaque composante de $\mathfrak{q}^n$, il est défini par la relation (4.92). Dans les tests effectués, une solution analytique est disponible. Nous mesurons l'erreur relative sur η en normes $\|\cdot\|_1$, $\|\cdot\|_2$ et $\|\cdot\|_\infty$ en calculant au temps t^n

$$e_l^n = \frac{\|\eta^n - \eta(t^n,\cdot)^*\|_l}{\|\eta(t^n,\cdot)^*\|_l}, \text{ avec } l \in \{1,2,\infty\}. \tag{6.13}$$

L'erreur sur $\mathbf{u}$ est calculée au temps t^n par

$$e_{\mathbf{u}}^n = \frac{\mathcal{N}(\mathfrak{u}^n - \mathbf{u}(t^n,\cdot)^*)}{\mathcal{N}(\mathbf{u}(t^n,\cdot)^*)}. \tag{6.14}$$

La norme $\mathcal{N}$ est donnée par l'équation (4.28). Lorsque S_η et $S_{\mathbf{u}}$ sont des fonctions nulles, le système d'équation (6.11) conserve la masse et l'énergie au sens de la proposition 6.1. On vérifie la conservation de ces quantités en mesurant

$$\frac{Q(\mathfrak{u}^n,\eta^n) - Q(\mathbf{u}(t^n,\cdot)^*,\eta(t^n,\cdot)^*)}{Q(\mathbf{u}(t^n,\cdot)^*,\eta(t^n,\cdot)^*)} \tag{6.15}$$

où Q représente la masse ou l'énergie numérique.

De plus, pour les tests qui suivent, nous choisissons les données physiques suivantes :

- le rayon terrestre : $a = 6371220$ mètres,

- la constante de gravité : $g = 9.80616\text{m} \cdot \text{s}^{-2}$,
- la hauteur de référence : $H = 10^5$ mètres,
- la force de Coriolis : $f = 2\Omega \sin\theta$ avec $\Omega = 7.292 \times 10^{-5}\text{s}^{-1}$ la vitesse angulaire de rotation de la Terre.

De plus, les tests numériques sont effectués en considérant une condition sur les pas de discrétisation de la forme suivante

$$\text{CFL} = \frac{c\Delta t}{a\Delta\xi} = C^{\text{ste}}, \tag{6.16}$$

c représente une vitesse caractéristique du système d'équation (6.11). On choisit

$$c = \max\left(c_{\text{grav}}, c_{\text{cor}}\right) \tag{6.17}$$

avec $c_{\text{grav}} = \sqrt{gH}$ et $c_{\text{cor}} = a\Omega$. La constante c_{grav} est la vitesse caractéristique des ondes de gravité, c_{cor} caractérise la vitesse de Coriolis et correspond à la vitesse de rotation le long de l'équateur.

6.1.3 Solution stationnaire zonale

Le premier test que nous considérons concerne la conservation d'une solution stationnaire zonale pour l'équation (6.11) sans forçage, c'est à dire que l'on a S_η et $S_{\mathbf{u}}$ des fonctions nulles. Donc le système d'équations considérées est (6.4). La masse et l'énergie sont conservées.

Définition 6.1. *On dit que η et $\mathbf{u}$ sont des solution zonales de* (6.11) *si $(\mathbf{u}, \eta)$ est solution de* (6.11) *et si en coordonnées longitude-latitude (λ, θ), on a*

- *η est indépendant de la latitude λ,*
- *$\mathbf{u}(t, \mathbf{x}) = u(t, \theta)\mathbf{e}_\lambda$ avec u une fonction définie sur la sphère indépendante de λ.*

Les solutions stationnaires zonales de (6.4) sont décrites dans la proposition suivante :

Proposition 6.2. *Les solution stationnaires zonales $(\mathbf{u}, \eta)$ du système d'équations* (6.4) *sont données par*

$$\begin{cases} \eta(\mathbf{x}) = \eta_{eq} - \dfrac{a}{g}\displaystyle\int_0^\theta f(s)u(s)ds \\ \mathbf{u}(\mathbf{x}) = u(\theta)\mathbf{e}_\lambda. \end{cases} \tag{6.18}$$

Démonstration. Le couple $(\mathbf{u}, \eta)$ est un couple de solutions stationnaires zonales de (6.4), donc il existe $u : \theta \in [-\pi/2, \pi/2] \mapsto u(\theta) \in \mathbb{R}$ tel que

$$\mathbf{u}(\mathbf{x}) = u(\theta)\mathbf{e}_\lambda \tag{6.19}$$

et η est indépendant du temps, donc

$$\begin{aligned} \frac{\partial \eta}{\partial t} + H\nabla_T \cdot \mathbf{u} &= H\nabla_T \cdot \mathbf{u} \\ &= \frac{H}{a\cos\theta}\frac{\partial u}{\partial \lambda} \\ &= 0. \end{aligned}$$

De plus, $\mathbf{u}$ est indépendant de t. En considérant la première équation de (6.4), on a

$$f\mathbf{n} \wedge \mathbf{u} + g\nabla_T \eta = 0. \tag{6.20}$$

Dans la base $(\mathbf{e}_\lambda, \mathbf{e}_\theta)$, l'équation (6.20) s'écrit

$$fu\mathbf{e}_\theta + \frac{g}{a}\left(\frac{1}{\cos\theta}\frac{\partial \eta}{\partial \lambda}\mathbf{e}_\lambda + \frac{\partial \eta}{\partial \theta}\mathbf{e}_\theta\right) = 0. \tag{6.21}$$

Or, η est indépendant de λ (car zonale), donc

$$fu + \frac{g}{a}\frac{\partial \eta}{\partial \theta} = 0. \tag{6.22}$$

Le paramètre f ne dépend que de θ donc par intégration de cette équation on obtient

$$\eta(\mathbf{x}) = \eta_{\text{eq}} - \frac{a}{g}\int_0^{\theta} f(s)u(s)ds, \tag{6.23}$$

et la proposition est démontrée. □

Dans la suite, nous considérons une solution stationnaire zonale donnée par la proposition 6.2 avec $u : \theta \in]-\pi/2, \pi/2[\mapsto u(\theta) \in \mathbb{R}$ donné par

$$u(\theta) = u_0 \psi(\theta)\mathbf{e}_\lambda, \tag{6.24}$$

où ψ désigne la fonction, de classe $\mathcal{C}^\infty$ et à support compact, définie pour $\theta \in [-\pi/2, \pi/2]$ par

$$\psi(\theta) = \begin{cases} 0 & \text{si } \theta \leq \theta_0 \\ \dfrac{1}{e_n}\exp\left[\dfrac{1}{(\theta-\theta_0)(\theta-\theta_1)}\right] & \text{si } \theta_0 \leq \theta \leq \theta_1 \\ 0 & \text{si } \theta \geq \theta_1. \end{cases} \tag{6.25}$$

La constante e_n est une constante de normalisation donnée par $e_n = \exp\left(\dfrac{-4}{(\theta_0-\theta_1)^2}\right)$. Elle permet d'assurer que $\max_{-\pi/2\leq\theta\leq\pi/2}\psi(\theta) = 1$.

L'intégrale, présente dans la condition initiale, est calculée par la méthode des trapèzes composites en utilisant un grand nombre de points d'interpolation (ici 1000). Ce choix permet de rapidement obtenir une erreur très faible de sorte que les erreurs numériques ne soient pas issues du calcul de la solution initiale.

Sur la figure 6.1, on représente l'historique de l'erreur relative en normes $\|\cdot\|_1$, $\|\cdot\|_2$ et $\|\cdot\|_\infty$ pour η au cours du temps ainsi que l'historique de l'erreur de conservation lorsque $N = 32$ sur 20 jours avec une condition CFL $= 0.9$. On constate que même sur un temps long, le comportement de l'erreur ainsi que celui de la conservation de la masse et de l'énergie est bon.

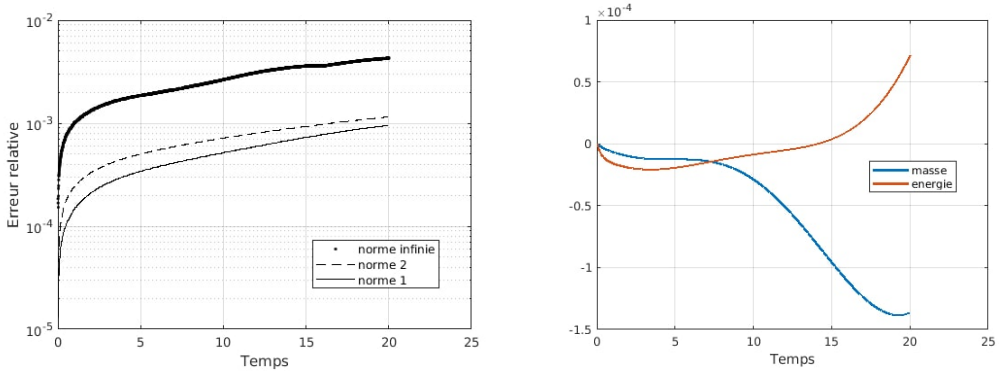


FIGURE 6.1 – Erreur pour la solution stationnaire zonale de (6.4). On mesure l'erreur en norme et erreur de conservation pour le test stationnaire de l'équation (6.4). Le pas de temps est issu de CFL $= 0.9$. Le temps final est $t = 20$ jours. Le paramètre de la Cubed-Sphere est $N = 32$.

L'analyse de convergence est fait sur la Table 6.1 et la Figure 6.2 pour des simulations sur $t = 5$ jours et CFL $= 0.9$. Le taux de convergence est supérieur à 3 pour toutes les normes. De plus, la conservation de la masse et de l'énergie est vérifiée à un ordre proche de 7, ce qui est excellent.

N	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_\infty$	$\mathbf{e}_\mathbf{u}$	**Masse**	**Énergie**
32	3.4013(−4)	5.0150(−4)	1.8541(−3)	1.0220(−2)	1.2153(−5)	1.9082(−5)
64	1.8727(−5)	3.9000(−5)	1.9349(−4)	1.9720(−3)	1.5016(−7)	1.1257(−6)
128	1.0394(−6)	1.8943(−6)	1.2232(−5)	2.4415(−4)	5.8722(−10)	7.6073(−9)
256	5.1235(−7)	4.7288(−7)	7.3839(−7)	1.3996(−5)	8.5125(−12)	1.8352(−11)
Ordre :	3.23	3.45	3.78	3.16	6.93	6.72

TABLE 6.1 – Table de convergence pour le test stationnaire de l'équation (6.4). Le pas de temps est donné par CFL = 0.9. Le temps final est $t = 5$ jours. On mesure également l'erreur sur la conservation de la masse et de l'énergie.

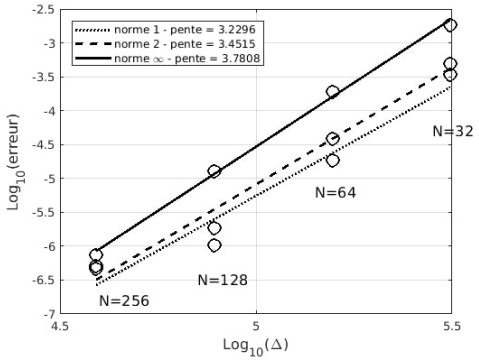


FIGURE 6.2 – Convergence pour le test stationnaire de l'équation (6.4) en fonction de $\Delta = a\Delta\xi$. Le pas de temps est donné par CFL = 0.9. Le temps final est $t = 5$ jours.

6.1.4 Solution à décroissance exponentielle

Le test précédent mesure le comportement d'une solution stationnaire. Nous étudions à présent le comportement d'une solution dépendant du temps de manière à mesurer les effets de la discrétisation en temps. Nous considérons des solutions à décroissance exponentielle :

$$\left\{ \begin{array}{rcl} \eta(t,\mathbf{x}) & = & \psi(\theta)\exp\left(-\sigma t\right) \\ \mathbf{u}(t,\mathbf{x}) & = & \dfrac{\sqrt{gH}}{10}\psi(\theta)\exp\left(-\sigma t\right)\mathbf{e}_\lambda, \end{array} \right. \tag{6.26}$$

la fonction ψ est donnée par (6.25), $\sigma > 0$ est le paramètre de décroissance. L'équation résolue est (6.11). Dans cette dernière $S_\mathbf{u}$ et S_η sont données et calculées pour que (6.26) représente le couple de solutions de (6.11).

Pour les simulations numériques, les paramètres choisis sont $\theta_0 = -\pi/3$, $\theta_1 = \pi/3$ et $\sigma = 10^{-5}\mathrm{s}^{-1}$. Le temps de demi-vie est $\ln(2)/\sigma \approx 0.8$ jour. On étudie le taux de convergence de la solution jusqu'à $t = 1.5$ heures avec une condition sur les pas de discrétisation CFL = 0.9 et différents paramètres de grilles. On se limite à un temps court car la décroissance exponentielle vers 0 de la solution rend le calcul d'erreurs délicat sur des temps plus longs. Au bout de 1.5 heure, la solution a diminué d'environ 5%. Les résultats sont donnés dans la Figure 6.3 et la Table 6.2. Le taux de convergence est supérieur à 4.

Nous ne mesurons pas l'erreur sur la conservation de la masse ou de l'énergie car ces dernières ne sont pas conservées lorsque $S_\mathbf{u}$ et S_η ne sont pas nulles.

Les tests effectués nous permettent de mesurer la précision et l'ordre de convergence en combinant les opérateurs divergence, gradient pour la discrétisation spatiale ainsi que la discrétisation RK4 couplée à l'opérateur de filtrage. Les ordres de convergence restent proches de 4 pour l'erreur sur η. L'ordre de convergence est supérieur à 3 pour l'erreur sur $\mathbf{u}$.

N	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_\infty$	$\mathbf{e}_\mathbf{u}$
32	2.2494(−2)	4.3922(−2)	2.0509(−1)	2.6585(−3)
64	1.0764(−3)	2.4775(−3)	1.7295(−2)	4.1696(−4)
128	4.7943(−5)	8.7464(−5)	6.2330(−4)	2.8324(−5)
Ordre :	4.44	4.49	4.18	3.28

TABLE 6.2 – Table de convergence pour le test à décroissance exponentielle de l'équation (6.11). Le pas de temps est donné par CFL = 0.9. Le temps final est $t = 1.5$ heures.

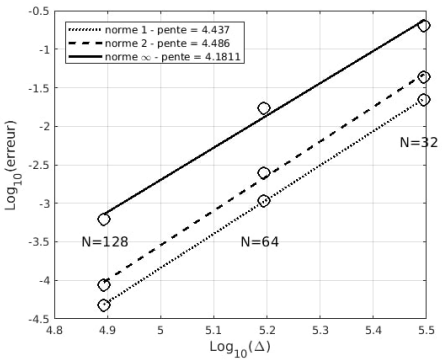


FIGURE 6.3 – Convergence pour le test à décroissance exponentielle de l'équation (6.11) en fonction de $\Delta = a\Delta\xi$. Le pas de temps est donné par CFL = 0.9. Le temps final est $t = 1.5$ heures.

6.2 Equation Shallow Water

6.2.1 Propriétés de l'équation Shallow Water

L'équation Shallow Water est déduite de l'équation de Navier-Stokes en dimension 3 en tenant compte d'une faible profondeur de fluide et de la faible viscosité. Si l'on note h_s la fonction décrivant les reliefs sur la sphère, le système d'équations s'écrit

$$\left\{\begin{array}{rcl} \dfrac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla_T)\,\mathbf{u} + f\mathbf{n}\wedge\mathbf{u} + g\nabla_T h &=& \mathbf{0} \\ \dfrac{\partial h^\star}{\partial t} + \nabla_T\cdot(h^\star\mathbf{u}) &=& 0 \end{array}\right. \tag{6.27}$$

avec $h^\star = h - h_s$. Dans la suite, nous supposons que les reliefs ne se déforment pas avec les mouvements du fluide. Ainsi h_s est indépendant du temps t. Afin d'éviter de discrétiser le terme $(\mathbf{u}\cdot\nabla)\,\mathbf{u}$, nous utilisons l'égalité suivante

$$(\mathbf{u}\cdot\nabla_T)\,\mathbf{u} = \nabla_T\left(\frac{1}{2}|\mathbf{u}|^2\right) + \zeta\mathbf{n}\wedge\mathbf{u} \tag{6.28}$$

où $\zeta = \mathbf{n}\cdot(\nabla\wedge\mathbf{u})$ la vorticité relative. On note la présence de $\nabla_T\wedge\mathbf{u}$ le rotationnel de $\mathbf{u}$ dont nous avons vu un opérateur de discrétisation dans la définition 4.6.

L'équation (6.27) s'écrit :

$$\left\{\begin{array}{rcl} \dfrac{\partial \mathbf{u}}{\partial t} + \nabla_T\left(gh + \dfrac{1}{2}|\mathbf{u}|^2\right) + (\zeta+f)\,\mathbf{n}\wedge\mathbf{u} &=& \mathbf{0} \\ \dfrac{\partial h^\star}{\partial t} + \nabla_T\cdot(h^\star\mathbf{u}) &=& 0, \end{array}\right. \tag{6.29}$$

où f est la fonction paramétrant la force de Coriolis. Sauf mention contraire, cette fonction est donnée par

$$f(\theta) = 2\Omega\sin\theta. \tag{6.30}$$

Dans les équations, les constantes physiques sont données par

- la constante de gravité : $g = 9.80616m \cdot \mathrm{s}^{-2}$,
- la vitesse angulaire de rotation de la sphère : $\Omega = 7.292 \times 10^{-5}\mathrm{s}^{-1}$,
- le rayon terrestre : $a = 6.37122 \times 10^{6}\mathrm{m}$.

Le système d'équations (6.29) est une loi de conservation. Les propriétés de conservations suivantes sont vérifiées

Proposition 6.3. *Si* $(\mathbf{u}, h)$ *est solution de* (6.29) *alors les relations de conservations suivantes sont vérifiées :*

- *Conservation de la masse totale :*
$$\frac{d}{dt}\int_{\mathbb{S}_a^2} h^\star(t,\mathbf{x})d\sigma(\mathbf{x}) = 0, \tag{6.31}$$
- *Conservation de l'énergie :*
$$\frac{d}{dt}\int_{\mathbb{S}_a^2} \frac{1}{2}h^\star(t,\mathbf{x})\mathbf{u}(t,\mathbf{x})^2 + \frac{1}{2}g\left(h^2(t,\mathbf{x}) - h_s^2(\mathbf{x})\right)d\sigma(\mathbf{x}) = 0, \tag{6.32}$$
- *Conservation de l'enstrophie potentielle :*
$$\frac{d}{dt}\int_{\mathbb{S}_a^2} \frac{\left(\zeta(t,\mathbf{x}) + f\right)^2}{h^\star(t,\mathbf{x})}d\sigma(\mathbf{x}) = 0, \tag{6.33}$$
- *Conservation de la vorticité :*
$$\frac{d}{dt}\int_{\mathbb{S}_a^2} \zeta(t,\mathbf{x})d\sigma(\mathbf{x}) = 0, \tag{6.34}$$
- *Conservation de la divergence*
$$\frac{d}{dt}\int_{\mathbb{S}_a^2} \nabla_T \cdot \mathbf{u}(t,\mathbf{x})d\sigma(\mathbf{x}) = 0, \tag{6.35}$$

avec $\zeta = (\nabla_T \wedge \mathbf{u}(t,\mathbf{x})d\sigma(\mathbf{x})) \cdot \mathbf{n}$.

Remarque 6.1. *Pour prouver qu'une quantité* δ *est conservée, il suffit de montrer qu'il existe* $\mathbf{F} : \mathbf{x} \in \mathbb{S}_a^2 \mapsto \mathbf{F}(\mathbf{x}) \in \mathbb{T}_{\mathbf{x}}\mathbb{S}_a^2$ *tel que :*
$$\frac{\partial \delta}{\partial t} = \nabla_T \cdot \mathbf{F} \tag{6.36}$$
puis d'intégrer.

Démonstration. La conservation de la masse est obtenue en intégrant la seconde équation de (6.29). La conservation de la divergence est immédiate car
$$\int_{\mathbb{S}_a^2} \nabla_T \cdot \mathbf{u}(t,\mathbf{x})d\sigma(\mathbf{x}) = 0. \tag{6.37}$$

On pose $q = \dfrac{\zeta + f}{h^\star}$. En appliquant $\mathbf{n} \cdot (\nabla_T \wedge \cdot)$ à l'équation Shallow Water (6.29), on obtient
$$\frac{\partial \zeta}{\partial t} + \nabla_T \wedge (qh^\star \mathbf{n} \wedge \mathbf{u}) \cdot \mathbf{n} + \underbrace{\nabla_T \wedge \nabla_T \left(gh + \frac{1}{2}\mathbf{u}^2\right) \cdot \mathbf{n}}_{=0} = 0. \tag{6.38}$$

Or, l'égalité suivante est vérifiée :

$$\nabla_T \wedge (qh^\star \mathbf{n} \wedge \mathbf{u}) = -(\nabla_T \wedge \mathbf{u}) \cdot (qh^\star \mathbf{n}), \tag{6.39}$$

d'où

$$\frac{\partial \zeta}{\partial t} + \nabla \cdot (qh^\star \mathbf{u}) = 0. \tag{6.40}$$

En intégrant cette dernière équation sur $\mathbb{S}_a^2$, on trouve la conservation de la vorticité.

La fonction f est indépendante du temps, donc :

$$\begin{aligned} \frac{\partial}{\partial t}(qh^\star) &= \frac{\partial}{\partial t}(\zeta + f) \\ &= \frac{\partial \zeta}{\partial t} \end{aligned}$$

est vérifiée. Donc :

$$\frac{\partial}{\partial t}(qh^\star) + \nabla_T \cdot (qh^\star \mathbf{u}) = 0 \tag{6.41}$$

ce qui prouve la conservation de l'enstrophie potentielle.

L'énergie est la somme de l'énergie cinétique E_c donnée par

$$E_c = \frac{1}{2} h^\star \mathbf{u}^2, \tag{6.42}$$

et de l'énergie potentielle E_p donnée par

$$E_p = \frac{1}{2} g(h^2 - h_s^2). \tag{6.43}$$

En dérivant E_c et E_p par rapport au temps et comme $(\mathbf{u}, h)$ est solution de (6.29), on obtient

$$\begin{aligned} \frac{\partial}{\partial t} E_c &= -\frac{1}{2}\mathbf{u}^2 \nabla_T \cdot (h^\star \mathbf{u}) - h^\star \mathbf{u} \cdot \nabla_T \left(\frac{1}{2}\mathbf{u}^2 + gh \right) \\ \frac{\partial}{\partial t} E_p &= -gh \nabla_T \cdot (h^\star \mathbf{u}) - gh_s \frac{\partial h_s}{\partial t}. \end{aligned} \tag{6.44}$$

Puisque h_s est indépendant du temps, on a

$$\frac{\partial}{\partial t}(E_c + E_p) = -\nabla_T \cdot \left(\frac{1}{2}\mathbf{u}^2 + gh \right). \tag{6.45}$$

En intégrant cette équation sur la sphère $\mathbb{S}_a^2$, on démontre l'équation de conservation de l'énergie. □

On considère à présent (λ', θ') les coordonnées longitudes latitude de $\mathbf{x} \in \mathbb{S}_a^2$ associées au pôle $\mathbf{P}$ de coordonnées $(\pi, \pi/2 - \alpha)$. Il est possible de passer de (λ, θ) à (λ', θ') en utilisant les équations (5.8) et inversement en utilisant les équations (5.9). Si la sphère tourne autour de l'axe (**OP**), alors la fonction f est donnée par

$$f(\mathbf{x}) = f(\theta') = 2\Omega \sin \theta'. \tag{6.46}$$

Dans ce cadre, les solutions stationnaires zonales de (6.29) font l'objet de la proposition suivante :

Proposition 6.4. *Les solutions stationnaires zonales dans le système de coordonnées longitude-latitude* (λ', θ') *du système d'équations* (6.29) *sont données par :*

$$\left\{ \begin{array}{rcl} \mathbf{u}(\theta') &=& u(\theta') \mathbf{e}_{\lambda'} \\ h(\theta') &=& h_0 - \dfrac{a}{g} \displaystyle\int^{\theta'} u(s) \left(u(s) \frac{\tan(s)}{a} + f(s) \right) ds, \end{array} \right. \tag{6.47}$$

avec $f(\theta') = 2\Omega \sin \theta'$.

Démonstration. Soit $(\mathbf{u}, h)$ solutions stationnaires zonales de (6.29). Alors $\mathbf{u}$ et h sont indépendants de λ' et du temps t. Il existe $u : \theta' \in [-\pi/2, \pi/2] \mapsto u(\theta') \in \mathbb{R}$ tel que

$$\mathbf{u}(\theta') = u(\theta')\mathbf{e}_{\lambda'}. \tag{6.48}$$

Ce champ de vecteurs $\mathbf{u}$ est à divergence nulle par construction, et h est indépendant de t, donc

$$\frac{\partial h^\star}{\partial t} + \nabla_T \cdot (h^\star \mathbf{u}) = 0. \tag{6.49}$$

En exprimant la seconde équation de (6.29) dans le système de coordonnées (λ', θ'), on a

$$\zeta + f = u(\theta')\frac{\tan\theta'}{a} - \frac{1}{a}u'(\theta') + f. \tag{6.50}$$

Ainsi :

$$(\zeta + f)\,\mathbf{k} \wedge \mathbf{u} = \left(u^2(\theta')\frac{\tan\theta'}{a} - \frac{1}{a}u(\theta')u'(\theta') + f(\theta')u(\theta')\right)\mathbf{e}_{\theta'} \tag{6.51}$$

De même, on obtient :

$$\nabla_T\left(gh + \frac{1}{2}|\mathbf{u}|^2\right) = \frac{g}{a\cos\theta'}\frac{\partial h}{\partial \lambda'}\mathbf{e}_{\lambda'} + \left[\frac{g}{a}\frac{\partial h}{\partial \theta'} + \frac{1}{a}u'(\theta')u(\theta')\right]\mathbf{e}_{\theta'} \tag{6.52}$$

Puisque la solution recherchée est stationnaire, h et $\mathbf{u}$ sont indépendants de t, d'où :

$$(\zeta + f)\,\mathbf{n} \wedge \mathbf{u} + \nabla\left(gh + \frac{1}{2}|\mathbf{u}|^2\right) = 0. \tag{6.53}$$

En traitant cette équation composante par composante, on déduit des informations sur h.

- **Composante en $\mathbf{e}_{\lambda'}$** :

$$\frac{g}{a\cos\theta'}\frac{\partial h}{\partial \lambda'} = 0 \tag{6.54}$$

  donc h est indépendant de λ', ce qui était attendu (h est zonale).

- **Composante en $\mathbf{e}_{\theta'}$** :

$$u^2(\theta')\frac{\tan\theta'}{a} + f(\theta')u(\theta') + \frac{g}{a}h'(\theta') = 0, \tag{6.55}$$

  d'où l'on déduit :

$$h'(\theta') = -u(\theta')\frac{a}{g}\left(u(\theta')\frac{\tan\theta'}{a} + f(\theta')\right). \tag{6.56}$$

  En intégrant cette dernière relation, on obtient :

$$h(\theta') = h_0 - \frac{a}{g}\int^{\theta'} u(s)\left(u(s)\frac{\tan(s)}{a} + f(s)\right)ds. \tag{6.57}$$

La proposition est prouvée. □

Les solutions stationnaires zonales servent de base à de nombreux tests. En particulier, le second test de [85] est un cas particulier de cette proposition. Dans le test 5 du même article, il s'agit d'une perturbation de ce cas par un relief. Le test de J. Galewsky et al. [38] est une perturbation d'une solution zonale stationnaire instable obtenue en perturbant h initialement.

6.2.2 Résolution numérique de l'équation Shallow Water

L'équation (6.29) est résolue numériquement en utilisant la méthode des lignes. Chaque opérateur différentiel est approché à l'aide des schémas hermitiens. La discrétisation temporelle se fait à l'aide de la méthode de Runge-Kutta d'ordre 4 couplée à un filtre $\mathcal{F}$ défini par (4.92). L'opérateur de filtrage utilisé est basé sur le filtre 1D d'ordre 10. On définit J_Δ la fonction agissant sur les fonctions de grille de la Cubed-Sphere $\mathfrak{q} = (\mathfrak{u}, \mathfrak{h})$ par

$$J_\Delta(t, \mathfrak{q}) = \begin{pmatrix} -\nabla_{T,\Delta}\left(g\mathfrak{h} + \frac{1}{2}|\mathfrak{u}|^2\right) - (\zeta_\Delta + f^*)\mathbf{n}^* \wedge \mathfrak{u} \\ -\nabla_{T,\Delta}(\mathfrak{h}^\star \mathfrak{u}) \end{pmatrix} \tag{6.58}$$

avec $\zeta_\Delta = \mathbf{n}^* \cdot (\nabla_{T,\Delta} \wedge \mathfrak{u}) = \mathrm{vort}_\Delta(\mathfrak{u})$. L'algorithme de résolution est analogue à l'algorithme 13, il est donné par l'algorithme 14.

Algorithme 14 : Systèmes d'équations (6.29)

$\mathfrak{q}^0 = (\mathfrak{u}(0,\cdot)^*, \eta(0,\cdot)^*)$ connu,
for $n = 0, 1, \ldots$ **do**
  $K^{(1)} = J_\Delta(t^n, \mathfrak{q}^n)$,
  $K^{(2)} = J_\Delta\left(t^n + \frac{\Delta t}{2}, \mathfrak{q}^n + \frac{\Delta t}{2}K^{(1)}\right)$,
  $K^{(3)} = J_\Delta\left(t^n + \frac{\Delta t}{2}, \mathfrak{q}^n + \frac{\Delta t}{2}K^{(2)}\right)$,
  $K^{(4)} = J_\Delta\left(t^n + \Delta t\mathfrak{q}^n + \Delta t K^{(3)}\right)$,
  $\mathfrak{q}^{n+1} = \mathcal{F}\left(\mathfrak{q}^n + \frac{\Delta t}{6}\left(K^{(1)} + 2K^{(2)} + 2K^{(3)} + K^{(4)}\right)\right)$.
end for

Lorsqu'une solution analytique est disponible, nous mesurons l'erreur

$$e_l = \frac{\|\mathfrak{h}^n - h(t^n,\cdot)^*\|_l}{\|h(t^n,\cdot)^*\|_l}, \text{ avec } l \in \{1, 2, \infty\}. \tag{6.59}$$

De plus, nous avons vu dans la proposition 6.3 que la solution du système (6.29) vérifie des propriétés de conservation. Pour la conservation de la masse, de l'énergie et de l'enstrophie potentielle, nous mesurons l'erreur de conservation relative :

$$\frac{Q(\mathfrak{u}(t^n,\cdot)^*, h(t^n,\cdot)^*) - Q(\mathfrak{u}^n, \mathfrak{h}^n)}{Q(\mathfrak{u}(t^n,\cdot)^*, h(t^n,\cdot)^*)}, \tag{6.60}$$

où Q désigne l'intégrale numérique à conserver. La divergence et la vorticité sont nulles dans le cadre continu, nous ne mesurons donc pas l'erreur de conservation relative de ces quantités mais l'erreur de conservation moyenne :

$$\frac{Q(\mathfrak{u}(t^n,\cdot)^*, h(t^n,\cdot)^*) - Q(\mathfrak{u}^n, \mathfrak{h}^n)}{4\pi a^2}, \tag{6.61}$$

où Q désigne une formule de quadrature sphérique appliquée à la divergence ou à la vorticité.

Dans les simulations numériques effectuées, le pas de temps Δt est proportionnel au pas d'espace $\Delta\xi$ par la relation

$$\mathrm{CFL} = \frac{c\Delta t}{a\Delta\xi} = C^{\mathrm{ste}} \tag{6.62}$$

avec $c = \max(c_{\mathrm{grav}}, c_{\mathrm{cor}}, u_0)$, $c_{\mathrm{grav}} = \sqrt{gh_0}$ et $c_{\mathrm{cor}} = a\Omega$. Les constantes h_0 et u_0 sont données par la condition initiale.

6.2.3 Solution stationnaire zonale

Dans le second test de [85], on considère une solution stationnaire zonale. D'après la proposition 6.4, $\mathbf{u}$ est de la forme

$$\mathbf{u}(t, \mathbf{x}) = u(\theta')\mathbf{e}_{\lambda'}. \tag{6.63}$$

On choisit $u(\theta') = u_0 \cos\theta'$ ce qui donne

$$\mathbf{u}(t, \lambda, \theta) = u_0 \left(\cos\theta\cos\alpha + \cos\lambda\sin\theta\sin\alpha\right)\mathbf{e}_\lambda - u_0 \sin\lambda\sin\alpha\mathbf{e}_\theta. \tag{6.64}$$

Le paramètre de Coriolis f est une fonction donnée par

$$f(\theta') = 2\Omega\sin\theta'. \tag{6.65}$$

En utilisant (5.9), on obtient

$$\begin{aligned} f(\theta') &= f(\lambda, \theta) \\ &= 2\Omega\sin\theta' \\ &= 2\Omega\left(-\cos\lambda\cos\theta\sin\alpha + \sin\theta\cos\alpha\right). \end{aligned}$$

La fonction h est zonale stationnaire associée à ce choix de $\mathbf{u}$. Elle est donnée par

$$h(\lambda', \theta') = h_0 - \frac{a}{g}\int^{\theta'} u_0 \cos s \left(u_0\frac{\tan s}{a} + 2\Omega\sin s\right) ds. \tag{6.66}$$

Après intégration, on obtient :

$$\begin{aligned} h(\lambda, \theta) &= h_0 - \frac{a}{g}\left(\Omega u_0 + \frac{u_0^2}{2}\right)\sin^2\theta' \\ &= h_0 - \frac{a}{g}\left(\Omega u_0 + \frac{u_0^2}{2}\right)\left(-\cos\lambda\cos\theta\sin\alpha + \sin\theta\cos\alpha\right)^2. \end{aligned}$$

La fonction h est donnée par :

$$h(\lambda, \theta) = h_0 - \frac{a}{g}\left(\Omega u_0 + \frac{u_0^2}{2}\right)\left(-\cos\lambda\cos\theta\sin\alpha + \sin\theta\cos\alpha\right)^2. \tag{6.67}$$

On utilise h et $\mathbf{u}$ comme données initiales avec différentes valeurs de α pour observer l'influence de ce paramètre. D'après la proposition 6.4, cette condition initiale est une solution stationnaire.

Les constantes h_0, u_0 ainsi que les reliefs sur la sphère sont donnés par

- $gh_0 = 2.94 \times 10^4 m^2 \cdot \mathrm{s}^{-2}$,
- $u_0 = 2\pi a/(12 \text{ jours})$,
- $h_s \equiv 0$ (absence de reliefs sur la sphère).

Sur les figures 6.4 et 6.6 nous représentons la solution calculée au temps $t = 5$ avec une grille Cubed-Sphere de paramètre $N = 32$ pour $\alpha = 0$ et $\alpha = \pi/4$. Les calculs sont effectués sous la condition sur le pas de temps CFL $= 0.9$. Nous représentons aussi la localisation spatiale de l'erreur à l'aide de la fonction de grille :

$$\frac{\mathfrak{h}^n - h(t^n, \cdot)^*}{\|h(t^n, \cdot)^*\|_\infty}. \tag{6.68}$$

Cette représentation permet de remarquer la faible influence des bords et des coins de la Cubed-Sphere. De plus, l'amplitude de l'erreur relative est très faible, proche de 2.5×10^{-6} pour les deux valeurs de l'angle α considérés.

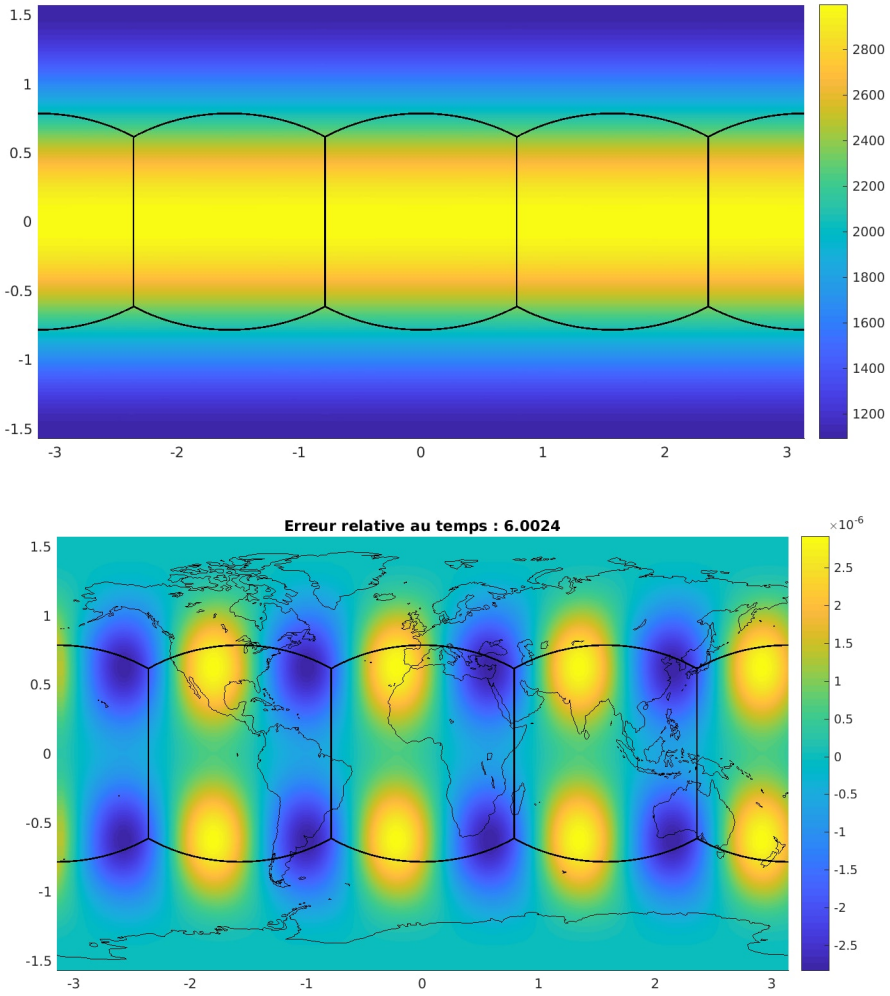


FIGURE 6.4 – Test stationnaire zonal (test numéro 2 de [85]) avec $\alpha = 0$, le paramètre de la Cubed-Sphere est $N = 32$. Le pas de temps est issu de CFL $= 0.9$. On représente h à $t = 6$ jours et l'erreur relative sur h. L'erreur n'est pas localisée aux coins de la Cubed-Sphere. De plus les niveaux d'erreur sont très bons.

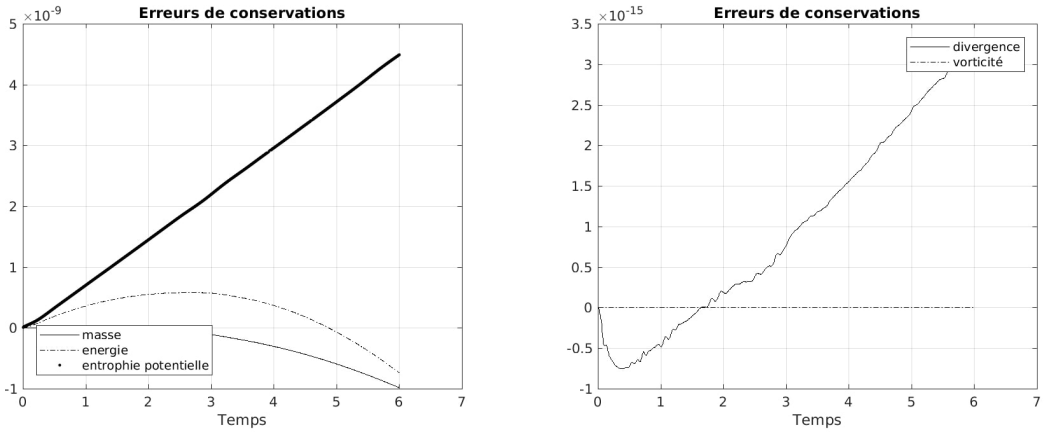


FIGURE 6.5 – Test stationnaire zonal (test numéro 2 de [85]) avec $\alpha = 0$, le paramètre de la Cubed-Sphère est $N = 32$. Le pas de temps est donné par CFL $= 0.9$. On représente l'historique de l'erreur relative sur la conservation de la masse, l'énergie et l'enstrophie (gauche), erreur sur la conservation de la divergence et de la vorticité (droite). Les ordres de grandeurs de ces erreurs sont excellents. La conservation de la vorticité est exacte ce qui est lié aux symétries de la solution sur la sphère.

Sur les Figures 6.5 et 6.7, on présente les résultats de conservation de la masse, de l'énergie, de l'enstrophie potentielle ainsi que la conservation de la divergence et de la vorticité.

L'historique de l'erreur au cours du temps est donnée Figure 6.8. Pour les normes $\|\cdot\|_1$, $\|\cdot\|_2$ et $\|\cdot\|_\infty$, ces erreurs restent très faibles et se comportent correctement.

Dans les Tables 6.3 et 6.4, on représente l'erreur pour différentes grilles lorsque $\alpha = 0$ et $\alpha = \pi/4$. Les valeurs sont obtenues avec CFL $= 0.9$. Les valeurs obtenues sur une grille donnée sont comparables à celles obtenues par la méthode des volumes finis dans [16] et meilleures que celles obtenues par la méthode de Galerkin d'ordre élevé dans [56]. Dans les deux cas de figure, la convergence se fait à l'ordre 4. Dans [80], la convergence obtenue par une méthode de volumes finis est plus rapide que l'ordre 4.

N	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_\infty$
32	1.1422(−6)	1.3885(−6)	2.4469(−6)
64	7.1216(−8)	8.6513(−8)	1.5229(−7)
128	4.4469(−9)	5.4018(−9)	9.5186(−9)
Ordre estimé :	4.00	4.00	4.00

TABLE 6.3 – Table de convergence pour le test stationnaire zonal de l'équation (6.29). Le pas de temps est donné par CFL $= 0.9$. On donne $\alpha = 0$. Le temps final est $t = 5$ jours. Le taux de convergence est proche de 4 et est excellent. De plus, les niveaux d'erreurs sont très faibles.

En norme infinie, e_∞, les valeurs obtenues au temps $t = 5$, sont proches de 2.75×10^{-6} et sont comparables à 5.86×10^{-6} sur une grille $32 \times 32 \times 6$ obtenue en utilisant un schéma volumes finis d'ordre 4 [16]. En extrapolant les données, le schéma volumes finis utilisé dans [80] donne une erreur en norme infinie proche de 1.47×10^{-6} en utilisant le flux numérique de type AUSM+.

6.2.4 Cas test de la montagne isolée

Le test 5 de [85] est une perturbation du précédent. On considère les données initiales (6.64) et (6.67) avec $\alpha = 0$. Il s'agit d'une solution stationnaire zonale lorsque $h_s \equiv 0$. On rappelle cette solution

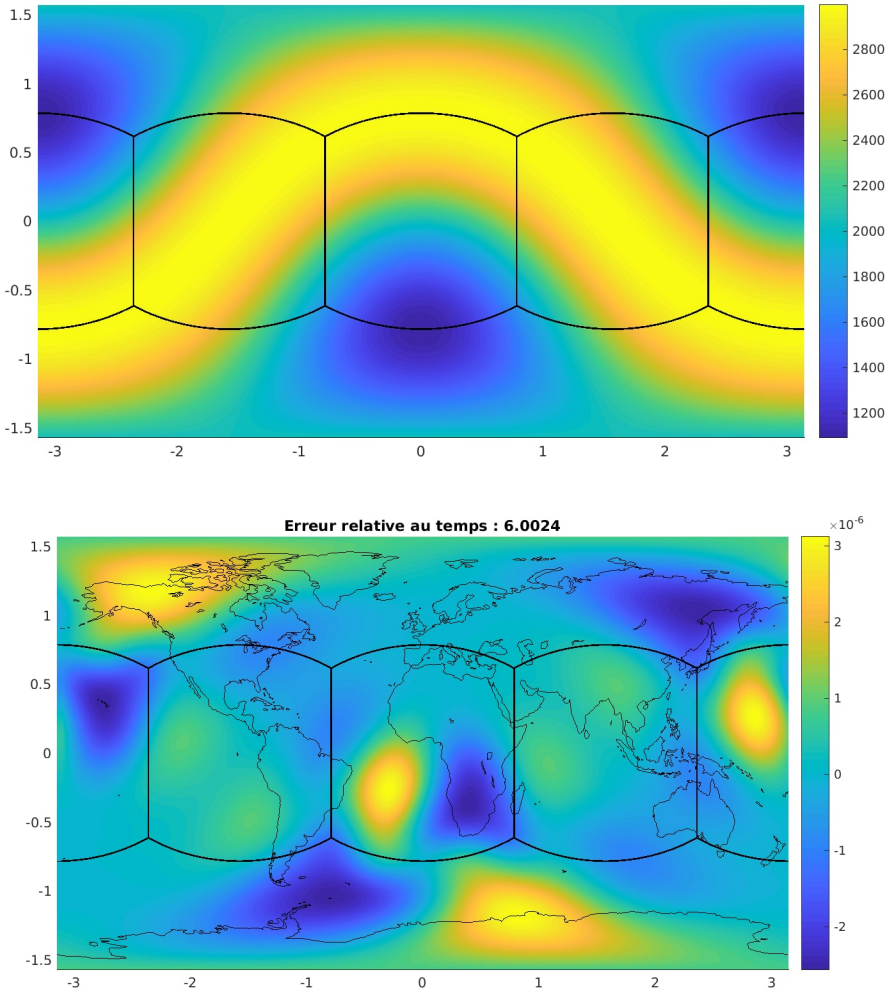


FIGURE 6.6 – Test stationnaire zonal (test numéro 2 de [85]) avec $\alpha = \pi/4$, le paramètre de la Cubed-Sphere est $N = 32$. Le pas de temps est issu de CFL $= 0.9$. On représente h à $t = 6$ jours et l'erreur relative sur h. L'erreur n'est pas localisée aux coins de la Cubed-Sphere. De plus les niveaux d'erreur sont très bons et les symétries de la solutions sont retrouvées.

N	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_\infty$
32	$7.5712(-7)$	$1.0446(-6)$	$2.7809(-6)$
64	$4.7213(-8)$	$6.5124(-8)$	$1.7387(-7)$
128	$2.9487(-9)$	$4.0672(-9)$	$1.0858(-8)$
Ordre estimé :	4.00	4.00	4.00

TABLE 6.4 – Table de convergence pour le test stationnaire zonal de l'équation (6.29). Le pas de temps est donné par la contrainte CFL $= 0.9$. On donne $\alpha = \pi/4$. Le temps final est $t = 5$ jours. Les taux d'erreurs sont proches de 4. L'erreur est très faible.

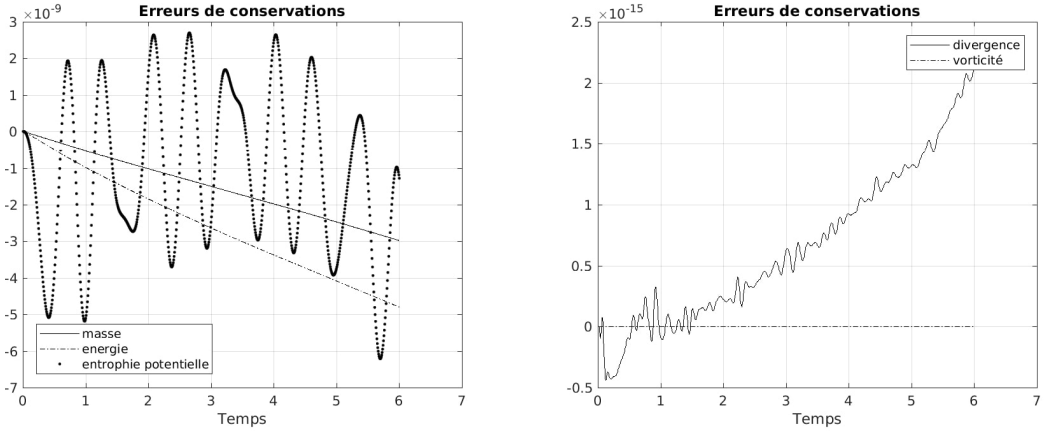


FIGURE 6.7 – Test stationnaire zonal (test numéro 2 de [85]) avec $\alpha = \pi/4$, le paramètre de la Cubed-Sphère est $N = 32$. Le pas de temps est donné par CFL $= 0.9$. On représente l'historique de l'erreurs relative sur la conservation de la masse, l'énergie et l'enstrophie (gauche), erreur sur la conservation de la divergence et de la vorticité (droite). Les ordres de grandeurs de ces erreurs sont excellents. Comme pour $\alpha = 0$, la vorticité est parfaitement conservée grâce aux symétries de la solution et du maillage.

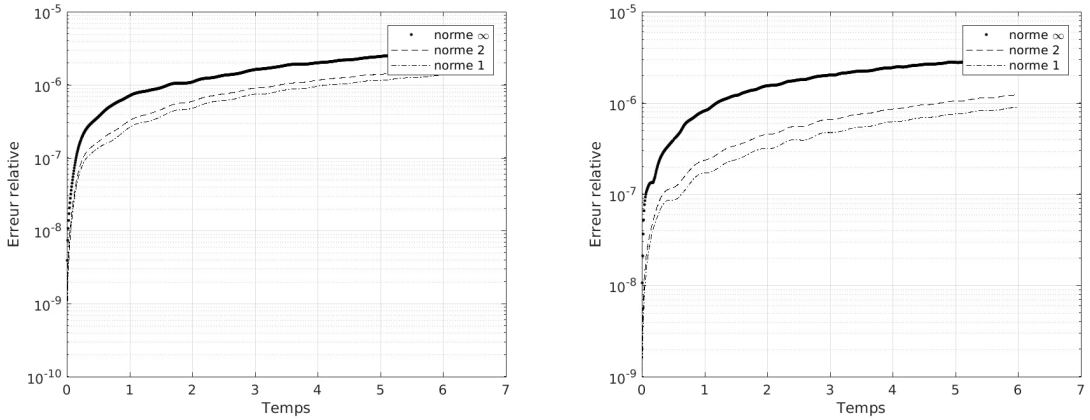


FIGURE 6.8 – Test stationnaire zonal (second test de [85]) avec le paramètre de la Cubed-Sphere $N = 32$ ainsi que $CFL = 0.9$. On représente l'historique de l'erreur relative au cours du temps $\alpha = 0$ (gauche) et $\alpha = \pi/4$ (droite). Les niveaux d'erreurs sont très faibles.

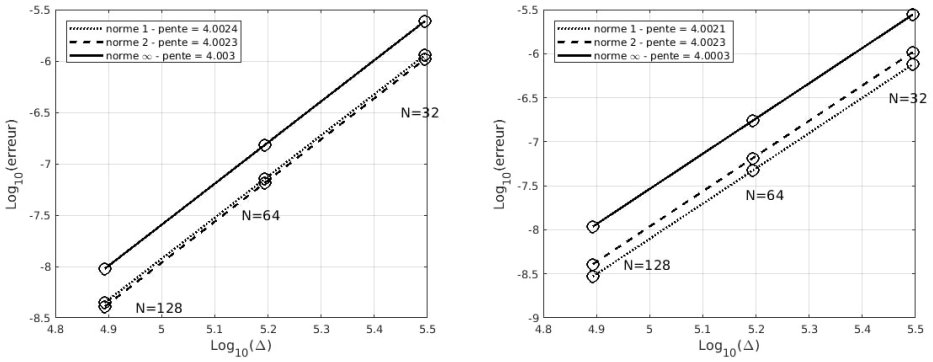


FIGURE 6.9 – Convergence pour le test stationnaire zonal de l'équation (6.29) en fonction de $\Delta = a\Delta\xi$. Le pas de temps est donné par la contrainte CFL $= 0.9$. On donne $\alpha = \pi/4$. Le temps final est $t = 5$ jours.

stationnaire :

$$\left\{ \begin{array}{rcl} \mathbf{u}(\lambda,\theta) & = & u_0 \cos\theta \mathbf{e}_\lambda \\ h(\lambda,\theta) & = & h_0 - \dfrac{a h_0}{g}\left(\Omega + \dfrac{u_0}{2}\right)\sin^2\theta. \end{array} \right. \tag{6.69}$$

La force de Coriolis est donnée par le paramètre $f(\theta) = 2\Omega \sin\theta$. Les données h_0 et u_0 sont

- $h_0 = 5960\text{m}$,
- $u_0 = 20\text{m} \cdot \text{s}^{-1}$.

Cette condition initiale est perturbée à l'aide d'un "relief". On considère la présence d'une montagne conique de hauteur $h_{s_0} = 2000$ mètres donnée par :

$$h_s = h_{s_0}\left(1 - \frac{r}{R}\right) \tag{6.70}$$

où $R = \pi/9$, $r^2 = min\left[R^2, (\lambda - \lambda_c)^2 + (\theta - \theta_c)^2\right]$. Le point $(\lambda_c, \theta_c) = (3\pi/2, \pi/6)$ correspondant à la position du sommet de la montagne. Il s'agit d'une perturbation importante, puisque la montagne représente environ 33% de l'épaisseur du fluide.

Pour ce test, aucune solution analytique n'est disponible. Nous comparons la solution obtenue aux temps $t = 5$ jours, 10 jours et 15 jours avec les résultats de la littérature [80, 56]. Pour un paramètre de Cubed-Sphere $N = 32$, on obtient les résultats des figures 6.10. Les résultats sont visuellement très similaires à ceux obtenus par des méthodes de volumes finis [51, 16] d'ordre élevé ainsi que ceux obtenus par des méthodes de Galerkin Discontinu [67]. On note en particulier l'absence d'oscillations parasites qui pourraient résulter de la forme conique de la montagne.

Les propriétés de conservation sont analysées sur la Figure 6.11. Les résultats sont très bons. Au jour 15, l'erreur sur l'enstrophie potentielle est proche de -0.9×10^{-4}, valeur comparable à celle de -1.0×10^{-4} obtenue lorsque le paramètre de la Cubed-Sphere est $N = 40$ avec un schéma volumes finis d'ordre 4 dans [81]. L'enstrophie potentielle est difficile à conserver, le schéma volumes finis utilisé dans [16] possède un historique de l'erreur de conservation similaire lorsque $N = 32$. Au temps $t = 15$ jour, l'erreur finale obtenue par [16] sur la conservation de l'enstrophie potentielle est de l'ordre de -1.1×10^{-4}.

On représente sur la Figure 6.11 l'erreur de conservation de la divergence et de la vorticité. Les erreurs sont très faibles, de l'ordre de 2×10^{-11} pour la divergence et 3×10^{-11} pour la vorticité. La valeur pour la divergence est meilleure que celle obtenue en utilisant des schémas compacts sur une grille longitude-latitude 128×64 dans [68]. En effet, avec ce schéma, l'erreur sur la conservation de la divergence est -1.1×10^{-9}, l'erreur de conservation pour la vorticité est en revanche meilleure et proche de 2.2×10^{-17}.

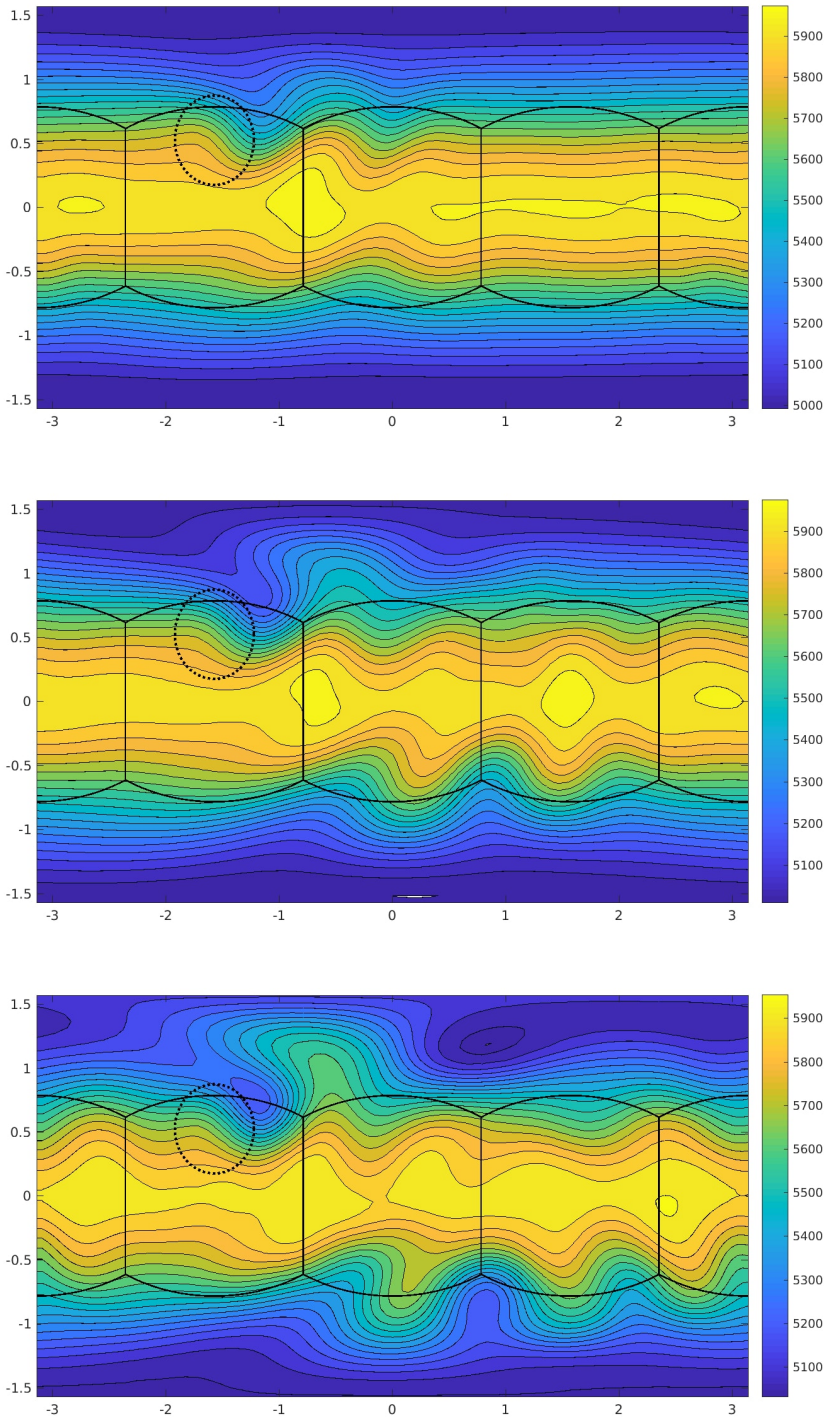


FIGURE 6.10 – Cas de la montagne isolée [85], le paramètre de la Cubed-Sphere est $N = 32$. On choisit CFL $= 0.9$. On représente $\mathfrak{h}$ aux temps $t = 5$, 10 et 15 jours (dans cet ordre, de haut en bas). Le cercle en pointillés désigne la position de la montagne.

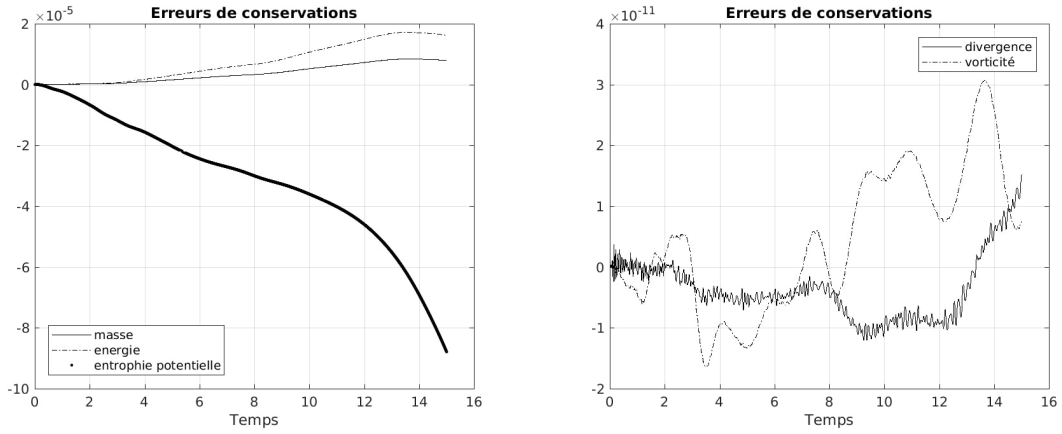


FIGURE 6.11 – Cas test de la montagne isolée [85] sur une grille $32 \times 32 \times 6$ avec CFL $= 0.9$. Erreurs relatives sur la conservation de la masse, de l'énergie et de l'enstrophie potentielle (gauche), erreur sur la conservation de la divergence et de la vorticité (droite). Les erreurs de conservation sont très faibles. L'enstrophie potentielle est la plus difficile à conserver mais l'erreur reste à un niveau acceptable.

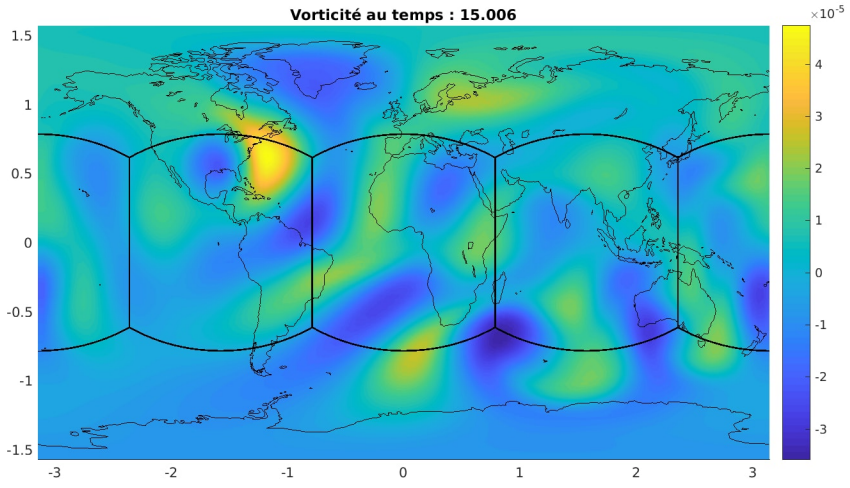


FIGURE 6.12 – Cas test de la Montagne isolée [85] sur une grille $32 \times 32 \times 6$ avec CFL $= 0.9$. On représente la vorticité à 15 jours.

6.2.5 Cas test barotrope avec instabilité

Introduit dans [38], ce test est similaire aux tests numéro 2 (Stationnaire zonale) et numéro 5 (Montagne isolée) de [85]. La condition initiale est donnée par l'état stationnaire de la proposition 6.4 avec $\alpha = 0$ que l'on perturbe.

Soit u la fonction définie par :

$$u(\theta) = u_0 \psi(\theta), \tag{6.71}$$

où ψ est la fonction à support compact définie par l'équation (6.25). Le champ de vitesse considéré est donné par $\mathbf{u} = u\mathbf{e}_\lambda$.

La donnée initiale pour h est donnée par :

$$h(\theta) = h_0 - \frac{a}{g} \int_{-\pi/2}^{\theta} u(s) \left(u(s) \frac{\tan(s)}{a} + f(s) \right) ds \tag{6.72}$$

Les valeurs des constantes sont les suivantes :

- $u_0 = 80\text{m} \cdot \text{s}^{-1}$,
- $\theta_0 = \pi/7$,
- $\theta_1 = \pi/2 - \theta_0$,
- h_0 est choisi de telle manière que h ait pour moyenne 10000m, soit approximativement $h_0 \approx 9841.8139\text{m}$.

Cette condition initiale est une solution stationnaire d'après la proposition 6.4. Cependant, cette solution stationnaire est instable et les erreurs numériques suffisent à la perturber. Le test repose sur cette instabilité. On ajoute à la condition initiale h une perturbation locale h' de la forme :

$$h'(\lambda, \theta) = \hat{h} \cos(\theta) \exp\left[-\left(\frac{\lambda_2 - \lambda}{\alpha} \right)^2 - \left(\frac{\theta_2 - \theta}{\beta} \right)^2 \right]. \tag{6.73}$$

Les constantes de la perturbation sont données par :

- $\hat{h} = 120\text{m}$, ce qui représente une perturbation de 1.2% de la condition initiale, la perturbation est donc faible,
- La perturbation est localisée en $(\lambda_2, \theta_2) = (0, \pi/4)$,
- $\alpha = 1/3$,
- $\beta = 1/15$.

Ce test consiste à observer la vorticité au fil du temps, en particulier au bout de 2 jours, 4 jours et 6 jours. La perturbation commence à être visible au bout d'environ 3 jours. Au bout de 6 jours, on compare la forme de la vorticité numérique avec celle donnée dans la littérature [38, 16]. Ce test est particulièrement difficile pour la Cubed-Sphere. En effet, la perturbation est concentrée sur le bord du Panel (V). De plus elle est localisée à l'intersection des panels (I) et (V) (voir Fig. 6.13).

La vorticité au bout de 2, 4 et 6 jours est donnée en Figure 6.14. On constate que les résultats sont bons et comparables à ceux de la littérature [16, 38, 67].

La Figure 6.15 est faite en utilisant le schéma compact à 3 points d'ordre 4 $\delta^H_{x,4}$ et celui d'ordre 4 non compact $\delta_{x,4}$ dans le calcul des opérateurs. On constate une différence sur la forme de la vorticité, en particulier dans la région de la Chine et du Japon sur la carte (intersection des panels (II), (III) et (V)). L'utilisation du schéma compact permet une meilleure représentation des hautes fréquences. Ce phénomène est visible ici où les tourbillons sont mieux représentés sur un maillage fixé $86 \times 86 \times 6$.

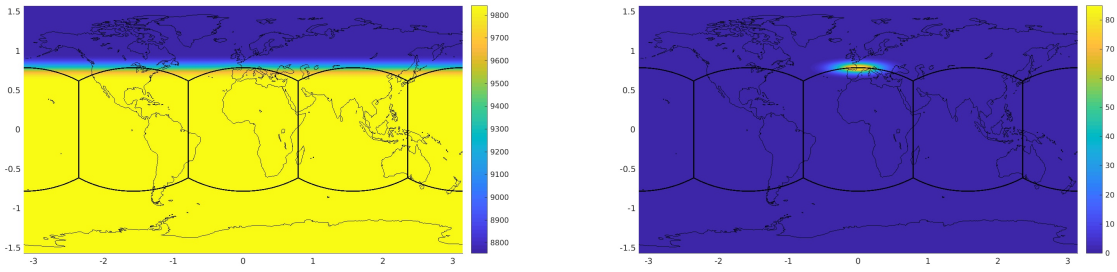


FIGURE 6.13 – Flux barotrope avec instabilité. On représente la condition initiale $h+h'$ (gauche) et perturbation initiale h' (droite) pour le Flux barotrope, [38]. La perturbation est localisée à l'intersection de différents panels.

En comparant ces résultats avec la Figure 6.17, on constate que le schéma converge plus vite lors de l'utilisation de $\delta^H_{4,x}$ que lors de l'utilisation de $\delta^H_{4,x}$.

Sur la Figure 6.16, on représente les erreurs sur la conservation sur la Cubed-Sphere $96 \times 96 \times 6$. Les propriétés de conservation sont bonnes. Comme cela avait déjà été observé pour la montagne isolée, l'enstrophie potentielle est difficile à conserver. On observe une perte importante de l'enstrophie potentielle lorsque la perturbation commence à être visible sur la vorticité, à partir du jour 4.

La convergence du schéma est visible en raffinant le maillage. C'est ce que nous représentons sur la Figure 6.17. La solution est clairement mal représentée sur la grille $32 \times 32 \times 6$, alors que nous ne distinguons pas de différences entre les grilles $96 \times 96 \times 6$ et $128 \times 128 \times 6$.

6.2.6 Cas test de type ondes de Rossby-Haurwitz

Les ondes de Rossby-Haurwitz sont des solutions analytiques de l'équation de la vorticité barotrope [46, 69]. Il s'agit du test 6 de [85]. Cependant, ce ne sont pas des solutions analytiques pour le système d'équations Shallow Water. On s'attend à observer un déplacement des ondes d'Ouest en Est.

Le champ de vitesse $\mathbf{u}$ au temps $t = 0$ est donné par :

$$\mathbf{u} = u\mathbf{e}_\lambda + v\mathbf{e}_\theta, \tag{6.74}$$

avec u et v données par

$$\left\{ \begin{array}{rcl} u & = & a\omega\cos\theta + aK\cos^{R-1}\theta\left(R\sin^2\theta - \cos^2\theta\right)\cos R\lambda \\ v & = & -aKR\cos^{R-1}\theta\sin\theta\sin R\lambda. \end{array} \right. \tag{6.75}$$

La fonction h est initialement donnée par :

$$gh = gh_0 + a^2A(\theta) + a^2B(\theta)\cos R\lambda + a^2C(\theta)\cos 2R\lambda, \tag{6.76}$$

avec

$$\left\{ \begin{array}{rcl} A(\theta) & = & \dfrac{\omega}{2}(2\Omega+\omega)\cos^2\theta + \dfrac{1}{4}K^2\cos^{2R}\theta\left[(R+1)\cos^2\theta + (2R^2-R-2) - 2R^2\cos^{-2}\theta\right] \\ B(\theta) & = & \dfrac{2(\Omega+\omega)K}{(R+1)(R+2)}\cos^R\theta\left[(R^2+2R+2) - (R+1)^2\cos^2\theta\right] \\ C(\theta) & = & \dfrac{1}{4}K^2\cos^{2R}\theta\left[(R+1)\cos^2\theta - (R+2)\right]. \end{array} \right. \tag{6.77}$$

Les constantes sont :

$$\left\{ \begin{array}{rcl} \omega & = & 7.848 \times 10^{-6}\mathrm{s}^{-1}, \\ K & = & 7.848 \times 10^{-6}\mathrm{s}^{-1}, \\ h_0 & = & 8 \times 10^3\mathrm{m}, \\ R & = & 4. \end{array} \right. \tag{6.78}$$

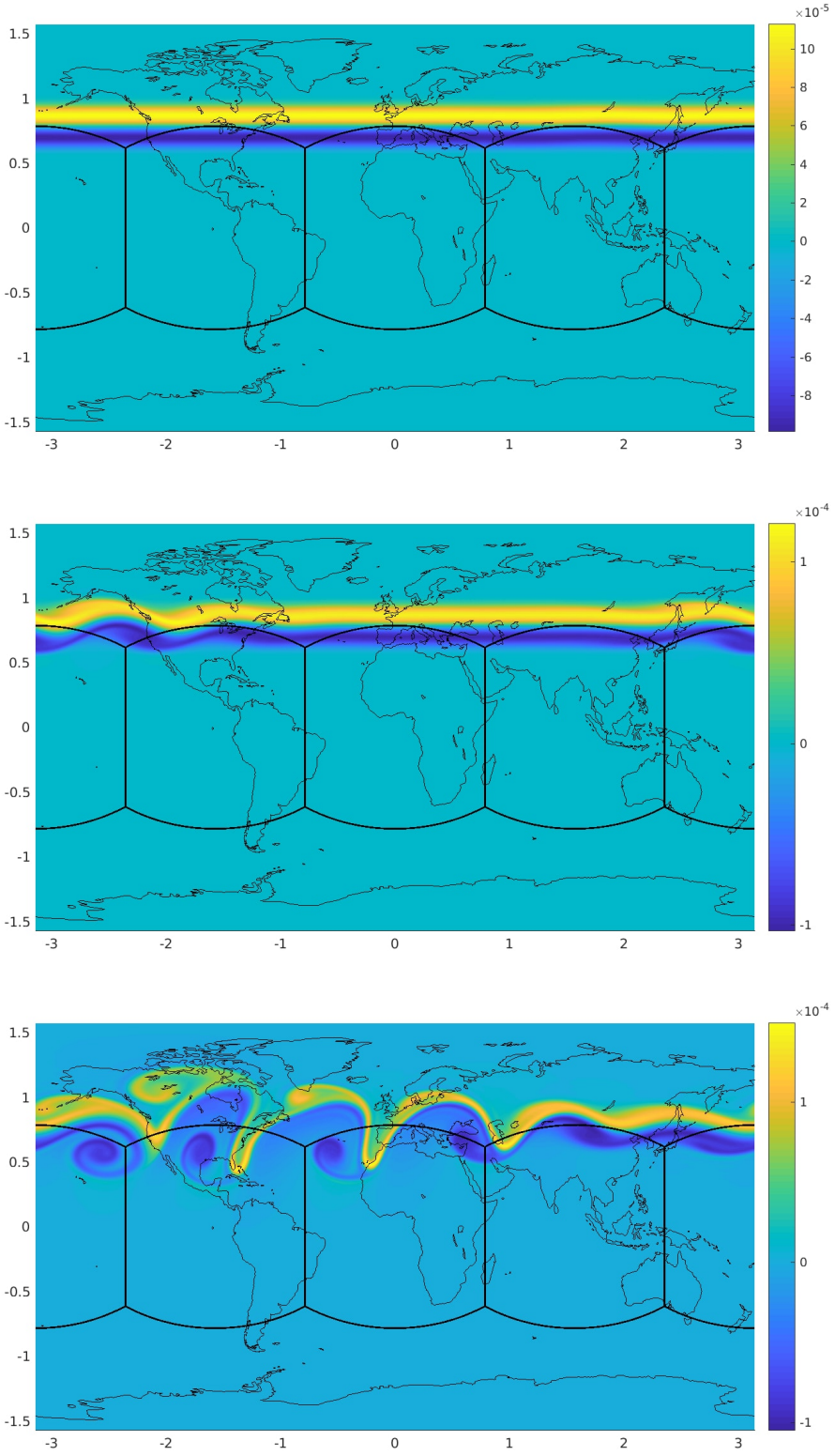


FIGURE 6.14 – Cas test du flux barotrope [38]. Au bout de 2, 4 et 6 jours (dans cet ordre, de haut en bas), on représente la vorticité. Le paramètre de la Cubed-Sphere est $N = 128$, le pas de temps est calculé grâce à la relation CFL $= 0.9$. La perturbation apparaît sur la vorticité au temps $t = 3$ jours. Au temps $t = 6$ jours, on observe le bon nombre de tourbillons ainsi que leur localisation.

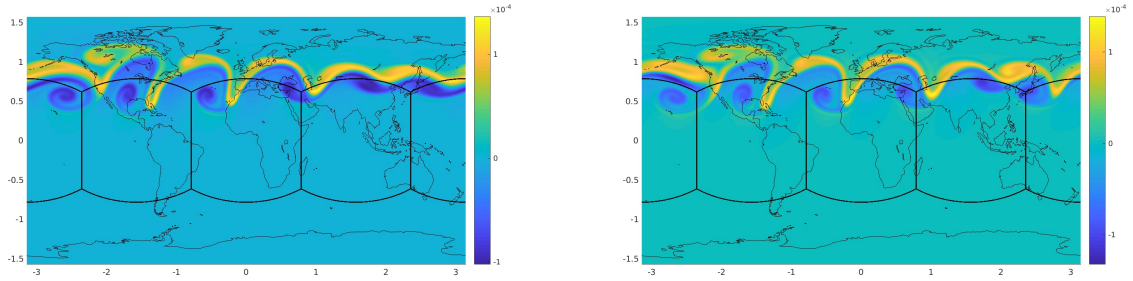


FIGURE 6.15 – Cas test barotrope instable [38] à 6 jours sur une grille $86 \times 86 \times 6$ avec CFL $= 0.9$. On représente la vorticité. A gauche, utilisation d'un schéma compact d'ordre 4 : $\delta^H_{4,x}$. A droite utilisation d'un schéma explicite d'ordre 4 $\delta_{4,x}$. Les deux solutions ne sont pas identiques, en particulier au niveau du Japon et de la Chine.

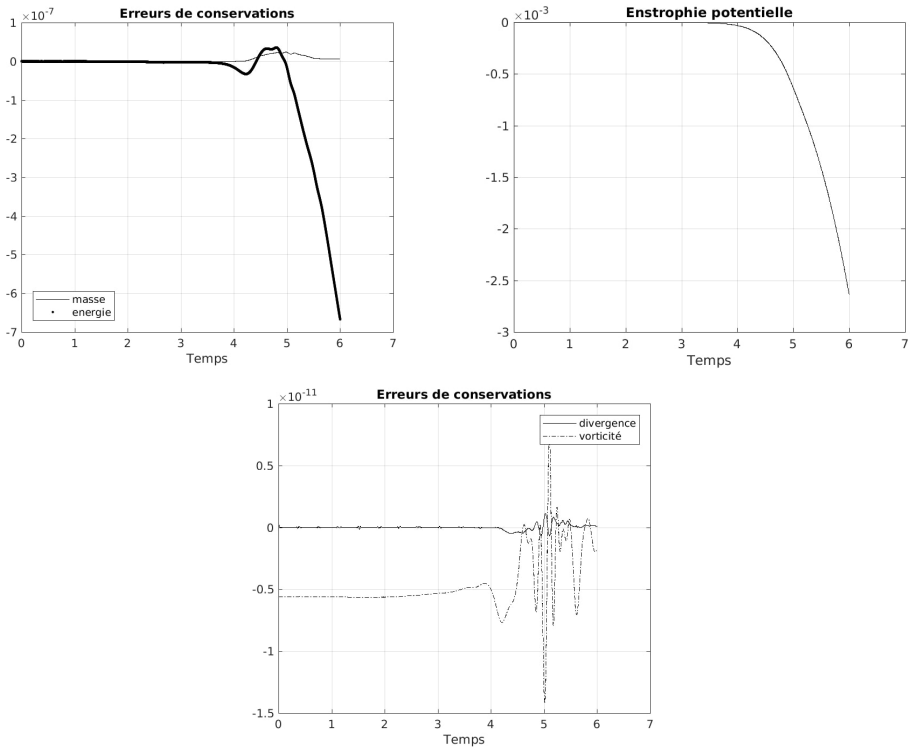


FIGURE 6.16 – Cas test barotrope instable, le paramètre de la Cubed-Sphere est $N = 128$, CFL $= 0.9$. Historique des erreurs relatives sur la conservation de la masse, de l'énergie et de l'enstrophie potentielle, erreur sur la conservation de la divergence et de la vorticité. Les niveaux d'erreurs sont très bons mais l'erreur croît autour du jour 4, lorsque l'instabilité devient visible.

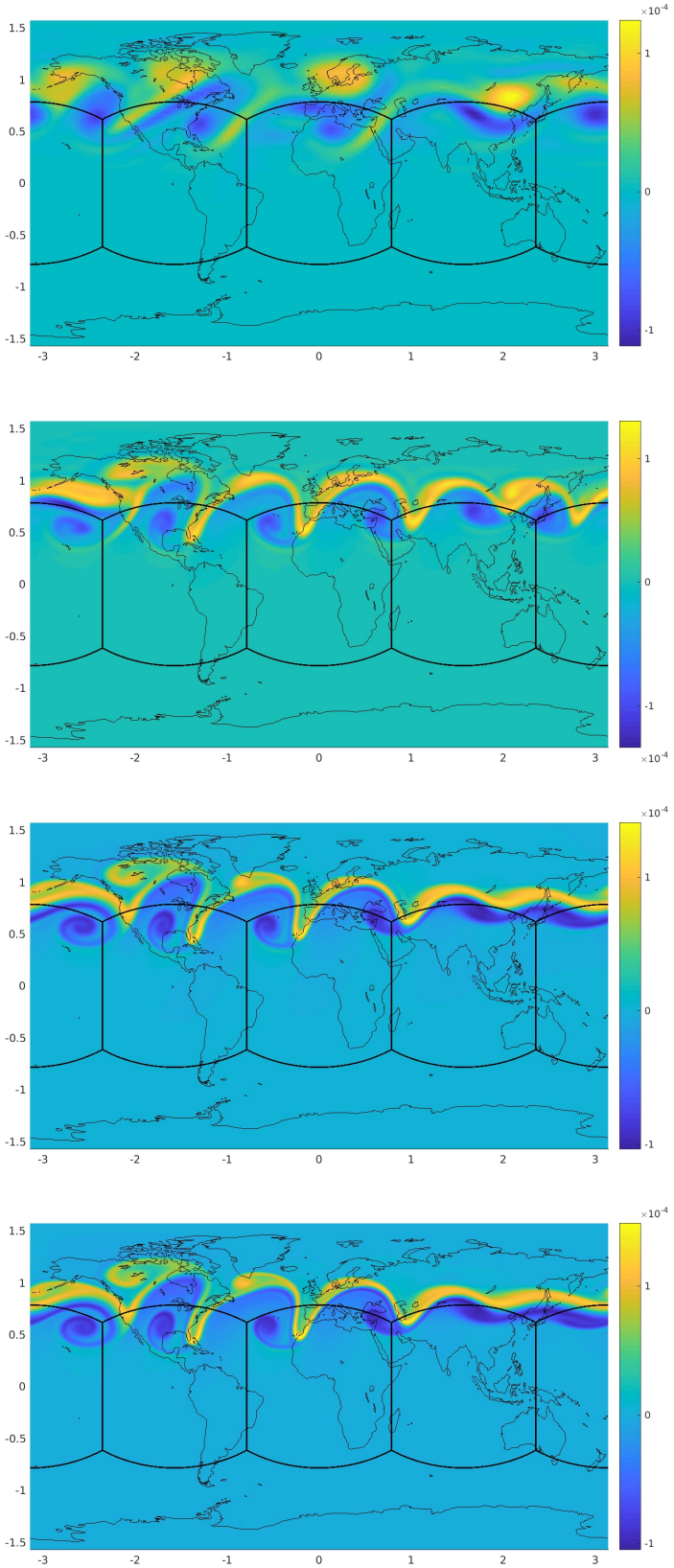


FIGURE 6.17 – Cas test barotrope avec différentes grilles $N \times N \times 6$. On représente la vorticité avec (de haut en bas) $N = 32$, $N = 64$, $N = 96$ et $N = 128$. La valeur de la condition CFL est 0.9. Les solutions lorsque $N = 96$ et $N = 128$ sont pratiquement identiques ce qui confirme la convergence du schéma.

Il est connu que cette condition initiale est instable [79]. Le comportement de "déplacement vers l'ouest" attendu est difficilement vérifié sur un temps long et les symétries de la condition initiale peuvent être perdues. C'est pour ces raisons qu'il est intéressant d'observer le comportement en temps long des simulations.

Sur la Figure 6.18, nous présentons la solution obtenue après résolution numérique aux temps $t = 7$ jours et $t = 14$ jours. Les résultats sont similaires avec ceux obtenus par éléments finis ou volumes finis [16, 38]. Les erreurs relatives de conservation pour la masse, l'énergie et l'enstrophie potentielle, ainsi que les erreurs de conservation pour la divergence et la vorticité sont données en Figure 6.19.

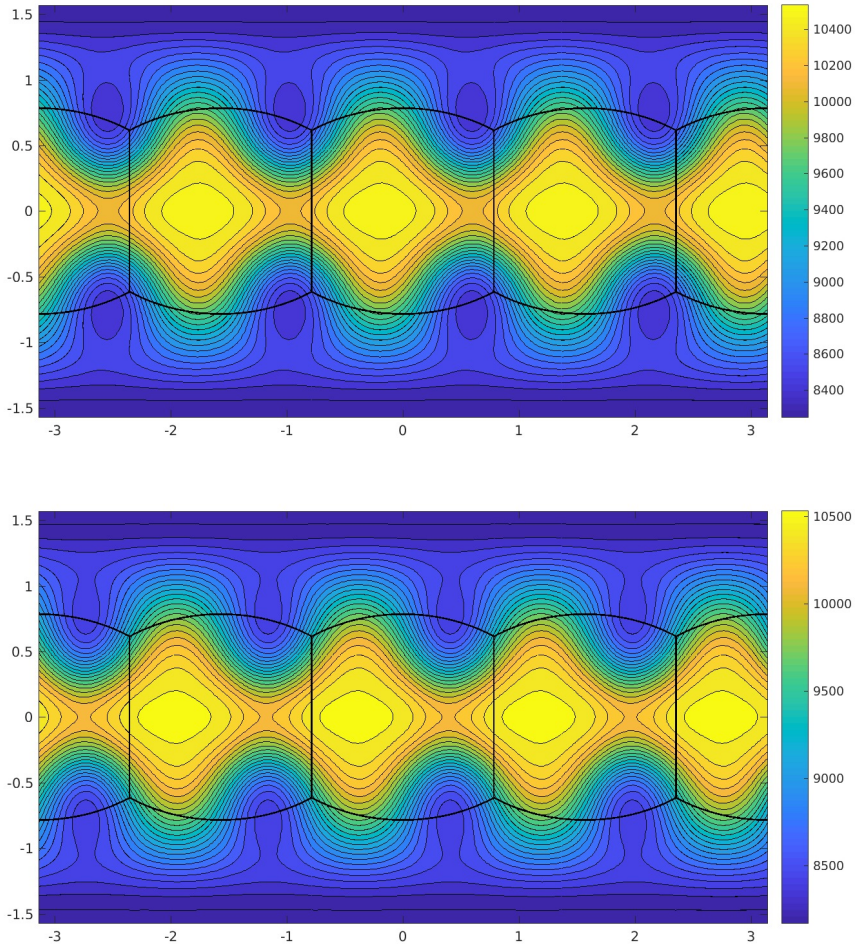


FIGURE 6.18 – Cas test de Rossby-Haurwitz à 7 (haut) et 14 jours (bas). Le paramètre de la Cubed-Sphere est $N = 80$. Le pas de temps est donné par CFL $= 0.9$. On représente $\mathfrak{h}$ à différents temps. Les résultats obtenus sont identiques à ceux obtenus dans la littérature [16, 80]

Sur la Figure 6.20, nous représentons le temps de transition entre la solution attendue pour le cas des ondes de Rossby-Haurwitz et l'instabilité. Comme cela a été également observé dans [80, 81], le temps de transition est lié aux paramètres numériques, en particulier à la dissipation numérique.

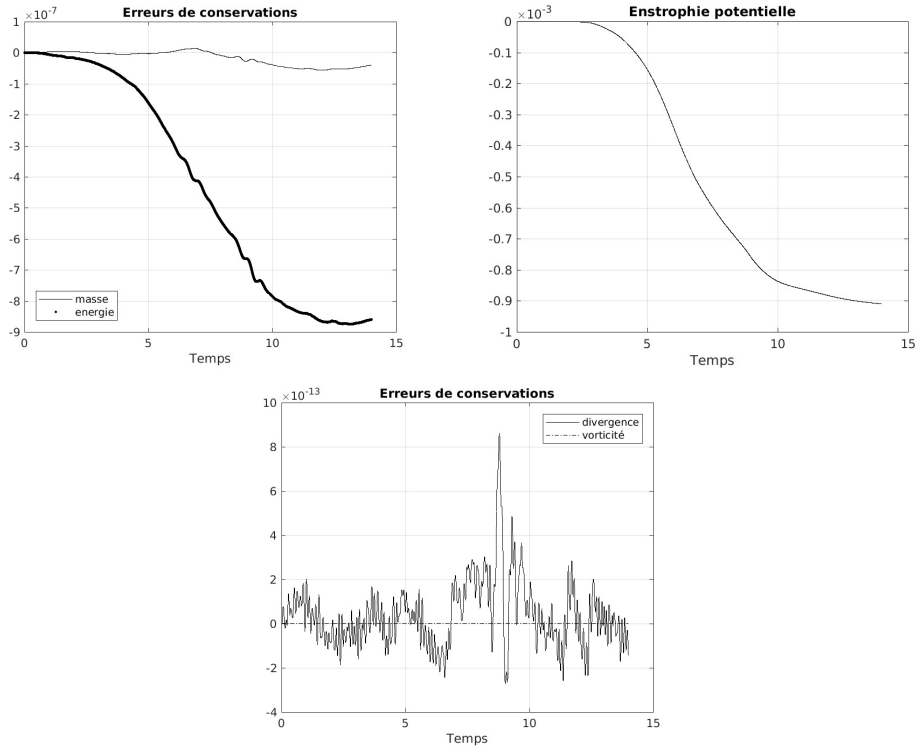


FIGURE 6.19 – Cas test des ondes de Rossby-Haurwitz, le paramètre de la Cubed-Sphere est $N = 80$, le pas de temps est donné par CFL $= 0.9$. On représente les historiques des erreurs relatives sur la conservation de la masse, de l'énergie et de l'enstrophie potentielle (haut). Historique des erreurs sur la conservation de la divergence et de la vorticité (bas). La conservation de la masse et de l'énergie est excellente, l'enstrophie potentielle est moins bien conservée mais l'erreur est semblable à l'erreur obtenue par la méthode de Galerkin ou par les volumes finis. La divergence et la vorticité sont très bien conservées grâce aux symétries de la solution.

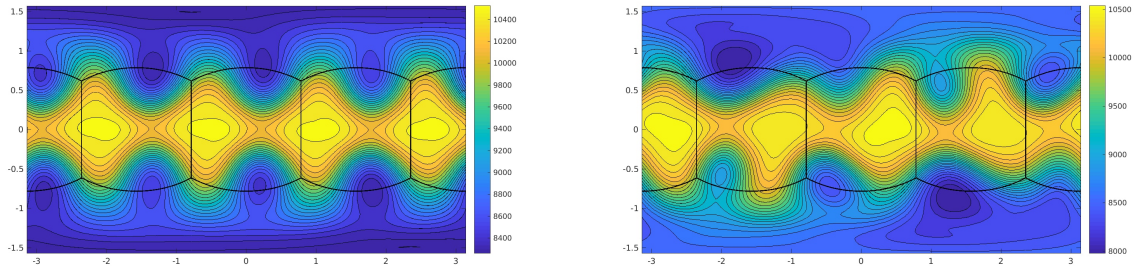


FIGURE 6.20 – Cas test des ondes de Rossby-Haurwitz à 45 et 50 jours sur une Cubed-Sphere de paramètre $N = 80$ avec CFL $= 0.9$. Au bout de 50 jours, la solution calculée $\mathfrak{h}^n$ a perdu une grande partie des symétries qui étaient présentes au temps $t = 45$.

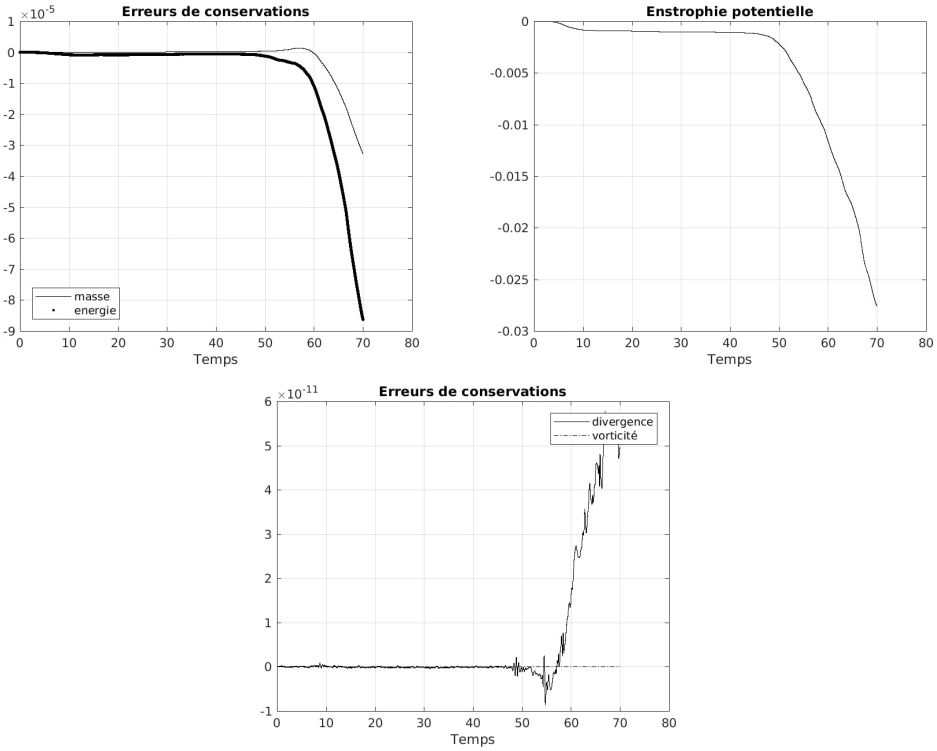


FIGURE 6.21 – Cas test des ondes de Rossby-Haurwitz, le paramètre de la Cubed-Sphere est $N = 80$, le pas de temps est donné par CFL $= 0.9$. On représente les historiques des relatives sur la conservation de la masse, de l'énergie et de l'enstrophie potentielle (haut). Historique des erreurs sur la conservation de la divergence et de la vorticité (bas). La conservation de la masse et de l'énergie sont bons même sur un temps long, l'enstrophie potentielle est moins bien conservée. La divergence et la vorticité sont très biens conservées grâce aux symétries de la solution. La perte de symétrie a lieu entre 45 et 50 jours.

Conclusion générale

Dans cette thèse, nous présentons un nouveau schéma aux différences finies pour la résolution d'équations aux dérivées partielles d'évolution sur la sphère en rotation.

Le schéma est d'abord étudié dans le cadre plan et périodique en dimensions 1 et 2. Notre schéma est centré en espace. L'approximation en temps est effectuée par un schéma de Runge-Kutta RK4. Un opérateur de filtrage est ajouté à chaque pas de temps. D'excellents résultats sont obtenus sur l'équation de transport, sur l'équation des ondes avec paramètre de Coriolis ainsi que sur l'équation de Burgers. On a observé que l'opérateur de filtrage est suffisant pour éviter l'apparition d'oscillations parasites. On observe une excellente conservation numérique de la masse. Le choix de l'ordre de précision est discuté. Un filtrage d'ordre 2, 4 ou 6 donne une importante perte de précision et une dissipation numérique de la solution. Le filtrage d'ordre 10 est un bon compromis entre stabilité, la précision et l'atténuation des oscillations parasites.

Différents types d'opérateurs différentiels discrets ont été mis en œuvre. On montre une convergence à l'ordre 3. Sur les essais numériques effectués, nous observons en pratique une convergence à l'ordre 4. Les niveaux d'erreurs observés sont très faibles.

D'autre part, nous avons utilisé ces opérateurs différentiels pour l'approximation de systèmes d'équations du type Shallow Water sphérique. Les tests effectués sur l'équation Shallow Water linéarisée et l'équation Shallow Water complète donnent des résultats comparables à ceux obtenus par des méthodes de Galerkin ou de volumes finis d'ordre élevé. Les niveaux d'erreurs observés avec notre schéma sont très bons. Bien que le schéma ne soit pas conservatif, les erreurs de conservation sont excellentes. Pour la masse le comportement est très satisfaisant. Pour l'énergie et l'enstrophie potentielle, les erreurs sont similaires à celles obtenues par d'autres méthodes y compris sur des tests difficiles tels que le test de la montagne isolée.

Les perspectives de ce travail sont les suivantes :

- l'utilisation de splines cubiques limite la montée en ordre du schéma. Nous avons observé que les splines cubiques représentent la principale source d'erreur des opérateurs différentiels discrets utilisés. L'utilisation d'harmoniques sphériques pour cette phase d'interpolation pourrait être intéressante.

- La conception d'un schéma implicite en temps est indispensable à la résolution des équations sur des temps longs ou la simulation de l'acoustique n'est pas prise en compte. Il s'agit de pouvoir effectuer les itérations en temps en utilisant des pas de temps plus grands. La conception d'un solveur rapide de type FFT serait aussi intéressante.

- Des méthodes de type zoom locaux peuvent permettre d'obtenir une excellente résolution pour des phénomènes de type tourbillon.

- La résolution de modèle en dimension 3 sur la Cubed-Sphere est également une perspective à moyen terme.

Annexe A

Opérateurs en coordonnées Longitude-Latitude

A.1 Coordonnées Longitude-Latitude

A.1.1 Système de coordonnées

Un point $\mathbf{x}$ de la sphère $\mathbb{S}_a^2 = \{(x, y, z) \in \mathbb{R}^3 \text{ tel que } x^2+y^2+z^2 = a^2\}$ est repéré par ses coordonnées longitude-latitude $(\lambda, \theta) \in]0, 2\pi] \times]-\pi/2, \pi/2[$. La donnée λ est la longitude du point donnée par l'angle équatorial et θ est l'angle latitudinal (Voir Fig. A.1). Il peut aussi être repéré par ses coordonnées cartésiennes $(x, y, z) \in \mathbb{R}^3$.

Les coordonnées cartésiennes (x, y, z) et longitude-latitude (λ, θ) sont liées par :

$$\left\{ \begin{array}{rcl} x & = & a \cos\theta \cos\lambda \\ y & = & a \cos\theta \sin\lambda \\ z & = & a \sin\theta. \end{array} \right. \tag{A.1}$$

On construit la base associée sur la sphère. La base $(\mathbf{g}_\lambda, \mathbf{g}_\theta)$ est donnée en $\mathbf{x}(x, y, z) \in \mathbb{S}_a^2$ par :

$$\mathbf{g}_\lambda = \frac{\partial \mathbf{x}}{\partial \lambda} = \begin{bmatrix} -a \cos\theta \sin\lambda \\ a \cos\theta \cos\lambda \\ 0 \end{bmatrix} \tag{A.2}$$

ainsi que

$$\mathbf{g}_\theta = \frac{\partial \mathbf{x}}{\partial \theta} = \begin{bmatrix} -a \sin\theta \cos\lambda \\ -a \sin\theta \sin\lambda \\ a \cos\theta. \end{bmatrix} \tag{A.3}$$

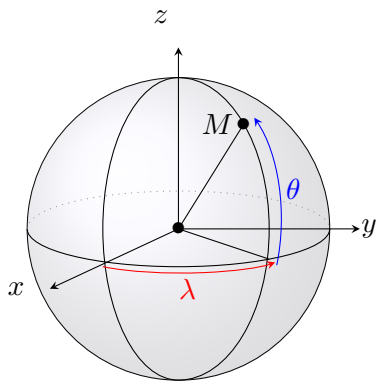


FIGURE A.1 – Longitude-Latitude

Remarque A.2. *On normalise cette base en* $\mathbf{e}_\lambda = \dfrac{1}{\|\mathbf{g}_\lambda\|}\mathbf{g}_\lambda$ *et* $\mathbf{e}_\theta = \dfrac{1}{\|\mathbf{g}_\theta\|}\mathbf{g}_\theta$. *Si* $\mathbf{F} \in \mathbb{TS}_a$ *alors il existe* F_λ *et* F_θ *tels que* $\mathbf{F} = F_\lambda \mathbf{e}_\lambda + F_\theta \mathbf{e}_\theta$.

Ainsi la métrique $\mathbf{G}$ est :

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}_\lambda \cdot \mathbf{g}_\lambda & \mathbf{g}_\lambda \cdot \mathbf{g}_\theta \\ \mathbf{g}_\theta \cdot \mathbf{g}_\lambda & \mathbf{g}_\theta \cdot \mathbf{g}_\theta \end{bmatrix} = \begin{bmatrix} a^2 \cos\theta & 0 \\ 0 & a^2 \end{bmatrix} \tag{A.4}$$

On pose $\overline{\mathbf{G}} = det(\mathbf{G}) = a^4 \cos^2\theta$. La base $(\mathbf{g}^\lambda, \mathbf{g}^\theta)$ est telle que :

$$\begin{cases} \mathbf{g}^\lambda &= \mathbf{G}^{1,1}\mathbf{g}_\lambda + \mathbf{G}^{1,2}\mathbf{g}_\theta \\ \mathbf{g}^\theta &= \mathbf{G}^{2,1}\mathbf{g}_\lambda + \mathbf{G}^{2,2}\mathbf{g}_\theta, \end{cases} \tag{A.5}$$

avec $\mathbf{G}^{i,j} = \left(\mathbf{G}^{-1}\right)_{i,j}$. De là, il découle :

$$\begin{cases} \mathbf{g}^\lambda = \dfrac{1}{a\cos\theta}\mathbf{e}_\lambda \\ \mathbf{g}^\theta = \dfrac{1}{a}\mathbf{e}_\theta. \end{cases} \tag{A.6}$$

A partir de ces vecteurs élémentaires, on peut obtenir des formules pour calculer des opérateurs différentiels sur la sphère ainsi que l'intégrale sur la surface d'une sphère.

A.1.2 Opérateurs sur la sphère

Soit une fonction $h : \mathbf{x} \in \mathbb{S}_a^2 \mapsto h(\mathbf{x})$ régulière. On peut calculer le gradient de h par :

$$\nabla_T h = \frac{\partial h}{\partial \lambda}\mathbf{g}^\lambda + \frac{\partial h}{\partial \theta}\mathbf{g}^\theta \tag{A.7}$$

ce qui se traduit en :

$$\nabla_T h = \frac{1}{a\cos\theta}\frac{\partial h}{\partial \lambda}\mathbf{e}_\lambda + \frac{1}{a}\frac{\partial h}{\partial \theta}\mathbf{e}_\theta. \tag{A.8}$$

Soit un champ de vecteurs sur la sphère $\mathbf{F} = F_\lambda \mathbf{e}_\lambda + F_\theta \mathbf{e}_\theta$ régulier. Alors

$$\begin{cases} \mathbf{F} \cdot \mathbf{g}^\lambda = \dfrac{1}{a\cos\theta}F_\lambda \\ \mathbf{F} \cdot \mathbf{g}^\theta = \dfrac{1}{a}F_\theta. \end{cases} \tag{A.9}$$

La divergence de $\mathbf{F}$ est donnée par :

$$\nabla_T \cdot \mathbf{F} = \frac{1}{\sqrt{\overline{\mathbf{G}}}}\left(\frac{\partial}{\partial \lambda}\left(\sqrt{\overline{\mathbf{G}}}\mathbf{F} \cdot \mathbf{g}^\lambda\right) + \frac{\partial}{\partial \theta}\left(\sqrt{\overline{\mathbf{G}}}\mathbf{F} \cdot \mathbf{g}^\theta\right)\right) \tag{A.10}$$

d'où :

$$\nabla_T \cdot \mathbf{F} = \frac{1}{a\cos\theta}\frac{\partial}{\partial \lambda}F_\lambda + \frac{1}{a\cos\theta}\frac{\partial}{\partial \theta}\left(\cos\theta F_\theta\right). \tag{A.11}$$

Le rotationnel de $\mathbf{F}$ se calcule grâce au produit vectoriel. Il est donné par :

$$\nabla_T \wedge \mathbf{F} = \mathbf{g}^\lambda \wedge \frac{\partial}{\partial \lambda}\mathbf{F} + \mathbf{g}^\theta \wedge \frac{\partial}{\partial \theta}\mathbf{F}. \tag{A.12}$$

Après calculs, on obtient :

$$\nabla_T \wedge \mathbf{F} = \left[F_\lambda \frac{\tan\theta}{a} + \frac{1}{a\cos\theta}\frac{\partial F_\theta}{\partial \lambda} - \frac{1}{a}\frac{\partial F_\lambda}{\partial \theta}\right]\mathbf{e}_R - \frac{F_\theta}{a}\mathbf{e}_\lambda, \tag{A.13}$$

avec $\mathbf{e}_R = (\cos\theta\cos\lambda, \cos\theta\sin\lambda, \sin\theta)^T$. Le Laplacien est la composition des deux opérateurs précédents : $\Delta_T h = \nabla_T \cdot \nabla_T h$:

$$\Delta_T h = \frac{1}{a^2\cos^2\theta}\frac{\partial^2 h}{\partial\lambda^2} + \frac{1}{a^2\cos\theta}\frac{\partial}{\partial\theta}\left(\cos\theta\frac{\partial h}{\partial\theta}\right). \tag{A.14}$$

En plus des opérateurs différentiels courants, en utilisant le changement de variable $\mathbf{x} \mapsto (\lambda, \theta)$ issu de (A.3), on peut donner une formule d'intégration :

$$\int_{\mathbb{S}^2_a} h(\mathbf{x}) d\sigma(\mathbf{x}) = \int_{\lambda=0}^{2\pi}\int_{\theta=-\pi/2}^{\pi/2} h(\lambda,\theta)\underbrace{a^2\cos\theta}_{\sqrt{\overline{\mathbf{G}}}} d\theta d\lambda. \tag{A.15}$$

Bibliographie

[1] A. Abbas. *Schémas boîte hermitiens - Algorithmes rapides pour la discrétisation des équations aux dérivées partielles.* PhD thesis, Université Paul Verlaine - Metz, 2011.

[2] J. H. Ahlberg, E. N. Nilson, and J. L. Walsh. *The Theory of Splines and Their Applications : Mathematics in Science and Engineering : A Series of Monographs and Textbooks*, volume 38. Elsevier, 2016.

[3] C. Ahrens and G. Beylkin. Rotationally invariant quadratures for the sphere. *Proceedings of the Royal Society of London A : Mathematical, Physical and Engineering Sciences*, 465(2110) :3103–3125, 2009.

[4] W. F. Ames. *Numerical methods for partial differential equations.* Academic press, 1992.

[5] K. Atkinson and W. Han. *Spherical harmonics and approximations on the unit sphere : an introduction*, volume 2044. Springer Science & Business Media, 2012.

[6] J. M. Augenbaum. An adaptive pseudospectral method for discontinuous problems. *Applied Numerical Mathematics*, 5(6) :459–480, 1989.

[7] J. M. Augenbaum, S. E. Cohn, E. Isaacson, D. P. Dee, and D. Marchesin. A factored implicit scheme for numerical weather prediction. *Communications on Pure and Applied Mathematics*, 38(5) :503–517, 1985.

[8] M. Ben-Artzi, J.-P. Croisille, and D. Fishelov. *Navier-Stokes equations in planar domains.* World Scientific, 2013.

[9] M. Ben-Artzi, J.-P. Croisille, D. Fishelov, and S. Trachtenberg. A pure-compact scheme for the streamfunction formulation of Navier–Stokes equations. *Journal of Computational Physics*, 205(2) :640–664, 2005.

[10] M. Ben-Artzi, J. Falcovitz, and P. G. LeFloch. Hyperbolic conservation laws on the sphere. A geometry-compatible finite volume scheme. *Journal of Computational Physics*, 228(16) :5650–5668, 2009.

[11] E. Blayo. Compact finite difference schemes for ocean models : 1. ocean waves. *Journal of Computational Physics*, 164(2) :241–257, 2000.

[12] C. Bogey and C. Bailly. A family of low dispersive and low dissipative explicit schemes for flow and noise computations. *Journal of Computational physics*, 194(1) :194–214, 2004.

[13] P. Bosler, L. Wang, C. Jablonowski, and R. Krasny. A lagrangian particle/panel method for the barotropic vorticity equations on a rotating sphere. *Fluid Dynamics Research*, 46(3) :031406, 2014.

[14] J. M. Burgers. A mathematical model illustrating the theory of turbulence. In *Advances in applied mechanics*, volume 1, pages 171–199. Elsevier, 1948.

[15] J. C. Butcher. *Numerical methods for ordinary differential equations.* John Wiley & Sons, 2016.

[16] C. Chen and F. Xiao. Shallow water model on cubed-sphere by multi-moment finite volume method. *Journal of Computational Physics*, 227(10) :5019–5044, 2008.

[17] P. C. Chu and C. Fan. A three-point combined compact difference scheme. *Journal of Computational Physics*, 140(2) :370–399, 1998.

[18] L. Collatz. *The numerical treatment of differential equations*, volume 60. Springer Science & Business Media, 2012.

[19] L. Comtet. *Advanced Combinatorics : The art of finite and infinite expansions*. Springer Science & Business Media, 2012.

[20] S. D. Conte and C. De Boor. *Elementary numerical analysis : an algorithmic approach*, volume 78. SIAM, 2017.

[21] J. H. Conway and R. Guy. *The book of numbers*. Springer Science & Business Media, 2012.

[22] A. W. Cook and W. H. Cabot. Hyperviscosity for shock-turbulence interactions. *Journal of Computational Physics*, 203(2) :379–385, 2005.

[23] R. Courant, K. Friedrichs, and H. Lewy. On The Partial Difference Equation of mathematical physics. *Mathematische Annalen*, 100, 1928.

[24] J.-P. Croisille. Hermitian compact interpolation on the cubed-sphere grid. *Journal of Scientific Computing*, 57(1) :193–212, 2013.

[25] J.-P. Croisille. Hermitian approximation of the spherical divergence on the Cubed-Sphere. *Journal of Computational and Applied Mathematics*, 280 :188–201, 2015.

[26] B. Cushman-Roisin and J.-M. Beckers. *Introduction to geophysical fluid dynamics : physical and numerical aspects*, volume 101. Academic Press, 2011.

[27] J.-P. Demailly. *Analyse numérique et équations différentielles-4ème Ed.* EDP sciences, 2016.

[28] G. Desquesnes. *Couplage par recouvrement de maillages curviligne/cartésien pour la simulation en aéroacoustique*. PhD thesis, Université Paris VI Pierre et Marie Curie, 2007.

[29] Q. Dubois. *Approximation volumes finis d'ordre élevé-Flux dissipatifs en maillage quelconque et applications à la LES en aérothermique cavité nacelle à l'arrêt moteur*. PhD thesis, Université de Lorraine, 2016.

[30] D. R Durran. *Numerical methods for wave equations in geophysical fluid dynamics*, volume 32. Springer Science & Business Media, 2013.

[31] C. Eldred. *Linear and nonlinear properties of numerical methods for the rotating shallow water equations*. PhD thesis, Colorado State University, 2015.

[32] R.J. Evans and I.M. Isaacs. Generalized Vandermonde determinants and roots of unity of prime order. *Proceedings of the American Mathematical Society*, 58(1) :51–54, 1976.

[33] N. Flyer and B. Fornberg. Radial basis functions : Developments and applications to planetary scale flows. *Computers & Fluids*, 46(1) :23–32, 2011.

[34] B. Fornberg and J. M. Martel. On spherical harmonics based numerical quadrature over the surface of a sphere. *Advances in Computational Mathematics*, 40(5-6) :1169–1184, 2014.

[35] B. Fornberg and C. Piret. On choosing a radial basis function and a shape parameter when solving a convective pde on a sphere. *Journal of Computational Physics*, 227(5) :2758–2780, 2008.

[36] M. S. Fox-Rabinovitz. Computational dispersion properties of horizontal staggered grids for atmospheric and ocean models. *Monthly weather review*, 119(7) :1624–1639, 1991.

[37] C. Frye and C. J. Efthimiou. Spherical Harmonics in p Dimensions. *arXiv preprint arXiv :1205.3548*, 2012.

[38] J. Galewsky, R. K. Scott, and L. M. Polvani. An initial-value problem for testing numerical models of the global shallow-water equations. *Tellus A*, 56(5) :429–440, 2004.

[39] M. Ghil and S. Childress. *Topics in geophysical fluid dynamics : atmospheric dynamics, dynamo theory, and climate dynamics*, volume 60. Springer Science & Business Media, 1987.

[40] F. X Giraldo, J. S Hesthaven, and T. Warburton. Nodal high-order discontinuous galerkin methods for the spherical shallow water equations. *Journal of Computational Physics*, 181(2) :499–525, 2002.

[41] H Golshahy, S Ghader, and F Ahmadi-Givi. Accuracy assessment of the super compact and combined compact schemes for spatial differencing of a two-layer oceanic model : Presentation of linear inertia-gravity and rossby waves. *Ocean Modelling*, 37(1-2) :49–63, 2011.

[42] E. Hairer, C. Lubich, and M. Roche. *The numerical solution of differential-algebraic systems by Runge-Kutta methods*, volume 1409. Springer, 2006.

[43] D. M Hall and R. D Nair. Discontinuous galerkin transport on the spherical yin–yang overset mesh. *Monthly Weather Review*, 141(1) :264–282, 2013.

[44] G. Hämmerlin and K-H. Hoffmann. *Numerical Mathematics*. Springer-Verlag, 1991.

[45] G. H. Hardy. *Divergent series*, volume 334. American Mathematical Soc., 2000.

[46] B. Haurwitz. The motion of atmospheric disturbances on the spherical earth. *J. of Marine Res.*, 3 :254–267, 1940.

[47] K. Hesse, I. H. Sloan, and R. S. Womersley. Numerical integration on the sphere. In *Handbook of Geomathematics*, pages 1185–1219. Springer, 2010.

[48] C. Hirsch. *Numerical computation of internal and external flows : The fundamentals of computational fluid dynamics*. Butterworth-Heinemann, 2007.

[49] W. Hundsdorfer and J. G. Verwer. *Numerical solution of time-dependent advection-diffusion-reaction equations*, volume 33. Springer Science & Business Media, 2013.

[50] L. Jiang, H. Shan, and C. Liu. Weighted compact scheme for shock capturing. *International Journal of Computational Fluid Dynamics*, 15(2) :147–155, 2001.

[51] K. K. Katta, R. D. Nair, and V. Kumar. High-order finite volume shallow water model on the cubed-sphere : 1D reconstruction scheme. *Applied Mathematics and Computation*, 266 :316–327, 2015.

[52] H. B. Keller. A new difference scheme for parabolic problems. In *Numerical Solution of Partial Differential Equations–II*, pages 327–350. Elsevier, 1971.

[53] J. W. Kim. Optimised boundary compact finite difference schemes for computational aeroacoustics. *Journal of Computational Physics*, 225(1) :995–1019, 2007.

[54] J. W. Kim and D. J. Lee. Optimized compact finite difference schemes with maximum resolution. *AIAA journal*, 34(5) :887–893, 1996.

[55] Z. Kopal. Numerical Analysis. *Chapman and Hull Ltd., London*, 1955.

[56] Y. Kuang, K. Wu, and H. Tang. Runge-Kutta discontinuous local evolution Galerkin methods for the shallow water equations on the cubed-sphere. *arXiv preprint arXiv :1608.06700*, 2016.

[57] R. Lagrange. Polynomes et fonctions de Legendre. *Gauthier-Villars*, 1939.

[58] M. Läuter, F. X Giraldo, D. Handorf, and K. Dethloff. A discontinuous galerkin method for the shallow water equations in spherical triangular coordinates. *Journal of Computational Physics*, 227(24) :10226–10242, 2008.

[59] D. Y Le Roux, V. Rostand, and B. Pouliot. Analysis of numerically induced oscillations in 2d finite-element shallow-water models part i : inertia-gravity waves. *SIAM Journal on Scientific Computing*, 29(1) :331–360, 2007.

[60] S. K. Lele. Compact Finite Difference Schemes with Spectral-like Resolution. *Journal of Computational Physics*, 103, 1991.

[61] X. Li, D. Chen, X. Peng, K. Takahashi, and F. Xiao. A multimoment finite-volume shallow-water model on the Yin–Yang overset spherical grid. *Monthly Weather Review*, 136(8) :3066–3086, 2008.

[62] A. D. McLaren. Optimal numerical integration on a sphere. *Mathematics of Computation*, 17(84) :361–383, 1963.

[63] G. Monegato and J. N. Lyness. The Euler-MacLaurin expansion and finite-part integrals. *Numerische Mathematik*, 81(2) :273–291, 1998.

[64] R. D. Nair and C. Jablonowski. Moving vortices on the sphere : A test case for horizontal advection problems. *Monthly Weather Review*, 136(2) :699–711, 2008.

[65] R. D. Nair and P. H. Lauritzen. A class of deformational flow test cases for linear transport problems on the sphere. *Journal of Computational Physics*, 229, 2010.

[66] R. D. Nair and B. Machenhauer. The mass-conservative cell-integrated semi-Lagrangian advection scheme on the sphere. *Monthly Weather Review*, 130(3) :649–667, 2002.

[67] R. D. Nair, S. J. Thomas, and R. D. Loft. A discontinuous Galerkin global shallow water model. *Monthly weather review*, 133(4) :876–888, 2005.

[68] T. Nihei and K. Ishii. A fast solver of the shallow water equations on a sphere using a combined compact difference scheme. *Journal of Computational Physics*, 187(2) :639–659, 2003.

[69] J. Pedlosky. *Geophysical fluid dynamics*. Springer Science & Business Media, 2013.

[70] B. Portelenelle and J.-P. Croisille. An efficient quadrature rule on the Cubed Sphere. *Journal of Computational and Applied Mathematics*, 328 :59–74, 2018.

[71] J. Qiu and C.-W. Shu. On the construction, comparison, and local characteristic decomposition for high-order central WENO schemes. *Journal of Computational Physics*, 183(1) :187–209, 2002.

[72] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical mathematics*, volume 37. Springer Science & Business Media, 2010.

[73] S. Redonnet. *Simulation de la propagation acoustique en présence d'écoulements quelconques et de structures solides par résolution numérique des équations d'Euler*. PhD thesis, Bordeaux 1, 2001.

[74] R. Sadourny. Conservative finite-difference approximations of the Primitive equations on Quasi-Uniform spherical grid. *Monthly Weather Review*, 100 (2), 1972.

[75] J. G. Simmonds. *A Brief on Tensor Analysis*. Undergraduate Texts in Math. Springer, 2cd edition, 1994.

[76] J. C. Strikwerda. *Finite difference schemes and partial differential equations*, volume 88. Siam, 2004.

[77] C. K. W. Tam and J. C. Webb. Dispersion-relation-preserving finite difference schemes for computational acoustics. *Journal of computational physics*, 107(2) :262–281, 1993.

[78] J. Thuburn, CJ. Cotter, and T. Dubos. A mimetic, semi-implicit, forward-in-time, finite volume shallow water model : comparison of hexagonal–icosahedral and cubed-sphere grids. *Geoscientific Model Development*, 7(3) :909–929, 2014.

[79] J. Thuburn and Y. Li. Numerical simulations of Rossby–Haurwitz waves. *Tellus A : Dynamic Meteorology and Oceanography*, 52(2) :181–189, 2000.

[80] P. A. Ullrich. *Atmospheric modeling with high-order finite-volume methods*. University of Michigan, 2011.

[81] P. A. Ullrich, C. Jablonowski, and B. Van Leer. High-order finite-volume methods for the shallow-water equations on the sphere. *Journal of Computational Physics*, 229(17) :6104–6134, 2010.

[82] G. K Vallis. *Atmospheric and oceanic fluid dynamics*. Cambridge University Press, 2017.

[83] C. Van Loan. *Computational frameworks for the fast Fourier transform*, volume 10. Siam, 1992.

[84] M. R. Visbal and D. V. Gaitonde. On the use of higher-order finite-difference schemes on curvilinear and deforming meshes. *Journal of Computational Physics*, 181(1) :155–185, 2002.

[85] D. L. Williamson, J. B. Drake, J. J. Hack, R. Jakob, and P. N. Swarztrauber. A standard test set for numerical approximations to the shallow water equations in spherical geometry. *Journal of Computational Physics*, 102(1) :211–224, 1992.

[86] G. B. Witham. Linear and nonlinear waves, 1974.

[87] H. C. Yee. A class of high resolution explicit and implicit shock-capturing methods. *NASA Technical Memorandum*, 1989.

[88] V. Zeitlin. *Nonlinear dynamics of rotating shallow water : Methods and advances*, volume 2. Elsevier, 2007.

Résumé

L'enjeu de la simulation de la dynamique atmosphérique et/ou océanographique a pris une importance accrue avec la question du réchauffement climatique. Le modèle mathématique complet à simuler s'obtient en couplant les équations de la mécanique des fluides avec les équations de la thermodynamique.

Au 19ième siècle, le mathématicien Adhémar Barré de Saint-Venant formule un système d'équations aux dérivées partielles décrivant les mouvements d'un fluide soumis à la gravité et de faible épaisseur. Cette simplification des équations de Navier-Stokes permet de transformer un problème 3D en un problème 2D. Dans le contexte de la sphère en rotation, elles décrivent la réponse d'une couche mince de fluide soumise aux forces de gravité et de Coriolis. Elles permettent de décrire de nombreux phénomènes (ondes de Kelvin, ondes de Rossby, ...). Bien que représentant un problème simplifié, ces équations sont complexes et leur résolution nécessite des méthodes numériques adaptées.

L'objectif de cette thèse est d'étudier une méthode numérique de type différences finies pour résoudre ces équations, les équations Shallow Water, grâce à la grille Cubed-Sphere.

Dans la première partie, on introduit les notations et les schémas utiles à la résolution d'équations aux dérivées partielles sur la Cubed-Sphere. On étudie les schémas aux différences finies dans le contexte périodique pour approcher la dérivée première à différents ordres. Le schéma utilisé sur la sphère est un schéma hermitien d'ordre 4. Nous introduisons aussi les schémas de filtrage. Ces derniers sont consistants avec l'identité et permettent de supprimer les modes oscillants. Ces schémas permettent des approximations en espace. La discrétisation en temps est faite par un algorithme de Runge-Kutta d'ordre 4 explicite couplé à l'opérateur de filtrage. Nous étudions l'algorithme pour la résolution de l'équation de transport et de l'équation des ondes.

La seconde partie est dédiée au maillage sur la sphère ainsi qu'à la construction des opérateurs approchés. Le maillage utilisé est la Cubed-Sphere, introduit en 1972 par Robert Sadourny. Il s'agit du maillage des faces d'un cube projeté sur la sphère. Chaque face est appelé panel. Il y a de nombreuses symétries entre les panels, ce qui permet de construire un produit scalaire vérifiant l'orthogonalité d'un grand nombre d'harmoniques sphériques. De plus, on construit des formules de quadrature précises sur ce maillage. Les points sur un panel sont des portions de grands cercles. En complétant les données sur les grands cercles à l'aide de splines cubiques, on construit des opérateurs approchés de la divergence, du gradient et de la vorticité. Ces opérateurs utilisent les schémas aux différences finis périodiques et sont analysés.

La troisième partie concerne la résolution numérique d'équations sur la sphère. Les expériences numériques concernent l'équation d'advection, l'équation Shallow Water linéarisée et l'équation Shallow Water. La résolution se fait par la méthode des lignes en couplant les opérateurs différentiels discrets avec un schéma de RK4 muni d'un opérateur de filtrage. Les tests sont issus de la littérature classique. Sur certains, une solution analytique est disponible. On compare la solution exacte et la solution donnée par l'algorithme. Les erreurs mesurées confirment la précision attendue. Lorsqu'il n'y a pas de solution analytique connue, nous comparons nos résultats numériques avec ceux obtenus par d'autres méthodes. Nous vérifions la conservation de la masse et de l'énergie.

Dans notre contexte, les simulations en temps long jouent un rôle important. Les résultats obtenus sur temps longs sont ceux attendus. Il serait intéressant d'utiliser un algorithme de résolution en temps implicite pour effectuer ce type de simulations. C'est l'une des perspectives de ce travail.

A plus long terme, l'objectif est de simuler un modèle en 3D. Il faudra coupler un tel modèle avec les équations de la thermodynamique (modèle GCM).

Mots-clés: Equation Shallow Water, Cubed-Sphere, schémas compacts, discrétisation en temps.

Abstract

The challenge to simulate the atmospheric and/or oceanic fluid dynamics has become crucial with the climate change problems. The full mathematical model to simulate consists in the coupling of fluid dynamics with thermodynamics.

In the 19-th century, Adhémar Barré de Saint-Venant first formulated the equations describing the dynamic of a fluid subject to gravity and bottom topography. This equation can be considered as a bidimensional simplification of the 3D Navier-Stokes system. When expressed in the context of the rotating sphere, this equation describes the reaction of a fluid thin layer subject to the gravitational and Coriolis forces. It permits to describe many wave phenomena (Kelvin waves, Rossby waves, ...). Although representing a simplified problem, this equation is difficult to solve and its numerical resolution requires suitable numerical schemes.

The goal of this thesis is to study a particular finite difference scheme to solve this equations, (also called Shallow Water equation) with the Cubed-Sphere grid.

In the first part, we introduce notations and schemes used to solve partial differential equation on the Cubed-Sphere. We study finite difference schemes in the periodic context. They allow us to approximate the first derivative with different order of accuracy. The scheme used on the sphere is the hermitian scheme of order 4. We introduce filtering schemes. These ones are consistent with the identity with high order of accuracy and allow us to remove oscillating modes. These schemes allow us space approximation. The time discretization is made with explicit 4-th order Runge Kutta algorithm coupled to a filtering operator. We study this algorithm to solve the advection equation and the wave equation. The accuracy is proved and the stability is studied. The filtering operator allows us to reduce parasitic oscillations which can appear during the resolution of hyperbolic equations by a centered scheme.

The second part is devoted to the spherical grid and to approximation operators. The mesh used is the Cubed-Sphere. This mesh was introduced in 1972 by Robert Sadourny. It is the mesh of faces of a cube projected on the sphere. Each face of this mesh, on the sphere, is named panel. There are multiple symetries between panels, that allow us to build a scalar product. This product permits us to check the orthogonality for a large number of spherical harmonics on the mesh. Further more, we build accurate quadrature formulas on this grid. Panel points are portions of great circles. We build approximated operators of the divergence, the gradient and the vorticity by completing the data on great circles with cubic spline. These operators use periodic finite difference schemes and are analyzed.

The third part is devoted to the numerical resolution of partial differential equations on the sphere. Experiments concerne advection equation, linearized Shallow Water equation and Shallow Water equation. The resolution is made with the lines method by coupling discrete differential operators, fourth order Runge-Kutta algorithm and filtering operator. Numerical tests are from the classical litterature. For some, an analytical solution is available. Errors are weak and allow us to confirm the expected accuracy. For numerical tests without analytical solution, we compare our numerical results with those obtained by other methods. Furthermore, we check the conservation of mass and energy.

Our context is the one of simulations over particularly long time. This is why we perform tests to predict a solution in a distant future. We are satisfied by the behaviours observed. It would be interesting to use implicit time algorithm. It is one of the perspectives of this work.

In the longer term, the goal is to simulate a model in dimension 3 coupled with the equations of thermodynamic (GCM model).

Keywords: Shallow Water equation, Cubed-Sphere, compact scheme, time discretisation.