

Modélisation des corrélations spatiales d'erreurs d'observation en assimilation de données variationnelle. Etudes sur des maillages non structurés.

Oliver Guillet

▶ **To cite this version:**

Oliver Guillet. Modélisation des corrélations spatiales d'erreurs d'observation en assimilation de données variationnelle. Etudes sur des maillages non structurés.. Océan, Atmosphère. INP Toulouse, 2019. Français. ⟨NNT : ⟩. ⟨tel-02361795⟩

HAL Id: tel-02361795

https://hal.science/tel-02361795v1

Submitted on 13 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization



THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Institut National Polytechnique de Toulouse

Présentée et soutenue par

Oliver GUILLET

Le 8 février 2019

Modélisation des corrélations spatiales d'erreurs d'observation en assimilation de données variationnelle. Étude sur des maillages non structurés.

Ecole doctorale : **SDU2E - Sciences de l'Univers, de l'Environnement et de l'Espace**

Spécialité : **Océan, Atmosphère, Climat**

Unité de recherche :
CECI - Climat, Environnement, Couplages et Incertitudes / CERFACS

Thèse dirigée par
Anthony WEAVER et Yann MICHEL

Jury

M. Marc BOCQUET, Rapporteur
M. Emmanuel COSME, Rapporteur
M. Arthur VIDARD, Rapporteur
M. Marcin CHRUST, Examinateur
Mme Selime GUROL, Examinatrice
M. Serge GRATTON, Examinateur
M. Anthony WEAVER, Directeur de thèse
M. Yann MICHEL, Co-directeur de thèse

M. Xavier VASSEUR, invité

Remerciements

Toi, lecteur, ce prologue ne t'est pas destiné. Du moins pas directement. Il est destiné aux grandes personnes qui m'ont accompagné lors de mon dernier voyage, et dont je me souviens dans les moindres détails. A la manière d'un explorateur, j'aurais voulu commencer cette histoire tel un conte de fées :

« Il était une fois un étudiant qui habitait une planète à peine plus grande que lui, et qui cherchait le nom du vent... ». Mais ce serait amoindrir les qualités humaines des protagonistes en leur confiant un rôle trop allégorique. D'autant plus que cette citation n'est tirée d'aucun livre, ni même de plusieurs. N'arrivant pas à les replacer dans ma mémoire, je préfère ne pas trop en faire.

Tu lis ces mots et tu as bien sûr compris qu'il s'agissait du premier chapitre de mon manuscrit, les « remerciements ». Mais ce terme est trop stérile, et puis j'aime parler, et j'aime écrire. Puisqu'il ne s'agit pas de science, je peux me permettre de divaguer. Il s'agit donc d'un avant-propos, un prologue, une pré-introduction... tant qu'il désigne les premiers mots qui introduiront les personnes les plus importantes de ce récit.

Alors commençons. Et nous savons que commencement rime traditionnellement avec précision. Je m'attache donc à commencer ma narration en bon temps, et en bon lieu. Pour moi, l'histoire ne commence pas en l'année 1993, quand mes adorables parents me donnèrent un nom et plus d'importance que je ne pourrais le décrire en quelques mots. Elle ne commence pas non plus le jour où mon oncle vint m'apprendre l'existence d'un pays où les nuages étaient une science, où le vent m'amènerait à faire mes études. Non plus quand, finissant mes études et m'apprêtant à trébucher dans l'aldut...age ? l'adultance ? Bref, à quitter l'enfance, je sauvais les meubles en trouvant un moyen de prolonger mon séjour au pays des nuages.

Non. L'histoire ne commence pas **là**, ni à ce **moment-là**. Elle commence avec un homme, un professeur du nom de...

... **Serge Gratton**. C'est un grand magicien qui, un jour, m'a fait un tour de magie. Le genre de tour expliqué à la craie et à la brosse qui vous laisse perplexe suffisamment longtemps pour vous faire passer pour un mouton. « Maiiss... ». Bon, je n'ai jamais fait le mouton, mais c'est tout comme. De

toute façon, lui se souvient de tout, et c'est ce qui importe. Je le sais car il me l'a dit. Et comme les mots s'effacent, il me l'a écrit. Et s'il oublie, je serai là pour lui rappeler.

J'entends une voix qui s'élève de la porte du fond. Une voix calme, non de résonnance grave et carverneuse, je n'irais pas jusqu'à la qualifier d'aigüe, mais disons qui vous surprend par sa douceur et vous invite à tendre l'oreille. Elle n'est nulle autre que celle de...

... **Xavier Vasseur**, un magicien qui a fait un pacte avec forcément-une-entité-très-puissante, en échange d'un savoir infini. Aucune exagération. Heureusement pour nous, il est resté du côté des gentils et nous en sommes tous reconnaissants.

C'est alors à mon tour de faire un pacte. Je vais porter l'anneau en Mordor (bizarrement, j'accepte cette mission) en l'échange de trois années supplémentaires à chercher le vent. Je prépare mon sac et je fais le pas. Je pars pour une contrée lointaine à la lisière du pays des nuages, à la limite de l'océan, où je rencontre pour la première fois le capitaine de mon expédition...

... **Anthony Weaver**. Je n'irai pas inventer une histoire de grand navigateur pour vous le présenter. Je me contenterai de vous dire qu'il connait la direction, et qu'il connait son navire. A ses côtés, je pars pour un voyage fantastique, et je regrette presque qu'il soit passé aussi vite. Merci pour tout.

Pendant que je parle au capitaine et que je plisse le front de concentration, une personne m'observe en souriant. Cette personne veille à la bienséance de l'expédition...

... **Selime Gürol**, car tel est son nom, attend son tour pour parler. Attendez... En plein élan de prose lyrique, prêt à me perdre dans la mélancolie que m'inspire cette aventure, j'allais décrire Selime sans mentionner sa vivacité et sa spontanéité. C'est la dernière des « premiers », cette équipe de choc qui m'aide à garder le cap.

Mais pourquoi parlé-je des « premiers » ? Une aventure n'est pas une recette de cuisine. On ne déclare pas tous les ingrédients avant de les mélanger (voyez les ingrédients comme des personnages). Beaucoup plus tard, alors que commence la fin de l'aventure, vient nous rejoindre un dernier magicien / navigateur / confrère / aventurier / simple curieux / ami ?

... **Yann Michel**, le premier et le dernier des « derniers ». Terrible, je n'ai pas fini de le présenter, qu'on me coupe la prose pour me demander son rôle. Eh bien lui aussi, c'est une personne et non un personnage, et comme les autres, mes circonvolutions ne sauraient rendre compte de ses qualités infinies. Que dire ? Peut-être que dans une histoire, le commencement n'importe pas seul.

Et le récit ? Et les méchants ? Et les rebondissements ? Venez me voir et je vous raconterai tout. Ou rien. Comprenez que ces détails n'ont pas d'importance. Mon ton devient plus grave et je le répète une dernière fois : les vagues, les nuages, le vent n'importent pas. L'essentiel, j'en ai déjà parlé. Non c'est faux, il en manque, sans quoi je poserais ma plume.

Ils sont venus, ils m'ont aidé, il m'ont porté... Je remercie (volontairement dans le désordre !) :

Brigitte, Loris, Loïk, Paul, Gwenn, Gerald, Rémi, Aristote, Vincent, Horace, Said, Anthony, Julien, Christophe, Philippe, Jess, Thibaud (pour son soutien indéfectible), Thibaut, Jean, Nacer, Nicolas, Florent, Mr Kdr...

Et puis la famille, bien que je préfère leur dire de vive voix.

Si j'en ai oublié, disons que c'est fait exprès, même si ce n'est pas vrai. On dira que la maladresse me sied. Je rajouterai ceux qui manquent au crayon.

Arrive la conclusion (alors que toi, lecteur, si tu m'as suivi, tu n'as pas encore lu l'introduction), et j'en profite pour te donner mes meilleurs mots :

Je souhaite qu'à l'avenir, l'enfant en moi ne cesse de grandir, les yeux grand ouverts, observant l'adulte qui s'efforce à mûrir. Et je te souhaite à toi, lecteur de ce manuscrit, de ne jamais cesser d'apprendre, d'apprendre à apprendre, à désapprendre sans jamais comprendre, car l'apprentissage est un commencement. Tout comme les personnages de mon histoire, je souhaite que ton regard soit celui d'un enfant, qu'il brille longtemps, de vivacité et de candeur.

Bonne lecture.

Résumé / Abstract

Dans cette thèse, nous proposons une classe de méthodes permettant de représenter numériquement les corrélations spatiales d'erreurs d'observation en assimilation de données variationnelle. Partant du lien existant entre les solutions de l'équation de diffusion implicite et les fonctions de corrélation de Matérn, nous construisons des opérateurs de corrélation et des opérateurs de corrélation inverses adaptés à la grande dimension. La discrétisation de l'équation de diffusion par la méthode des éléments finis permet de manipuler des données qui ne reposent pas nécessairement sur des maillages structurés, comme c'est le cas pour les observations satellites assimilées dans les modèles de météorologie. Les expériences sont menées en utilisant des données de l'imageur infrarouge Seviri dont les images contiennent notamment de fortes corrélations horizontales. Nous montrons que la qualité de notre modèle de corrélation peut dépendre localement de la distribution spatiale des observations. Néanmoins, l'introduction d'un maillage auxiliaire pour effectuer les calculs en éléments finis permet de s'affranchir de cette dépendance. Dans ce cas, la précision de la méthode s'acquiert au prix d'un opérateur inverse plus difficile à appliquer. Finalement, on propose des stratégies pour appliquer cet inverse efficacement.

In this thesis, we propose a class of methods to represent spatial observation error correlations numerically in variational data assimilation. Based on the existing link between solutions of the time-implicit diffusion equation and Matérn correlation functions, we design correlation operators and inverse correlation operators that are appropriate for large datasets. Discretizing the diffusion equation with the finite element method allows us to account for data that do not necessarily lie on a structured mesh, as is the case with satellite observations assimilated in meteorology. Experiments are carried out using data from the infrared imager Seviri, the images of which are known for containing strong horizontal correlations. We show that the quality of our correlation model may depend locally on the spatial distribution of the observations. Nevertheless, by introducing an auxiliary mesh to perform the finite element computations, we can control this dependency to a large extent. Improving the accuracy of the method this way comes at the expense of making the inverse correlation operator more complicated. Finally, strategies for efficiently modelling the inverse of the correlation operator are proposed.

Table des matières

Remerciements iii

Résumé / Abstract vii

Table des matières xii

Introduction 1

I Aspects d'analyse fonctionnelle 7

1 Opérateurs de corrélation en assimilation de données 9
- 1.1 Opérateurs de covariance . . . 10
  - 1.1.1 Espaces de fonctions et de distributions . . . 11
  - 1.1.2 Fonctions et opérateurs de covariance . . . 13
  - 1.1.3 Fonctions de corrélation de Matérn . . . 16
  - 1.1.4 Corrélation inverse et diffusion . . . 19
- 1.2 Assimilation de données . . . 24
  - 1.2.1 Optimisation en dimension infinie . . . 25
  - 1.2.2 Du continu au discret . . . 27
  - 1.2.3 Fenêtres d'assimilation et aspects temporels . . . 30
  - 1.2.4 Formulation incrémentale . . . 32
- 1.3 Résolution du problème variationnel . . . 33
  - 1.3.1 Formulations primale et duale . . . 33
  - 1.3.2 Formulation point-selle . . . 34
- 1.4 L'essentiel du chapitre . . . 37

2 Traitement de la diffusion par la méthode des éléments finis 39
- 2.1 Généralisations de l'équation de diffusion homogène . . . 41
  - 2.1.1 Diffusion hétérogène . . . 41
  - 2.1.2 Diffusion généralisée . . . 44
- 2.2 Formulation faible de l'équation de diffusion . . . 45
  - 2.2.1 Formulation faible . . . 46
  - 2.2.2 Cyclage et propriétés mathématiques . . . 48
- 2.3 Familles d'élements finis et degrés de liberté . . . 51
  - 2.3.1 Eléments finis de type $\mathbb{P}_k$ . . . 51
  - 2.3.2 Formules d'intégration $\mathbb{P}_1$ . . . 52

2.4 Aspects de convergence 55
2.4.1 Stabilité de la diffusion implicite 55
2.4.2 Convergence en espace 57
2.5 Assimilation des dérivées du champ d'observation 60
2.5.1 Du champ dérivé à la diffusion : formulation continue . 61
2.5.2 De la diffusion au champ dérivé : formulation discrète . 62
2.5.3 Discussion 64
2.6 L'essentiel du chapitre 65

II Vers les maillages non structurés 67

3 Etude sur des maillages structurés 69
3.1 Structure de la réponse 70
3.1.1 Profils des matrices en éléments finis 70
3.1.2 Comparaison avec le modèle théorique 72
3.2 Introduction de la condensation de masse 74
3.2.1 Principe de la condensation de masse 74
3.2.2 Diagonalisation par quadrature 77
3.2.3 Comparaison avec le modèle théorique 77
3.3 Corrections analytiques aux frontières 78
3.3.1 Solutions analytiques en temps continu 78
3.3.2 Corrections analytiques 81
3.4 Synthèse sur la spécificité des maillages structurés 83
3.5 L'essentiel du chapitre 84

4 Application aux données du sondeur Seviri 85
4.1 Présentation des données de Seviri 86
4.2 Estimation des paramètres de corrélation 89
4.2.1 Estimation à partir des innovations 89
4.2.2 Exploitation des dérivées 91
4.2.3 Ajustement de modèle 92
4.3 Géométrie du maillage des observations 97
4.3.1 Génération de maillage 98
4.3.2 Qualité du maillage 102
4.3.3 Noeuds de frontière 105
4.4 Validation des opérateurs de corrélation 107
4.4.1 Fonctions de corrélation 107
4.4.2 Test de l'adjoint 113
4.5 Méthodes de correction de l'amplitude 115
4.5.1 Normalisation exacte et randomisation 115

4.5.2 Méthodes avancées 117
4.5.3 Régression à partir de la qualité du maillage 117
4.6 L'essentiel du chapitre 120

III Contrôle de la précision des opérateurs 123

5 Stratégies de raffinement de maillage 125
5.1 Aspects pratiques du raffinement de maillage 126
5.1.1 Types de raffinement 127
5.1.2 Types de déraffinement 132
5.1.3 Maillage sur SEVIRI 138
5.2 Opérateurs de transfert 143
5.2.1 Transfert par interpolation linéaire 144
5.2.2 Transfert par injection 149
5.3 Opérateurs de corrélations augmentés 150
5.3.1 Etablissement de la formule générale 151
5.3.2 Application aux données SEVIRI 153
5.4 Interpolation de l'opérateur de corrélation inverse 156
5.5 Effets de la condensation de masse 157
5.6 L'essentiel du chapitre 161

6 Inversion dans l'espace des observations 163
6.1 Méthodes itératives 164
6.1.1 Méthodes de point fixe 164
6.1.2 Méthodes de Krylov 167
6.1.3 Cas des matrices définies positives 170
6.1.4 Cas des matrices non symétriques définies positives . . 174
6.1.5 Iterations de Chebyshev 177
6.1.6 Discussion 179
6.2 Préconditionnement et résolution 180
6.2.1 Principe du préconditionnement 181
6.2.2 Approximation grossière de l'inverse 183
6.2.3 Approximation fine de l'inverse 186
6.3 Reformulation comme point-selle 187
6.3.1 Formulation point-selle 187
6.3.2 Performances 189
6.4 Préconditionnement du deuxième ordre 190
6.4.1 Déflation 190
6.4.2 Mise à jour du préconditionneur 194
6.4.3 Sensibilité au second membre 197

6.5 Troncature des itérations 198
6.6 L'essentiel du chapitre 200

7 Raffinement et stratégies alternatives 201
7.1 Scission des opérateurs 202
7.1.1 Opérateur scindé 202
7.1.2 Inversion et troncature 204
7.2 Redéfinition des matrices de masse et de raideur 205
7.2.1 Critère d'interpolation 205
7.2.2 Performances 207
7.2.3 Déraffinement et perspectives 209
7.3 Somme d'opérateurs 211
7.3.1 Inversion sur le maillage initial 211
7.3.2 Inversion en présence de raffinement 213
7.3.3 Inversion en présence de déraffinement 214
7.4 L'essentiel du chapitre 216

Bilan et perspectives 217
8.1 Bilan de l'étude 217
8.1.1 Réponses aux objectifs principaux 217
8.1.2 Contributions notables 219
8.2 Perspectives de recherche 220
8.2.1 Axes de développements privilégiés 220
8.2.2 Enjeu de l'insertion dans l'opérationnel 222

Bibliographie 225

A Méthode des volumes finis 239

B Méthode des éléments virtuels 241

C Diffusion sur des graphes généraux 245

D Autres jeux de données 249

E Structure du code 257

F Publication 261

Introduction

L'assimilation de données répond au problème de la spécification des conditions initiales des modèles de prévision numérique du temps [Daley, 1991, Kalnay, 2003]. Faisant usage d'observations de natures variées, elle s'inscrit dans une démarche qui consiste à croiser plusieurs types d'informations pour estimer au mieux l'état de l'atmosphère au temps de l'analyse. Populaire et robuste, l'assimilation de données variationnelle est la méthode qui a su s'imposer au cours des dernières décennies dans la plupart des centres nationaux de météorologie [Rabier et al., 2000, Desroziers et al., 2003, Rawlins et al., 2007, Gauthier et al., 2007]. Plus récemment, l'émergence des méthodes hybrides (voir par exemple Lorenc [2015]) a permis de tirer le meilleur profit des informations des ensembles, tout en gardant le formalisme variationnel.

L'assimilation de données variationnelle est basée sur la recherche de l'état, du vecteur décrivant l'atmosphère, qui minimise à la fois l'écart à la prédiction du modèle (on parle d'« ébauche ») et l'écart aux observations. L'ébauche étant jugée inexacte et les observations imprécises, la prise en compte de ces incertitudes apparaît comme un enjeu majeur dans la définition des algorithmes. Aussi, pour des raisons d'efficacité numérique, suppose-t-on que lesdits écarts, ou « erreurs », sont entièrements décrits par leurs deux premiers moments statistiques : la moyenne et la (co-)variance.

Sous réserve d'agrémenter le contexte de quelques hypothèses mathématiques, on en arrive naturellement à la formulation du 3DVar. Le lien entre l'ébauche et les observations est rendu possible par l'intermédiaire d'un opérateur d'observation, à tendance fortement non-linéaire. En résulte un problème de minimisation non quadratique, qu'on traite habituellement par itérations, donnant lieu à l'approche incrémentale du 3DVar. Le 4DVar diffère du 3DVar par la prise en compte de la distribution temporelle des observations, le changement du « 3 » en « 4 » faisant référence à l'addition d'une quatrième dimension au problème.

Ainsi, l'assimilation de données variationnelle se réduit schématiquement à la minimisation d'une fonctionnelle, composées de normes et d'écarts, dont chacun des termes est pondéré par l'inverse d'une matrice de covariance. Ces matrices de covariance atteignent des tailles colossales, ce qui rend leur manipulation délicate et incommode. Cela explique les nombreux sujets de

recherche qui concernent leur estimation et leur modélisation.

En pratique, il n'est pas nécessaire de connaître explicitement les contenus des matrices de covariance ou de leurs inverses. En effet, les algorithmes itératifs employés pour résoudre le problème variationnel ne requierent pas l'accès aux coefficients des matrices (voir par exemple Saad [2003]). A la place, ces dernières sont représentées comme des opérateurs, partant du principe qu'il suffit de connaître leur action sur n'importe quel vecteur pour savoir les utiliser. Cette astuce permet d'éviter le stockage de milliards de milliards de coefficients qui ne tiendraient pas même dans le plus gros des ordinateurs actuels.

Par ailleurs, les vraies statistiques des erreurs d'ébauche et d'observation ne sont pas connues. Leur connaissance supposerait de disposer de l'état réel de l'atmosphère, celui-là même qu'on cherche à estimer. La modélisation des erreurs fait donc l'objet d'un compromis entre physique, propriétés statistiques et aspects numériques. Par exemple, la matrice de covariance d'erreurs d'ébauche contient à la fois des structures dynamiques, une partie de climatologie et des informations de la prévision ensembliste.

En termes de développements récents, beaucoup de travaux se sont concentrés sur l'estimation et la modélisation de la matrice de covariance d'erreurs d'ébauche [Bannister, 2008a,b], fait en partie justifié par le rôle central de cette dernière dans la résolution du problème variationnel. Conséquemment, peu d'études ont concerné la modélisation des corrélations dans la matrice de covariance d'erreurs d'observation, à l'heure même où les données abondent et que les instruments d'observation atteignent des résolutions jusqu'alors inégalées.

La période actuelle connaît un intérêt grandissant pour l'étude des covariances d'erreurs d'observation. Jusqu'alors, on supposait en effet que la matrice associée était diagonale, c'est-à-dire qu'elle contenait des variances, mais pas de corrélations. Autrement dit, on considérait que les erreurs d'observation étaient indépendantes les unes des autres. La réalité est pourtant légèrement différente. Pour justifier l'hypothèse d'indépendance des erreurs, les données d'observation doivent faire l'objet d'une sélection agressive [Rabier, 2006]. En effet, dans certains cas, ce seul procédé est à lui seul responsable de l'élimination de plus de 90% des données télédétectées. C'est une solution qui ne peut être que temporaire compte-tenu de l'augmentation régulière de la résolution des instruments et du nombre de données assimilables. A titre d'exemple, le futur sondeur infrarouge MTG-IRS combinera la résolution

horizontale et temporelle de SEVIRI avec les capacités spectrales de IASI !

D'un point de vue scientifique, des études récentes ont mis en évidence l'existence de corrélations non négligeables contenues dans les données des radars et des satellites. Ne pas tenir compte de ces corrélations en assimilation de données contrarie la notion d'optimalité du problème variationel qu'on cherche à résoudre [Liu and Rabier, 2002, Dando et al., 2007]. C'est ainsi que les premiers travaux de modélisation des corrélations d'erreurs d'observation ont vu le jour.

Au départ, l'accent fut mis sur les corrélations verticales ou « intercanaux » des données satellitaires [Bormann et al., 2010, Bormann and Bauer, 2010, Stewart et al., 2014, Weston et al., 2014, Waller et al., 2016a]. Ces corrélations relient les erreurs commises par les différents canaux d'acquisition d'un même instrument, comme les sondeurs embarqués sur les satellites. En parallèle, des travaux se sont attelés à la représentation des corrélations temporelles de ces erreurs [Järvinen et al., 1999]. Plus récemment, Waller et al. [2016a,b,c] se sont intéressés à l'estimation des corrélations spatiales d'erreurs d'observation des radars et des satellites. La modélisation de ces corrélations est un domaine de recherche en pleine expansion, avec par exemple les travaux de Brankart et al. [2009], Ruggiero et al. [2016], Michel [2018].

On décide ici de s'intéresser aux observations satellitaires. Les satellites ont l'avantage d'observer la quasi-intégralité du globe terrestre, ce qui leur confère une importance cruciale dans la correction des modèles de prévision (voir Rabier [2006] pour une revue complète). Cependant, leur manipulation n'est pas évidente, en particulier quand il s'agit de représenter les corrélations spatiales de leurs erreurs de mesure. Cette difficulté a pour origine plusieurs constats. Le premier est que les images des satellites se composent de millions, voire de milliards de pixels, ce qui contraint la modélisation des matrices de corrélation par des opérateurs. Le deuxième constat est que les images, initialement structurées en lignes et colonnes de pixels, perdent leur structure suite aux prétraitements météorologiques imposés en amont de l'assimilation de données. Conséquence immédiate : la définition des opérateurs de corrélation gagne en complexité, et les approches suggérées dans la littérature, comme celles de Brankart et al. [2009] et Michel [2018] deviennent difficile à mettre en oeuvre. En cause, la difficulté de calculer les gradients des erreurs d'observations, qui interviennent dans la plupart des formulations.

Le défi ne s'arrête pas là. Les algorithmes populaires d'assimilation de données requierent idéalement l'inverse de l'opérateur de corrélation d'er-

reurs d'observation, que cet inverse intervienne dans leur formulation primale (Bannister [2008b] ou encore Gratton and Tshimanga [2009] dans la version RPCG du 4Dvar) ou dans l'expression de leur préconditionnement (Courtier [1997] et, plus récemment Fisher and Gürol [2017] qui formule le 4Dvar comme un point-selle). Or les travaux précédents se sont heurtés à cet écueil. Il s'agit donc d'arriver à modéliser les opérateurs de corrélation de sorte que leurs inverses soient aisément calculables. Pour ce faire, on exploite un formalisme proche de celui que proposent Lindgren et al. [2011] et Bui-Thanh et al. [2013], qui appliquent un opérateur de corrélation en résolvant une équation aux dérivées partielles stochastique (voir aussi Simpson et al. [2012] et Bolin and Lindgren [2013]).

Voici un bref récapitulatif du travail attendu dans la thèse :

Motivation principale :

- Les données du réseau d'observation terrestre sont sous-employées ;
- On souhaite augmenter la quantité de données assimilées, en particulier les observations à haute densité spatiale, afin de mieux représenter les phénomènes atmosphériques de petites échelles ;
- On améliorerait ainsi l'analyse aux échelles convectives, et par conséquent la qualité des prévisions.

Objectifs :

- Proposer une modélisation de l'opérateur de corrélation d'erreurs d'observation et de son inverse ;
- Le coût d'application de l'opérateur doit être raisonnable, sa modélisation robuste et précise ;
- Préparer une futur implantation en opérationnel en travaillant si possible à partir de données réalistes.

Pour répondre à ces objectifs, nous proposons une étude en trois parties.

Plan du manuscrit :

La partie I introduit les outils d'analyse fonctionnelle qui sont utilisés tout au long du manuscrit. Dans le chapitre 1, on présente l'assimilation de données, depuis la définition des objets mathématiques continus jusqu'à la résolution du problème variationnel discret. Le chapitre 2 est consacré au traitement de l'équation de diffusion par la méthode des éléments finis. Cette technique est utilisée pour modéliser les corrélations spatiales d'erreurs d'observation. Le chapitre 2 contient en outre une section dédiée au lien entre l'équation de diffusion et la méthode de Brankart et al. [2009].

La partie II revêt un caractère plus expérimental, avec pour objectif l'évaluation des performances de l'opérateur de corrélation en éléments finis sur différents types de maillages, structurés, puis non structurés. Les résultats sur maillages cartésiens sont contenus dans le chapitre 3, qui offre l'occasion d'introduire la stratégie de validation de ces résultats par comparaison avec le modèle théorique. La distribution hétérogène des observations satellitaires est prise en compte dans le chapitre 4, à partir duquel toutes les expériences sont réalisées avec des données réelles.

Dans la partie III, on discute de la précision de la modélisation en éléments finis et on propose plusieurs approches pour assurer la robustesse des opérateurs de corrélation. L'étude du chapitre 5 démontre qu'il est possible de gagner en précision et en robustesse en ayant recours au procédé de raffinement de maillage. Ce procédé vient toutefois compliquer l'inversion de l'opérateur de corrélation, qui fait l'objet du chapitre 6. Enfin, le chapitre 7 propose un certain nombre de modélisations alternatives, qui réutilisent le formalisme et les résultats des chapitres précédents. Il s'agit du dernier chapitre avant la conclusion.

Les annexes A, B et C présentent trois alternatives à la méthode des éléments finis pour résoudre l'équation de diffusion. Pour chacune, une discussion explique pourquoi la méthode n'a pas été retenue lors de l'étude. L'annexe D reproduit les expériences-clefs du manuscrit à partir d'autres situations météorologiques. L'annexe E porte sur la structure du code formant le cadre d'expérimentation numérique de la thèse. Enfin, l'article portant sur la représentation des opérateurs de corrélation en éléments finis est placé en annexe F. Il recoupe en grande partie les développements de la partie II.

Première partie

Aspects d'analyse fonctionnelle

Chapitre 1

Opérateurs de corrélation en assimilation de données

Dans les modèles de prévision numérique du temps, l'état de l'atmosphère est représenté comme solution d'un ensemble d'équations aux dérivées partielles [Malardel, 2005]. Ces équations étant continues en espace et en temps, leurs solutions appartiennent à des espaces de fonctions de dimension infinie [Tarantola, 2005]. Bien que la discrétisation de ces équations conduise naturellement à la résolution de systèmes linéaires dans des espaces de dimension finie, il est utile d'étudier les propriétés des opérateurs continus pour s'assurer de la consistence des opérateurs discrétisés. Dans cette étude, on s'intéresse particulièrement aux opérateurs de covariance et leurs inverses, dont le rôle est essentiel en assimilation de données [Lorenc, 1986]. En effet, l'assimilation de donnée a pour objectif de produire une estimation de l'état de l'atmosphère à partir de différentes sources d'information. Les opérateurs de covariance servent à pondérer ces différentes sources d'information en fonction de leur importance relative et de l'incertitude qu'on leur attribue [Courtier et al., 1998].

La section 1.1 est ainsi dédiée à la définition des opérateurs de covariance et de corrélation en dimension infinie. Le cadre de l'analyse fonctionnelle est utile à bien des égards. Tout d'abord, il permet d'introduire rigoureusement la définition de ces opérateurs, en utilisant les notions d'espaces primal et dual. C'est également l'occasion de tisser des liens entre les espaces fonctionnels, objets mathématiques, et l'interprétation physique des champs météorologiques comme étant lisses ou bruités. Enfin, il est utile pour souligner la correspondance entre les équations aux dérivées partielles et certaines familles de fonctions de corrélation communément utilisés en géostatistiques.

Dans les chapitres suivants, cette correspondance est exploitée pour représenter des corrélations en grande dimension, s'appuyant sur des données dont la distribution spatiale est hétérogène. La section 1.2 introduit les principes de l'assimilation de données variationnelle. La formulation discrète est déduite de la formulation continue en supposant que les champs météorologiques discrets ne sont autres que l'échantillonnage de champs continus à des endroits privilégiés. Enfin, la résolution du problème variationnel est évoquée dans la section 1.3. L'importance relative de l'opérateur de corrélation, de son inverse et de sa factorisation de Cholesky est abordée, permettant de justifier les choix de modélisation effectués dans les prochains chapitres.

Les développements de ce chapitre relèvent de l'analyse fonctionnelle classique, même si les outils utilisés se retrouvent plutôt dans les cours de mécanique quantique. Il propose une introduction concise à l'assimilation de données, allant du continu au discret sans manquement théorique, faisant le lien entre des domaines trop souvent traités séparément. L'idée est de présenter le problème de l'assimilation de données de ses fondements jusqu'à sa résolution avec un niveau de détail suffisant pour le physicien comme pour le numéricien.

1.1 Opérateurs de covariance

Le cadre adéquat pour la définition des opérateurs de covariance est celui de l'analyse fonctionnelle. Comme l'objet de ce chapitre n'est pas de fournir un cours complet dans ce domaine, on choisit d'introduire les définitions minimales permettant de comprendre et donner du sens aux propriétés mises en avant. Selon le contexte, l'accent est mis sur la nature physique des différents objets, afin de permettre à l'intuition de retrouver les résultats à partir d'élements logiques simples.

Cette section contient ainsi les prérequis pour définir l'assimilation de données en dimension infinie. C'est aussi le cadre approprié pour modéliser les opérateurs de corrélation à partir d'opérateurs différentiels. Cette construction étant à la base de la présente étude, il convient de l'introduire dès le début du manuscrit.

1.1.1 Espaces de fonctions et de distributions

Soit $L^2(\mathbb{R}^d)$ l'espace des fonctions de $\mathbb{R}^d$ à valeurs dans $\mathbb{R}$ de carré intégrable. Ici, d est un entier strictement positif correspondant à la dimension de l'espace physique étudié. En géophysique, on considère habituellement $d = 2$ ou $d = 3$ selon le type de champ (2D ou 3D). Par convention, lorsque le contexte ne prête pas à confusion, on s'autorise à noter simplement $L^2(\mathbb{R}^d) = L^2$. Dire que $f \in L^2$ signifie que l'intégrale :

$$\int_{\mathbb{R}^d} f(\boldsymbol{z})^2 \mathrm{d}\boldsymbol{z} \tag{1.1}$$

existe et admet une valeur dans $\mathbb{R}$.

L'espace L^2 est un espace vectoriel normé, complet pour la norme associée au produit scalaire [1], donc un espace de Hilbert [Bourbaki et al., 1987]. Il est l'un des espaces de fonctions les plus connus en physique et en mathématiques car il permet de représenter des signaux d'énergie finie. Sur L^2, on peut définir le produit scalaire en posant :

$$\langle f, g \rangle_{L^2} = \int_{\mathbb{R}^d} f(\boldsymbol{z}) g(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}, \tag{1.2}$$

cela $\forall f \in L^2$ et $\forall g \in L^2$.

Les espaces de Sobolev sont des sous-espaces de L^2, dont les éléments satisfont des conditions supplémentaires de régularité. Ils offrent ainsi un cadre spécialement adapté à la résolution des équations aux dérivées partielles [Brezis, 2010].

Notons H^m, où m désigne un entier positif, l'espace de Sobolev d'ordre m. On dit que $f \in H^m$ si $f \in L^2$ et si toutes les dérivées partielles de f, jusqu'à l'ordre m, sont encore dans L^2. En d'autres termes :

$$H^m = \{f \in L^2, \forall \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d \text{ avec } |\boldsymbol{\alpha}| \leqslant m, \mathrm{D}^{\boldsymbol{\alpha}} f \in L^2\}, \tag{1.3}$$

avec

$$\mathrm{D}^{\boldsymbol{\alpha}} f = \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial^{\alpha_1} z_1 \ldots \partial^{\alpha_d} z_d} \tag{1.4}$$

1. On peut rajouter que, par définition des classes d'équivalence dans L^2, deux fonctions égales presque partout sont égales. Voilà qui règle l'existence du produit scalaire et de la norme dont on fait usage dans la suite.

et $\boldsymbol{z} = (z_1, \ldots, z_d) \in \mathbb{R}^d$. On peut montrer que H^m est lui-même un espace de Hilbert lorsqu'on le munit du produit scalaire :

$$\begin{aligned} \langle f, g \rangle_{H^m} &= \int_{\mathbb{R}^d} f(\boldsymbol{z}) g(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} + \int_{\mathbb{R}^d} \mathrm{D}f(\boldsymbol{z}) \mathrm{D}g(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} + \ldots + \\ &\int_{\mathbb{R}^d} \mathrm{D}^{\boldsymbol{m}} f(\boldsymbol{z}) \mathrm{D}^{\boldsymbol{m}} g(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \end{aligned} \tag{1.5}$$

avec $f \in H^m$ et $g \in H^m$.

La régularité des fonctions de H^m augmente lorsque m augmente [2]. Pour la suite, on aimerait également définir des espaces de fonctions irrégulières. Ces espaces nous permettront de donner un sens au crochet de distribution et à la distribution de Dirac, très utile pour valider numériquement les opérateurs de covariance.

On remarque que l'extension des définitions ci-dessus au cas des fonctions définies sur un ouvert $\mathcal{D} \subset \mathbb{R}^d$ est immédiate. Si de plus, $\mathcal{D}$ est un ouvert borné de classe $\mathcal{C}^1$, alors il existe une injection continue de $H^m(\mathcal{D})$ dans $\mathcal{C}^k(\bar{\mathcal{D}})$ dès que $m > k + d/2$. Cette propriété ne doit pas être confondue avec le fait que $\mathcal{C}^1(\bar{\mathcal{D}}) \subset H^1(\mathcal{D})$.

Notons H^{-m} l'ensemble des formes linéaires continues sur H^m [Schwartz and Melese, 1951]. On dit que H^{-m} est le dual topologique de H^m, ce qui s'écrit [3] :

$$H^{-m} = (H^m)^\star. \tag{1.6}$$

Pour appliquer un élément $f \in H^{-m}$ à un élément $g \in H^m$, on introduit le crochet de distribution défini par :

$$\langle f, g \rangle_{H^{-m}, H^m} \stackrel{d}{=} f(g) \in \mathbb{R}. \tag{1.7}$$

A noter que le double indiçage permet de ne pas le confondre avec le produit scalaire sur H^m.

2. En fait, les espaces de Sobolev admettent plusieurs définitions équivalentes. La définition (1.3) peut s'étendre au cas où f est une distribution régulière et où la dérivation est comprise au sens des distributions. Il est également possible de construire les espaces de Sobolev par complétion d'espaces de fonctions-tests. On s'autorisera donc à utiliser cette dernière dénomination pour désigner les éléments de H^m.

3. Selon l'ouvrage, les notations peuvent varier et la notation $E^\star$ est parfois réservée au dual algébrique, qui n'est pas le dual topologique de E. Comme la notion de dual algébrique ne sera pas utile dans cette étude, il ne sera pas nécessaire de distinguer les notations.

On souhaite désormais relier la définition (1.7) au produit scalaire dans L^2 (1.2). Pour ce faire, remarquons que pour tout entier m positif :

$$H^m \subset L^2, \tag{1.8}$$

avec la convention que $H^0 = L^2$. Cette inclusion est dense et continue, et permet de justifier l'inclusion entre espaces duaux [Gelfand and Vilenkine, 1964] :

$$(L^2)^\star \subset (H^m)^\star = H^{-m}. \tag{1.9}$$

On identifie alors L^2 à son dual grâce au théorème de Riesz et on se retrouve en présence du triplet de Gelfand :

$$H^m \subset L^2 \subset H^{-m}. \tag{1.10}$$

Dans la relation (1.10), L^2 joue le rôle d'espace « pivot » entre l'espace de fonctions H^m et l'espace de distributions H^{-m}. En particulier, le produit scalaire de L^2 et le crochet de distribution sont compatibles, ce qui signifie que pour tout $f \in L^2$ et tout $g \in H^m$:

$$\langle f, g \rangle_{H^{-m}, H^m} = \langle f, g \rangle_{L^2}. \tag{1.11}$$

La structure offerte par le triplet de Gelfand est beaucoup utilisée en analyse fonctionnelle et en mécanique quantique. En effet, toutes les opérations ne sont pas possibles sur les fonctions de L^2. Dans la suite, on souhaite considérer plus particulièrement des opérateurs de dérivation. Ces derniers n'agissent pas sur toutes les fonctions de L^2. Il est donc naturel d'introduire le sous-espace H des fonctions suffisamment régulières pour qu'on puisse les manipuler. L'espace dual $H^\star$ est quant à lui traditionnellement employé pour définir les vecteurs propres des opérateurs au spectre continu.

1.1.2 Fonctions et opérateurs de covariance

Un opérateur de covariance Υ sur un espace de Hilbert H est une application linéaire symétrique définie positive de $H^\star$ dans H [Tarantola, 2005]. Physiquement, Υ associe à toute distribution de $H^\star$ une unique fonction de H. C'est un « lisseur », comme illustré sur la figure 1.1. Inversement, l'opérateur Υ^{-1} est une application de H dans $H^\star$ et définit un nouveau produit scalaire sur H :

$$\langle f, g \rangle_H \stackrel{d}{=} \langle \Upsilon^{-1}(f), g \rangle_{H^\star, H}, \tag{1.12}$$

cela pour toutes fonctions $f \in H$ et $g \in H$.

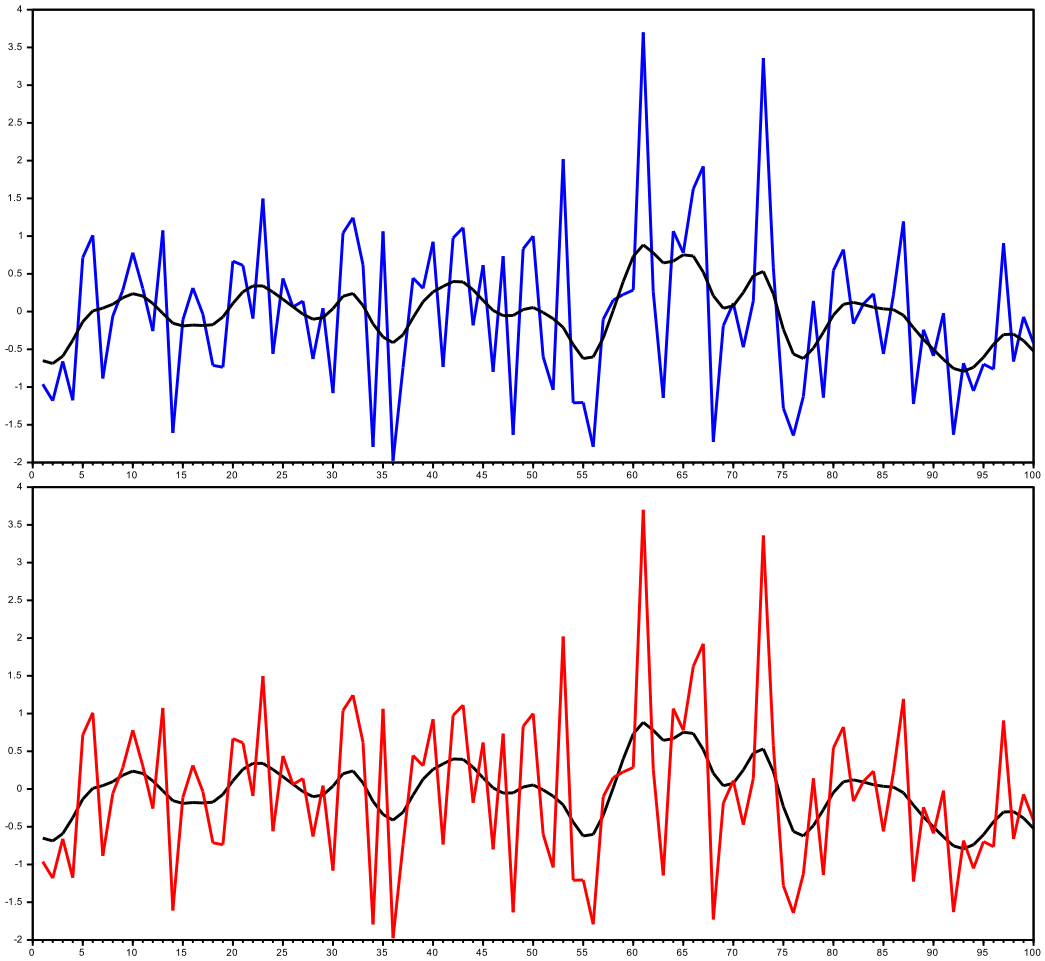


Figure 1.1 – En haut : action de Υ. Signal d'entrée en bleu, signal de sortie en noir. On constate l'effet de lissage de l'opérateur de covariance. En bas : action de Υ^{-1}. Signal d'entrée en noir, signal de sortie en rouge. L'opérateur de covariance inverse introduit des hautes fréquences dans la réponse.

Lorsque l'on dispose d'une fonction de covariance ρ du type :

$$\begin{aligned} \rho \quad : \quad & \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \\ & (\boldsymbol{z}, \boldsymbol{z}') \mapsto \rho(\boldsymbol{z}, \boldsymbol{z}'), \end{aligned} \tag{1.13}$$

il est possible de définir un opérateur de covariance en posant pour tout $f \in H^\star$:

$$\begin{aligned} \Upsilon(f) \quad : \quad & \mathbb{R}^d \to \mathbb{R} \\ & \boldsymbol{z} \mapsto \langle f, \rho(\boldsymbol{z}, \cdot) \rangle_{H^\star, H}, \end{aligned} \tag{1.14}$$

où $\rho(\boldsymbol{z}, \cdot)$ désigne l'application $\boldsymbol{z}' \mapsto \rho(\boldsymbol{z}, \boldsymbol{z}')$. Si f est suffisamment régulière (typiquement L^2), l'expression (1.14) prend la forme d'une intégrale, grâce à la relation (1.11). On a donc :

$$\begin{aligned} \Upsilon(f) \quad &: \quad \mathbb{R}^d \to \mathbb{R} \\ &\boldsymbol{z} \mapsto \int_{\mathbb{R}^d} \rho(\boldsymbol{z}, \boldsymbol{z}') f(\boldsymbol{z}') \mathrm{d}\boldsymbol{z}'. \end{aligned} \tag{1.15}$$

Dans ce cas, on dit que ρ est le noyau de Υ. En général, il n'est pas possible d'exprimer Υ^{-1} à partir d'un noyau ρ^{-1} qui soit une fonction classique. Toutefois, à condition d'autoriser ρ^{-1} à être une somme de distributions, on peut dériver une expression similaire à (1.15) pour Υ^{-1} [Jones, 1982].

Les fonctions de covariance contiennent globalement deux types d'informations : la variance et les corrélations. Pour la suite, il est utile de pouvoir distinguer les deux.

Notons $\sigma(\boldsymbol{z}) = \sqrt{\rho(\boldsymbol{z}, \boldsymbol{z})}$ la fonction écart-type évaluée en $\boldsymbol{z} \in \mathbb{R}^d$ et $\sigma^2(\boldsymbol{z})$ la variance. La fonction de corrélation c est définie par la relation :

$$\rho(\boldsymbol{z}, \boldsymbol{z}') = \sigma(\boldsymbol{z}) \sigma(\boldsymbol{z}') c(\boldsymbol{z}, \boldsymbol{z}'), \tag{1.16}$$

cela pour tout couple $(\boldsymbol{z}, \boldsymbol{z}') \in \mathbb{R}^d \times \mathbb{R}^d$. Cette définition permet d'assurer que la fonction de corrélation est d'amplitude unitaire, c'est-à-dire que :

$$\forall \boldsymbol{z} \in \mathbb{R}^d \quad : \quad c(\boldsymbol{z}, \boldsymbol{z}') = 1. \tag{1.17}$$

L'opérateur de corrélation $\mathcal{C}$ se définit ensuite de la même façon que Υ en posant :

$$\mathcal{C}(f)(\boldsymbol{z}) = \int_{\mathbb{R}^d} c(\boldsymbol{z}, \boldsymbol{z}') f(\boldsymbol{z}') \mathrm{d}\boldsymbol{z}'. \tag{1.18}$$

L'opérateur de corrélation inverse est quant à lui défini dans Mirouze and Weaver [2010] par :

$$\mathcal{C}^{-1}(f)(\boldsymbol{z}) = \int_{\mathbb{R}^d} c^{-1}(\boldsymbol{z}, \boldsymbol{z}') f(\boldsymbol{z}') \mathrm{d}\boldsymbol{z}', \tag{1.19}$$

avec

$$c^{-1}(\boldsymbol{z}, \boldsymbol{z}') = \sigma(\boldsymbol{z}) \sigma(\boldsymbol{z}') \rho^{-1}(\boldsymbol{z}, \boldsymbol{z}'). \tag{1.20}$$

En pratique, les variances d'erreurs d'observation sont faciles à estimer et à modéliser. En effet, ces variances constituent un champ de scalaires positifs dont on dispose, dans le cas discret, d'un estimateur non biaisé prenant la

forme d'une matrice diagonale. Dans cette étude, l'estimation et la modélisation des variances en assimilation de données sont donc simplement mentionnées. Le choix est fait de se concentrer sur le cas difficile de la modélisation des fonctions de corrélation, des opérateurs de corrélation et de leurs inverses. On fait donc peu usage des définitions sur les covariances (à l'exception de la section 2.5, où corrélations et variances ne sont pas séparées). Néanmoins, tous les résultats concernant les opérateurs de covariance sont applicables aux fonctions de corrélations. En effet, il suffit de voir qu'un opérateur de corrélation n'est qu'un opérateur de covariance d'amplitude unitaire. Il en hérite donc de toutes les propriétés.

1.1.3 Fonctions de corrélation de Matérn

On dispose maintenant de tous les outils théoriques permettant de définir l'assimilation de données en dimension infinie. Cependant, on souhaite profiter de ce cadre théorique pour mettre en avant une classe spécifique de fonctions de corrélations : les fonctions de la famille Matérn [Whittle, 1963]. Cette famille de fonctions est une des plus communément utilisées en sciences appliquées, en raison de sa dépendance en un petit nombre de paramètres et de son lien direct avec les équations aux dérivées partielles. Ce dernier lien est une des principales motivations de cette étude, puisqu'il permet de faire appel à toutes les techniques de traitement des équations différentielles dans la modélisation des opérateurs de corrélation. La pertinence des fonctions de Matérn pour les applications en géosciences est discutée et illustrée dans la section 4.2.

La fonction de Matérn de paramètres (m, l) allant de $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, notée $c_{m,l}$, est définie en posant [Stein, 1999, Guttorp and Gneiting, 2006] :

$$c_{m,l}(\boldsymbol{z}, \boldsymbol{z}') = \frac{2^{1-m+d/2}}{\Gamma(m-d/2)} \left(\frac{\|\boldsymbol{z}-\boldsymbol{z}'\|_2}{l} \right)^{m-d/2} K_{m-d/2} \left(\frac{\|\boldsymbol{z}-\boldsymbol{z}'\|_2}{l} \right), \quad (1.21)$$

où $\|\cdot\|_2$ désigne la norme euclidienne dans $\mathbb{R}^d$, Γ la fonction gamma et K_v la fonction de Bessel modifiée de second type et d'ordre v. Par commodité, on pose dans la suite $r = \|\boldsymbol{z}-\boldsymbol{z}'\|_2$ et $c_{m,l}(\boldsymbol{z}, \boldsymbol{z}') = c_{m,l}(r)$.

On vérifie sur la figure 1.2 que $c_{m,l}$ est d'amplitude unitaire.

La régularité de $c_{m,l}$ dépend du paramètre m. A condition que $m > d/2$, on peut montrer que $c_{m,l}$ est $2m$ fois dérivable, cela pour tout $l > 0$. D'après

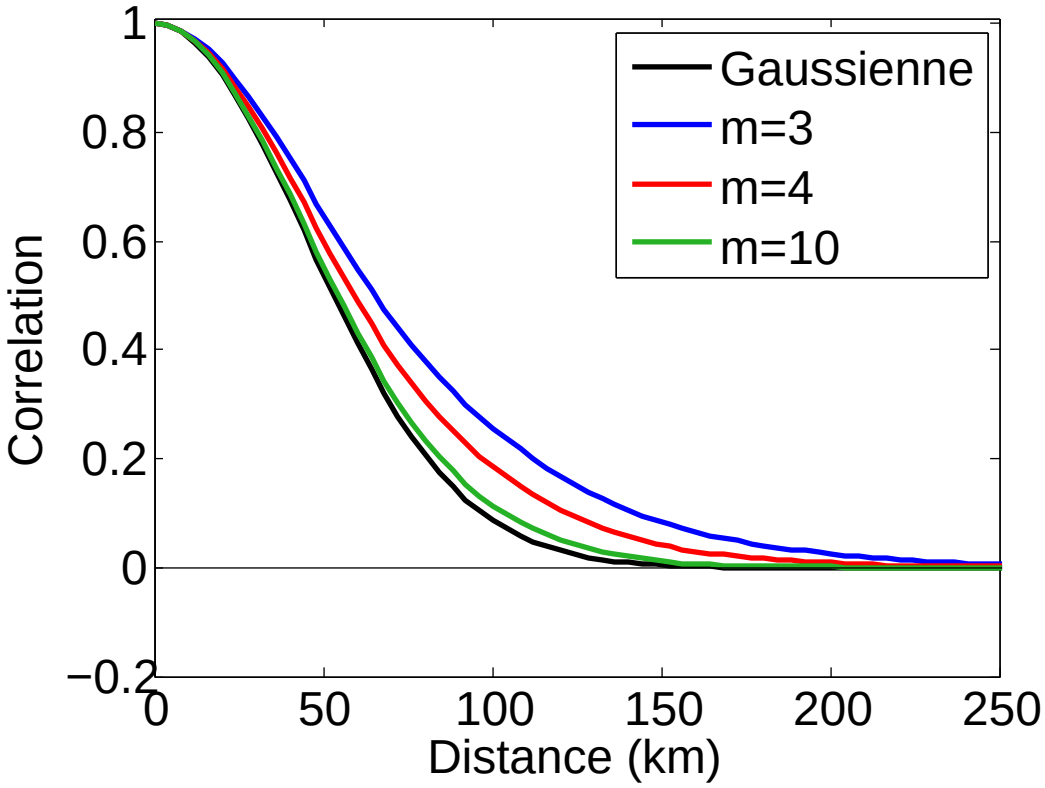


FIGURE 1.2 – Représentation de la fonction de Matérn $c_{m,l}$ pour différentes valeurs de m. La longueur de portée de Daley $D = 33.5$km est fixée. Le paramètre l est calculé d'après la formule (1.24). La courbe noire est une fonction gaussienne correspondant à la limite du cas $m \to +\infty$.

Stein [1999], lorsque $m \to +\infty$, $c_{m,l}$ converge vers la fonction gaussienne c^g d'expression :

$$c^g(r) = \frac{1}{\gamma_g} \exp\left(-\frac{r^2}{4l^2}\right). \tag{1.22}$$

A l'inverse, les petites valeurs de m correspondent à des fonctions de corrélation connues, comme la fonction exponentielle ($m = 1$, $d = 1$) ou la fonction autorégressive d'ordre deux ($m = 2$, d quelconque). La figure 1.3 donne l'allure des fonction de Matérn en dimension 2 lorsque $m = 2$.

Le paramètre l, quant à lui, est un paramètre d'échelle contrôlant l'intensité avec laquelle deux points éloignés de $\mathbb{R}^d$ sont liés entre eux. Soit D la

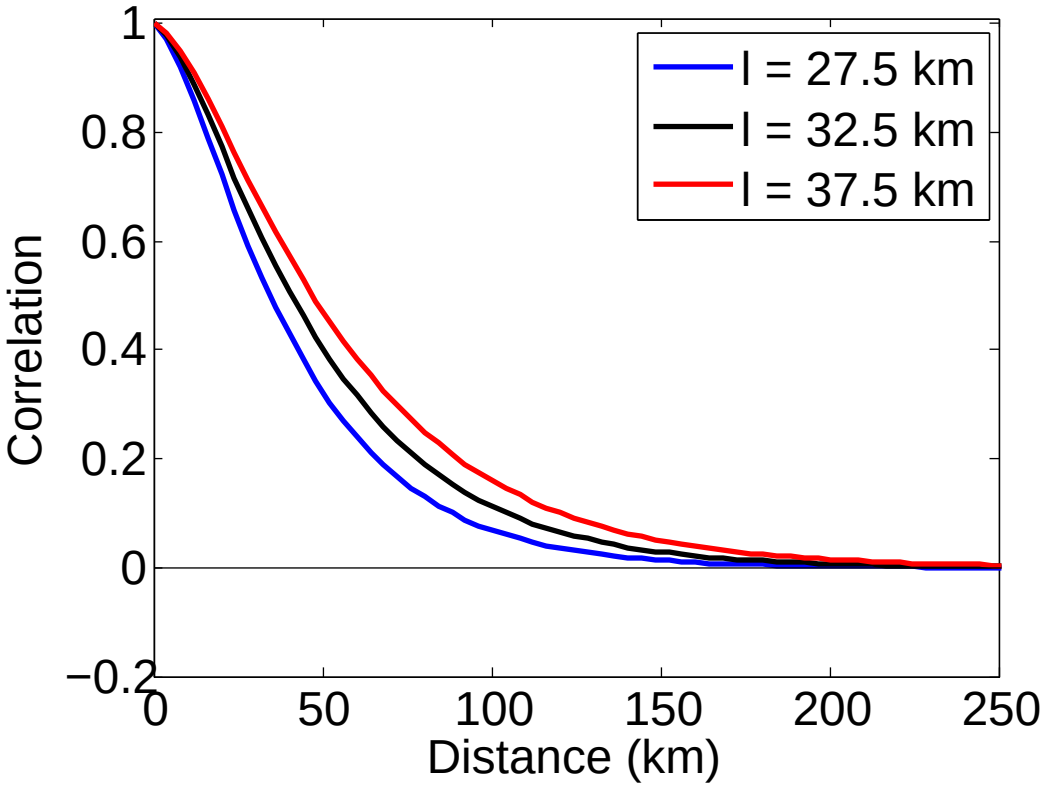


Figure 1.3 – Représentation de la fonction de Matérn $c_{m,l}$ pour différentes valeurs de l, le paramètre de régularité m étant fixé.

longueur de portée de Daley définie d'après Daley [1991] par

$$D = \sqrt{-\frac{c_{m,l}(0) \times d}{c''_{m,l}(0)}}, \tag{1.23}$$

où $c''_{m,l}$ est la dérivée seconde de c. La longueur D représente graphiquement la largeur de la parabole osculatrice de c à mi-hauteur (figure 1.4). Le lien entre l et D est alors donné par la relation :

$$D \approx \sqrt{2m - 2 - d} \times l. \tag{1.24}$$

Pour constater l'influence de chacun des paramètres sur l'aspect des fonctions de corrélation, le lecteur peut se référer aux figures 1.2, 1.3 et 1.4.

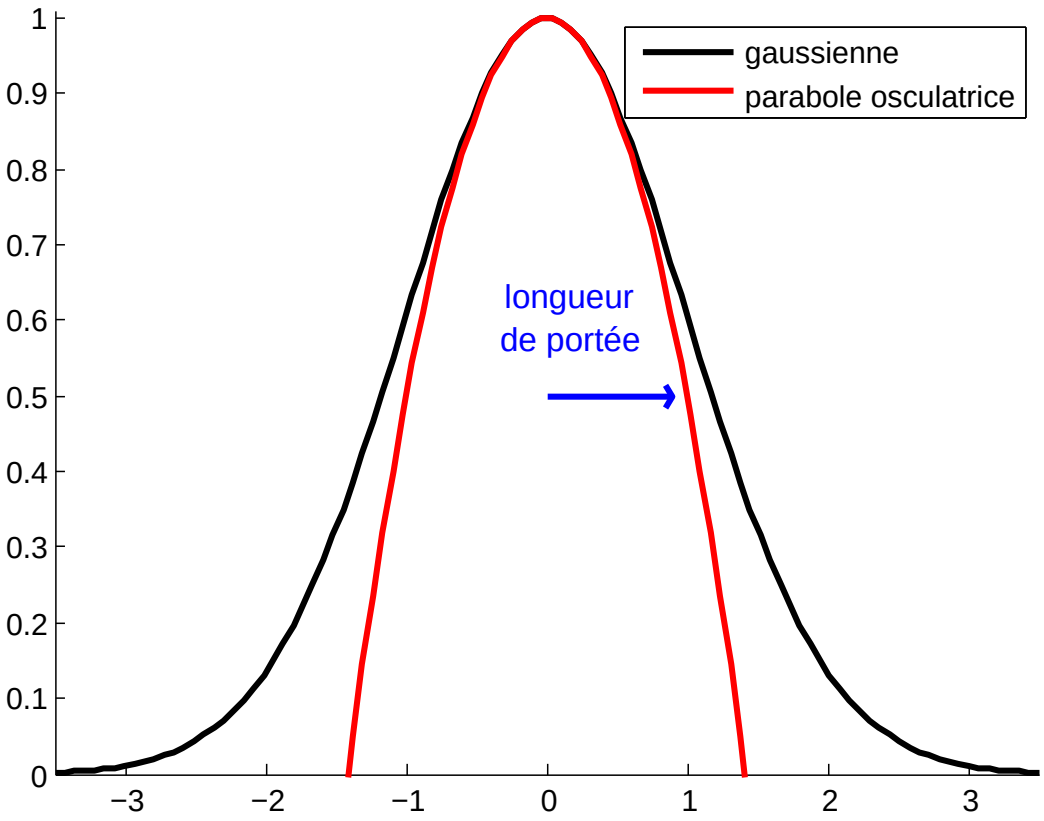


FIGURE 1.4 – Définition graphique de la longueur de portée de Daley D. Elle correspond à largeur de la parabole osculatrice (en rouge) à mi-hauteur de la fonction de Matérn (en noir).

1.1.4 Corrélation inverse et diffusion

La famille de fonctions de Matérn a bien des particularités, parmis lesquelles son lien structurel avec les processus autorégressifs, ce qui permet de représenter l'action des opérateurs de corrélation au travers d'équations différentielles. Dans cette sous-section, on montre comment les opérateurs de corrélation inverses sont dérivés à partir de ces fonctions en utilisant la transformée de Fourier (figure 1.5).

Soit $L^1(\mathbb{R})$ l'espace des fonctions intégrables de $\mathbb{R}$ à valeurs dans $\mathbb{R}$. Considérons c, une fonction de $L^2(\mathbb{R}) \cap L^1(\mathbb{R})$. On définit la transformée de Fourier $\hat{c}$ de c par

$$\hat{c}(\hat{r}) = \int_{\mathbb{R}} c(r) \exp(-2\mathrm{i}\pi r\hat{r}) \mathrm{d}\boldsymbol{z}, \tag{1.25}$$

où $\hat{r}$ désigne la variable du domaine de Fourier et i le nombre imaginaire pur.

La transformée de Fourier inverse permet de reconstituer c à partir de $\hat{c}$:

$$c(r) = \int_{\mathbb{R}} \hat{c}(\hat{r}) \exp(+2\mathrm{i}\pi r\hat{r}) \mathrm{d}\boldsymbol{z}. \tag{1.26}$$

D'après Gaspari and Cohn [1999], Mirouze and Weaver [2010], la transformée de Fourier de (1.21) s'écrit :

$$\hat{c}_{m,l}(\hat{r}) = \frac{\gamma_{m,l}}{[1 + (l\hat{r})^2]^m}, \tag{1.27}$$

où $\gamma_{m,l}$ est un facteur d'amplitude d'expression :

$$\gamma_{m,l} = 2^d \pi^{d/2} \frac{\Gamma(m)}{\Gamma(m - d/2)} l^d. \tag{1.28}$$

Posons

$$\hat{d}_{m,l}(\hat{r}) = \frac{1}{\hat{c}_{m,l}(\hat{r})} = \frac{[1 + (l\hat{r})^2]^m}{\gamma_{m,l}} \tag{1.29}$$

l'inverse de la transformée de Fourier de $c_{m,l}$. En réécrivant $(l\hat{r})^2 = -(\mathrm{i}l\hat{r})^2$ et en utilisant la transformée de Fourier inverse, Jones [1982] montre que la fonction $d_{m,l}$ peut s'écrire :

$$d_{m,l} = \frac{1}{\gamma_{m,l}} \sum_{k=0}^{m} \begin{pmatrix} m \\ k \end{pmatrix} (-1)^k l^{2k} \delta^{(2k)}, \tag{1.30}$$

où $\delta^{(k)}$ désigne la dérivée généralisée d'ordre k de la distribution de Dirac définie pour toute fonction f régulière par :

$$\langle \delta^{(k)}, f \rangle_{H^{-(k+1)}, H^{(k+1)}} = (-1)^k f^{(k)}(0) \tag{1.31}$$

et $f^{(k)}$ est la k-ième dérivée de f.

Dès lors, on construit $\mathcal{C}^{-1}$ en posant pour toute fonction $f \in H^{2m}$:

$$\begin{aligned}
\mathcal{C}^{-1}(f)(\boldsymbol{z}) &= (d_{m,l} * f)(\boldsymbol{z}) \\
&= \frac{1}{\gamma_{m,l}} \sum_{k=0}^{m} \begin{pmatrix} m \\ k \end{pmatrix} (-1)^k l^{2k} (\delta^{(2k)} * f)(\boldsymbol{z}) \\
&= \frac{1}{\gamma_{m,l}} \sum_{k=0}^{m} \begin{pmatrix} m \\ k \end{pmatrix} (-1)^k l^{2k} f^{(2k)}(\boldsymbol{z}) \\
&= \frac{1}{\gamma_{m,l}} \sum_{k=0}^{m} \begin{pmatrix} m \\ k \end{pmatrix} (-1)^k l^{2k} \Delta^k f(\boldsymbol{z}) \\
&= \frac{1}{\gamma_{m,l}} [1 - l^2 \Delta]^m f(\boldsymbol{z}).
\end{aligned} \tag{1.32}$$

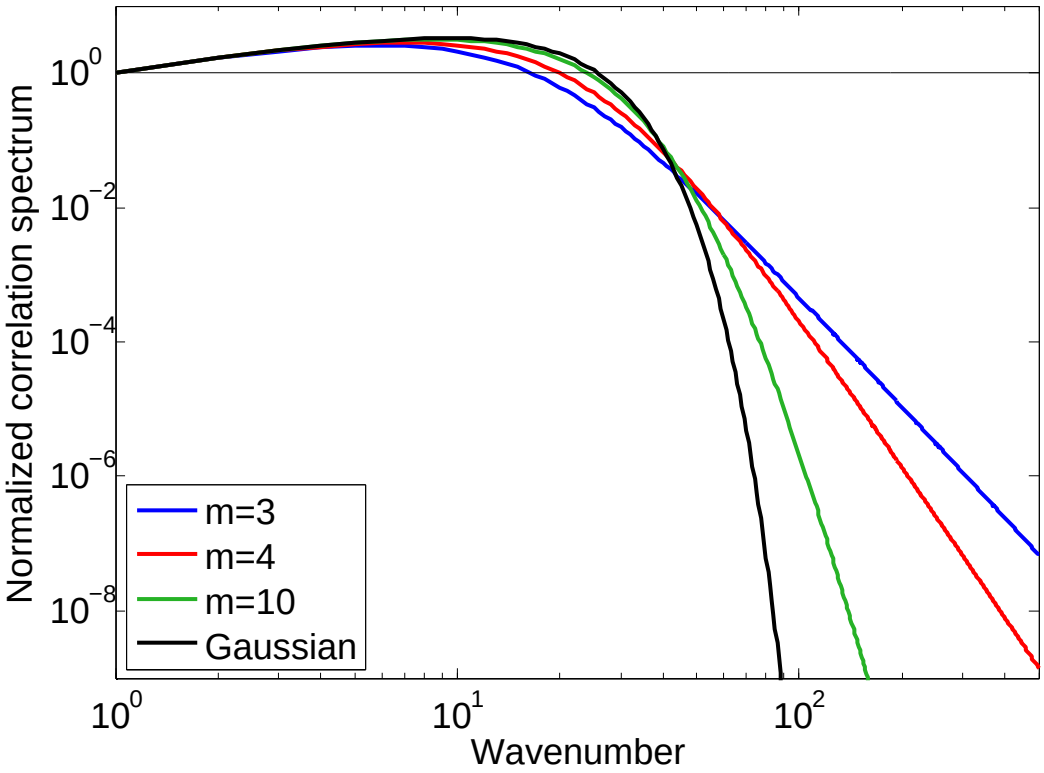


FIGURE 1.5 – Transformée de Fourier de la fonction de Matérn $c_{m,l}$ pour différentes valeurs de m et à longueur de portée de Daley D fixée. Plus m est grand, plus le spectre correspond à un filtre passe-bas.

Dans cette équation, $(T * f)(\boldsymbol{z})$ désigne le produit de convolution des distributions défini pour toute distribution $T \in H^\star$, toute fonction $f \in H$ et tout $\boldsymbol{z} \in \mathbb{R}^d$ par

$$(T * f)(\boldsymbol{z}) = \langle T, f(\boldsymbol{z} - \cdot)\rangle_{H^\star,H}. \tag{1.33}$$

L'opérateur Laplacien Δ est quant à lui défini comme

$$\Delta = \nabla \cdot \nabla, \tag{1.34}$$

où ∇ désigne l'opérateur gradient et $\nabla\cdot = -\nabla^{\mathrm{T}}$ l'opérateur de divergence. La transposition est à comprendre au sens du crochet de dualité [Tarantola, 2005]. Soient E et F deux espaces de Hilbert, u un élément de E, v un élément de F et Z un opérateur de E dans F. L'opérateur transposé de Z est l'unique opérateur de $F^\star$ dans $E^\star$ vérifiant

$$\langle v, Z[u]\rangle_{F^\star,F} = \langle Z^{\mathrm{T}}[v], u\rangle_{E^\star,E}. \tag{1.35}$$

A noter que l'opérateur Laplacien est symétrique par construction [Brezis, 2010, Evans, 2010], c'est-à-dire que pour tout $u \in H^2$ et tout $v \in H^2$ (le

domaine de Δ),

$$\langle v, \Delta u \rangle_{L^2} = \langle \Delta v, u \rangle_{H^{-2}, H^2}. \tag{1.36}$$

D'après (1.11), l'équation (1.36) s'écrit sous la forme du produit scalaire dans l'espace pivot L^2 :

$$\int_{\mathcal{D}} (\Delta f)(\boldsymbol{z}) g(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \int_{\mathcal{D}} f(\boldsymbol{z}) (\Delta g)(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}, \tag{1.37}$$

ce qu'on peut vérifier à l'aide de la formule de Green lorsque $\mathcal{D}$ est infini ou lorsqu'on impose des conditions aux limites de Neumann à ses frontières. De même, $[1 - l^2\Delta]^m$ est symétrique, ce qui est en accord avec la symétrie de $\mathcal{C}^{-1}$:

$$\langle v, (1 - l^2\Delta)^m u \rangle_{L^2} = \langle (1 - l^2\Delta)^m v, u \rangle_{H^{-2m}, H^{2m}}, \tag{1.38}$$

ce qui se démontre facilement par récurrence sur m.

On remarque que la définition (1.31) implique que l'opérateur (1.32) agit sur les fonctions de H^{2m} et non de H^m. On peut contourner ce problème en définissant :

$$\begin{aligned} \mathcal{C}^{-1/2} &: H^m \to L^2 \\ &= [1 - l^2\Delta]^{m/2} \end{aligned} \tag{1.39}$$

et son application transposée :

$$\begin{aligned} (\mathcal{C}^{-1/2})^T &: L^2 \to H^{-m} \\ &= [1 - l^2\Delta]^{m/2}, \end{aligned} \tag{1.40}$$

où l'opération de dérivation est cette fois à prendre au sens des distributions, de telle sorte que :

$$\begin{aligned} \mathcal{C}^{-1} &: H^m \to H^{-m} \\ &= (\mathcal{C}^{-1/2})^{\mathrm{T}} (\mathcal{C}^{-1/2}). \end{aligned} \tag{1.41}$$

En inversant (1.32), on arrive à formuler l'opérateur de corrélation $\mathcal{C}$ comme la résolution d'une équation de diffusion sur m itérations :

$$\left\{ \begin{array}{l} \frac{1}{\gamma_{m,l}} [1 - l^2\Delta]^m f_m(\boldsymbol{z}) = f_0(\boldsymbol{z}) \\ f_0(\boldsymbol{z}) = f(\boldsymbol{z}) \end{array} \right. \quad \text{et} \quad f_m(\boldsymbol{z}) = \mathcal{C}(f)(\boldsymbol{z}). \tag{1.42}$$

Cette forme a l'avantage de ne plus faire apparaître le noyau de corrélation $c(\boldsymbol{z}, \boldsymbol{z}')$ et on peut résoudre cette équation par étapes successives en ne traitant qu'un seul « $[1 - l^2\Delta]$ » à la fois.

De fait, il est possible d'affaiblir le cadre fonctionnel de telle sorte que $\mathcal{C}$ devienne un opérateur de H^{-1} dans H^1, cela peu importe le nombre d'itérations m. Pour ce faire, il suffit de constater que (H^1, L^2, H^{-1}) est lui-même un triplet de Gelfand, en tant que cas particulier de (1.10), et de considérer l'injection continue $\Phi : H^1 \to L^2$ et son application transposée comme moyen de « remonter » de l'espace d'arrivée de $[1 - l^2\Delta]$ vers son espace de départ. On rappelle que l'application transposée agit « du dual vers le dual » et qu'on a donc :

$$\begin{aligned} \Phi &: H^1 \to L^2 \\ \Phi^{\mathrm{T}} &: L^2 = (L^2)^\star \to (H^1)^\star = H^{-1}. \end{aligned} \tag{1.43}$$

Pour mnémotechnique, il convient d'introduire un diagramme de dualité qui expose les espaces fonctionnels et leurs liens (figure 1.6).

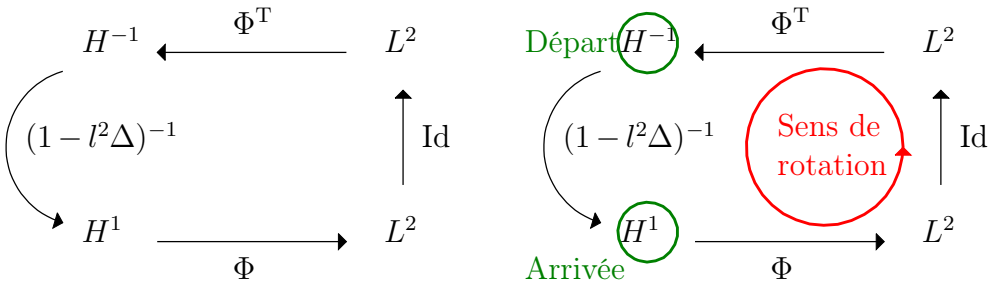


Figure 1.6 – Diagramme de dualité. A gauche le diagramme simple contenant les informations sur les espaces. A droite le même diagramme avec le sens de lecture.

Avec ce nouveau formalisme, l'expression de $\mathcal{C}$ se complexifie d'apparence, puisqu'elle fait intervenir le produit $\Phi^{\mathrm{T}}\Phi$:

$$\mathcal{C} = \gamma_{m,l} \times [1 - l^2\Delta]^{-1} \times \Phi^{\mathrm{T}}\Phi \times \ldots \times \Phi^{\mathrm{T}}\Phi \times [1 - l^2\Delta]^{-1}, \tag{1.44}$$

où la multiplication est à prendre au sens de la composition des fonctions. De même, on a

$$\mathcal{C}^{-1} = \frac{1}{\gamma_{m,l}} \times [1 - l^2\Delta] \times (\Phi^{\mathrm{T}}\Phi)^{-1} \times \ldots \times (\Phi^{\mathrm{T}}\Phi)^{-1} \times [1 - l^2\Delta]. \tag{1.45}$$

En réalité, le produit $\Phi^{\mathrm{T}}\Phi$ est transparent lorsqu'on reste dans le domaine continu. Toutefois, il sera d'importance capitale de le prendre en compte lors

de la discrétisation des opérateurs dans des espaces de dimension finie. En effet, l'espace H^1 sera approximé par un espace vectoriel de dimension finie. La matrice de représentation de $\Phi^{\mathrm{T}}\Phi$ sera la matrice de Gram de la base choisie dans cet espace.

1.2 Assimilation de données

En météorologie, l'état de l'atmosphère est estimé grâce à un modèle de prévision numérique du temps. Néanmoins, cette estimation est imprécise et il est nécessaire de la corriger régulièrement à l'aide d'observations de l'atmosphère indépendantes du modèle de prévision. Ce procédé cyclique de prédiction/correction est appelé « assimilation de données ». L'assimilation de données est un domaine scientifique transverse qui rassemble des concepts issus à la fois de la physique et des mathématiques appliquées : paramétrisations physiques variées, optimisation, statistiques, probabilités, contrôle optimal, analyse fonctionnelle, algèbre linéaire, techniques de parallélisation, couplage de code... et bien d'autres. Ainsi, il existe plusieurs manières d'introduire l'assimilation de données, suivant le sujet de l'étude.

Le choix est fait d'introduire l'assimilation de données comme un problème d'optimisation [Lorenc, 1986]. Le cycle de prédiction/correction donne lieu à la minimisation d'une fonctionnelle dont l'expression fait intervenir des opérateurs de covariance, comme définis au début du chapitre. A noter que ce choix n'est pas unique. Nombre d'ouvrages introduiront l'assimilation de données à partir de la théorie des probabilités [Burgers et al., 1998, Hamill et al., 2000]. Toutefois, ce point de vue n'est pas exploité dans la suite de cette étude, qui s'attache principalement à la construction des opérateurs de covariance et de corrélation, et non aux propriétés statistiques des champs considérés. Pour une présentation de l'assimilation de données probabiliste, le lecteur peut se référer à Houtekamer and Mitchell [1998] ou Whitaker and Hamill [2002].

Dans cette section, on montre comment faire le lien entre la théorie continue des opérateurs de corrélation et la formulation discrète classique. La discrétisation est vue comme une approximation du problème continu en dimension finie. Ces développements vont de paire avec la lecture du diagramme de dualité. On retrouve ainsi la distinction entre dimension finie et dimension infinie d'une part, et espaces fonctionnels et espaces de coordonnées d'autre part.

1.2.1 Optimisation en dimension infinie

Supposons que l'on dispose d'un modèle qui prévoit l'état x^b de l'atmosphère. Cet état x^b, appelé « ébauche », est une fonction de la variable spatiale $\boldsymbol{z}$ définie sur un domaine Ω à la surface de la Terre et représente une approximation du « vrai » état de l'atmosphère x^t. Cette approximation peut être corrigée à l'aide d'« observations » dont on suppose l'existence d'une représentation continue notée y^o (figure 1.7). Notons $\mathcal{X}$ l'espace du modèle et $\mathcal{Y}$ l'espace des observations :

$$(x^a, x^b) \in \mathcal{X}^2 \tag{1.46}$$

$$y^o \in \mathcal{Y}. \tag{1.47}$$

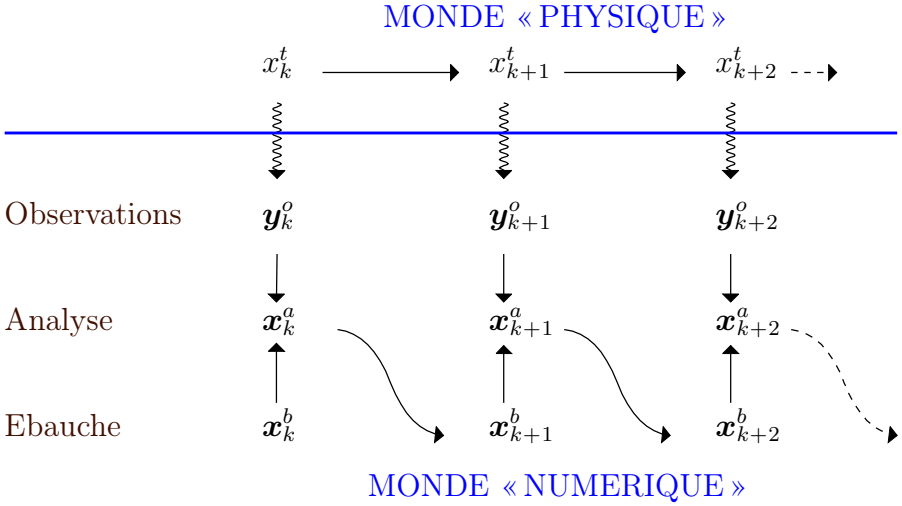


Figure 1.7 – Représentation du cycle d'estimation-correction en assimilation de données. L'indice k est une variable temporelle. A chaque pas de temps, l'analyse est calculée à partir de l'ébauche et des observations. Ces dernières sont reliées à l'état vrai x^t au travers de l'opérateur d'observation $\mathcal{H}$.

Sauf cas particulier, il n'y a pas de raison pour que les espaces $\mathcal{X}$ et $\mathcal{Y}$ contiennent les mêmes types de fonctions. En effet, les variables habituellement manipulées par le modèle sont des champs météorologiques comme le vent, la température, la pression de surface ou le géopotentiel. La prise en compte des liens entre ces variables est expliquée dans Derber and Bouttier [1999], Dee et al. [2011]. A l'inverse, les observations contiennent des informations très diverses, allant des champs météorologique aux radiances des

satellites. On a donc en général $\mathcal{X} \neq \mathcal{Y}$. De plus, cette distinction permettra d'inclure $\mathcal{X}$ et $\mathcal{Y}$ dans des espaces de Sobolev de régularités différentes.

L'enjeu de l'assimilation de données variationnelle est alors de trouver l'état x^a, appelé « analyse », qui soit le meilleur compromis entre les deux sources d'information x^b et y^o. On cherche donc à minimiser à la fois l'écart à l'ébauche $(x^a - x^b)$ et l'écart aux observations $(\mathcal{H}(x^a) - y^o)$. Dans cette dernière expression, $\mathcal{H}$ désigne l'opérateur d'observation qui associe à tout état $x \in \mathcal{X}$ le champ $y \in \mathcal{Y}$ qui serait observé par le réseau d'observation si x était le vrai état de l'atmosphère :

$$\mathcal{H} \quad : \quad \mathcal{X} \to \mathcal{Y}. \tag{1.48}$$

Pour chercher ce compromis, on munit respectivement $\mathcal{X}$ et $\mathcal{Y}$ des produits scalaires

$$(f, g) \in \mathcal{X}^2 \mapsto \langle f, g \rangle_{\mathcal{B}^{-1}} = \langle \mathcal{B}^{-1/2} f, \mathcal{B}^{-1/2} g \rangle_{L^2} \in \mathbb{R} \tag{1.49}$$

et

$$(f, g) \in \mathcal{Y}^2 \mapsto \langle f, g \rangle_{\mathcal{R}^{-1}} = \langle \mathcal{R}^{-1/2} f, \mathcal{R}^{-1/2} g \rangle_{L^2} \in \mathbb{R} \tag{1.50}$$

où $\mathcal{B}$ et $\mathcal{R}$ sont des opérateurs de covariance définis comme en partie 1.1. Les normes $\|\cdot\|_{\mathcal{B}^{-1}}$ et $\|\cdot\|_{\mathcal{R}^{-1}}$ sont les normes respectivements issues des produits scalaires (1.49) et (1.50). Le problème de l'assimilation de données se formule alors comme un problème de minimisation sur $\mathcal{X}$ [Courtier et al., 1998] et s'écrit :

$$x^a = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \mathcal{J}(x) \tag{1.51}$$

avec

$$\mathcal{J}(x) = \frac{1}{2} \|x - x^b\|^2_{\mathcal{B}^{-1}} + \frac{1}{2} \|\mathcal{H}(x) - y^o\|^2_{\mathcal{R}^{-1}}. \tag{1.52}$$

On remarque que l'existence d'un minimum de la fonctionnelle (1.52) dépend fortement de la non-linéarité de l'opérateur d'observation $\mathcal{H}$. En théorie, il suffit de montrer que $\mathcal{J}$ est fortement convexe [Allaire, 2005]. En pratique, il n'est pas possible de montrer que cette condition est vérifiée et l'existence d'un minimum de $\mathcal{J}$ est admise.

La formulation de l'assimilation de données comme la minimisation de (1.52) peut s'interpréter comme la recherche d'un maximum de vraisemblance dans l'espace $\mathcal{X}$ des paramètres du modèle [Tarantola, 2005]. Toutefois, la définition de la distribution *a posteriori* caractérisant l'état de l'atmosphère sachant les observations n'est pas immédiate en dimension infinie. Considérant l'ébauche comme l'information *a priori* et les observations comme source

d'information secondaire, le théorème de Bayes sur les densités de probabilités doit être généralisé aux mesures en utilisant la dérivée de Radon-Nykodin [Nikodym, 1930]. Dans ce cas, les deux approches sont équivalentes.

1.2.2 Du continu au discret

La représentation continue des champs météorologiques est utile d'un point de vue théorique pour définir les opérateurs de corrélations et leurs inverses. Cependant, on ne connaît en pratique la valeur des champs x^b et y^o qu'en un nombre fini de points d'échantillonnage (figure 1.8). En effet, les observations sont des mesures ponctuelles (réseau d'observation au sol, en mer, avions ...) ou des pixels d'images télédétectées (radars, satellites, ...). Le modèle, quant à lui, effectue sa prédiction sur une grille et sa résolution maximale dépend principalement des moyens en calcul disponibles. Il est donc nécessaire d'exprimer l'assimilation de données sous forme discrète.

On construit le paradigme discret à partir du continu. Pour ce faire, supposons que les espaces $\mathcal{X}$ et $\mathcal{Y}$ soient de dimension finie et notons $n = \dim(\mathcal{X})$ et $p = \dim(\mathcal{Y})$. On peut considérer une base $(\phi_k)_{k\in[\![1,n]\!]}$ sur laquelle représenter tout élément de $\mathcal{X}$ et une base $(\varphi_l)_{l\in[\![1,p]\!]}$ sur laquelle représenter tout élément de $\mathcal{Y}$. On obtient ainsi les décompositions :

$$\forall x \in \mathcal{X}, \quad \exists (x_k)_{k\in[\![1,n]\!]} \in \mathbb{R}^n, \quad x = \sum_{k\in[\![1,n]\!]} x_k \phi_k \tag{1.53}$$

et

$$\forall y \in \mathcal{Y}, \quad \exists (y_l)_{l\in[\![1,p]\!]} \in \mathbb{R}^p, \quad y = \sum_{l\in[\![1,p]\!]} y_l \varphi_l. \tag{1.54}$$

Suivant ces notations, on désigne par $\boldsymbol{x}^b = (x_k^b)_{k\in[\![1,n]\!]}$ l'ensemble des valeurs de champs prédites par le modèle aux points $(\boldsymbol{z}_k^b)_{k\in[\![1,n]\!]}$ et par $\boldsymbol{y}^o = (y_l^o)_{l\in[\![1,p]\!]}$ l'ensemble des mesures aux points $(\boldsymbol{z}_l^o)_{l\in[\![1,p]\!]}$ issues du réseau d'observation. Chaque point $\boldsymbol{z}$ est un vecteur de coordonnées dans le domaine d'étude. *A fortiori*, $\boldsymbol{z} \in \mathbb{R}^d$ où d est la dimension du domaine d'étude. Dès lors, l'équation (1.52) se transforme en posant

$$y^s = \mathcal{H}(x). \tag{1.55}$$

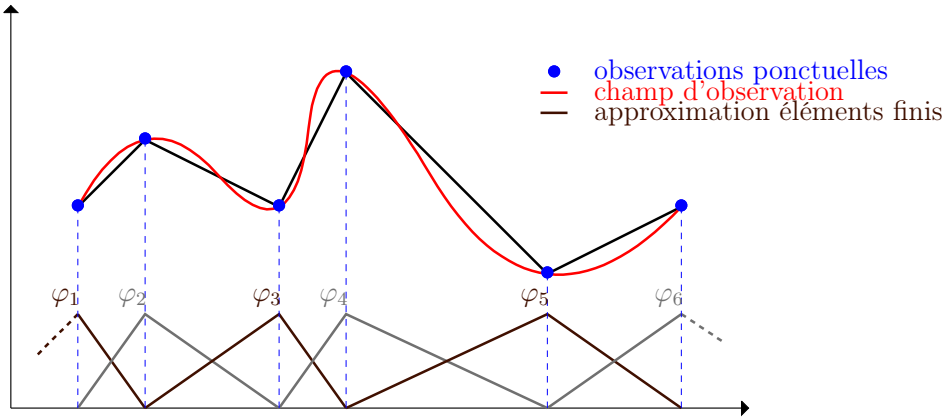


FIGURE 1.8 – Lien entre le champ d'observation (en rouge), sa représentation en éléments finis (en noir) et les observations ponctuelles (en bleu). Sont représentées en bas les fonctions de pondération (φ_l) associées aux données (y_l).

On obtient :

$$\begin{aligned}
\mathcal{J}(x) &= \frac{1}{2}\|x - x^b\|^2_{\mathcal{B}^{-1}} + \frac{1}{2}\|y^s - y^o\|^2_{\mathcal{R}^{-1}} \\
&= \frac{1}{2}\left\|\sum_{k\in[\![1,n]\!]}(x_k - x^b_k)\phi_k\right\|^2_{\mathcal{B}^{-1}} + \frac{1}{2}\left\|\sum_{l\in[\![1,p]\!]}(y^s_l - y^o_l)\varphi_l\right\|^2_{\mathcal{R}^{-1}} \\
&= \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^b\|^2_{\boldsymbol{B}^{-1}} + \frac{1}{2}\|\boldsymbol{y}^s - \boldsymbol{y}^o\|^2_{\boldsymbol{R}^{-1}}, \qquad (1.56)
\end{aligned}$$

où $\boldsymbol{B}^{-1}$ et $\boldsymbol{R}^{-1}$ sont les matrices de précision de termes généraux :

$$\begin{aligned}
(\boldsymbol{B}^{-1})_{\text{ij}} &= \langle \mathcal{B}^{-1/2}\phi_i, \mathcal{B}^{-1/2}\phi_j\rangle_{L^2} \qquad (1.57)\\
(\boldsymbol{R}^{-1})_{\text{ij}} &= \langle \mathcal{R}^{-1/2}\varphi_i, \mathcal{R}^{-1/2}\varphi_j\rangle_{L^2}. \qquad (1.58)
\end{aligned}$$

Une expression plus détaillée des termes (1.57) et (1.58) sera donnée dans le chapitre 2 sur les éléments finis, lorsqu'on spécifiera la forme des fonctions de pondération.

L'analyse du système, x^a, est elle aussi déterminée par le vecteur $\boldsymbol{x}^a = (x^a_k)_{k\in[\![1,n]\!]}$ de ses coordonnées sur la base des $(\phi_k)_{k\in[\![1,n]\!]}$. On peut donc minimiser $\mathcal{J}(x)$ de manière équivalente dans l'espace des fonctions $\mathcal{X}$ ou dans l'espace des coordonnées, qu'on note X. Dans ce cas, on choisit d'adopter la

notation

$$\min_{x\in\mathcal{X}} \mathcal{J}(x) \Leftrightarrow \min_{\boldsymbol{x}\in X} \tilde{\mathcal{J}}(\boldsymbol{x}). \tag{1.59}$$

On suppose que le contexte permet de dissocier les deux notations $\mathcal{J}$ et $\tilde{\mathcal{J}}$ en cas de besoin et on décide d'abandonner le tilde. De la même manière, on note Y l'espace des coordonnées des éléments de $\mathcal{Y}$ sur la base $(\varphi_l)_{l\in[\![1,p]\!]}$.

En météorologie, il est courant que $n \gg p$ [Bouttier and Courtier, 1999]. Si on considère l'assimilation de données comme un problème de régression à partir des observations, le terme d'écart à l'ébauche dans (1.56) peut naturellement s'interpréter comme un terme de régularisation [Engl et al., 2000].

D'autre part, la non-linéarité de $\mathcal{H}$ se retrouve dans l'expression du terme y^s et se cache dans l'expression (1.56). Pour la rendre explicite, il suffit de considérer l'équation (1.55) comme une contrainte d'égalité. Le problème de minimisation s'écrit ainsi plus précisément :

$$\min_{\substack{\boldsymbol{x}\in X,\boldsymbol{y}\in Y \\ \boldsymbol{y}=\tilde{\mathcal{H}}(\boldsymbol{x})}} \mathcal{J}(\boldsymbol{x},\boldsymbol{y}), \tag{1.60}$$

où $\tilde{\mathcal{H}}$ désigne cette fois-ci l'opérateur qui associe les coordonnées entre elles. Notons Φ_X l'application qui associe x à $\boldsymbol{x}$ et Φ_Y l'application qui associe y à $\boldsymbol{y}$:

$$\begin{aligned} \Phi_X \quad : \quad & X \to \mathcal{X} \\ & \boldsymbol{x} \mapsto x = \sum_{k\in[\![1,n]\!]} x_k \phi_k \end{aligned} \tag{1.61}$$

et

$$\begin{aligned} \Phi_Y \quad : \quad & Y \to \mathcal{Y} \\ & \boldsymbol{y} \mapsto y = \sum_{l\in[\![1,p]\!]} y_l \varphi_l. \end{aligned} \tag{1.62}$$

L'opérateur d'observation $\tilde{\mathcal{H}}$ discret s'écrit :

$$\tilde{\mathcal{H}} = \Phi_Y^{-1} \times \mathcal{H} \times \Phi_X \tag{1.63}$$

De nouveau, on décide d'abandonner le tilde lorsque le contexte permet d'éviter la confusion entre $\mathcal{H}$ et $\tilde{\mathcal{H}}$.

Les applications Φ_X et Φ_Y sont au coeur de la notion de discrétisation. Elles sont l'analogue de l'application Φ définie en partie 1.1.4 (voir équation

1.43) qui permet de changer d'espace pour définir le crochet de dualité à partir du produit scalaire L^2. Dans la partie 2 sur les éléments finis, un sens sera donné au dual de Y (c'est aussi possible pour X, mais on s'intéresse à l'espace des observations). L'utilisation du théorème de Riesz et la spécification de Φ_Y^{T} permettront d'exprimer la métrique dans l'espace des coordonnées en fonction de $\Phi_Y^{\mathrm{T}}\Phi_Y$.

1.2.3 Fenêtres d'assimilation et aspects temporels

Dans l'approche précédente, le modèle fait une prévision à un temps t_0. Cette ébauche est combinée aux observations recueillies entre t_0 et un temps final t_1 pour produire une analyse à $t_{1/2}$, centre de la fenêtre d'assimilation $[t_0, t_1]$. Cette approche ne prend pas en compte la distribution temporelle des observations dans la fenêtre d'assimilation, ce qui donne lieu à une approche de type 3DVAR [Courtier et al., 1998]. Dans l'approche 4DVAR, en revanche, on assimile chaque observation au temps où elle est valide [Courtier et al., 1994]. On introduit pour cela l'indice temporel $i \in [\![0, N_f]\!]$, où N_f désigne le nombre de sous-intervalles de $[t_0, t_1]$, et on note respectivement $\boldsymbol{y}_i^o$, $\mathcal{H}_i$ et $\boldsymbol{R}_i$ le vecteur d'observations, l'opérateur d'observation et la matrice de covariance d'erreurs d'observation au temps t_i. La différence entre 3DVAR et 4DVAR est illustrée sur la figure 1.9. On introduit également le modèle dynamique non-linéaire $\mathcal{M}_i$ intégré entre les temps t_0 et t_i, de sorte que la fonctionnelle à minimiser s'écrit :

$$\mathcal{J}(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^b\|^2_{\boldsymbol{B}^{-1}} + \frac{1}{2}\sum_{i=0}^{N_f}\|\mathcal{H}_i(\mathcal{M}_i(\boldsymbol{x})) - \boldsymbol{y}_i^o\|^2_{\boldsymbol{R}_i^{-1}} \tag{1.64}$$

où $\boldsymbol{x}_b$ est défini à l'instant t_0. Dans cette expression, on suppose que les erreurs d'observation ne sont pas corrélées temporellement.

Dans sa formulation forte, le problème de l'assimilation 4DVAR peut s'écrire sous la forme :

$$\min_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}) \tag{1.65}$$

où $\mathcal{J}(\boldsymbol{x})$ est définie par l'équation (1.64). On souhaite compacifier les notations pour la suite. On pose donc :

$$\mathcal{G}_i(\boldsymbol{x}) = \mathcal{H}_i(\mathcal{M}_i(\boldsymbol{x})) \quad ; \quad \mathcal{G}(\boldsymbol{x}) = \begin{pmatrix} \mathcal{G}_1(\boldsymbol{x}) \\ \vdots \\ \mathcal{G}_{N_f}(\boldsymbol{x}) \end{pmatrix} \tag{1.66}$$

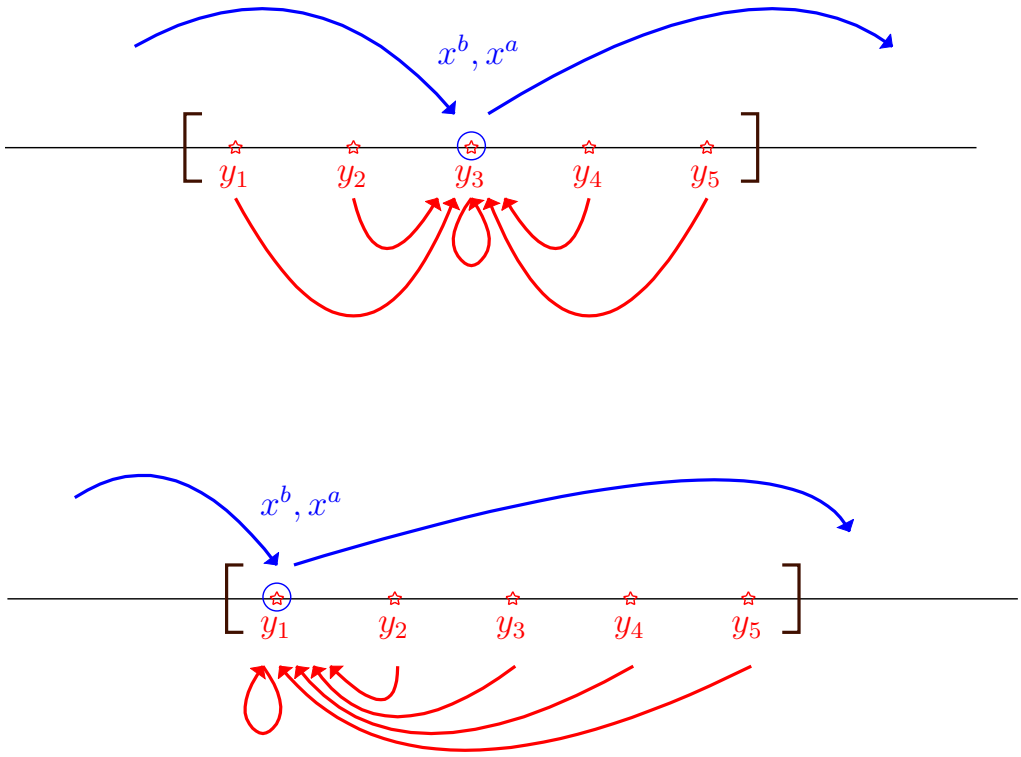


FIGURE 1.9 – En haut : 3DVAR. L'analyse est produite au centre de la fenêtre d'assimilation. En bas : 4DVAR. L'analyse est produite au début de la fenêtre d'assimilation.

puis

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1^o \\ \vdots \\ \boldsymbol{y}_{N_f}^o \end{pmatrix} \quad \text{et} \quad \boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_1 & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{R}_{N_f} \end{pmatrix}. \tag{1.67}$$

La fonctionnelle (1.64) se réécrit :

$$\mathcal{J}(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^b\|^2_{\boldsymbol{B}^{-1}} + \frac{1}{2}\|\mathcal{G}(\boldsymbol{x}) - \boldsymbol{Y}\|^2_{\boldsymbol{R}^{-1}}. \tag{1.68}$$

La similarité des formulations (1.56) et (1.68) justifie qu'on se contente d'étudier le problème 3DVAR dans la suite de ce manuscrit. De plus, « l'opérateur $\mathcal{R}$ » ou « la matrice $\boldsymbol{R}$ » feront systématiquement mention aux opérateurs $\mathcal{R}_i$ ou $\boldsymbol{R}_i$ dans le cas du 4DVAR.

1.2.4 Formulation incrémentale

Pour trouver la solution qui minimise $\mathcal{J}(\boldsymbol{x})$, on utilise l'algorithme de Gauss-Newton qui consiste à résoudre une séquence de problèmes quadratiques qui approximent localement le problème initial [Lawless et al., 2005, Gratton et al., 2007]. On parle alors de 4DVAR incrémental [Courtier et al., 1994].

On pose $\boldsymbol{x}^{(0)} = \boldsymbol{x}_b$. Soit $\boldsymbol{x}^{(k)}$ la solution à l'itération k. Le vecteur $\boldsymbol{x}^{(k+1)}$ est déterminé comme $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{s}^{(k)}$ où $\boldsymbol{s}^{(k)}$ est la quantité qui minimise :

$$\mathcal{J}^{(k)}(\boldsymbol{s}^{(k)}) = \|\boldsymbol{s}^{(k)} + \boldsymbol{x}^{(k)} - \boldsymbol{x}_b\|^2_{\boldsymbol{B}^{-1}} + \|\boldsymbol{G}^{(k-1)}\boldsymbol{s}^{(k)} - \boldsymbol{d}\|^2_{\boldsymbol{R}^{-1}}. \tag{1.69}$$

Ici, $\boldsymbol{G}^{(k-1)}$ est la matrice jacobienne de $\mathcal{G}$ au voisinage de $\boldsymbol{x}^{(k-1)}$ et $\boldsymbol{d} = \mathcal{G}(\boldsymbol{x}^{(k-1)}) - \boldsymbol{Y}$ le vecteur des innovations. La séquence des itérations de Gauss-Newton est communément appelée « boucle externe ». Elle est représentée sur la figure 1.10.

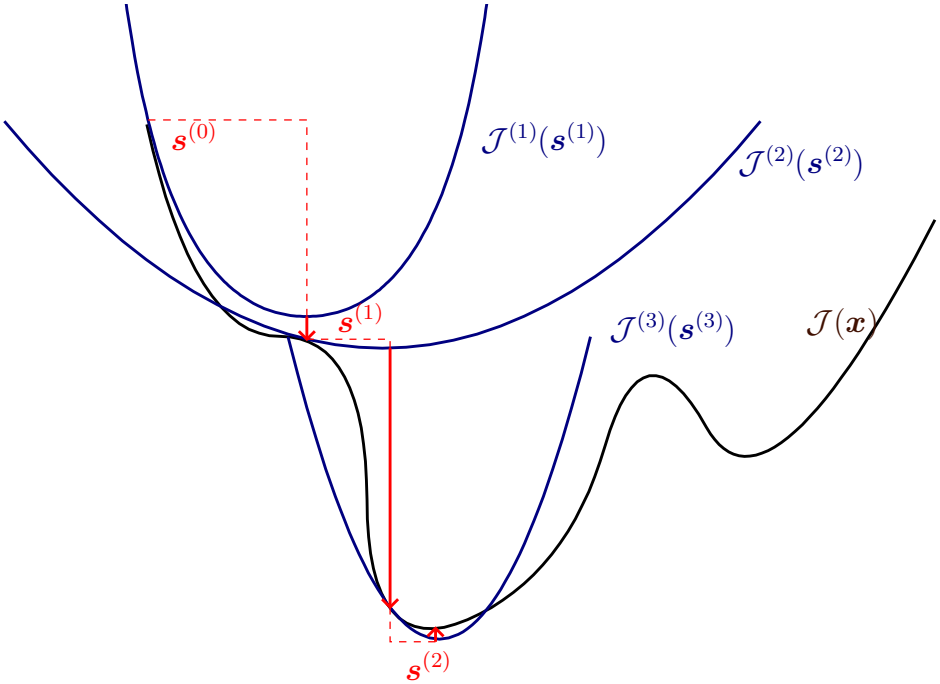


FIGURE 1.10 – Minimisation d'une fonction (en noir) par l'algorithme de Gauss-Newton. A chaque itération, $\boldsymbol{s}$ (en rouge) détermine le minimum de la parabole (en bleu) approximant localement la fonction à minimiser.

1.3 Résolution du problème variationnel

Le problème de l'assimilation de données admet de multiples formulations. Chaque formulation admet plusieurs méthodes de résolution. La multitude de ces méthodes fait naître des besoins variés en terme de modélisation et de spécification des opérateurs. En particulier, se pose la question de l'importance relative des matrices $\boldsymbol{R}$ et $\boldsymbol{R}^{-1}$. Le cadre statistique donne un sens précis à la matrice $\boldsymbol{R}$. A l'inverse, l'analyse des dépendances entre les variables fait plutôt intervenir la matrice $\boldsymbol{R}^{-1}$. Mais à quel prix faut-il chercher une modélisation flexible permettant de spécifier aussi bien les opérateurs de corrélation que leurs inverses ? Cette section expose les raisons algorithmiques qui motivent le développement d'une méthode robuste permettant d'accéder aussi bien à $\boldsymbol{R}$ qu'à $\boldsymbol{R}^{-1}$. Les multiples facettes de cette approche sont mises à profit dans toute l'étude pour fournir des interprétations aux expériences, des procédés de validation et des tests unitaires.

1.3.1 Formulations primale et duale

A chaque itération de la boucle externe, on résout le problème quadratique par une méthode de Krylov (voir chapitre 6). Ainsi, pour trouver le minimum de $\mathcal{J}(\boldsymbol{s})$, on cherche à annuler son gradient. Après omission des exposants, ce dernier s'écrit

$$\nabla\mathcal{J}(\boldsymbol{s}) = \boldsymbol{B}^{-1}\boldsymbol{s} + \boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{G}\boldsymbol{s} - \boldsymbol{d}). \tag{1.70}$$

En faisant l'hypothèse que $(\boldsymbol{B}^{-1} + \boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{G})$ est inversible, ce qui est vérifié en pratique, la condition $\nabla\mathcal{J}(\boldsymbol{s}) = 0$ conduit à

$$\boldsymbol{s} = (\boldsymbol{B}^{-1} + \boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{d}. \tag{1.71}$$

Comme la taille du problème est très grande, et puisque les matrices sont pour la plupart disponibles seulement sous forme d'opérateurs, $\boldsymbol{s}$ est calculé en résolvant le système linéaire

$$(\boldsymbol{B}^{-1} + \boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{G})\boldsymbol{s} = \boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{d} \tag{1.72}$$

à l'aide d'une méthode de gradient conjugué préconditionné par $\boldsymbol{B}$ (PCG) [Courtier, 1997, Derber and Bouttier, 1999], ce qui donne

$$\begin{aligned}&(\boldsymbol{I} + \boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{G}\boldsymbol{B})\hat{\boldsymbol{s}} = \boldsymbol{G}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{d}\\ &\boldsymbol{s} = \boldsymbol{B}\hat{\boldsymbol{s}}.\end{aligned}$$

D'autres approches dites « duales » sont possibles [Egbert et al., 1994, Da Silva et al., 1995, Courtier, 1997, Cohn et al., 1998, Daley and Barker,

2001, Bennett, 2002]. La formule de Sherman-Morrison-Woodbury permet d'obtenir $\boldsymbol{s}$ comme solution du système linéaire

$$(\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}} + \boldsymbol{R})\boldsymbol{\mu} = \boldsymbol{d} \quad ; \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{\mu}. \tag{1.73}$$

Cette équation est alors résolue par une méthode de gradient conjugué. L'avantage est que la taille du problème est réduite, puisque $\boldsymbol{\mu}$ évolue dans l'espace des observations, qui est en général de dimension plus petite que l'espace d'état de $\boldsymbol{x}$. En préconditionnant par $\boldsymbol{R}$, on obtient

$$(\boldsymbol{I} + \boldsymbol{R}^{-1}\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}})\boldsymbol{\mu} = \boldsymbol{R}^{-1}\boldsymbol{d}. \tag{1.74}$$

Les différentes méthodes se distinguent alors par le choix du produit scalaire :

- la méthode PSAS (*Physical-space Statistical Analysis System*) travaille avec le produit scalaire associé à $\boldsymbol{R}$ [Courtier, 1997] ;
- la méthode RPCG (*Restricted Preconditionned Conjugate Gradient*) travaille avec le produit scalaire associé à $\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}$ [Gratton and Tshimanga, 2009].

Bien que ces deux méthodes convergent vers la même solution, elles ne le font pas de la même façon. Avec les préconditionneurs donnés précédemment, la méthode du PCG et RPCG sont strictement équivalentes, puisqu'elles produisent les mêmes itérés [El Akkraoui et al., 2008, El Akkraoui and Gauthier, 2010]. Elles s'interprètent statistiquement comme un maximum de vraisemblance, à l'inverse de la méthode PSAS qui est purement algébrique. La méthode PSAS ne produisant pas les mêmes itérés, il est donc dangereux de l'arrêter avant d'avoir atteint la convergence.

1.3.2 Formulation point-selle

Deux nouvelles variables sont introduites comme suit :

$$\delta\boldsymbol{p} = \boldsymbol{s} \tag{1.75}$$
$$\delta\boldsymbol{w} = \boldsymbol{G}\boldsymbol{s}. \tag{1.76}$$

Théoriquement, la variable $\delta\boldsymbol{p}$ est différente de $\boldsymbol{s}$ quand on prend en compte l'erreur du modèle, ce qui n'est pas fait ici [Gratton et al., 2017]. C'est pourquoi on s'autorise à séparer $\delta\boldsymbol{p}$ de $\boldsymbol{s}$. Avec ces nouvelles notations, la fonction-coût du problème d'optimisation s'écrit

$$\mathcal{J}(\delta\boldsymbol{p}, \delta\boldsymbol{w}) = \frac{1}{2}\|\delta\boldsymbol{p}\|^2_{\boldsymbol{B}^{-1}} + \frac{1}{2}\|\delta\boldsymbol{w} - \boldsymbol{d}\|^2_{\boldsymbol{R}^{-1}} \tag{1.77}$$

et la solution au problème de minimisation se calcule en en annulant Lagrangien :

$$\nabla \mathcal{L}(\delta \boldsymbol{p}, \delta \boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{0}, \tag{1.78}$$

où

$$\mathcal{L}(\delta \boldsymbol{p}, \delta \boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathcal{J}(\delta \boldsymbol{p}, \delta \boldsymbol{w}) + \boldsymbol{\lambda}^{\mathrm{T}}(\delta \boldsymbol{p} - \boldsymbol{s}) + \boldsymbol{\mu}^{\mathrm{T}}(\delta \boldsymbol{w} - \boldsymbol{G}\boldsymbol{s}). \tag{1.79}$$

On se retrouve ainsi en présence du système d'équations :

$$\begin{array}{rcccl} \frac{\partial \mathcal{L}}{\partial \delta \boldsymbol{p}} & = & \boldsymbol{B}^{-1}\delta \boldsymbol{p} + \boldsymbol{\lambda} & = & \boldsymbol{0} \\ \frac{\partial \mathcal{L}}{\partial \delta \boldsymbol{w}} & = & \boldsymbol{R}^{-1}(\delta \boldsymbol{w} - \boldsymbol{d}) + \boldsymbol{\mu} & = & \boldsymbol{0} \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{s}} & = & -\boldsymbol{\lambda} - \boldsymbol{G}^{\mathrm{T}}\boldsymbol{\mu} & = & \boldsymbol{0} \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} & = & \delta \boldsymbol{p} - \boldsymbol{s} & = & \boldsymbol{0} \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} & = & \delta \boldsymbol{w} - \boldsymbol{G}\boldsymbol{s} & = & \boldsymbol{0} \end{array}, \tag{1.80}$$

qui conduit à l'équation matricielle

$$\begin{pmatrix} \boldsymbol{B} & \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{R} & \boldsymbol{G} \\ \boldsymbol{I} & \boldsymbol{G}^{\mathrm{T}} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\mu} \\ \boldsymbol{s} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{d} \\ \boldsymbol{0} \end{pmatrix}, \tag{1.81}$$

laquelle ne fait plus intervenir explicitement l'opérateur $\boldsymbol{R}^{-1}$.

Lors de la résolution de (1.81), les produits matrice-vecteurs par $\boldsymbol{G}$ et $\boldsymbol{G}^{\mathrm{T}}$ peuvent être traités en parallèle. De plus, dans le cas du 4DVAR à contrainte faible, le modèle $\boldsymbol{M}$ et son adjoint $\boldsymbol{M}^{\mathrm{T}}$, $\boldsymbol{G}$ et $\boldsymbol{G}^{\mathrm{T}}$ peuvent être parallélisés pour chaque sous-fenêtre d'assimilation. Cela permet d'introduire une forme de parallélisme en temps lors de la résolution [Fisher and Gürol, 2017].

La formulation à partir du lagrangien permet de donner un sens à la variable $\boldsymbol{\mu}$ de l'algorithme PSAS en l'interprétant comme l'un des multiplicateurs de Lagrange du système point-selle. En effet, en travaillant par substitutions sur le système (1.80), on obtient :

$$(\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}} + \boldsymbol{R})\boldsymbol{\mu} = \boldsymbol{d} \quad \text{et} \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{\mu}. \tag{1.82}$$

Notons

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{R} & \boldsymbol{G} \\ \boldsymbol{I} & \boldsymbol{G}^{\mathrm{T}} & \boldsymbol{0} \end{pmatrix} \tag{1.83}$$

la matrice du système point-selle. Cette matrice est définie comme un opérateur, ce qui en plus de la taille du problème, impose de résoudre le système par une méthode itérative. Cependant, on ne peut pas utiliser la méthode du gradient conjugué puisque $\boldsymbol{A}$ n'est pas définie positive. Il faut donc recourir à une méthode plus générale, comme celle du GMRES décrite dans Saad [2003].

Enfin, un préconditionneur possible pour le système point-selle se trouve en simplifiant la matrice $\mathbf{A}$ en enlevant les termes en $\boldsymbol{G}$ et $\boldsymbol{G}^{\mathrm{T}}$ [Golub and Van Loan, 1996]. Ainsi, on trouve :

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{B} & \mathbf{0} & \boldsymbol{I} \\ \mathbf{0} & \boldsymbol{R} & \mathbf{0} \\ \boldsymbol{I} & \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{et} \quad \boldsymbol{P}^{-1} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \boldsymbol{I} \\ \mathbf{0} & \boldsymbol{R}^{-1} & \mathbf{0} \\ \boldsymbol{I} & \mathbf{0} & -\boldsymbol{B} \end{pmatrix}. \tag{1.84}$$

On remarque que le préconditionnement nécessite de spécifier $\boldsymbol{R}^{-1}$. Toutefois, pour accélérer la convergence, on peut se contenter d'une approximation de $\boldsymbol{R}^{-1}$.

1.4 L'essentiel du chapitre

Les espaces de distributions forment un cadre naturel pour définir les opérateurs de covariance. La notion d'espaces primal et dual permet d'associer des objets mathématiques aux champs physiques lisses ou bruités. Ce cadre diffère légèrement de celui des espaces de Hilbert, habituellement utilisé dans le traitement des équations aux dérivées partielles. L'introduction du triplet de Gelfand pour faire le lien entre espaces de distributions, fonctions de carrés intégrable et opérateurs de covariance est nouveau en météorologie. La représentation des opérateurs et de leurs espaces de départ et d'arrivée dans un diagramme de dualité est une nouveauté qui permet de synthétiser l'ensemble des définitions mathématiques sous une forme graphique et compacte. L'intérêt pratique de cet outil est d'autant plus grand qu'il est facile de le modifier pour intégrer de nouveaux développements ou de nouvelles définitions. C'est le cas dans la partie III ou l'introduction de nouveaux espaces de discrétisation se traduit par l'addition de nouveaux espaces et de nouveaux liens dans le diagramme de dualité.

Ce qu'il faut retenir :

- Le triplet de Gelfand $H \subset L^2 \subset H^\star$ est la structure adaptée à la modélisation des fonction de corrélation dans des espaces de Hilbert.
- L'opérateur de corrélation est un opérateur « du primal vers le dual ».
- Le modèle de Matérn dépend de deux paramètres. Le paramètre m contrôle la régularité de $c_{m,l}$ et l contrôle sa longueur de portée.
- Au lieu d'effectuer le produit de convolution par une fonction de Matérn, on peut résoudre une équation de diffusion implicite.
- Les étapes pour appliquer l'opérateur de corrélation déduit de l'équation de diffusion implicite se résument sur un diagramme de dualité.
- En interprétant le passage du continu au discret comme une projection sur un espace de dimension finie, on arrive à faire apparaître les matrices de corrélation de façon naturelle à partir de la formulation continue de l'asimialation de données.
- Dans la pratique, il n'y a pas besoin de modéliser $\boldsymbol{B}^{-1}$. En revanche, on souhaite mettre au point un modèle donnant accès à la fois à $\boldsymbol{R}$, $\boldsymbol{R}^{-1}$ et $\boldsymbol{R}^{1/2}$.

Chapitre 2

Traitement de la diffusion par la méthode des éléments finis

L'équation de diffusion est un outil permettant de modéliser l'inverse d'un opérateur de corrélation. La discrétisation de cette équation de diffusion dépend de la nature discrète des champs auxquels on souhaite appliquer l'opérateur de corrélation. En effet, les champs discrets manipulés en assimilation de données sont supposés représenter une réalité continue. Cette réalité continue étant inconnue, on choisit d'associer à chaque valeur de champ ponctuelle une extension spatiale spécifiée par une fonction de pondération. Ce choix mène à résoudre les équations continues dans des espaces fonctionnels de dimension finie, ce qui aboutit naturellement à la résolution d'un système linéaire. Les propriétés mathématiques de ce système linéaire dépendent directement du type de fonctions de pondération utilisées pour la discrétisation.

Les méthodes spectrales et les ondelettes ont recours à des familles de fonctions orthonormales [Courtier et al., 1998, Parrish and Derber, 1992]. A chaque valeur de champ est associée une fonction de pondération dont le support non compact recouvre tout le domaine d'étude. Ainsi, l'espace des coordonnées, dit « espace spectral » est bien séparé de l'espace des champs météorologiques, parfois appelé « espace physique ». Cela signifie qu'une formule de reconstruction est nécessaire pour reconstituer un champ à partir de ses coordonnées. Le caractère orthonormal des fonctions de pondérations assure quant à lui que la résolution de l'équation dans l'espace spectral est facile. En particulier, les modèles de corrélation admettent une représentation diagonale dans l'espace spectral [Deckmyn and Berre, 2005, Bannister, 2008b, Pannekoucke et al., 2007].

Les méthodes en point de grille diffèrent des méthodes spectrales par leur choix de fonctions de pondération. En effet, elles favorisent l'utilisation de fonctions à support compact [Canuto et al., 1987, Ern and Guermond, 2010] pour résoudre les équations différentielles, donnant lieu à des systèmes linéaires creux. C'est le cas de la méthode des éléments finis, présentée dans la suite de ce chapitre. Le coût de résolution du système linéaire est compensé par la facilité de reconstruction des champs à partir des coordonnées. En effet, la plupart des méthodes assurent que les coordonnées dans la base choisie sont égales aux valeurs des champs physiques aux points de grille. Les méthodes en point de grille constituent un choix naturel pour la résolution d'équations aux dérivées partielles sur des grilles non structurées, ce qui justifie notre choix pour cette étude.

La section 2.1 présente deux généralisations de l'équation de diffusion. Dans la première, le coefficient de diffusion dépendant de l'espace permet de modéliser des fonctions de corrélation anisotropes. Cette forme de la diffusion anisotrope sera utilisée tout au long du chapitre par soucis de généralisation. La seconde généralisation revient sur le développement de l'opérateur de diffusion comme une somme de puissances du Laplacien. Elle sera utile pour faire le lien avec la méthode de Brankart et al. [2009]. La section 2.2 présente la formulation dite « faible » de l'équation de diffusion. Elle constitue la base des méthodes de Galerkin, dont fait partie la méthode des éléments finis. La méthode des éléments finis fait l'objet de la section 2.3. La section 2.4 donne des aspects de convergence de cette méthode de discrétisation, en introduisant la notion de sensibilité au maillage. Cette notion sera approfondie sur des cas pratiques dans les chapitres suivants. Enfin, un lien formel est tissé entre l'équation de diffusion et la méthode de Brankart et al. [2009] dans la section 2.5, qui consiste à assimiler le champ d'observation et ses dérivées successives.

Remarque : dans la littérature, le traitement de l'équation de diffusion par la méthode des éléments finis est un grand classique. Ce chapitre retranscrit les grandes lignes de cette méthodologie en mettant l'accent sur la cohérence avec la théorie introduite dans le chapitre 1. Quelques digressions font le lien avec d'autres méthodes existantes. Notamment, le lien avec l'assimilation des dérivées du champ d'observation est une nouveauté, appréciable aux yeux de la communauté de l'assimilation de données.

2.1 Généralisations de l'équation de diffusion homogène

L'équation de diffusion présentée dans la section 1.1.4 est homogène et isotrope. Sa paramétrisation dépend de deux paramètres, m et l, qui contrôlent la régularité et la longueur de portée des fonctions de corrélation. Il est néanmoins possible de modifier l'équation de diffusion lorsqu'on souhaite modéliser un opérateur de corrélation dont l'action fluctue dans l'espace, ou lorsqu'on souhaite utiliser des fonctions de corrélation possédant des lobes négatifs.

On considère que l'étude porte sur le domaine $\mathcal{D}$, qui est supposé non-vide, ouvert, connexe et borné dans $\mathbb{R}^2$. Son adhérence est notée $\bar{\mathcal{D}}$.

2.1.1 Diffusion hétérogène

Le paramètre l^2 de l'équation de diffusion homogène est relié à la longueur de portée, notée D au travers de la relation (1.24). Cette longueur de portée représente la largeur à mi-hauteur de la parabole osculatrice de la fonction de corrélation (voir figure 1.4). Autrement dit, la longueur de portée règle la largeur de la fonction de corrélation en chaque point du domaine et, ce faisant, l'intensité avec laquelle les mesures en deux points distincts du domaine d'étude sont corrélées. Utiliser un paramètre de portée l homogène en espace revient à considérer que la corrélation entre deux points ne dépend que de la distance qui les sépare. Quand cette propriété devient fausse, il convient d'introduire une dépendance du paramètre l à la variable d'espace $\boldsymbol{z}$. L'opérateur de diffusion sur un pas de temps s'écrit alors

$$1 - \nabla \cdot [l(\boldsymbol{z})^2 \nabla] \overset{\text{notation}}{=} 1 - \nabla \cdot l(\boldsymbol{z})^2 \nabla. \tag{2.1}$$

On a en chaque point $\boldsymbol{z} \in \mathcal{D}$ et pour toute fonction $f \in H^{2m}$:

$$\mathcal{C}^{-1}(f)(\boldsymbol{z}) = \gamma(\boldsymbol{z})^{-1/2}[1 - \nabla \cdot l(\boldsymbol{z})^2 \nabla]^m \gamma(\boldsymbol{z})^{-1/2} f(\boldsymbol{z}) \tag{2.2}$$

et

$$\mathcal{C}(f)(\boldsymbol{z}) = \gamma(\boldsymbol{z})^{1/2}[1 - \nabla \cdot l(\boldsymbol{z})^2 \nabla]^{-m} \gamma(\boldsymbol{z})^{1/2} f(\boldsymbol{z}). \tag{2.3}$$

Dans ce cas-là, on parle de diffusion hétérogène (figure 2.1). Les facteurs $\gamma(\boldsymbol{z})^{1/2}$ de l'équation (2.3) servent à normaliser $\mathcal{C}$ de telle sorte que son noyau c soit une fonction de corrélation (*i.e.* d'amplitude unitaire, $c(\boldsymbol{z}, \boldsymbol{z}) = 1$). En revanche, contrairement au cas homogène, on ne dispose pas de formule

exacte pour le calcul de $\gamma(\boldsymbol{z})$. Pour son calcul, la valeur de la fonction de corrélation $c(\boldsymbol{z}, \boldsymbol{z}')$ associée à l'opérateur (2.3) est approximée par

$$c(\boldsymbol{z}, \boldsymbol{z}') \simeq c_{m,\bar{l}}(\boldsymbol{z}, \boldsymbol{z}'), \tag{2.4}$$

où $c_{m,\bar{l}}$ est la fonction de Matérn de paramètres m et $\bar{l}$ et

$$\bar{l}(\boldsymbol{z}, \boldsymbol{z}')^2 = \frac{l(\boldsymbol{z})^2 + l(\boldsymbol{z}')^2}{2} \tag{2.5}$$

est la moyenne arithmétique des longueurs de portée en $\boldsymbol{z}$ et $\boldsymbol{z}'$ [Paciorek and Schervish, 2006]. On mentionne également la moyenne géométrique $\bar{l}(\boldsymbol{z}, \boldsymbol{z}')^2 = l(\boldsymbol{z})l(\boldsymbol{z}')$ utilisée dans Cummings [2005], mais cette dernière ne garantit pas le caractère positif défini de (2.3). Une discussion approfondie est disponible dans Mirouze and Weaver [2010].

De même, l'anisotropie se modélise en introduisant un tenseur de corrélation dans l'équation de diffusion, noté $\boldsymbol{\kappa}(\boldsymbol{z})$ et homogène à $l(\boldsymbol{z})^2$. On a ainsi

$$\mathcal{C}^{-1}(f)(\boldsymbol{z}) = \gamma(\boldsymbol{z})^{-1/2}[1 - \nabla \cdot \boldsymbol{\kappa}(\boldsymbol{z})\nabla]^m \gamma(\boldsymbol{z})^{-1/2} f(\boldsymbol{z}). \tag{2.6}$$

et

$$\mathcal{C}(f)(\boldsymbol{z}) = \gamma(\boldsymbol{z})^{1/2}[1 - \nabla \cdot \boldsymbol{\kappa}(\boldsymbol{z})\nabla]^{-m} \gamma(\boldsymbol{z})^{1/2} f(\boldsymbol{z}). \tag{2.7}$$

En dimension d, le tenseur $\boldsymbol{\kappa}$ est une matrice symétrique ($\boldsymbol{\kappa}_{\mathrm{ij}} = \boldsymbol{\kappa}_{\mathrm{ji}}$), définie positive ($\boldsymbol{\kappa}_{\mathrm{ii}}\boldsymbol{\kappa}_{\mathrm{jj}} > \boldsymbol{\kappa}_{\mathrm{ij}}^2$) de dimension $d \times d$, condition sans laquelle l'opérateur $\nabla \cdot \boldsymbol{\kappa}(\boldsymbol{z})\nabla$ ne serait lui-même plus symétrique [Weaver and Mirouze, 2013]. On a cette fois :

$$c(\boldsymbol{z}, \boldsymbol{z}') \simeq \frac{2^{1-m+d/2}}{\Gamma(m - d/2)} (\|\boldsymbol{z} - \boldsymbol{z}'\|_{\boldsymbol{\kappa}^{-1}})^{m-d/2} K_{m-d/2}(\|\boldsymbol{z} - \boldsymbol{z}'\|_{\boldsymbol{\kappa}^{-1}}), \tag{2.8}$$

où

$$\|\boldsymbol{z} - \boldsymbol{z}'\|_{\boldsymbol{\kappa}^{-1}} = \sqrt{(\boldsymbol{z} - \boldsymbol{z}')^{\mathrm{T}} \left(\frac{\boldsymbol{\kappa}(\boldsymbol{z}) + \boldsymbol{\kappa}(\boldsymbol{z}')}{2} \right)^{-1} (\boldsymbol{z} - \boldsymbol{z}')}. \tag{2.9}$$

Le terme $-\boldsymbol{\kappa}(\boldsymbol{z})\nabla f$ peut s'interpréter comme un « flux d'information », l'information étant ici représentée par la fonction f. En physique, $f(\boldsymbol{z})$ représenterait la température en un point $\boldsymbol{z}$ du domaine, et $-\boldsymbol{\kappa}(\boldsymbol{z})\nabla f(\boldsymbol{z})$ serait un flux de chaleur passant chaque seconde à travers une section du domaine au voisinage de $\boldsymbol{z}$. Considérer la divergence de ce flux, $\nabla \cdot \boldsymbol{\kappa}(\boldsymbol{z})\nabla f(\boldsymbol{z})$, permet de représenter la propagation de l'information/la chaleur, à partir du point $\boldsymbol{z}$ et vers son voisinage. D'où le nom d'équation de « diffusion ». Cette

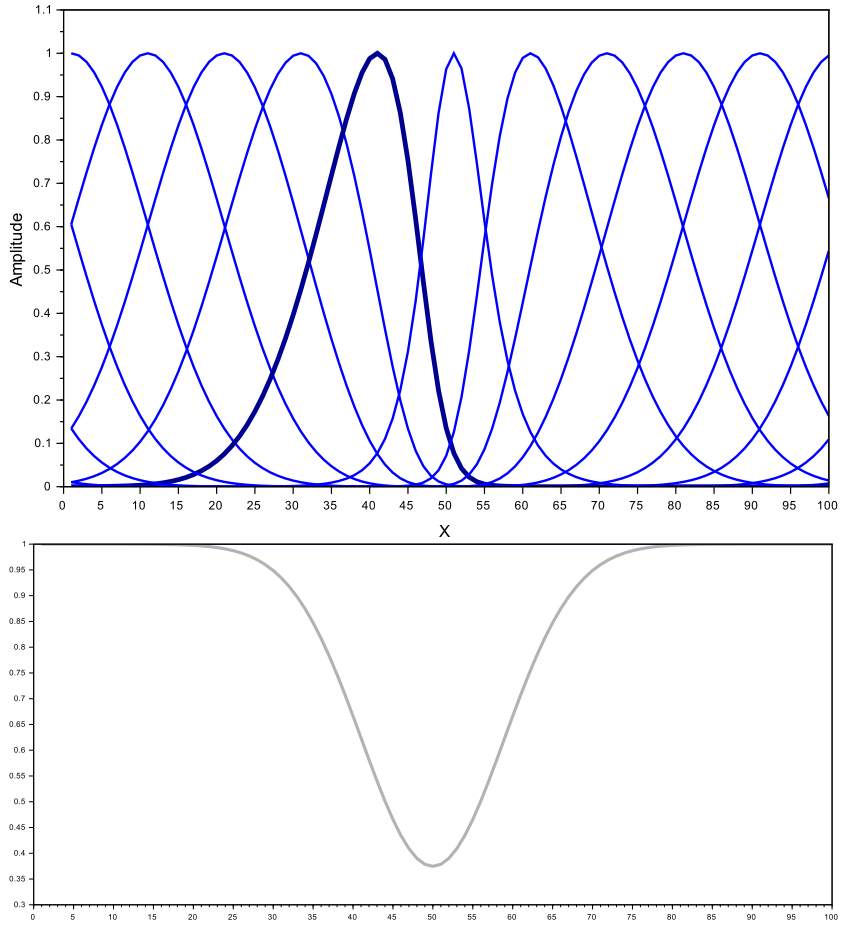


FIGURE 2.1 – En haut : fonctions de corrélation anisotropes. On a mis en évidence une fonction en trait gras à titre d'exemple. En bas : perturbation du champ de portée.

analogie donne lieu à une interprétation géométrique du tenseur de diffusion. Lorsque $\boldsymbol{\kappa}(\boldsymbol{z})$ admet deux valeurs propres égales à λ, alors il est égal à

$$\boldsymbol{\kappa}(\boldsymbol{z}) = \lambda \boldsymbol{I} \tag{2.10}$$

et on retrouve le cas de la diffusion homogène et isotrope. En 2D, on représente $\boldsymbol{\kappa}$ par un cercle de rayon λ. Lorsque les valeurs propres $(\lambda_i)_{i\in[\![1,d]\!]}$ de $\boldsymbol{\kappa}$ sont distinctes, il existe toujours une base de vecteurs propres $(\boldsymbol{v}_i)_{i\in[\![1,d]\!]}$ dans laquelle $\boldsymbol{\kappa}$ est diagonal. On a alors

$$\boldsymbol{\kappa}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i. \tag{2.11}$$

On représente alors $\boldsymbol{\kappa}$ par une ellipse dont les directions des axes principaux

sont données par les vecteurs propres $(\boldsymbol{v}_i)_{i\in[\![1,d]\!]}$ et l'étirement dans chaque direction est proportionnel à $(\lambda_i)_{i\in[\![1,d]\!]}$ (figure 2.2).

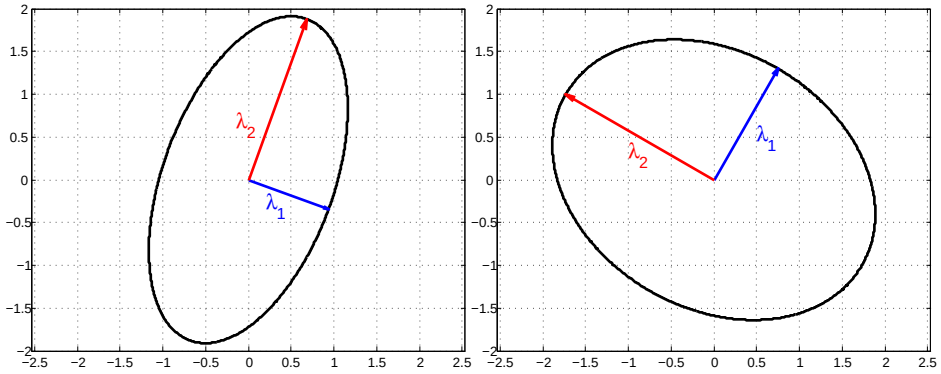


FIGURE 2.2 – Représentation de deux tenseurs de corrélation en dimension 2 par leurs axes principaux.

Dans la suite, on se réfèrera le plus souvent possible à la formulation (2.7) pour dériver les formules des éléments finis dans le cadre général de la diffusion hétérogène et anisotrope. En revanche, le cas homogène et isotrope servira de référence pour la validation de la modélisation des opérateurs de corrélations sur des maillages non structurés obtenus à partir de données réelles.

2.1.2 Diffusion généralisée

Il existe multiples façons de généraliser l'équation de diffusion qui mériteraient tout aussi bien le nom de diffusion « généralisée ». Dans cette sous-section, on présente une version homogène de l'équation de diffusion qui permet de modéliser des fonctions de corrélations avec des lobes négatifs.

On remarque tout d'abord que l'opérateur $[\mathbf{1} - l^2\Delta]^m$ peut se développer en une somme alternée de puissances du Laplacien :

$$\begin{aligned}[\mathbf{1} - l^2\Delta]^m &= \sum_{k=0}^{m} \binom{m}{k} (-1)^k l^{2k} \Delta^k \\ &= \mathbf{1} - ml^2\Delta + \ldots + (-1)^m \Delta^m \end{aligned} \tag{2.12}$$

où la notation $\binom{m}{k}$ désigne le coefficient binomial. En posant

$$a_k = \binom{m}{k} l^{2k}, \tag{2.13}$$

on a

$$[\mathbf{1} - l^2\Delta]^m = \sum_{k=0}^{m} (-1)^k a_k \Delta^k. \tag{2.14}$$

Pour généraliser l'équation (2.14), il suffit de considérer tout opérateur pouvant s'exprimer comme une somme alternée de puissances du Laplacien, où les coefficients $(a_k)_{k\in[\![0,m]\!]}$ désignent des nombres réels positifs [Weaver and Courtier, 2001]. On obtient alors un opérateur symétrique (grâce aux propriétés du Laplacien) et défini positif (grâce à l'opérateur identité et à la positivité des coefficients) de la forme

$$\sum_{k=0}^{m} (-1)^k a_k \Delta^k, \tag{2.15}$$

où $a_k \geqslant 0$ pour tout $k \in [\![1, m]\!]$.

Bien qu'on n'utilise pas cette formulation générale dans la suite du manuscrit, elle nous permettra de faire le lien entre l'équation de diffusion et la méthode de Brankart et al. [2009] dans la section 2.5.

2.2 Formulation faible de l'équation de diffusion

Le prototype le plus simple de l'équation de diffusion est l'équation au Laplacien $-\Delta u = f$, où $f \in L^2(\mathcal{D})$. Sa solution u appartient à l'espace $C^2(\mathcal{D})$ des fonctions deux fois dérivables dont les dérivées sont continues. De manière analogue, dans le cas des équations de diffusion plus générales faisant intervenir des puissances du Laplacien d'ordre élevé, les solutions sont recherchées dans un espace « C^{CIF} » des fonctions « comme il faut », c'est-à-dire continues, suffisamment dérivables et aux dérivées successives elles-mêmes continues. Cette condition sur la régularité des solutions est forte et ne correspond pas toujours à la réalité physique des solutions. L'approche variationnelle a pour idée d'affaiblir ces conditions de régularité dans la recherche des solutions en employant une formulation intégrale de l'équation aux dérivées partielles.

2.2.1 Formulation faible

On présente dans cette sous-section la formulation variationnelle, dite « formulation faible » de l'équation de diffusion. Par souci de généricité, on décide de traiter le cas général de la diffusion anisotrope.

Ainsi, considérons l'équation sur un pas de temps, de condition initiale $f_n \in V^\star$ et d'inconnue $f_{n+1} \in V$:

$$(\mathbf{1} - \nabla \cdot \boldsymbol{\kappa}(\boldsymbol{z})\nabla) f_{n+1} = f_n, \tag{2.16}$$

où $\boldsymbol{\kappa}(\boldsymbol{z})$ est un tenseur de diffusion variable dans l'espace, comme défini en section 2.1.1 et où $n \in [\![0, m]\!]$ est maintenant un indice temporel. Ici, V désigne un espace de Sobolev de type H^1, comme présenté en section 1.1. L'approche variationnelle [Lax and Milgram, 1954, Babuska, 1971] consiste à comprendre l'équation (2.16) au sens des distributions et rechercher la solution $f_{n+1} \in V$ vérifiant pour toute fonction de pondération $\varphi \in V$:

$$\langle (\mathbf{1} - \nabla \cdot \boldsymbol{\kappa}(\boldsymbol{z})\nabla) f_{n+1}, \varphi \rangle_{V^\star, V} = \langle f_n, \varphi \rangle_{V^\star, V}. \tag{2.17}$$

La discrétisation de (2.17) se fait alors en plusieurs étapes [Ciarlet, 2002, Brenner and Scott, 2013].

Tout d'abord, on restreint la recherche de f_{n+1} à un espace de dimension finie $V_c \subset V$, tel que :

$$\dim(V_c) = p < +\infty. \tag{2.18}$$

Soit $(\varphi_i)_{i \in [\![1,p]\!]}$ une base de V_c. Les $(\varphi_i)_{i \in [\![1,p]\!]}$ sont appelées « fonctions de forme ». La projection de f_{n+1} sur cette base s'écrit :

$$f_{n+1} = \sum_{i \in [\![1,p]\!]} \alpha^i_{n+1} \varphi_i, \tag{2.19}$$

où les $(\alpha^i_{n+1})_{i \in [\![1,p]\!]}$ désignent les coordonnées de f_{n+1} dans la base $(\varphi_i)_{i \in [\![1,p]\!]}$.

On impose ensuite la relation (2.17) pour toute fonction de pondération dans V_c. On parle dans ce cas d'approximation conforme, en opposition au cas où l'espace des fonctions de pondération est différent de celui des fonctions de forme [Crouzeix and Raviart, 1973]. Dans l'approximation de Galerkin, on impose de plus que les fonctions de forme et de pondération soient les mêmes. On cherche donc les coordonnées $(\alpha^i_{n+1})_{i \in [\![1,p]\!]}$ vérifiant

$$\sum_{i \in [\![1,p]\!]} \alpha^i_{n+1} \langle (\mathbf{1} - \nabla \cdot \boldsymbol{\kappa}(\boldsymbol{z})\nabla) \varphi_i, \varphi_j \rangle_{V^c, V_c} = \langle f_n, \varphi_j \rangle_{V^c, V_c}, \tag{2.20}$$

cela $\forall j \in [\![1,p]\!]$ et où $V^c = (V_c)^\star$ désigne l'espace dual de V_c.

Soit $\mathbf{1}$ l'injection canonique de V_c dans V^c et $(\varphi^i)_{i\in[\![1,p]\!]}$ une base de V^c (noter l'indice haut pour les éléments du dual) telle que $\forall i \in [\![1,p]\!]$, $\varphi^i = \mathbf{1}(\varphi_i)$. On projette la condition initiale f_n sur la base de V^c pour obtenir

$$f_n \simeq \sum_{i\in[\![1,p]\!]} \alpha_i^n \varphi^i, \tag{2.21}$$

où les $(\alpha_i^n)_{i\in[\![1,p]\!]}$ désignent les coordonnées de f_n dans la base $(\varphi_i)_{i\in[\![1,p]\!]}$. On a alors

$$\sum_{i\in[\![1,p]\!]} \alpha_{n+1}^i \langle (\mathbf{1} - \nabla\cdot\boldsymbol{\kappa}(\boldsymbol{z})\nabla)\varphi_i, \varphi_j\rangle_{V^c,V_c} = \sum_{i\in[\![1,p]\!]} \alpha_i^n \langle \varphi^i, \varphi_j\rangle_{V^c,V_c}. \tag{2.22}$$

Comme, par définition de l'injection, $\varphi^i = \mathbf{1}(\varphi_i)$, on peut transformer le deuxième membre de l'équation (2.22) en écrivant

$$\langle \varphi^i, \varphi_j\rangle_{V^c,V_c} = \langle \mathbf{1}(\varphi_i), \varphi_j\rangle_{V^c,V_c}. \tag{2.23}$$

On introduit une nouvelle fois l'espace $L^2(\mathcal{D})$ des fonctions de carré intégrable sur $\mathcal{D}$ et on définit le crochet de dualité $\langle\cdot,\cdot\rangle_{V^c,V_c}$ à partir du produit scalaire dans L^2. Pour ce faire, on utilise le fait que $V_c \subset L^2$ et donc l'ensemble des formes linéaires de L^2 agissant sur V_c est lui-même inclus dans V^c : $(L^2)^\star_{|V_c} = L^2_{|V_c} \subset V^c$. On est ainsi en présence du triplet (qui n'est pas un triplet de Gelfand, puisque la première inclusion n'est pas dense) :

$$V_c \subset L^2_{|V_c} \subset V^c. \tag{2.24}$$

Dès lors, l'équation (2.22) se réécrit uniquement avec des produits scalaires dans L^2. En utilisant la formule de Green sur le terme de divergence et en changeant la notation α_i^n pour α_n^i, on obtient :

$$\sum_{i\in[\![1,p]\!]} \alpha_{n+1}^i (\langle \varphi_i, \varphi_j\rangle_{L^2} - \langle \boldsymbol{\kappa}(\boldsymbol{z})\nabla\varphi_i, \nabla\varphi_j\rangle_{L^2}) = \sum_{i\in[\![1,p]\!]} \alpha_n^i \langle \varphi_i, \varphi_j\rangle_{L^2}, \tag{2.25}$$

cela $\forall j \in [\![1,p]\!]$.

L'équation (2.25) correspond à la formulation de Galerkin traditionnelle, n'employant pas le crochet de dualité. La formulation matricielle de cette

équation est définie en posant $\boldsymbol{M}_c = ((\boldsymbol{M}_c)_{\mathrm{ij}})$ la matrice de masse et $\boldsymbol{K}_c = ((\boldsymbol{K}_c)_{\mathrm{ij}})$ la matrice de raideur de termes généraux

$$\begin{aligned}(\boldsymbol{M}_c)_{\mathrm{ij}} &= \langle \varphi_i, \varphi_j \rangle_{L^2} \\ &= \int_{\mathcal{D}} \varphi_i(\boldsymbol{z})\varphi_j(\boldsymbol{z})\mathrm{d}\boldsymbol{z}\end{aligned} \tag{2.26}$$

et

$$\begin{aligned}(\boldsymbol{K}_c)_{\mathrm{ij}} &= \langle \boldsymbol{\kappa}(\boldsymbol{z})\nabla\varphi_i, \nabla\varphi_j \rangle_{L^2} \\ &= \int_{\mathcal{D}} \boldsymbol{\kappa}(\boldsymbol{z})\nabla\varphi_i(\boldsymbol{z}) \cdot \nabla\varphi_j(\boldsymbol{z})\mathrm{d}\boldsymbol{z}.\end{aligned} \tag{2.27}$$

Soit $\boldsymbol{\alpha}_n$ le vecteur colonne des (α_n^i) et $\boldsymbol{\alpha}_{n+1}$ le vecteur colonne des (α_{n+1}^i). La formule (2.25) est équivalente au système linéaire d'inconnue $\boldsymbol{\alpha}_{n+1}$

$$(\boldsymbol{M}_c + \boldsymbol{K}_c)\boldsymbol{\alpha}_{n+1} = \boldsymbol{M}_c\boldsymbol{\alpha}_n. \tag{2.28}$$

A noter que la présence de la matrice de masse résulte de l'existence de l'opérateur $\mathbf{1}$ dans la formulation continue de la diffusion, alors que la matrice de raideur $\boldsymbol{K}_c$ provient de l'opérateur Laplacien généralisé (ou, dans le cas présent, de l'opérateur symétrique « divergence-kappa-gradient »).

2.2.2 Cyclage et propriétés mathématiques

Le systéme linéaire (2.28) correspond à un pas de temps de l'équation de diffusion. En faisant une analogie avec le diagramme de dualité de la figure 1.6, on voit qu'il est nécessaire d'exprimer des fonctions Φ_c et Φ_c^{T} qui permettent de revenir de l'espace d'arrivée à l'espace de départ, afin de composer plusieurs itérations de l'opérateur de diffusion discret.

Soit Φ_c l'application de $\mathbb{R}^p$ dans V_c permettant de recomposer toute fonction de V_c à partir de ses coordonnées :

$$\begin{aligned}\Phi_c \quad &: \quad \mathbb{R}^p \to V_c \\ &\quad \boldsymbol{\alpha} = (\alpha^i)_{i\in[\![1,p]\!]} \mapsto f = \sum_{i\in[\![1,p]\!]} \alpha^i\varphi_i.\end{aligned} \tag{2.29}$$

L'application Φ_c^{T} est l'application du dual qui associe à tout $f \in V^c$ le vecteur de ses coordonnées :

$$\begin{aligned}\Phi_c^{\mathrm{T}} \quad &: \quad V^c \to \mathbb{R}^p \\ &\quad f = \sum_{i\in[\![1,p]\!]} \alpha_i\varphi^i \mapsto \boldsymbol{\alpha} = (\alpha_i)_{i\in[\![1,p]\!]}.\end{aligned} \tag{2.30}$$

Elle vérifie pour toute fonction $\varphi \in V^c$ et pour tout vecteur $\boldsymbol{\alpha} \in \mathbb{R}^p$:

$$\begin{aligned} \langle \Phi_c^{\mathrm{T}}(\varphi), \boldsymbol{\alpha} \rangle_{\mathbb{R}^p} &= \langle \varphi, \Phi_c(\boldsymbol{\alpha}) \rangle_{V^c, V_c} \\ &= \langle \varphi, \Phi_c(\boldsymbol{\alpha}) \rangle_{L^2} \\ &= \sum_{i \in [\![1,p]\!]} \alpha^i \langle \varphi, \varphi_i \rangle_{L^2}, \end{aligned} \tag{2.31}$$

cela d'après (2.24). Puisque la relation (2.31) est vérifiée pour tout vecteur $\boldsymbol{\alpha} \in \mathbb{R}^p$, elle est vérifiée en particulier pour le vecteur défini par $\alpha^j = \delta_{ij}$, avec δ_{ij} le symbole de Kronecker et $(i, j) \in [\![1, p]\!]^2$. On a donc :

$$[\Phi_c^{\mathrm{T}}(\varphi)]_i = \langle \varphi, \varphi_i \rangle_{L^2}, \tag{2.32}$$

cela $\forall i \in [\![1, p]\!]$.

Calculons maintenant le produit $\Phi_c^{\mathrm{T}} \circ \Phi_c$. Soit $\boldsymbol{\alpha} \in \mathbb{R}^p$ et $i \in [\![1, p]\!]$,

$$\begin{aligned} [(\Phi_c^{\mathrm{T}} \circ \Phi_c)(\boldsymbol{\alpha})]_i &= \langle \Phi_c(\boldsymbol{\alpha}), \varphi_i \rangle_{L^2} \\ &= \sum_{i \in [\![1,p]\!]} \alpha^j \langle \varphi_j, \varphi_i \rangle_{L^2} \\ &= \sum_{i \in [\![1,p]\!]} \alpha^j (\boldsymbol{M}_c)_{\mathrm{ij}} \end{aligned} \tag{2.33}$$

On trouve donc que le produit $\Phi_c^{\mathrm{T}} \circ \Phi_c$ est en fait la multiplication par la matrice de masse $\boldsymbol{M}_c$, ce qui s'écrit :

$$(\Phi_c^{\mathrm{T}} \circ \Phi_c)(\boldsymbol{\alpha}) = \boldsymbol{M}_c \boldsymbol{\alpha}. \tag{2.34}$$

Les relations et les définitions clefs de cette sous-section peuvent se résumer en un diagramme de dualité (figure 2.3).

L'application de l'opérateur de corrélation construit à partir de l'équation de difusion est équivalente à l'application séquentielle de l'opérateur $(\boldsymbol{M}_c + \boldsymbol{K}_c)^{-1}$ sur m itérations. Etant donné un vecteur initial $\boldsymbol{\alpha}_0 \in \mathbb{R}^p$ appartenant à l'espace dual (voir diagramme 2.3), la première itération produit un vecteur $\boldsymbol{\alpha}_1 \in \mathbb{R}^p$ appartenant à l'espace primal. Avant de procéder à la seconde itération, il est nécessaire de « remonter » dans l'espace dual à l'aide des fonctions Φ_c et Φ_c^{T}. On applique donc l'opérateur $\boldsymbol{M}_c = \Phi_c^{\mathrm{T}} \circ \mathrm{Id} \circ \Phi_c$ à $\boldsymbol{\alpha}_1$. On reprend ensuite le processus en procédant à l'itération suivante en multipliant par $(\boldsymbol{M}_c + \boldsymbol{K}_c)^{-1}$ et en répétant ces opérations m fois.

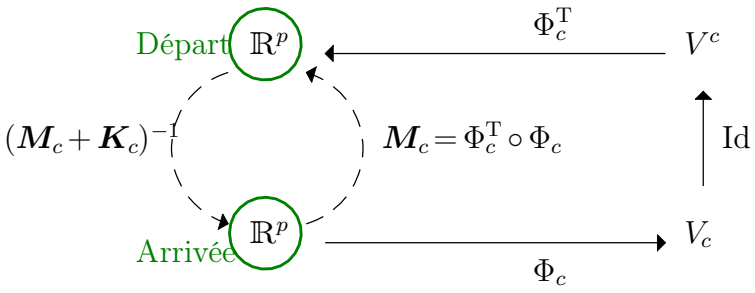


FIGURE 2.3 – Diagramme de dualité. Le sens de lecture suit celui des flèches.

Ce faisant, l'opérateur de corrélation discret prend la forme d'un produit de matrices alternant les multiplications par $(\boldsymbol{M}_c+\boldsymbol{K}_c)^{-1}$ et par $\boldsymbol{M}_c$. Cette séquence conduit à une expression de l'opérateur de corrélation analogue à (1.44), c'est-à-dire

$$\boldsymbol{C}=\gamma_{m,l}(\boldsymbol{M}_c+\boldsymbol{K}_c)^{-1}\times\boldsymbol{M}_c\times\ldots\times\boldsymbol{M}_c\times(\boldsymbol{M}_c+\boldsymbol{K}_c)^{-1}. \tag{2.35}$$

L'expression de son inverse est quant à elle donnée par

$$\boldsymbol{C}^{-1}=\frac{1}{\gamma_{m,l}}(\boldsymbol{M}_c+\boldsymbol{K}_c)\times\boldsymbol{M}_c^{-1}\times\ldots\times\boldsymbol{M}_c^{-1}\times(\boldsymbol{M}_c+\boldsymbol{K}_c). \tag{2.36}$$

Cette formule est symétrique et peut s'interpréter au choix comme un opérateur agissant sur un vecteur $\boldsymbol{\alpha}_0\in\mathbb{R}^p$ ou bien comme une matrice dans $\mathbb{R}^{p\times p}$. On remarque que, par construction, les matrices $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$ sont elles-mêmes symétriques, en vertu des équations (2.27) et (2.28). Le rang de $\boldsymbol{C}$ dépend également de celui de $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$. Comme les éléments de $\boldsymbol{M}_c$ sont des produits scalaires de fonctions de base, la matrice $\boldsymbol{M}_c$ est une matrice de Gram (*i.e.* de produits scalaires) et est par conséquent de rang p. La matrice $\boldsymbol{K}_c$ est, quant à elle, de rang $p-1$ [Ern and Guermond, 2010]. Intuitivement, les conditions aux limites de Neumann contraignent les flux aux interfaces, mais n'imposent pas de valeurs aux solutions dans le domaine, la matrice $\boldsymbol{K}_c$ seule n'est donc pas inversible. En revanche, l'inversibilité de $(\boldsymbol{M}_c+\boldsymbol{K}_c)$ assure l'existence de $\boldsymbol{C}$.

2.3 Familles d'élements finis et degrés de liberté

L'enjeu de la discrétisation d'une équation aux dérivées partielles partant de sa formulation faible réside dans le choix des fonctions de pondération formant l'espace de projection V_c. Selon le type de fonctions choisi, (harmoniques, fonctions locales à support compact, ondelettes), le schéma de discrétisation et la structure des matrices $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$ diffèrent. Dans cette section, on présente les éléments finis de Lagrange et on donne les formules d'intégrations utiles pour le reste de l'étude. Le choix des éléments finis pour la modélisation des opérateurs de corrélation en assimilation de données sera détaillé en partie II. Son intérêt sur la méthode des volumes finis tient à l'emplacement des degrés de liberté dans la définition des éléments (voir sous-section 2.3.1 ci-dessous). Pour plus de détails sur la méthode des volumes finis, le lecteur peut de référer à l'annexe A.

2.3.1 Eléments finis de type $\mathbb{P}_k$

La méthode des éléments finis classique s'appuie sur une partition de $\mathcal{D}$, c'est-à-dire un ensemble de parties non-vides et disjointes dont l'union recouvre $\bar{\mathcal{D}}$. Lorsque les éléments de cette partition sont des simplexes (*i.e.* des triangles en dimension 2 ou des tétraèdres en dimension 3) qui s'intersectionent selon leurs faces, on parle de triangulation. La méthode des éléments finis de Lagrange d'ordre k s'appuyant sur une triangulation est notée $\mathbb{P}_k$. A l'inverse, elle est notée $\mathbb{Q}_k$ quand elle s'appuie sur des quadrangles.

Soit $\mathcal{T}$ une triangulation de $\mathcal{D}$ et $\mathcal{P}_k$ l'espace des polynômes à d variables de degré k :

$$\mathcal{P}_k = \left\{ P : \mathbb{R}^d \to \mathbb{R}, \begin{pmatrix} \mathbb{R} \to \mathbb{R} \\ X \mapsto P(X, \dots, X) \end{pmatrix} \in \mathbb{R}^k[X] \right\}, \tag{2.37}$$

où $\mathbb{R}^k[X]$ est l'ensemble des polynômes à une variable de degré k. En dimension deux, les éléments de $\mathcal{T}$ sont des triangles et sont typiquement représentés par des triplets de points. On définit l'espace d'approximation $\mathcal{P}_k(\mathcal{T})$ par [Canuto et al., 1987] :

$$\mathcal{P}_k(\mathcal{T}) = \{\varphi \in \mathcal{C}^0(\mathcal{D}), \varphi_{|\tau} \in \mathcal{P}_k, \forall \tau \in \mathcal{T}\}, \tag{2.38}$$

avec $\mathcal{C}^0(\mathcal{D})$ l'ensemble des fonctions continues sur $\mathcal{D}$. On peut vérifier que $\mathcal{P}_k(\mathcal{T})$ est $H^1(\mathcal{D})$-conforme, c'est-à-dire que $\mathcal{P}_k(\mathcal{T}) \subset H^1(\mathcal{D})$. La méthode

des éléments finis de type $\mathbb{P}_k$ consiste à adopter $\mathcal{P}_k(\mathcal{T})$ comme espace d'approximation, et à définir une famille de fonctions de forme, base de $\mathcal{P}_k(\mathcal{T})$, qui va servir à décomposer les solutions de l'équation étudiée.

Dans la méthode des éléments finis de Lagrange, on choisit les fonctions de base de $\mathcal{P}_k(\mathcal{T})$ de telle sorte qu'elles soient entièrement caractérisées par leurs valeurs en certains points du domaine. Ces valeurs sont appelées « degrés de liberté » et les points correspondant les « noeuds des degrés de liberté ». Ces derniers coïncident avec les noeuds de la triangulation quand $k = 1$, et comprennent des points supplémentaires lorsque $k > 1$. Par exemple, pour $d = 2$ et $k = 2$, les noeuds des degrés de liberté correspondent aux sommets des triangles, auquels s'additionnent les points milieux de leurs arêtes. A noter qu'on ne compte qu'une seule fois un même point en cas de redondance. A l'inverse, dans la méthode des éléments finis d'Hermite, les degrés de liberté sont constitués des valeurs des fonctions en certains points, ainsi que les valeurs de leurs dérivées [Allaire, 2005].

Naturellement, le nombre de degrés de liberté nécessaire à la caractérisation des fonctions de forme augmente avec l'ordre de la méthode. Le système linéaire ainsi obtenu diffère donc par sa taille et sa structure en fonction de degré k des polynômes de l'espace d'approximation. En dimension 2 (qui est notre cas d'étude) et pour une triangulation contenant p points et n_t triangles, le nombre de degrés de liberté N_{dof} et donné par [Brenner and Scott, 2013, Canuto et al., 1987] :

$$N_{\text{dof}} = \dim \mathcal{P}_k = \frac{k(k-1)}{2} n_t + k(p-1) + 1. \tag{2.39}$$

On retrouve ainsi que si $k = 1$, $N_{\text{dof}} = p$.

2.3.2 Formules d'intégration $\mathbb{P}_1$

La suite de l'étude fait principalement usage des éléments finis de type $\mathbb{P}_1$. En effet, on souhaite modéliser des opérateurs de corrélation qui agissent sur des vecteurs à p éléments, où p désigne le nombre d'observations à assimiler. Il est donc avantageux d'exploiter la méthode des éléments finis $\mathbb{P}_1$, dont le nombre de degrés de liberté est exactement égal à la taille du vecteur d'observation pourvu en entrée de l'opérateur de corrélation. Néanmoins, la résolution de l'équation de diffusion sur des maillages fortement non structurés motive l'usage de techniques de raffinement. Les techniques de raffinement ont pour but de résoudre l'équation de diffusion dans un espace contenant plus de degrés de liberté que le vecteur d'observation. Cet espace peut être

construit de diverses façons, l'une d'elles exploitant un raffinement en ordre faisant intervenir des éléments finis de type $\mathbb{P}_k$. Cependant, d'autres méthodes (de raffinement en espace) semblent plus appropriées à nos applications et sont donc privilégiées pour cette étude. Le lecteur pourra se référer à la partie III pour une discussion approfondie sur le sujet.

Les fonctions de base $\mathbb{P}_1$ sont affines sur chaque triangle du maillage (figure 2.4) et entièrement déterminées par leurs valeurs en chaque noeud [Ern and Guermond, 2010]. Pour tout $i \in [\![1, p]\!]$, on a :

$$(\varphi_i)_{|\tau} \in \mathcal{P}_1 \quad , \quad \forall \tau \in \mathcal{T} \tag{2.40}$$

$$\varphi_i(\boldsymbol{z}_j) = \delta_{ij} \quad , \quad \forall j \in [\![1, p]\!]. \tag{2.41}$$

Ainsi définie, la famille $(\varphi_i)_{i \in [\![1,p]\!]}$ forme une base de $\mathcal{P}_1(\mathcal{T})$.

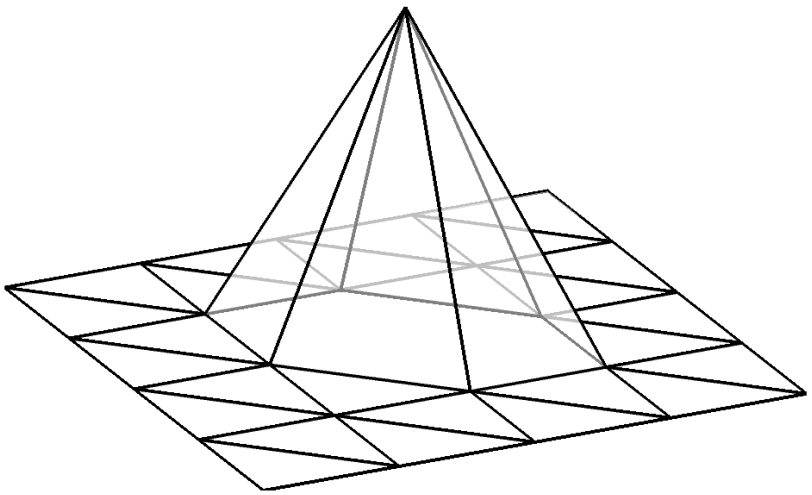


Figure 2.4 – Fonction de base $\mathbb{P}_1$ et son support en dimension 2.

On souhaite maintenant détailler le calcul des coefficients des matrices de masse et de raideur donnés par les équations (2.27) et (2.28). La stratégie consiste d'une part à utiliser le fait que $\mathcal{D} = \coprod_{\tau \in \mathcal{T}} \tau$ pour découper l'intégrale sur $\mathcal{D}$ en plusieurs intégrales indépendantes sur $\tau \in \mathcal{T}$. De cette manière, les calculs sont opérés sur des triangles sur lesquels les fonctions de base sont linéaires. D'autre part, on associe à chaque élément de la triangulation un élément de référence et on utilise la transformation géométrique permettant de passer de l'un à l'autre pour calculer les coefficients $(\boldsymbol{M}_c)_{\mathrm{ij}}$ et $(\boldsymbol{K}_c)_{\mathrm{ij}}$.

Soit τ le triangle de référence de sommets $\boldsymbol{A} = (0,0)$, $\boldsymbol{B} = (1,0)$ et $\boldsymbol{C} = (0,1)$. Tout point $\boldsymbol{K}$ de τ peut être repéré par ses coordonnées barycentriques :

$$\boldsymbol{K} = \lambda_1 \boldsymbol{A} + \lambda_2 \boldsymbol{B} + \lambda_3 \boldsymbol{C}, \tag{2.42}$$

et tout triangle τ_k de sommets $(\boldsymbol{A}_k, \boldsymbol{B}_k, \boldsymbol{C}_k)$ peut être vu comme la déformation du triangle de référence τ au travers de l'application $F_k : \mathbb{R}^d \to \mathbb{R}^d$ définie par :

$$F_k(\boldsymbol{A}) = \boldsymbol{A}_k, F_k(\boldsymbol{A}) = \boldsymbol{B}_k, F_k(\boldsymbol{C}) = \boldsymbol{C}_k. \tag{2.43}$$

L'intégrale sur τ_k se calcule à partir de l'intégrale sur τ par un changement de variable faisant intervenir la jacobienne de l'application F_k. On trouve alors que, en dimension $d = 2$ et pour tout $(i,j) \in [\![1,p]\!]^2$, les coefficients de la matrice de masse sont donnés par :

$$\int_{\tau_k} \varphi_i \varphi_i \mathrm{d}\boldsymbol{z} = \frac{\mathcal{A}(\tau_k)}{6}, \tag{2.44}$$

$$\int_{\tau_k} \varphi_i \varphi_j \mathrm{d}\boldsymbol{z} = \frac{\mathcal{A}(\tau_k)}{12}, \tag{2.45}$$

où $\mathcal{A}(\tau_k)$ désigne l'aire de τ_k. Les coefficients diagonaux de la matrice de raideur s'obtiennent quant à eux par :

$$\int_{\tau_k} \boldsymbol{\kappa} \nabla\varphi_1 \cdot \nabla\varphi_1 \mathrm{d}\boldsymbol{z} = \frac{(\boldsymbol{C}_k - \boldsymbol{B}_k)^{\mathrm{T}} \boldsymbol{\kappa} (\boldsymbol{C}_k - \boldsymbol{B}_k)}{4\mathcal{A}(\tau_k)}, \tag{2.46}$$

$$\int_{\tau_k} \boldsymbol{\kappa} \nabla\varphi_2 \cdot \nabla\varphi_2 \mathrm{d}\boldsymbol{z} = \frac{(\boldsymbol{C}_k - \boldsymbol{A}_k)^{\mathrm{T}} \boldsymbol{\kappa} (\boldsymbol{C}_k - \boldsymbol{A}_k)}{4\mathcal{A}(\tau_k)}, \tag{2.47}$$

$$\int_{\tau_k} \boldsymbol{\kappa} \nabla\varphi_3 \cdot \nabla\varphi_3 \mathrm{d}\boldsymbol{z} = \frac{(\boldsymbol{B}_k - \boldsymbol{A}_k)^{\mathrm{T}} \boldsymbol{\kappa} (\boldsymbol{B}_k - \boldsymbol{A}_k)}{4\mathcal{A}(\tau_k)}, \tag{2.48}$$

et les coefficients non-diagonaux par :

$$\int_{\tau_k} \boldsymbol{\kappa} \nabla\varphi_1 \cdot \nabla\varphi_2 \mathrm{d}\boldsymbol{z} = -\frac{(\boldsymbol{C}_k - \boldsymbol{B}_k)^{\mathrm{T}} \boldsymbol{\kappa} (\boldsymbol{C}_k - \boldsymbol{A}_k)}{4\mathcal{A}(\tau_k)}, \tag{2.49}$$

$$\int_{\tau_k} \boldsymbol{\kappa} \nabla\varphi_1 \cdot \nabla\varphi_3 \mathrm{d}\boldsymbol{z} = -\frac{(\boldsymbol{A}_k - \boldsymbol{B}_k)^{\mathrm{T}} \boldsymbol{\kappa} (\boldsymbol{C}_k - \boldsymbol{B}_k)}{4\mathcal{A}(\tau_k)}, \tag{2.50}$$

$$\int_{\tau_k} \boldsymbol{\kappa} \nabla\varphi_2 \cdot \nabla\varphi_3 \mathrm{d}\boldsymbol{z} = -\frac{(\boldsymbol{C}_k - \boldsymbol{A}_k)^{\mathrm{T}} \boldsymbol{\kappa} (\boldsymbol{B}_k - \boldsymbol{A}_k)}{4\mathcal{A}(\tau_k)}. \tag{2.51}$$

La construction des matrices $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$ s'appelle l'« assemblage ». En pratique, l'assemblage n'est pas effectué en parcourant les noeuds du maillage,

mais en parcourant la liste des éléments de la triangulation [Ern and Guermond, 2010, Canuto et al., 1987]. Cette étape nécessite de stocker la liste des triangles dans un tableau séparé du tableau contenant les coordonnées des points. L'assemblage se fait alors de la manière suivante : pour chaque triangle, on calcule les quantités (2.44) à (2.51), puis on les affecte aux noeuds (i, j) correspondant dans les matrices $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$. Quand on revisite un noeud à partir d'un autre triangle, les contributions sont additionnées.

2.4 Aspects de convergence

Dans cette section, on discute des aspects de convergence de la résolution de l'équation de diffusion par la méthode des éléments finis. Les résultats sont énoncés pour introduire les notions d'erreur et de qualité du maillage, mais ne sont pas démontrés. Ces notions sont mises à profit dans les chapitres suivants pour évaluer et contrôler la précision de notre modélisation des opérateurs de corrélation.

Il est important de remarquer que notre modèle de corrélation fait intervenir un nombre fini d'applications de l'opérateur de diffusion. Il est donc exclu d'aller vers des valeurs élevées du paramètre de régularité m dans la pratique. Toutefois, on présente la stabilité de la diffusion implicite dans la sous-section 2.4.1, qui permet de discuter d'autres choix de modélisation que nous n'avons pas retenus.

La convergence en espace est quant à elle étudiée de manière standard : le maillage est supposé suffisamment régulier et la taille de ses éléments est décrue de sorte d'obtenir une approximation de plus en plus précise de la solution de l'équation de diffusion. Cependant, la recherche de la convergence en espace se fera plus tard au travers d'un raffinement de maillage, qui se révèlera plus complexe à analyser et évaluer. On présente ces résultats pour introduire les paramètres géométriques qui permettent de quantifier localement la qualité du maillage et d'analyser les erreurs qui apparaissent au cours de la résolution.

2.4.1 Stabilité de la diffusion implicite

Dans la littérature, l'équation (2.16) est souvent appelée « diffusion implicite ». Cette dénomination provient du fait que (2.16) (et, *a fortiori*, 1.42) peut s'interpréter comme la semi-discrétisation implicite en temps de l'équa-

tion

$$\frac{\partial f}{\partial t} - \nabla \cdot \boldsymbol{\kappa} \nabla f = 0, \tag{2.52}$$

avec un pas de temps égal à 1 et une condition initiale $f = f_0$ [Mirouze and Weaver, 2010]. Les conditions aux limites de Neumann sont imposées aux bords du domaine.

De nombreux schémas de semi-discrétisation en temps de (2.52) existent. On présente les discrétisations d'Euler implicite et explicite, qui s'écrivent respectivement :

$$(\mathbf{1} - \nabla \cdot \boldsymbol{\kappa} \nabla) f_{n+1} = f_n \tag{2.53}$$

$$f_{n+1} = (\mathbf{1} + \nabla \cdot \boldsymbol{\kappa} \nabla) f_n, \tag{2.54}$$

où le pas de temps de la discrétisation a été fixé égal à $\delta t = 1$.

Comme l'opérateur $-\nabla \cdot \boldsymbol{\kappa} \nabla$ est symétrique semi défini positif, la plus petite valeur propre de $(\mathbf{1} - \nabla \cdot \boldsymbol{\kappa} \nabla)$ est supérieure à 1 :

$$\lambda_{\min}\{\mathbf{1} - \nabla \cdot \boldsymbol{\kappa} \nabla\} \geqslant 1. \tag{2.55}$$

En conséquence, la valeur propre maximale de l'opérateur inverse est inférieure à 1 :

$$\lambda_{\max}\{(\mathbf{1} - \nabla \cdot \boldsymbol{\kappa} \nabla)^{-1}\} \leqslant 1, \tag{2.56}$$

ce qui implique la stabilité de (2.53) au sens de la norme $L^2(\mathcal{D})$ [Mirouze and Weaver, 2010] :

$$\|f_{n+1}\|_{L^2(\mathcal{D})} \leqslant \|f_n\|_{L^2(\mathcal{D})} \leqslant \ldots \leqslant \|f_0\|_{L^2(\mathcal{D})}. \tag{2.57}$$

Le schéma implicite est donc inconditionnellement stable.

De même, on montre que (2.54) est stable à condition que la valeur propre maximale de $-\nabla \cdot \boldsymbol{\kappa} \nabla$ soit inférieure à la valeur du pas de temps, soit :

$$\lambda_{\max}\{-\nabla \cdot \boldsymbol{\kappa} \nabla\} \leqslant 1. \tag{2.58}$$

Cette condition contraint l'implémentation pratique du schéma explicite, notamment lorsque les valeurs propres du tenseur $\boldsymbol{\kappa}$ (*i.e.* les longueurs de portée contenues dans $\boldsymbol{\kappa}$) sont élevées. C'est la raison principale pour laquelle le schéma explicite est souvent rejeté en faveur du schéma implicite [Mirouze and Weaver, 2010]. On rajoute que la famille de fonctions de corrélation modélisées par cette dernière méthode est plus grande que la famille des fonctions gaussiennes qui sont solutions du schéma explicite.

Les principes de stabilité évoqués précédemment restent valables pour l'analyse du schéma discret en temps et en espace. En effet, l'équation (2.28) peut se refactoriser sous la forme

$$(\boldsymbol{I} + \boldsymbol{M}_c^{-1}\boldsymbol{K}_c)\boldsymbol{\alpha}_{n+1} = \boldsymbol{\alpha}_n. \tag{2.59}$$

La stabilité de (2.59) dépend donc de l'opérateur $(\boldsymbol{I}+\boldsymbol{M}_c^{-1}\boldsymbol{K}_c)$ ou, de manière équivalente, de $(\boldsymbol{I} + \boldsymbol{M}_c^{-1/2}\boldsymbol{K}_c\boldsymbol{M}_c^{-1/2})$, qui possède les mêmes propriétés spectrales. En effet, puisque $\boldsymbol{M}_c$ est symétrique définie positive,

$$\mathrm{Sp}\{\boldsymbol{M}_c^{-1}\boldsymbol{K}_c\} = \mathrm{Sp}\{\boldsymbol{M}_c^{1/2}\boldsymbol{M}_c^{-1}\boldsymbol{K}_c\boldsymbol{M}_c^{-1/2}\} = \mathrm{Sp}\{\boldsymbol{M}_c^{-1/2}\boldsymbol{K}_c\boldsymbol{M}_c^{-1/2}\}, \tag{2.60}$$

où $\mathrm{Sp}\{\boldsymbol{A}\}$ désigne l'ensemble des valeurs propres de $\boldsymbol{A}$. On a par conséquent

$$\lambda_{\min}\{\boldsymbol{I} + \boldsymbol{M}_c^{-1}\boldsymbol{K}_c\} \geqslant 1 \tag{2.61}$$

et

$$\lambda_{\max}\{(\boldsymbol{I} + \boldsymbol{M}_c^{-1}\boldsymbol{K}_c)^{-1}\} \leqslant 1. \tag{2.62}$$

Cette transformation permet donc de montrer que l'approximation de Galerkin en espace et implicite en temps est inconditionnellement stable.

2.4.2 Convergence en espace

Afin d'étudier la convergence de la méthode des éléments finis en espace, il convient de définir en premier lieu un certain nombre de notions géométriques. Plaçons-nous dès maintenant dans le cas où $\mathcal{D}$ est un domaine de dimension 2, c'est-à-dire que $\dim(\mathcal{D}) = d = 2$. Pour tout triangle $\tau \in \mathcal{T}$, on définit le diamètre $\mathrm{diam}(\tau)$ et la rondeur $\rho(\tau)$ comme

$$\mathrm{diam}(\tau) = \max_{x,y\in\tau}\|x - y\|_2, \tag{2.63}$$

où la norme Euclidienne est notée $\|\cdot\|_2$ et

$$\rho(\tau) = \max_{B_\mu\subset\tau}(2\mu), \tag{2.64}$$

où $B_\mu \subset \tau$ désigne l'ensemble des boules inclues dans τ et de rayon $\mu \geqslant 0$.

La relation $\mathrm{diam}(\tau)/\rho(\tau) \geqslant 1$ est toujours vérifiée. Ce rapport augmente soit quand $\mathrm{diam}(\tau)$ augmente indépendamment de $\rho(\tau)$, soit quand $\rho(\tau)$ diminue indépendamment de $\mathrm{diam}(\tau)$. Il augmente donc lorsque τ devient « aplati » (figure 2.5). Idéalement, en éléments finis, on souhaiterait que ce

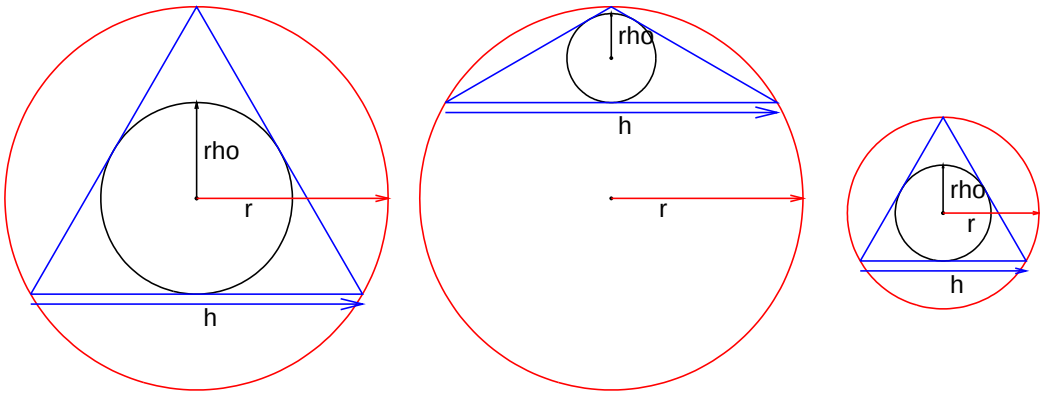


FIGURE 2.5 – Triangles et grandeurs géométriques. On numérote les triangles de gauche à droite. Les triangles 1 et 3 ont le même rapport d'aspect (invariance par homothétie), qui est inférieur à celui du triangle 2. En revanche, le triangle 1 a le même rayon de cercle circonscrit que le triangle 2, bien que ce dernier soit plus petit.

rapport soit borné sur l'ensemble de la triangulation [Shewchuk, 2002]. Autrement dit, on souhaite s'assurer qu'il existe une constante $C > 0$ telle que

$$\frac{\mathrm{diam}(\tau)}{\rho(\tau)} \leqslant C. \tag{2.65}$$

De l'équation (2.65) et en supposant que $f \in H^{k+1}(\mathcal{D})$, on peut montrer que la borne sur l'erreur s'écrit

$$\|f - f_h\|_{H^1(\mathcal{D})} \leqslant Ch^k \|f\|_{H^{k+1}(\mathcal{D})},$$

où $h = \max_{\tau \in \mathcal{T}}(\mathrm{diam}(\tau))$ est la taille maximale des arêtes de la triangulation, C est une constante indépendante de h et f_h est la fonction solution approximée par la méthode des éléments finis [Allaire, 2005].

Ainsi, la méthode des éléments finis converge quand la taille du maillage tend vers 0. Quand ce n'est pas le cas, la présence de triangles aplatis peut dégrader localement la qualité de la solution. En effet, la représentation des gradients sur un triangle comportant un angle large/obtus peut s'avérer médiocre [Shewchuk, 2002]. C'est dû à la propriété des fonctions de formes $\mathbb{P}_1$ exprimée dans la relation (2.41). La pente décrite par la fonction linéaire φ_i sur τ peut s'avérer très forte lorsque le sommet $\boldsymbol{z}_i$ se trouve proche de l'arrête opposée dans τ (figure 2.6). Cette dégénérescence se retrouve dans le contenu de la matrice de raideur dont le conditionnement se retrouve affecté.

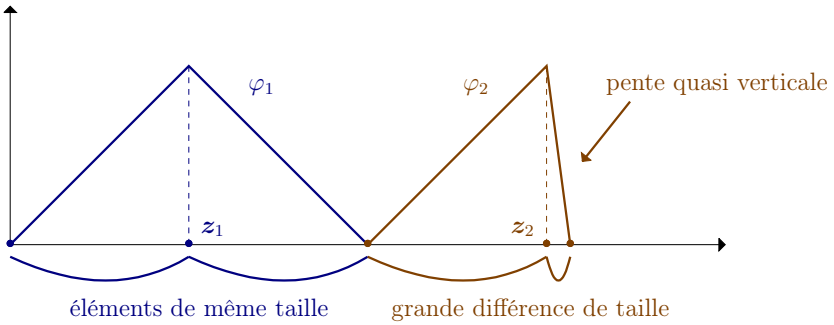


FIGURE 2.6 – A gauche : fonction $\mathbb{P}_1$ sur éléments réguliers. A droite : dégénérescence due à la présence d'un élément de petit taille. Le gradient de φ_2 sur cet élement est très fort.

La matrice de raideur $\boldsymbol{K}_c$ est assemblée à partir des matrices de raideur élémentaires $\boldsymbol{K}_\tau$ à l'intérieur de chaque élément τ. Soient $\omega_1(\boldsymbol{z}), \omega_2(\boldsymbol{z}), \omega_3(\boldsymbol{z})$ les coordonnées barycentriques d'un point $\boldsymbol{z}$ à l'intérieur de τ. D'après Shewchuk [2002], la matrice $\boldsymbol{K}_\tau$ peut s'exprimer comme :

$$\begin{aligned} \boldsymbol{K}_\tau &= \mathcal{A}(\tau) \begin{pmatrix} \nabla\omega_1 \cdot \nabla\omega_1 & \nabla\omega_1 \cdot \nabla\omega_2 & \nabla\omega_1 \cdot \nabla\omega_3 \\ \nabla\omega_2 \cdot \nabla\omega_1 & \nabla\omega_2 \cdot \nabla\omega_2 & \nabla\omega_2 \cdot \nabla\omega_3 \\ \nabla\omega_3 \cdot \nabla\omega_1 & \nabla\omega_3 \cdot \nabla\omega_2 & \nabla\omega_3 \cdot \nabla\omega_3 \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} \cot\theta_2 + \cot\theta_3 & -\cot\theta_3 & -\cot\theta_2 \\ -\cot\theta_3 & \cot\theta_2 + \cot\theta_3 & -\cot\theta_1 \\ -\cot\theta_2 & -\cot\theta_1 & \cot\theta_2 + \cot\theta_3 \end{pmatrix}, \quad (2.66) \end{aligned}$$

où $\mathcal{A}(\tau)$ est l'aire de τ et $\theta_1, \theta_2, \theta_3$ sont ses trois angles intérieurs. On voit que lorsqu'un des angles est proche de 0, sa cotangente tend vers l'infini, tout comme la plus grande valeur propre de $\boldsymbol{K}_\tau$.

La notion de « qualité d'un maillage » dépend en réalité de l'application. En l'absence d'information précise sur l'équation à résoudre, on se contente d'énoncer des résultats sur le cas isotrope. En pratique, les meilleurs résultats pour résoudre une équation dont les coefficients sont anisotropiques, par exemple, sont obtenus en utilisant un maillage lui-même anisotropique (*i.e.* non structuré).

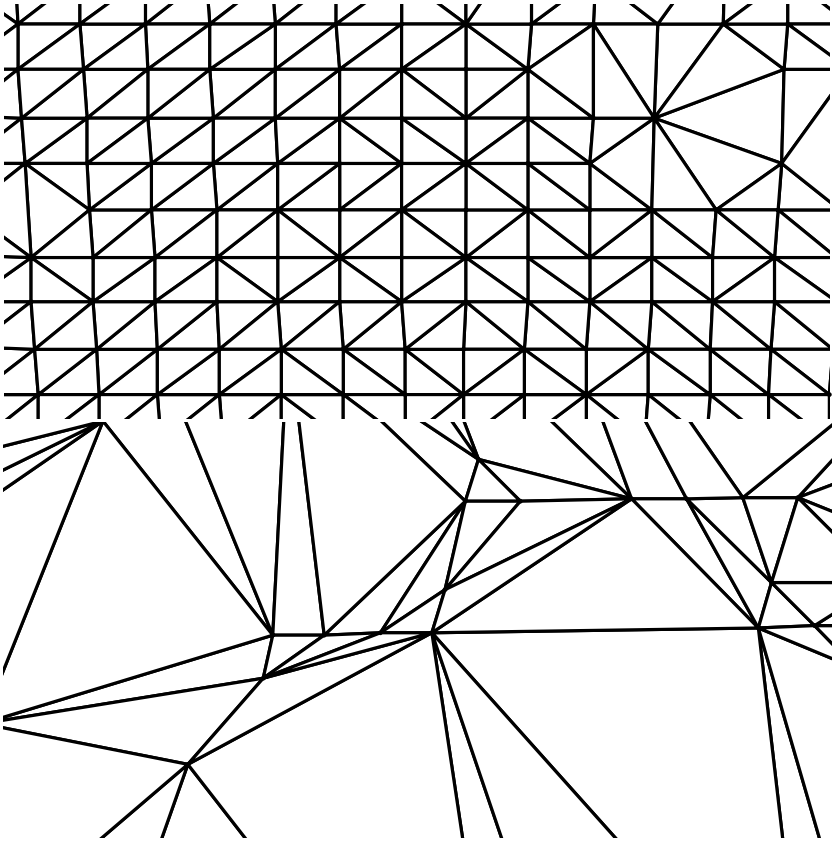


Figure 2.7 – En haut : portion de maillage « régulier ». En bas : portion de maillage « irrégulier ».

2.5 Assimilation des dérivées du champ d'observation

Dans cette section, on détaille le lien entre la modélisation des opérateurs de corrélation par l'équation de diffusion en assimilation de données et la méthode de Brankart et al. [2009] consistant à assimiler à la fois le champ d'observation et ses dérivées consécutives. La méthode de Brankart et al. [2009] est en effet populaire en assimilation de données. Son analyse mathématique révèle une connexion directe avec l'équation de diffusion. Le cadre théorique développé dans ce manuscrit est donc utile à l'analyse conjointe des deux méthodes. Bien que la formulation basée sur l'utilisation de la diffusion semble la plus flexible et générale des deux, il est certain que la mise en exergue d'un lien théorique entre les deux approches constitue une vraie étape vers la meilleure compréhension de la modélisation des opérateurs de

corrélation.

2.5.1 Du champ dérivé à la diffusion : formulation continue

Dans la méthode de Brankart et al. [2009], une transformation linéaire T est appliquée aux observations y avant qu'elles ne soient assimilées. Cette transformation linéaire peut prendre des formes variées, mais il est commun que le résultat $T(y)$ contienne à la fois le champ d'observation d'origine et ses dérivées successives. Dans son approche initiale, Brankart et al. [2009] présente principalement l'assimilation des dérivées premières du champ d'observation. Néanmoins, il est tout à fait possible d'assimiler des dérivées d'ordre supérieur. Par soucis de complétion, mais aussi de clarté, on choisit donc de présenter la méthode permettant d'assimiler à la fois y, ses dérivées premières et ses dérivées secondes. L'utilisation de ces trois moments est suffisante pour déduire toute généralisation (ou simplification) ultérieure.

On suppose dans toute la suite de l'étude que $\dim(\mathcal{D}) = d = 2$. Néanmoins, on précise que la généralisation à une dimension supérieure est aisée.

Notons ∂_{z_i} l'opérateur de dérivation selon la i-ème direction de l'espace et $\partial^2_{z_i}$ la dérivée seconde dans cette même direction. On suppose que l'espace des observations est inclus dans $L^2(\mathcal{D})$. Soit $T : \mathcal{Y} \to \mathcal{Y}^5$ l'opérateur défini par

$$y \mapsto T(y) = \begin{pmatrix} y \\ \partial_{z_1} y \\ \partial_{z_2} y \\ \partial^2_{z_1} y \\ \partial^2_{z_2} y \end{pmatrix} \tag{2.67}$$

pour toute fonction y vérifiant $\partial_{z_1} y \in \mathcal{Y}$, $\partial_{z_2} y \in \mathcal{Y}$ et $\partial^2_{z_1} y \in \mathcal{Y}$, $\partial^2_{z_2} y \in \mathcal{Y}$. On définit $D : \mathcal{Y}^5 \to \mathcal{Y}^5$ contenant les coefficients $(a_i)_{i=0,1,2}$ par

$$w \mapsto D(w) = \begin{pmatrix} a_0 w \\ a_1 w \\ a_1 w \\ a_2 w \\ a_2 w \end{pmatrix}. \tag{2.68}$$

L'opérateur de précision $\mathcal{C}^{-1}$ compris comme un opérateur de $\mathcal{Y}$ dans $\mathcal{Y}^\star = \mathcal{Y}$ est défini dans la méthode de Brankart et al. [2009] comme

$$\mathcal{C}^{-1} = T^{\mathrm{T}} D T, \tag{2.69}$$

où T^{T} désigne le transposé de T. En développant $\mathcal{C}^{-1}(y)$ et en utilisant la relation $(\partial_{z_i})^{\mathrm{T}} = -\partial_{z_i}$ (au sens des distributions), on trouve ainsi que

$$\begin{aligned} \mathcal{C}^{-1}(y) &= a_0 y - a_1 \partial^2_{z_1} y - a_1 \partial^2_{z_2} y + a_2 \partial^2_{z_1} y + a_2 \partial^2_{z_2} y \\ &= a_0 y - a_1 \Delta y + a_2 \Delta^2 y. \end{aligned} \tag{2.70}$$

On remarque qu'après multiplication par $\frac{1}{a_0}$ (pour $a_0 \neq 0$), l'expression (2.70) se ramène à l'équation de diffusion généralisée (2.14). En particulier, le choix

$$\begin{aligned} a_0 &= 1/\gamma_{m,l} \\ a_1 &= 2l^2/\gamma_{m,l} \\ a_2 &= l^4/\gamma_{m,l} \end{aligned} \tag{2.71}$$

conduit à l'expression de la diffusion homogène (1.42) avec $m = 2$.

On précise que la méthode de Brankart et al. [2009] ne sépare pas les variances des corrélations. Ainsi, dans sa méthode, le facteur de normalisation $\gamma_{m,l}$ est absent. Les coefficients $(a_i)_{i=0,1,2}$ sont directement estimés pour prendre en compte les variances.

2.5.2 De la diffusion au champ dérivé : formulation discrète

On souhaite maintenant montrer le lien entre l'opérateur de corrélation discrétisé en éléments finis et l'implémentation pratique de la méthode de Brankart et al. [2009]. L'expression complète de l'opérateur de covariance en éléments finis s'écrit

$$\begin{aligned} \boldsymbol{R}^{-1} &= \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma}^{-1/2}(\boldsymbol{M}+\boldsymbol{K})\boldsymbol{M}^{-1}(\boldsymbol{M}+\boldsymbol{K})\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} \\ &= \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma}^{-1/2}(\boldsymbol{M}+2\boldsymbol{K}+\boldsymbol{K}\boldsymbol{M}^{-1}\boldsymbol{K})\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}, \end{aligned} \tag{2.72}$$

où $\boldsymbol{\Gamma}$ est une matrice diagonale contenant des facteurs de normalisation et permettant d'assurer que les éléments diagonaux de la matrice de corrélation sont égaux à 1, et où $\boldsymbol{\Sigma}$ est une matrice diagonale contenant les écart-types de $\boldsymbol{R}$. Cette formule générale provenant du cas anisotrope se simplifie légèrement dans le cas homogène isotrope en prenant $\boldsymbol{\Gamma} = \gamma_{m,l}\boldsymbol{I}$.

On met l'équation (2.72) sous la forme

$$\boldsymbol{R}^{-1} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma}^{-1/2} \times \boldsymbol{T}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{T} \times \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}, \tag{2.73}$$

avec

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{I} & & \\ & 2\boldsymbol{I} & \\ & & \boldsymbol{I} \end{pmatrix} \tag{2.74}$$

et

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{M}_c^{1/2} \\ \boldsymbol{K}_c^{1/2} \\ \boldsymbol{M}_c^{-1/2}\boldsymbol{K}_c \end{pmatrix}. \tag{2.75}$$

Le tenseur de diffusion $\boldsymbol{\kappa}$ est caché dans la définition de $\boldsymbol{K}_c$. Lorsque $\boldsymbol{\kappa} = l^2\boldsymbol{I}$ (cas de la diffusion homogène), on pose $\boldsymbol{K}_c = l^2\widetilde{\boldsymbol{K}}_c$, ce qui permet de faire apparaître explicitement la longueur de portée dans la définition des matrices. En effet, on introduit

$$\tilde{\boldsymbol{D}} = \begin{pmatrix} \boldsymbol{I} & & \\ & 2l^2\boldsymbol{I} & \\ & & l^4\boldsymbol{I} \end{pmatrix} \tag{2.76}$$

et

$$\tilde{\boldsymbol{T}} = \begin{pmatrix} \boldsymbol{M}_c^{1/2} \\ \widetilde{\boldsymbol{K}}_c^{1/2} \\ \boldsymbol{M}_c^{-1/2}\widetilde{\boldsymbol{K}}_c \end{pmatrix}. \tag{2.77}$$

L'expression (2.73) se réécrit

$$\boldsymbol{R}^{-1} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma}^{-1/2} \times \tilde{\boldsymbol{T}}^{\mathrm{T}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{T}} \times \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}. \tag{2.78}$$

Pour interpréter (2.76-2.78), il suffit de constater que $\boldsymbol{M}_c$ ne contient aucune information sur les dérivées des fonctions de forme, et que $\boldsymbol{K}_c$ contient des produits de gradients. Ainsi, les multiplications par $\boldsymbol{M}_c^{1/2}$, $\widetilde{\boldsymbol{K}}_c^{1/2}$ et $\boldsymbol{M}_c^{-1/2}\widetilde{\boldsymbol{K}}_c$ correspondent respectivement à des dérivations d'ordre 0, 1 et 2. Etant donné un vecteur d'observations $\boldsymbol{y}$, le produit $\tilde{\boldsymbol{T}} \times \boldsymbol{y}$ contient donc les dérivées successives de $\boldsymbol{y}$, comme dans la méthode de Brankart et al. [2009].

A titre d'information, on donne l'expression de $\tilde{\boldsymbol{D}}$ et $\tilde{\boldsymbol{T}}$ dans le cas où $m = 4$:

$$\tilde{\boldsymbol{D}} = \begin{pmatrix} \boldsymbol{I} & & & & \\ & 4l^2\boldsymbol{I} & & & \\ & & 6l^4\boldsymbol{I} & & \\ & & & 4l^6\boldsymbol{I} & \\ & & & & l^8\boldsymbol{I} \end{pmatrix} \tag{2.79}$$

et

$$\tilde{\boldsymbol{T}} = \begin{pmatrix} \boldsymbol{M}_c^{1/2} \\ \widetilde{\boldsymbol{K}}_c^{1/2} \\ \boldsymbol{M}_c^{1/2}(\boldsymbol{M}_c^{-1}\widetilde{\boldsymbol{K}}_c) \\ \widetilde{\boldsymbol{K}}_c^{1/2}(\boldsymbol{M}_c^{-1}\widetilde{\boldsymbol{K}}_c) \\ \boldsymbol{M}_c^{1/2}(\boldsymbol{M}_c^{-1}\widetilde{\boldsymbol{K}}_c)(\boldsymbol{M}_c^{-1}\widetilde{\boldsymbol{K}}_c) \end{pmatrix}. \tag{2.80}$$

Toute généralisation à des valeurs de m supérieures se retrouve selon le même principe en développant le produit $(\boldsymbol{M}_c + \boldsymbol{K}_c)\boldsymbol{M}_c^{-1} \dots \boldsymbol{M}_c^{-1}(\boldsymbol{M}_c + \boldsymbol{K}_c)$.

2.5.3 Discussion

La normalisation de l'opérateur (2.69) dépend des valeurs attribuées aux coefficients $(a_i)_{i=0,1,2}$. Si ces derniers vérifient une relation du type (2.71) qui permet de les relier à un modèle théorique de corrélation, comme celui de Matérn, alors il est possible d'estimer analytiquement le coefficient de normalisation γ_c. En revanche, si les $(a_i)_{i=0,1,2}$ estimés ne vérifient pas une telle relation, alors il n'existe pas d'estimation analytique de γ_c, et il devient délicat de vouloir séparer les variances des corrélations. En conséquence, il n'est pas possible de procéder à une analyse d'erreurs fine, comme présentée dans le chapitre 3, puisque cette analyse d'erreurs compare directement la solution de l'équation de diffusion normalisée au modèle de corrélation théorique.

Modéliser des corrélations variables dans l'espace est également possible avec la méthode de Brankart et al. [2009]. Il suffit pour cela d'estimer les dérivées locales du champ d'entrée (opérateur T) et de les multiplier par des coefficients $(a_i(\boldsymbol{z}))_{i=0,1,2}$ dont les valeurs dépendent de l'emplacement géographique. Toutefois, il n'est pas décrit dans la littérature de procédure permettant d'automatiser cette étape. A l'inverse, l'estimation du tenseur de corrélation variable dans l'espace peut être faite grâce à la méthode de Desroziers et al. [2005].

Enfin, la formulation de la méthode de Brankart et al. [2009] suppose que la distribution spatiale des observation est suffisamment structurée pour pouvoir évaluer localement les dérivées spatiales dans le système de coordonnées choisi. Or, ce n'est pas le cas lorsque la distribution spatiale des observations comporte des « trous », comme c'est le cas pour les données SEVIRI présentées dans le chapitre 4. Généralement, la représentation des gradients sur une grille non-structurée n'est pas triviale. Cette raison motive l'utilisation des éléments finis plutôt que (ou en combinaison avec !) la méthode de Brankart et al. [2009], et ainsi les développements des prochains chapitres.

2.6 L'essentiel du chapitre

Comme on peut le voir, la traitement de l'équation de diffusion par la méthode des éléments finis s'inscrit naturellement dans le cadre fonctionnel Hilbertien. Dans la partie II, nous verrons que c'est également un choix pertinent quand il est question de maillages non structurés. Pour qu'il n'y ait pas de méprise sur la versatilité de la méthode, tous les développements utilisent la forme anisotrope de l'équation de diffusion. Une forme alternative, homogène, permet quant à elle de prouver que la méthode de Brankart et al. [2009] n'est autre qu'une équation de diffusion revisitée. La formulation faible et la méthode des éléments finis apportent dans ce cas une réponse à la question de la modélisation des gradients sur des maillages hétérogènes.

Ce qu'il faut retenir :

- La formulation faible de l'équation de diffusion permet d'affaiblir les hypothèses de régularité sur les solutions recherchées.
- Discrétiser l'équation de diffusion en éléments finis revient à résoudre le système
$$(\boldsymbol{M}_c + \boldsymbol{K}_c)\boldsymbol{\alpha}_{n+1} = \boldsymbol{M}_c\boldsymbol{\alpha}_n$$
sur m pas de temps.
- L'équation ci-dessus est appelée « équation de diffusion implicite » et est inconditionnellement stable.
- L'assimilation des dérivées du champ d'observation peut s'interpréter comme une équation de diffusion généralisée. La méthode des éléments finis permet de généraliser cette méthode en représentant les gradients sur des maillages non structurés.

Deuxième partie

Vers les maillages non structurés

Chapitre 3

Etude sur des maillages structurés

Dans ce chapitre, on s'intéresse aux aspects pratiques et expérimentaux de la modélisation des opérateurs de corrélation par la méthode des éléments finis. Les expériences sont menées dans un premier temps avec des données artificielles et structurées, de façon à présenter les résultats de la méthode dans un cas idéalisé. Ainsi, la section 3.1 fait un tour d'horizon des propriétés des opérateurs discrétisés, avec une discussion concernant la structure des matrices et la représentation des fonctions de corrélations.

La section 3.2 est une introduction à la condensation de masse, procédé de diagonalisation de la matrice de masse, populaire en éléments finis et très utile en pratique. Plusieurs approches sont présentées, équivalentes dans le cas des éléments finis $\mathbb{P}_1$, mais qui ont l'avantage de fournir des interprétations différentes, physiques ou mathématiques.

La section 3.3 s'intéresse à l'effet des conditions aux bords du domaine sur la représentation des corrélations. L'utilisation de conditions aux limites de Neumann (ou de Dirichlet) influence en effet l'amplitude des fonctions de corrélation modélisées par les méthodes de diffusion. Il est donc primordial de présenter ce comportement avant d'introduire les maillages non structurés, dont les éventuels effets de bords ont une origine distincte des conditions aux limites.

Finalement, une synthèse des spécificités des maillages structurés est présentée dans la section 3.4.

Ce chapitre sert de transition entre l'étude théorique et le contenu expérimental. Petit à petit, on met en avant les détails de l'implémentation, préparant ainsi le terrain pour le chapitre 4, l'un des piliers de ce manuscrit.

L'annexe E présente la structure du code développé au cours de la thèse.

3.1 Structure de la réponse

Dans cette section, on présente des propriétés élémentaires des opérateurs de diffusion discrétisés sur des maillages réguliers. Ces résultats font office de référence dans la suite de l'étude et permettent d'introduire de nouvelles discussions comme l'intérêt de la condensation de masse (section 3.2) et l'effet des conditions aux bords du domaine (section 3.3).

3.1.1 Profils des matrices en éléments finis

On débute cette section en donnant quelques propriétés essentielles des matrices de masse et de raideur apparaissant dans la méthode des éléments finis.

Les matrices $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$ sont creuses grâce au choix de fonctions à support compact dans l'approximation de Galerkin. En éléments finis de type $\mathbb{P}_1$, un noeud $\boldsymbol{z}_i$ n'est en relation qu'avec ses voisins directs, c'est-à-dire les noeuds $\boldsymbol{z}_j$ avec lesquels il forme une arête. Les éléments (i, j) tels que $i \neq j$ des matrices $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$ sont donc nuls dès lors que les noeuds $\boldsymbol{z}_i$ et $\boldsymbol{z}_j$ ne forment pas une arête. A titre d'exemple, une triangulation régulière compte en moyenne 6 voisins pour chaque noeud (en dimension 2). Cela signifie que $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$ comptent en moyenne 7 non-zéros par ligne (figure 3.1).

En assimilation de données, la matrice de corrélation d'erreurs d'observation $\boldsymbol{R}$ est de taille $p \times p$, où p est le nombre d'observations. C'est une matrice *a priori* dense qui ne peut être stockée en mémoire pour des grandes valeurs de p (typiquement $p > 10^7$). La méthode des éléments finis permet de ne pas stocker ces coefficients. A la place, la matrice $\boldsymbol{R}$ est modélisée à partir des deux matrices $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$ dont la taille augmente comme $\mathcal{O}(p)$. Plus précisément, $\boldsymbol{R}^{-1}$ est représentée par la formule (2.36), c'est-à-dire comme une série de produits par $(\boldsymbol{M}_c + \boldsymbol{K}_c)$ et $\boldsymbol{M}_c^{-1}$. Si la matrice $(\boldsymbol{M}_c + \boldsymbol{K}_c)$ est facile à stocker en mémoire et à appliquer à un vecteur, la matrice $\boldsymbol{M}_c^{-1}$ s'obtient en revanche en inversant la matrice $\boldsymbol{M}_c$. Heureusement, la structure symétrique définie positive de $\boldsymbol{M}_c$ la rend adaptée à la factorisation de Cholesky [Golub and Van Loan, 1996]. Si toutefois cette factorisation n'est pas souhaitée, il est toujours possible de diagonaliser $\boldsymbol{M}_c$ en utilisant la condensation de masse. Cette technique est présentée dans la section 3.2.

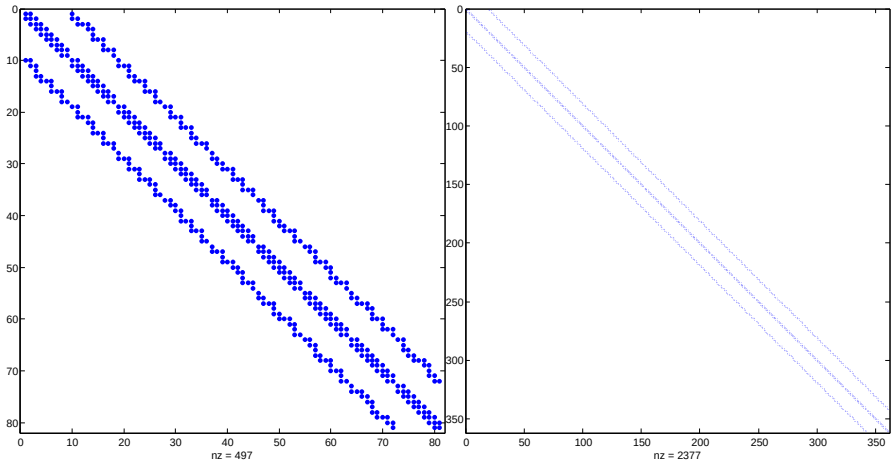


FIGURE 3.1 – Profil de la matrice de masse $\mathbb{P}_1$. A gauche : pour $p = 81$ points. A droite : pour $p = 360$ points.

Les opérateurs $\boldsymbol{R}$ et $\boldsymbol{R}^{-1}$ sont représentés grâce à des matrices creuses. Néanmoins, cela ne signifie pas qu'elle ne sont pas elles-mêmes denses (figure 3.2). En particulier, la matrice de corrélation $\boldsymbol{R}$ est totalement dense, même en recourant à une matrice de masse diagonale. Son application sous la forme du produit (2.35) nécessite d'appliquer m fois $(\boldsymbol{M}_c + \boldsymbol{K}_c)^{-1}$ et $m - 1$ fois $\boldsymbol{M}_c$. Pour appliquer $(\boldsymbol{M}_c + \boldsymbol{K}_c)^{-1}$, il est possible de recourir à une factorisation de Cholesky. Une solution alternative consiste à utiliser une méthode itérative de type gradient conjugué ou les itérations de Chebyshev (chapitre 6). Cependant, le conditionnement de $\boldsymbol{M}_c + \boldsymbol{K}_c$ est susceptible d'introduire des erreurs numériques dans les deux cas, en particulier lorsque l'assemblage des matrices s'appuie sur un maillage irrégulier. Il est donc utile de préconditionner le système linéaire par $\boldsymbol{M}_c^{-1}$ afin d'exploiter la propriété (2.61). Dans ce cas, il est commun de réécrire $\boldsymbol{R}$ sous la forme

$$\boldsymbol{R} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Gamma}^{1/2} \times \boldsymbol{C} \times \boldsymbol{\Gamma}^{1/2}\boldsymbol{\Sigma}^{1/2}, \tag{3.1}$$

avec

$$\boldsymbol{C} = [(\boldsymbol{M}_c + \boldsymbol{K}_c)^{-1}\boldsymbol{M}_c]^m \times \boldsymbol{M}_c^{-1}. \tag{3.2}$$

On remarque que la relation (3.2), avec l'introduction de $\boldsymbol{M}_c\boldsymbol{M}_c^{-1}$ à la fin de l'expression, se rapproche de la forme présentée dans Weaver and Courtier [2001].

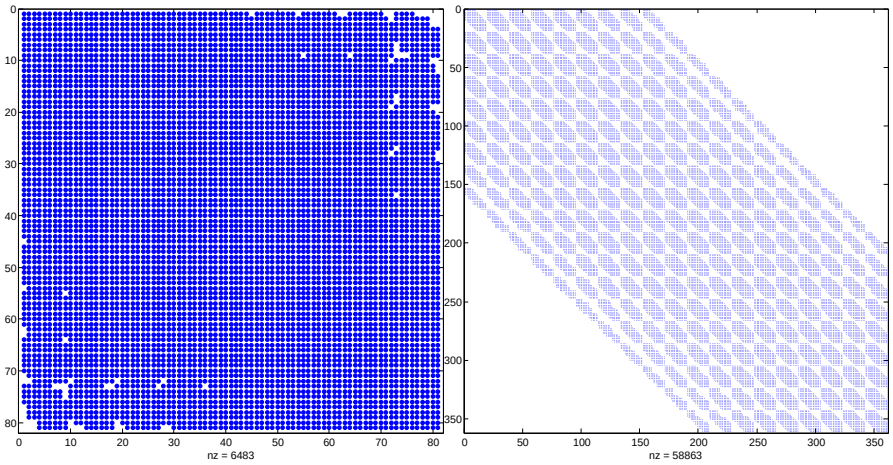


FIGURE 3.2 – Profil de la matrice $\boldsymbol{R}^{-1}$. A gauche : pour $p = 81$ points. A droite : pour $p = 360$ points.

3.1.2 Comparaison avec le modèle théorique

Considérons l'opérateur de corrélation (3.2) discrétisé sur un maillage régulier par la méthode des éléments finis. La théorie introduite dans le chapitre 1 permet de comparer les fonctions de corrélation représentées par cet opérateur discret aux fonctions de Matérn, solutions de l'équation de diffusion dans le cadre continu. Dans cette section, on étudie les fonctions représentées dans l'intérieur du domaine d'étude $\mathcal{D}$, à l'inverse de la section 3.3 qui traite des aspects numériques près des bords du domaine.

Notons $\delta_{\boldsymbol{z}}$ la distribution de Dirac centrée en $\boldsymbol{z} \in \mathcal{D}$ définie pour toute fonction $f \in H^1(\mathcal{D})$ par [Strichartz, 2003, Hewitt and Stromberg, 1975, Schwartz and Melese, 1951] :

$$\langle \delta_{\boldsymbol{z}}, f \rangle_{H^{-1},H^1} = f(\boldsymbol{z}). \tag{3.3}$$

Soit $\boldsymbol{z} \in \mathcal{D}$ un point situé suffisamment loin des frontières de $\mathcal{D}$. La notion d'éloignement sera définie plus précisément dans la section 3.3. Pour l'instant, supposons simplement que les expériences présentées ici ne s'inquiètent pas d'éventuels effets de bords.

La fonction de Matérn (1.21) est la réponse impulsionnelle de l'équation de diffusion normalisée (1.42) avec la condition initiale $f_0 = \delta_{\boldsymbol{z}}$. En d'autre termes, l'expression (1.21) résulte de l'application de l'opérateur de corrélation à la distribution $\delta_{\boldsymbol{z}}$ [Green, 1828, Hazewinkel, 2012].

Supposons maintenant que $\boldsymbol{z} = \boldsymbol{z}_i$ coïncide avec un noeud du maillage. La projection de $\delta_{\boldsymbol{z}}$ sur l'espace d'éléments finis V^c donne lieu à l'approximation $\delta_{\boldsymbol{z}} = \delta_{\boldsymbol{z}_i} \simeq \varphi^i$. Le vecteur de coordonnées $\mathfrak{d}$ correspondant comporte ainsi un « 1 » en position i et des « 0 » ailleurs. En multipliant ce vecteur $\mathfrak{d}$ par la matrice de corrélation $\boldsymbol{C}$, on obtient donc un vecteur $\boldsymbol{\alpha}$ dont les éléments sont exactement ceux de la i-ème colonne de $\boldsymbol{C}$. En faisant usage de la formule de reconstitution $f = \Phi(\boldsymbol{\alpha})$, on visualise une approximation linéaire par morceaux de la fonction de corrélation (1.21) centrée en $\boldsymbol{z}_i$. Ce résultat est représenté sur la figure 3.3.

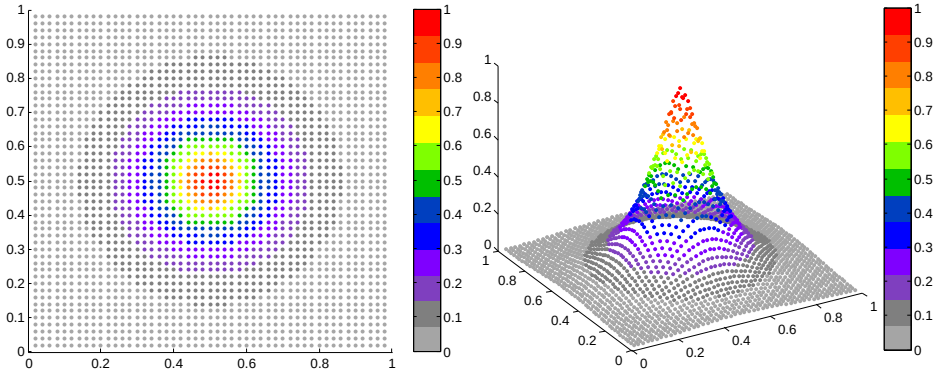


Figure 3.3 – Fonctions de corrélation de Matérn sur maillage cartésien. A gauche : en 2D. A droite : en perspective.

On remarque que l'amplitude de la réponse impulsionnelle n'est pas exactement égale à 1. Ceci est dû à l'approximation numérique des éléments finis. Néanmoins, lorsque le maillage est raffiné, c'est-à-dire que le diamètre du plus grand élément de $\mathcal{T}$ diminue :

$$\max_{\tau \in \mathcal{T}} \operatorname{diam}(\tau) \to 0, \tag{3.4}$$

l'amplitude tend également vers 1 et l'approximation de la fonction de corrélation se rapproche de sa valeur théorique $c_{m,l}$.

3.2 Introduction de la condensation de masse

On présente ici plusieurs procédés visant à transformer la matrice de masse $\boldsymbol{M}_c$ (habituellement nommée *consistent mass matrix* en anglais) en une matrice de masse diagonale $\boldsymbol{M}_c^{\#}$ (*lumped mass matrix*). Ces procédés sont couramment utilisés dans la résolution d'équations aux dérivées partielles pour éviter de factoriser la matrice de masse. La construction de $\boldsymbol{M}_c^{\#}$ passe soit par un raisonnement physique en faisant allusion à la répartition de la masse dans un système mécanique, soit par un raisonnement mathématique en faisant usage de formules de quadratures [Zienkiewicz et al., 2005]. Dans cette section, trois interprétations de la condensation de masse sont présentées : la distribution de la masse globale sur les noeuds du maillages et la condensation de masse ligne par ligne d'une part, et les formules de quadrature d'autre part. Enfin, la représentation numérique des fonctions de corrélation est comparée à l'attente théorique et à l'approche classique.

3.2.1 Principe de la condensation de masse

On considère dans la suite que $\boldsymbol{M}_c$ est issue de la discrétisation en éléments finis $\mathbb{P}_1$. Présentons d'abord la construction de $\boldsymbol{M}_c^{\#}$ par redistribution de masse.

A l'intérieur de chaque triangle du maillage et pour tout $i \in [\![1, p]\!]$, la répartition de la masse associée au noeud $\boldsymbol{z}_i$ est spécifiée par la fonction de pondération φ_i. En éléments finis $\mathbb{P}_1$, cette répartition n'est pas uniforme puisque les $(\varphi_i)_{i\in[\![1,p]\!]}$ sont des « fonctions chapeaux » (figure 2.4).

Soit $f \in V_c$. La valeur $f(\boldsymbol{z})$, pour $\boldsymbol{z} \in \tau = (\boldsymbol{z}_i, \boldsymbol{z}_j, \boldsymbol{z}_k)$ dépend linéairement des valeurs $f(\boldsymbol{z}_i), f(\boldsymbol{z}_j)$ et $f(\boldsymbol{z}_k)$. La force de cette dépendance est spécifiée par les fonctions φ_i, φ_j et φ_k. En particulier, une partie de la masse est associée à l'**intérieur** du triangle. Il en résulte que les termes croisés $(\boldsymbol{M}_c)_{\text{ij}}$ de la matrice de masse ne sont pas nuls lorsque $i \neq j$, et donc que la matrice de masse n'est pas diagonale (figure 3.4).

Pour remédier à ce caractère non diagonal, il est possible de changer la distribution de masse de façon équivalente, de telle sorte que les termes croisés soient nuls. Cela revient à substituer de nouvelles fonctions de pondération aux fonctions $(\varphi_i)_{i\in[\![1,p]\!]}$ dans le calcul de $(\boldsymbol{M}_c)_{\text{ij}}$. Dans chaque triangle et pour chaque fonction de base φ_i, on choisit donc de faire porter la masse

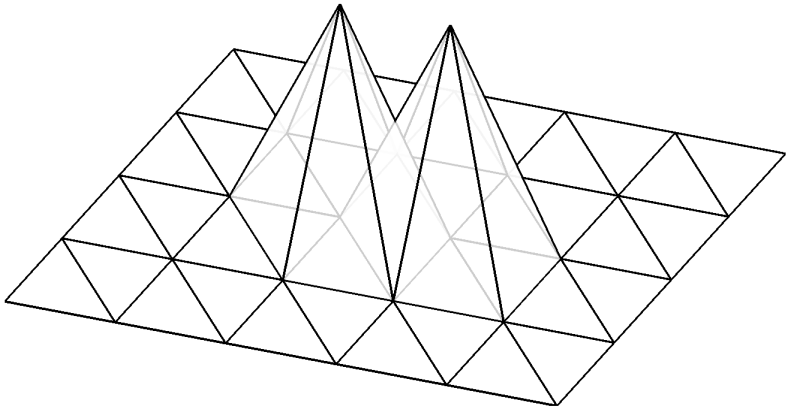


FIGURE 3.4 – Chevauchement des supports en éléments finis $\mathbb{P}_1$, responsable du caractère non-diagonal de la matrice de masse.

élémentaire

$$(\boldsymbol{M}_c^{\#})_{\mathrm{ii}} = \sum_{j\in[\![1,p]\!]} (\boldsymbol{M}_c)_{\mathrm{ij}} = \sum_{j\in[\![1,p]\!]} \int_{\mathcal{D}} \varphi_i(\boldsymbol{z}')\varphi_j(\boldsymbol{z}')\mathrm{d}\boldsymbol{z}' = \int_{\mathcal{D}} \varphi_i(\boldsymbol{z}')\mathrm{d}\boldsymbol{z}' \qquad (3.5)$$

par le noeud $\boldsymbol{z}_i$ uniquement (voir figure 3.5). Du point de vue global, cela revient à affecter à chaque noeud l'aire de sa cellule de Voronoï modifiée. La cellule de Voronoï modifiée est calculée en reliant les centres de gravité des triangles voisins. Elle est différente de la cellule de Voronoï obtenue en reliant les centres des cercles circonscrits (voir section 4.3.1). Cette différence a des conséquences dans le schéma de corrélation à deux niveaux présenté dans le chapitre 5 (section 5.5). La matrice $\boldsymbol{M}_c^{\#}$ ainsi obtenue est diagonale.

Une autre approche de la redistribution de la masse, globale cette fois, consiste à considérer la masse totale $\pi = \sum_{(i,j)\in[\![1,p]\!]^2} (\boldsymbol{M}_c)_{\mathrm{ij}}$ et la répartir sur les noeuds du maillage. Une stratégie connue est de créer une matrice $\boldsymbol{M}_c^{\#}$ diagonale proportionnelle à la diagonale de $\boldsymbol{M}_c$. Toutefois, cette approche globale n'est pas fondée sur un critère physique rigoureux.

Terminons cette sous-section en exposant la procédure de condensation de masse ligne par ligne (figure 3.6). Le procédé est simple, la formule (3.5) est implémentée ligne par ligne, sans se préoccuper de l'interprétation fonctionnelle citée précédemment. Dans le cas des éléments finis de type $\mathbb{P}_1$, cette méthode est en fait la même que la redistribution locale de la masse décrite précédemment, où la masse de chaque élément est portée uniquement par

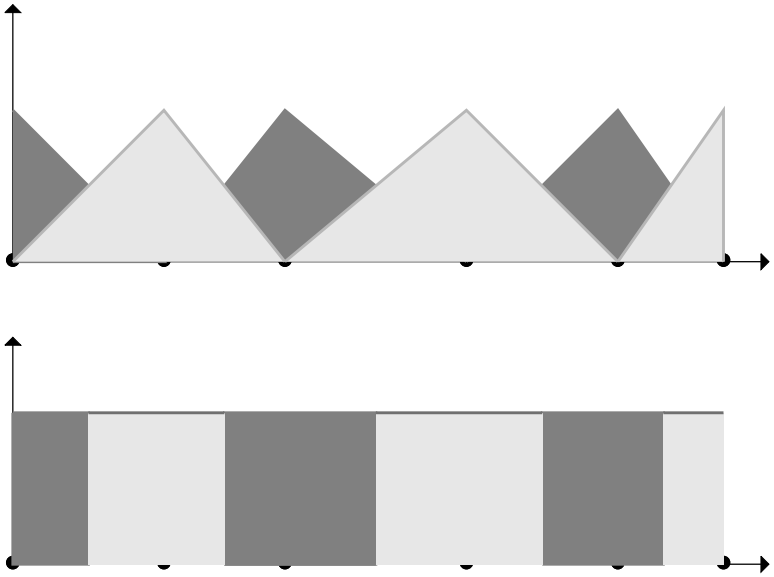


Figure 3.5 – Illustration en dimension 1. En haut : fonctions $\mathbb{P}_1$ sur maillage irrégulier (points noirs). En bas : fonctions condensées sur le même maillage.

les noeuds du maillage. Dans le cas des éléments finis d'ordre supérieur, les méthodes ne sont équivalentes que si les degrés de liberté sont situés aux points de condensation (*i.e.*, les $(\boldsymbol{z}_i)_{i\in[\![1,p]\!]}$).

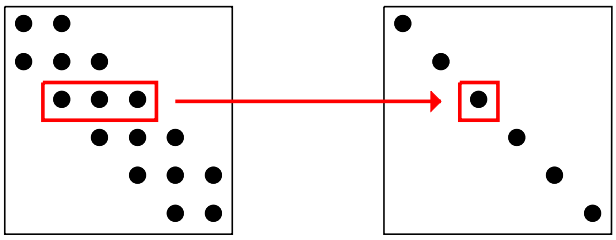


Figure 3.6 – Condensation de masse ligne par ligne.

3.2.2 Diagonalisation par quadrature

L'utilisation de formules de quadrature permet de fournir un cadre mathématique à la condensation de masse [Zienkiewicz et al., 2005]. L'idée est de calculer les éléments de $\boldsymbol{M}_c$ de manière approchée, au travers d'une formule de quadrature du type

$$(\boldsymbol{M}_c)_{\mathrm{ij}} = \int_{\mathcal{D}} \varphi_i(\boldsymbol{z}')\varphi_j(\boldsymbol{z}')\mathrm{d}\boldsymbol{z}' \simeq \sum_{\alpha} \omega_\alpha \varphi_i(\boldsymbol{z}_\alpha)\varphi_j(\boldsymbol{z}_\alpha). \tag{3.6}$$

Dans cette somme, les $\boldsymbol{z}_\alpha$ sont appelés « points de quadrature », tandis que les ω_α sont des poids à déterminer. Tout l'enjeu est donc de trouver les $(\boldsymbol{z}_\alpha)_\alpha$ et les $(\omega_\alpha)_\alpha$ de telle sorte que la matrice ${\boldsymbol{M}_{\mathrm{c}}}^{\#}$ ainsi estimée soit diagonale, sans pour autant dégrader l'ordre de la méthode de discrétisation. On retrouve donc le fait que les degrés de libertés doivent être situés aux points de quadrature [Ern and Guermond, 2010].

Dans le cas des méthodes spectrales, la formule de quadrature approxime l'intégrale sur tout le domaine $\mathcal{D}$. Dans le cas des éléments finis et des éléments finis spectraux en revanche, la formule est appliquée au calcul de l'intégrale sur l'élément triangulaire τ. On notera que dans le cas des éléments de type $\mathbb{P}_1$, l'utilisation d'une formule de quadrature revient au même que la re-distribution locale de la masse ou encore la sommation ligne par ligne. Ce n'est plus le cas quand l'ordre augmente. En particulier, la grande difficulté réside en la détermination des points de quadrature.

En pratique, on utilise la condensation de masse en remplaçant $\boldsymbol{M}_c$ par la matrice condensée $\boldsymbol{M}_c^{\#}$, mais sans toucher à la matrice $\boldsymbol{K}_c$. Sous ces hypothèses, le système (2.28) se réécrit

$$(\boldsymbol{M}_c^{\#} + \boldsymbol{K}_c)\boldsymbol{\alpha}_{n+1} = \boldsymbol{M}_c^{\#}\boldsymbol{\alpha}_n. \tag{3.7}$$

L'avantage de cette formulation est l'accessibilité de $(\boldsymbol{M}_c^{\#})^{-1}$ qui est une matrice diagonale.

3.2.3 Comparaison avec le modèle théorique

Pour illustrer les performances de la condensation de masse pour la représentation des fonctions de corrélation sur un maillage structuré, on reprend les expériences de la sous-section 3.1.2. Cette fois-ci, l'opérateur de corrélation usant de la condensation de masse est noté $\boldsymbol{C}^{\#}$. Il est égal à

$$\begin{aligned} \boldsymbol{C}^{\#} &= [(\boldsymbol{M}_c^{\#} + \boldsymbol{K}_c)^{-1}\boldsymbol{M}_c^{\#}]^m \times (\boldsymbol{M}_c^{\#})^{-1} \\ &= (\boldsymbol{M}_c^{\#} + \boldsymbol{K}_c)^{-1} \times \boldsymbol{M}_c^{\#} \times \ldots \times \boldsymbol{M}_c^{\#} \times (\boldsymbol{M}_c^{\#} + \boldsymbol{K}_c)^{-1}. \end{aligned} \tag{3.8}$$

La réponse impulsionnelle de l'opérateur (3.8) est représentée par la figure 3.7.

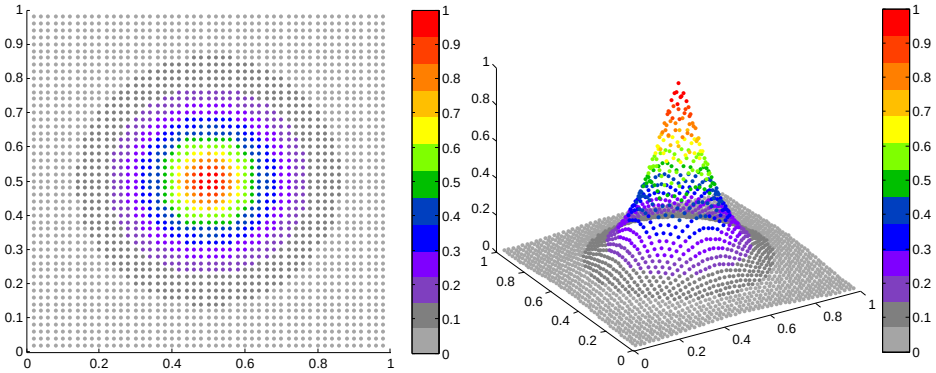


FIGURE 3.7 – Fonctions de corrélation de Matérn sur maillage cartésien. Utilisation de la condensation de masse. A gauche : en 2D. A droite : en perspective. On constate la forte ressemblance avec la figure 3.3, ce qui indique que la condensation de masse ne dégrade pas la qualité de l'approximation.

3.3 Corrections analytiques aux frontières

Pour simplifier l'étude des conditions aux limites sans perte de généralité, on choisit de se placer dans le cas unidimensionnel $d = 1$. De même, l'équation de diffusion (1.42) pouvant être vue comme la discrétisation implicite de l'équation de diffusion en temps continu (2.52), on se contente d'étudier les solutions de cette dernière. Les phénomènes de bords étant dus aux propriétés spatiales de l'équation et du domaine, les conclusions tirées dans cette section s'appliquent directement au cas de la diffusion semi-implicite.

3.3.1 Solutions analytiques en temps continu

Quand on modélise des opérateurs de corrélation, deux types de conditions aux limites sont couramment envisagés : les conditions de Dirichlet (valeurs imposées, figure 3.8) et les conditions de Neumann (flux imposés, figure 3.9). Dans les deux cas, la réponse impulsionnelle de l'équation de diffusion n'est pas d'amplitude unitaire près des bords du domaine [Mirouze and Weaver, 2010]. On attire l'attention sur le fait que ce phénomène est indépendant de la méthode de discrétisation du Laplacien. Ces résultats sont

exploités dans la sous-section 3.3.2 pour proposer une correction analytique de l'amplitude près des bords du domaine.

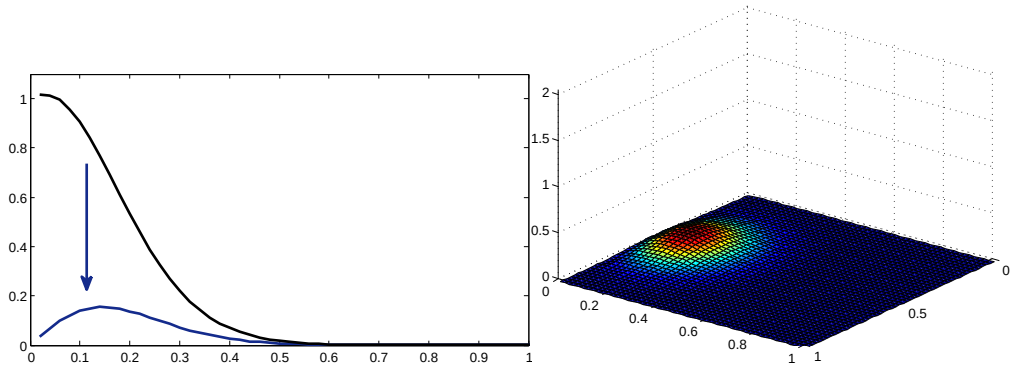


FIGURE 3.8 – Illustration de l'effet absorbant des conditions aux limites de Dirichlet. A gauche : principe schématique. A droite : Réponse impulsionnelle sur maillage cartésien en dimension 2.

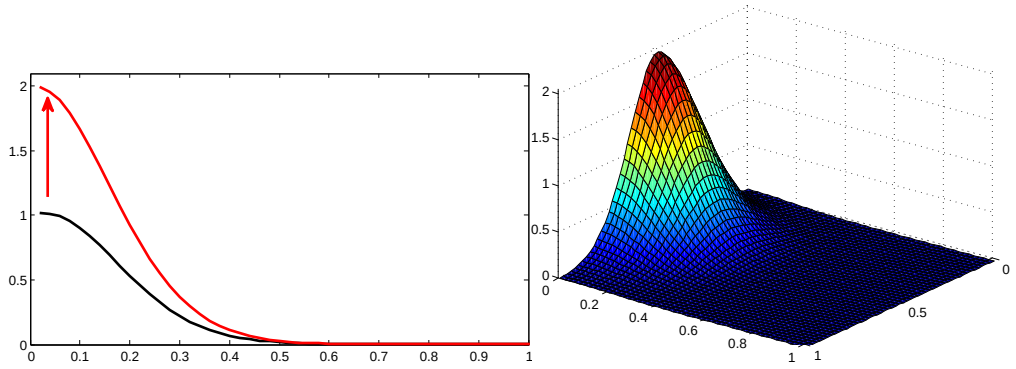


FIGURE 3.9 – Illustration de la réflexion par les conditions aux limites de Neumann. A gauche : principe schématique. A droite : Réponse impulsionnelle sur maillage cartésien en dimension 2.

Soit l'équation de diffusion en pseudo-temps continu, d'inconnue $u = u(\boldsymbol{z}, s)$, définie sur le domaine compact $[0, \theta] \times [0, +\infty[$:

$$\begin{cases} \left(\frac{\partial}{\partial s} - l^2\Delta\right) u = 0 \\ u(\boldsymbol{z}, 0) = \delta(\boldsymbol{z} - \boldsymbol{z}_0) \end{cases}, \tag{3.9}$$

où $\boldsymbol{z}_0 \in [0, \theta]$ et $\theta > 0$. La variable s est une convenance de modélisation, mais ne porte aucun sens physique, ainsi la qualifie-t-on de « pseudo-temps ».

Conformément à Mirouze and Weaver [2010], la forme générale de la solution de (3.9) est donnée par :

$$u(\boldsymbol{z}, s) = [a\cos(\xi \boldsymbol{z}) + b\sin(\xi \boldsymbol{z})]e^{-\xi^2 s l^2}. \tag{3.10}$$

Les constantes a, b et ξ sont déterminées à l'aide des conditions aux limites et des conditions initiales. Dans le cas des conditions de Neumann, on obtient

$$\frac{\partial u_N}{\partial n} = 0 \quad \text{en} \quad \{0, \theta\}, \tag{3.11}$$

où $\frac{\partial \cdot}{\partial n}$ désigne la dérivée par rapport à la normale sortante aux bord de $\mathcal{D}$. La solution impulsionnelle u_N est alors donnée par :

$$\begin{aligned} u_N(\boldsymbol{z}, s) &= \frac{1}{2\theta} + \sum_{n=1}^{\infty} \frac{1}{\theta} e^{-(n\pi/\theta)^2 s l^2} \cos\left[\frac{n\pi}{\theta}(\boldsymbol{z} - \boldsymbol{z}_0)\right] \\ &+ \frac{1}{2\theta} + \sum_{n=1}^{\infty} \frac{1}{\theta} e^{-(n\pi/\theta)^2 s l^2} \cos\left[\frac{n\pi}{\theta}(\boldsymbol{z} + \boldsymbol{z}_0)\right]. \end{aligned} \tag{3.12}$$

De même, avec les conditions de Dirichlet :

$$u_D = 0 \quad \text{en} \quad \{0, \theta\} \tag{3.13}$$

et

$$\begin{aligned} u_D(\boldsymbol{z}, s) &= \frac{1}{2\theta} + \sum_{n=1}^{\infty} \frac{1}{\theta} e^{-(n\pi/\theta)^2 s l^2} \cos\left[\frac{n\pi}{\theta}(\boldsymbol{z} - \boldsymbol{z}_0)\right] \\ &- \frac{1}{2\theta} - \sum_{n=1}^{\infty} \frac{1}{\theta} e^{-(n\pi/\theta)^2 s l^2} \cos\left[\frac{n\pi}{\theta}(\boldsymbol{z} + \boldsymbol{z}_0)\right]. \end{aligned} \tag{3.14}$$

On remarque la présence des signes « $-$ » dans cette dernière formule. Les solutions aux deux problèmes ont donc des formes très proches qui se simplifient quand on les somme. Mais d'abord, étudions la fonction gaussienne $h(z, t)$, solution asymptotique de notre modèle de diffusion :

$$h(\boldsymbol{z}, s) = \frac{1}{\sqrt{4\pi l^2 s}} \exp\left(\frac{-z^2}{4sl^2}\right). \tag{3.15}$$

Sous la condition que la fonction soit 2θ-périodique et que $\sqrt{2sl^2} \ll l$ [Mirouze and Weaver, 2010], h admet le développement en série :

$$\begin{aligned} h(\boldsymbol{z}, s) &= a_0 + \sum_{n=1}^{\infty} \left[a_n(s) \cos\left(\frac{n\pi}{\theta}\boldsymbol{z}\right) + b_n(s) \sin\left(\frac{n\pi}{\theta}\boldsymbol{z}\right) \right] \\ &= \frac{1}{2\theta} + \sum_{n=1}^{\infty} \frac{1}{\theta} e^{-(n\pi/\theta)^2 sl^2} \cos\left[\frac{n\pi}{\theta}\boldsymbol{z}\right]. \end{aligned} \tag{3.16}$$

Ainsi, on peut reconstruire h à l'aide des fonctions u_N et u_D. Trois égalités sont à disposition :

$$u_N(\boldsymbol{z}, s) = h(\boldsymbol{z} - \boldsymbol{z}_0, s) + h(\boldsymbol{z} + \boldsymbol{z}_0, s), \tag{3.17}$$

$$u_D(\boldsymbol{z}, s) = h(\boldsymbol{z} - \boldsymbol{z}_0, s) - h(\boldsymbol{z} + \boldsymbol{z}_0, s), \tag{3.18}$$

$$h(\boldsymbol{z} - \boldsymbol{z}_0, s) = \frac{1}{2}[u_N(\boldsymbol{z}, s) + u_D(\boldsymbol{z}, s)]. \tag{3.19}$$

De ces égalités vont naître deux idées pour corriger la valeur des fonctions de corrélation près des frontières du domaine.

3.3.2 Corrections analytiques

D'après l'équation

$$u_N(\boldsymbol{z}) = h(\boldsymbol{z} - \boldsymbol{z}_0) + h(\boldsymbol{z} + \boldsymbol{z}_0) \tag{3.20}$$

(la dépendance en s est omise car elle n'est pas une variable d'espace), on constate que l'amplitude de u_N est exactement deux fois égale à celle de h près des bords de $\mathcal{D} = [0, \theta]$. Cette dégénérescence s'observe en pratique, comme l'illustre la figure 3.9.

La fonction de corrélation gaussienne $g(\boldsymbol{z})$, de longueur de portée l, est définie par :

$$g(\boldsymbol{z} - \boldsymbol{z}_0) = e^{-(\boldsymbol{z} - \boldsymbol{z}_0)^2 / 2l^2}. \tag{3.21}$$

Lorsque la longueur de portée l est grande devant la distance $\|\boldsymbol{z} - \boldsymbol{z}_0\|_2$, la fonction de corrélation modélisée est additionnée d'une contribution due au terme $g(\boldsymbol{z} + \boldsymbol{z}_0)$. Ainsi, au lieu d'être égale à 1, la normalisation en $\boldsymbol{z}_0$ est égale à :

$$c(\boldsymbol{z}_0) = g(0) + g(2\boldsymbol{z}_0) = 1 + g(2\boldsymbol{z}_0). \tag{3.22}$$

On peut donc corriger (3.22) en divisant $c(\boldsymbol{z}_0)$ par $1 + g(2z_0)$ en chaque $\boldsymbol{z}_0$. Une analyse similaire menée avec les conditions de type Dirichlet mène à une

correction analogue en $1-g(2\boldsymbol{z}_0)$. Cependant, deux écueils sont à mentionner : cette renormalisation n'est valide que si la fonction de corrélation est proche d'une gaussienne (donc pour des valeurs de m élevées) et uniquement si la longueur de portée l ne dépend pas de z_0.

Une deuxième idée consiste à résoudre (3.9) avec les deux types de conditions aux limites (figure 3.10), puis d'effectuer la moyenne des solutions obtenues, en vertu de l'équation :

$$h(\boldsymbol{z}-\boldsymbol{z}_0)=\frac{1}{2}[u_N(\boldsymbol{z})+u_D(\boldsymbol{z})]. \tag{3.23}$$

La méthode est précisément deux fois plus coûteuse et présente un inconvénient majeur : on perd *a priori* la possibilité d'inverser l'opérateur de corrélation, alors que c'est la motivation principale qui pousse à utiliser un opérateur de diffusion implicite [Mirouze and Weaver, 2010].

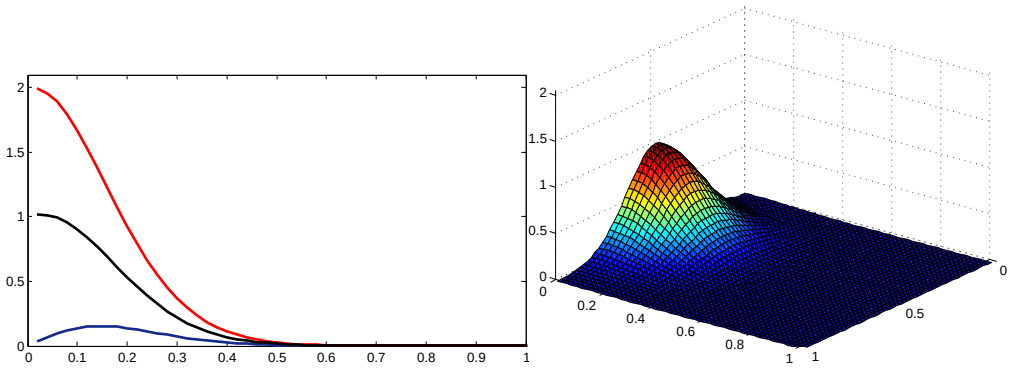


Figure 3.10 – Illustration de la somme Dirichlet + Neumann. A gauche : principe schématique. A droite : Réponse impulsionnelle sur maillage cartésien en dimension 2.

3.4 Synthèse sur la spécificité des maillages structurés

Avant de conclure ce chapitre, on fait le point sur la spécificité des maillages structurés, afin de mieux préparer le lecteur au chapitre 4.

Le premier point abordé est celui de la condensation de masse. Nous avons montré que ce procédé permettait d'éviter l'inversion d'une matrice de masse non diagonale, cela sans dégrader la précision de l'aproximation en éléments finis. En pratique, cette propriété reste vraie tant que le maillage reste homogène (la taille des éléments varie peu) et ne contient pas de triangles aplatis. Il faudra donc être très attentif quand on utilisera la condensation de masse sur des maillages non structurés.

Le deuxième point abordé concerne les conditions aux limites du domaine de résolution de l'équation de diffusion. En effet, imposer des conditions aux limites est obligatoire pour que l'équation soit bien posée, et a pour effet de déformer la réponse impulsionnelle proche des frontières du domaine. Ce phénomène est indépendant du type de maillage, car il est de nature analytique. Ainsi, on retrouvera le même effet sur des maillages non structurés.

3.5 L'essentiel du chapitre

Le premier moyen de vérifier le contenu d'une matrice de corrélation est de visualiser ses colonnes. C'est ce qu'on effectue en étudiant la réponse impulsionnelle de l'opérateur de diffusion. Cette première expérience permet de montrer en pratique la proximité entre le modèle de diffusion et le modèle de Matérn en pratique. C'est aussi un moyen d'illustrer les effets de bords et la condensation de masse, appoximation couramment utilisée en conjonction avec la méthode des éléments finis.

Ce qu'il faut retenir :

- Les matrices d'éléments finis sont creuses. Pourtant, leurs inverses ne le sont pas forcément.
- Il est donc exclu de calculer les coefficients de ces inverses.
- Pour inverser la matrice de masse, il est possible de la diagonaliser en faisant appel à la condensation de la masse.
- La qualité de la modélisation est soumise à l'effet des conditions aux limites du domaine.
- Dans la pratique, on utilise les conditions de Neumann, conservatives et faciles à normaliser.

Chapitre 4

Application aux données du sondeur SEVIRI

Ce chapitre est l'un des plus importants du manuscrit. Le chapitre 3 a montré comment la théorie développée dans la partie I s'adaptait aux données reposant sur des maillages réguliers. Il s'agit maintenant de valider cette modélisation sur des maillages non structurés. Pour ce faire, on met en place des expériences construites à partir d'observations satellites réelles. On montre ainsi les limites de l'approche basée sur la discrétisation de l'équation de diffusion en éléments finis, et on met en avant les éventuels écueils qui feront l'enjeu de la partie III.

La section 4.1 décrit les données utilisées dans les expériences. Ces données subissent un certain nombre de prétraitements en amont de leur assimilation, qui sont la cause de leur structure spatiale hétérogène. Dans la section 4.2, on estime les paramètres de corrélation en ajustant le modèle de Matérn au diagnostic de Desroziers et al. [2005]. On évoque aussi la possibilité de superposer plusieurs modèles pour mieux décrire les données.

La section 4.3 expose les grands principes de la génération de maillage. On y précise la notion de « bon maillage », cruciale dans l'interprétation de nos résultats. Ces aspects de géométrie plane sont repris dans la section 4.4, lors de la validation de notre méthode sur le maillage construit à partir des données satellitaires. L'objectif est de retrouver des résultats similaires à ceux du chapitre 3, tout en s'efforçant d'expliquer l'origine des erreurs de modélisation observées dans certaines régions du domaine d'étude. Le traitement de ces erreurs fait l'objet de la section 4.5, laquelle s'attache en particulier aux méthodes visant à corriger l'amplitude des fonctions de corrélations

modélisées.

4.1 Présentation des données de SEVIRI

Le capteur SEVIRI (Spinning Enhanced Visible and InfraRed Imager) est un radiomètre à balayage embarqué sur les satellites géostationnaires MSG (Meteosat Second Generation). Il dispose de 12 canaux d'acquisition permettant d'observer le rayonnement émis par la Terre dans 12 bandes spectrales quasi-distinctes (figure 4.1), conformément aux objectifs de la mission MSG [Aminou et al., 2003]. En particulier, 8 de ces canaux sont situés dans le domaine infrarouge (d'où le nom du sondeur), ce qui permet d'obtenir une grande quantité de données concernant la température de surface de la mer, de la terre et des nuages. La résolution spatiale du sondeur est égale à 3 kilomètres au nadir [1] (le point situé sous la satellite, directement à la verticale du sondeur), et se dégrade avec la latitude. Une image complète de SEVIRI comporte approximativement 10^7 pixels par canal.

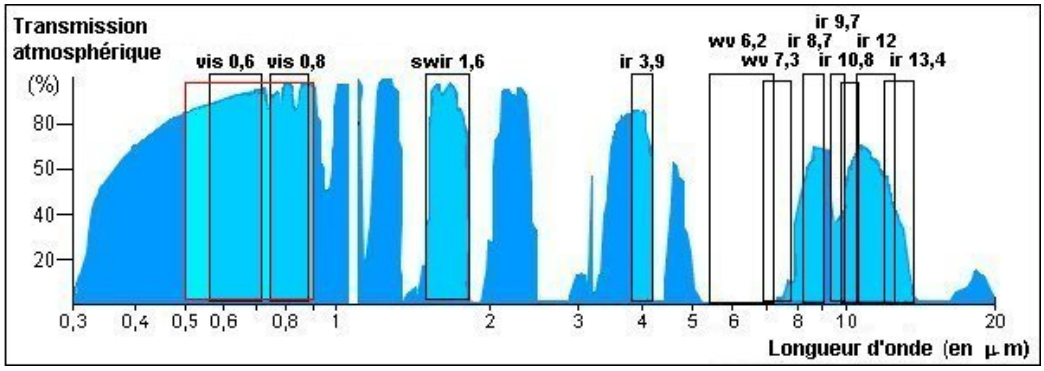


FIGURE 4.1 – Canaux d'observation du sondeur SEVIRI superposés au spectre de transmission de l'atmosphère.

Le sondeur SEVIRI couvre un domaine large de 11000 km (figure 4.2). Ses observations sont assimilées dans les modèles de prévision numérique du temps, tels que ARPÈGE et ARÔME. Nos expériences s'appuient sur le domaine ARÔME, qui couvre la France métropolitaine et son entourage proche (4.3). Le nombre typique d'observations disponibles par canal dans ce domaine est de l'ordre de 10^4-10^5.

1. Pour tous les canaux sauf HRV, pour lequel la résolution est de 1 km au nadir. Le nombre de pixels par image est également différent pour ce canal.

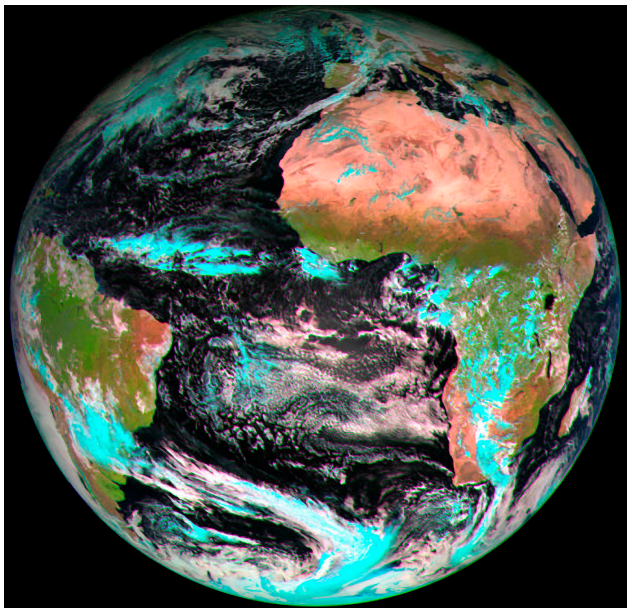


FIGURE 4.2 – Domaine spatial vu par SEVIRI.

On fait l'hypothèse que les données [2] provenant d'instruments différents ne sont pas corrélées entre elles. De même, on suppose que les données issues de canaux différents d'un même sondeur peuvent être corrélées, mais que ces corrélations peuvent être prises en compte indépendamment des corrélations spatiales [Michel, 2018] [3]. Ainsi, on fait l'hypothèse que la matrice de corrélation d'erreurs d'observation est diagonale par bloc. Chaque bloc correspond à un canal d'un instrument. Dans la suite du document, quand on fait référence à « la matrice $\boldsymbol{R}$ », on fait en fait référence à un bloc de cette matrice, sous-entendant que chaque bloc peut être traité indépendamment et de manière analogue (au choix des paramètres près).

Dans leur étude, Waller et al. [2016a], Michel [2018] ont montré que les observations de SEVIRI comportaient d'importantes corrélations horizontales d'erreurs d'observation dans le canal 2. On choisit donc d'utiliser ces mêmes données pour réaliser nos expériences. Les paramètres de corrélation correspondant sont estimés dans la section 4.2.

2. Terme abusif. Comprendre « les erreurs contenues dans les données ».

3. En pratique, la situation est légèrement plus compliquée. Pour pouvoir faire la séparation entre les corrélations verticales (intercanaux) et les corrélations horizontales, il faut supposer que le *thinning* est indépendant du niveau vertical. Lorsque ce n'est pas le cas, il est envisageable de modéliser des corrélations 3D, toujours avec des techniques de diffusion.

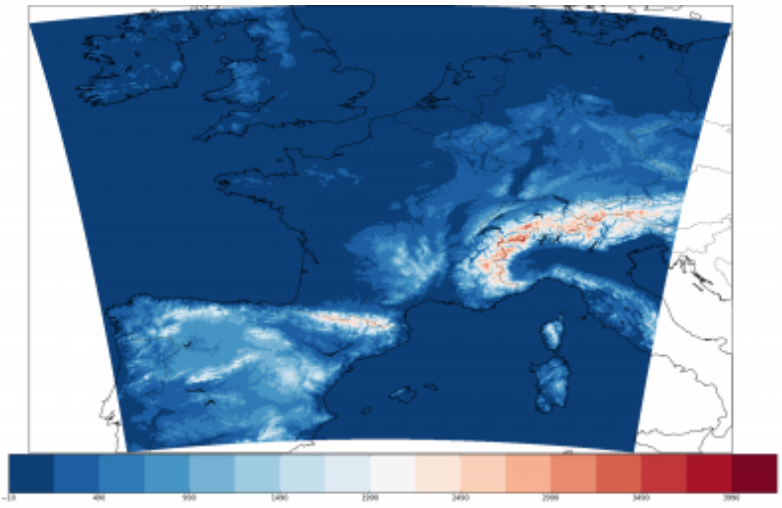


FIGURE 4.3 – Domaine spatial observé par SEVIRI, restreint au domaine d'AROME-FRANCE

Chaque image de SEVIRI est soumise à un certain nombre de prétraitements avant d'être assimilée dans le modèle de prévision numérique du temps. La première étape est celle du *thinning*. L'objectif est d'assurer que les données sont suffisamment éloignées les unes des autres pour qu'on puisse les considérer non-corrélées entre elles [Bergman and Bonner, 1976, Liu and Rabier, 2002]. Typiquement, si la distance de corrélation entre les erreurs d'observation est de 50km, seule une donnée tous les 50km est retenue dans chaque direction de l'espace. Cet artifice est responsable du rejet d'un grand nombre de données profitables à l'assimilation de données. Une des motivations de la présente étude est de pouvoir prendre en compte les corrélations spatiales d'erreurs d'observation, afin de réduire la nécessité de pré-sélectionner les données au travers de l'étape de *thinning*. Similaire au *thinning* dans ses conséquences, on mentionne le *superobing*, consistant à rassembler des observations proches en « macro-observations » [Berger and Forsythe, 2004]. Encore une fois, ce procédé a pour effet de réduire la densité spatiale des données.

Le second type de prétraitement intéressant pour notre étude est le *screening*. Il consiste à rejeter les données non pertinentes d'un point de vue météorologique. Par exemple, si une image provient d'un canal dont la longueur d'onde est caractéristique des phénomènes de basse couche, alors les pixels se trouvant dans des zones recouvertes de nuages hauts sont rejetés.

Cette sélection météorologique introduit de grandes disparités dans la répartition spatiale des données (figure 4.4). En effet, c'est à cause du *screening* que certaines zones géographiques contiennent très peu de données. Nous verrons dans la sous-section 4.3.1 que c'est la raison principale pour laquelle les maillages construits à partir des données satellitaires sont non-structurés.

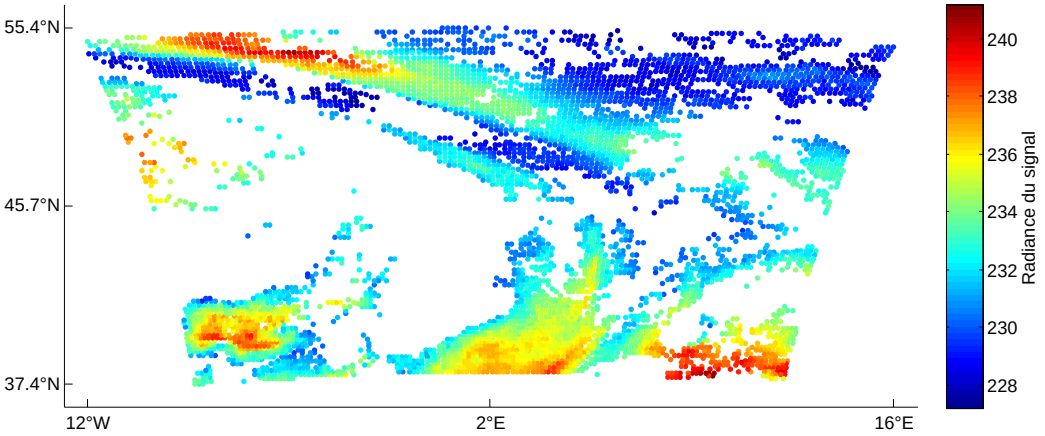


Figure 4.4 – Exemple d'image du sondeur Seviri dans le canal 2. Les couleurs représentent les valeurs de radiance. On voit que le pré-traitement des données a causé l'apparition de grands « trous » dans les données.

4.2 Estimation des paramètres de corrélation

Cette deuxième section est dédiée à l'estimation des paramètres du modèle de corrélation. Il peut s'agir des paramètres de régularité et portée m et l dans le modèle de Matérn, ou encore du tenseur de diffusion $\boldsymbol{\kappa}$ dans le modèle de la diffusion hétérogène. Ces paramètres sont estimés à partir des statistiques réelles des erreurs d'observation, selon la méthode de Desroziers et al. [2005].

4.2.1 Estimation à partir des innovations

En assimilation de données (voir section 1.2), l'expression de l'analyse est donnée par

$$\boldsymbol{x}^a = \boldsymbol{x}^b + \boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}(\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}} + \boldsymbol{R})^{-1}\boldsymbol{d}. \tag{4.1}$$

Dans cette expression, $\boldsymbol{d}$ désigne le vecteur des innovations, caractérisant l'écart entre les observations et l'équivalent de l'ébauche dans l'espace des

observations. Conformément à Desroziers et al. [2005], nous spécifions sa notation en notant $\boldsymbol{d} = \boldsymbol{d}_b^o$. On a

$$\begin{aligned} \boldsymbol{d}_b^o &= \boldsymbol{y}^o - \mathcal{G}(\boldsymbol{x}^b) \\ &= \boldsymbol{y}^o - \mathcal{G}(\boldsymbol{x}^t) + \mathcal{G}(\boldsymbol{x}^t) - \mathcal{G}(\boldsymbol{x}^b). \end{aligned} \tag{4.2}$$

On définit l'erreur d'observation et l'erreur d'ébauche respectivement par

$$\boldsymbol{\epsilon}^o = \boldsymbol{y}^o - \mathcal{G}(\boldsymbol{x}^t) \quad \text{et} \quad \boldsymbol{\epsilon}^b = \boldsymbol{x}^b - \boldsymbol{x}^t. \tag{4.3}$$

On trouve ainsi qu'au premier ordre du développement limité de $\mathcal{G}$,

$$\boldsymbol{d}_b^o \simeq \boldsymbol{\epsilon}^o - \boldsymbol{G}\boldsymbol{\epsilon}^b. \tag{4.4}$$

D'un point de vue statistique, $\boldsymbol{d}_b^o$ est supposé de moyenne nulle. En utilisant (4.4), on détermine sa matrice de covariance,

$$\begin{aligned} \mathbb{E}[\boldsymbol{d}_b^o(\boldsymbol{d}_b^o)^{\mathrm{T}}] &= \mathbb{E}[\boldsymbol{\epsilon}^o(\boldsymbol{\epsilon}^o)^{\mathrm{T}}] + \boldsymbol{G}\mathbb{E}[\boldsymbol{\epsilon}^b(\boldsymbol{\epsilon}^b)^{\mathrm{T}}]\boldsymbol{G}^{\mathrm{T}} \\ &= \boldsymbol{R} + \boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}. \end{aligned} \tag{4.5}$$

De même, on définit l'erreur d'analyse comme

$$\boldsymbol{\epsilon}^a = \boldsymbol{x}^a - \boldsymbol{x}^t \tag{4.6}$$

et on obtient que

$$\begin{aligned} \boldsymbol{d}_a^o &= \boldsymbol{y}^o - \mathcal{G}(\boldsymbol{x}^a) \\ &\simeq \boldsymbol{\epsilon}^o - \boldsymbol{G}\boldsymbol{\epsilon}^a. \end{aligned} \tag{4.7}$$

La quantité $\boldsymbol{d}_a^o$ désigne l'écart entre les observations et l'équivalent de l'analyse dans l'espace des observations. Son lien avec $\boldsymbol{d}_b^o$ vient de la relation

$$\begin{aligned} \boldsymbol{d}_a^o &= \boldsymbol{y}^o - \mathcal{G}(\boldsymbol{x}^a) \\ &= \boldsymbol{y}^o - \mathcal{G}(\boldsymbol{x}^b + \delta\boldsymbol{x}^a) \\ &= \boldsymbol{y}^o - \mathcal{G}(\boldsymbol{x}^b) - \boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}(\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}} + \boldsymbol{R})^{-1}\boldsymbol{d}_b^o \\ &= (\boldsymbol{I} - \boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}}(\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}} + \boldsymbol{R})^{-1})\boldsymbol{d}_b^o \\ &= \boldsymbol{R}(\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}} + \boldsymbol{R})^{-1}\boldsymbol{d}_b^o. \end{aligned} \tag{4.8}$$

Par conséquent, la covariance croisée de $\boldsymbol{d}_b^o$ et $\boldsymbol{d}_a^o$ est donnée par

$$\begin{aligned} \mathbb{E}[\boldsymbol{d}_b^o\boldsymbol{d}_a^o] &= \boldsymbol{R}(\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{\mathrm{T}} + \boldsymbol{R})^{-1}\mathbb{E}[\boldsymbol{d}_b^o\boldsymbol{d}_b^o] \\ &= \boldsymbol{R}. \end{aligned} \tag{4.9}$$

La formule (4.9) suppose que les matrices $\boldsymbol{B}$ et $\boldsymbol{R}$ utilisées dans l'évaluation de $\boldsymbol{x}^a$ sont conformes aux « vraies » matrices de covariances de $\boldsymbol{\epsilon}^b$ et $\boldsymbol{\epsilon}^a$. Quand ce n'est pas le cas, Desroziers et al. [2005] suggèrent de recourir à un algorithme de point fixe pour estimer $\boldsymbol{B}$ et $\boldsymbol{R}$.

On peut ainsi utiliser (4.9) pour estimer la matrice $\boldsymbol{R}$. Toutefois, la méthode se heurte à plusieurs écueils. Tout d'abord, la formule (4.9) ne garantit pas le calcul d'une matrice symétrique définie positive. Pour y remédier, on peut modifier la formule (4.9) ou bien rechercher la matrice symétrique définie positive la plus proche à l'aide d'une méthode d'optimisation (voir, par exemple, Higham [2002]). D'autre part, la méthode de Desroziers et al. [2005] demande d'exprimer tous les coefficients de la matrice $\boldsymbol{R}$, ce qui s'avère impossible pour un grand nombre de données. Enfin, il est nécessaire de disposer d'un ensemble de très grande taille pour estimer $\boldsymbol{R}$ sans erreur d'échantillonnage. On réserve ainsi cette méthode pour l'estimation des variances contenues dans $\boldsymbol{R}$. Pour estimer des corrélations, deux approches sont possibles. Soit on décide d'exploiter les dérivées du champ d'erreurs (sous-section 4.2.2), soit on modifie la méthode de Desroziers et al. [2005] en introduisant des moyennes temporelles et spatiales pour réduire le bruit d'échantillonnage (sous-section 4.2.3).

4.2.2 Exploitation des dérivées

L'approche décrite dans la sous-section 4.2.1 est mal adaptée à l'estimation des structures de corrélation contenues dans la matrice $\boldsymbol{R}$. Dans cette sous-section, on présente une méthode basée sur les dérivées du champ d'observation. Habituellement utilisée pour estimer les paramètres du modèle de covariances d'erreurs d'ébauche [Weaver and Mirouze, 2013, Belo Pereira and Berre, 2006, Michel, 2013], on la présente dans le cadre de l'estimation des covariances d'erreurs d'observation.

La définition (1.16) relie la fonction de covariance ρ à l'écart σ et à la fonction de corrélation c. D'un point de vue statistique, l'écart-type $\sigma(\boldsymbol{z})$ en un point $\boldsymbol{z} \in \mathcal{D}$ peut être relié à l'erreur d'observation calculée à partir d'un ensemble au travers de la formule

$$\sigma(\boldsymbol{z}) = \sqrt{\mathbb{E}[\epsilon(\boldsymbol{z})^2]}. \tag{4.10}$$

En supposant que la fonction de corrélation est deux fois différentiable, la longueur de portée de Daley est définie par la relation (1.23). Cette définition se généralise au cas de la diffusion généralisée où D dépend du tenseur de

corrélation $\boldsymbol{\kappa}$ en posant

$$D(\boldsymbol{z}) = \sqrt{\frac{d}{\mathrm{Tr}(\mathbb{H})}}, \tag{4.11}$$

où $\mathbb{H}$ est la Hessienne de la fonction de corrélation homogène $\tilde{c}(r)$ approximant localement c en $\boldsymbol{z}$, d'expression

$$\mathbb{H} = -\nabla^2 \tilde{c}(r)_{|r=0}. \tag{4.12}$$

Pour estimer $D(\boldsymbol{z})$, il convient donc d'évaluer la limite de la dérivée seconde de $\tilde{c}(r)$ en 0. Partant de (1.16), on peut montrer que $\mathbb{H}$ est reliée au champ dérivé $\nabla\left(\frac{\epsilon}{\sigma}\right)$ par la relation

$$\mathbb{H} = \mathbb{E}\left[\nabla\left(\frac{\epsilon(\boldsymbol{z})}{\sigma(\boldsymbol{z})}\right)\nabla\left(\frac{\epsilon(\boldsymbol{z})}{\sigma(\boldsymbol{z})}\right)^{\mathrm{T}}\right]. \tag{4.13}$$

En pratique, l'estimation du tenseur de diffusion $\boldsymbol{\kappa}$ à partir de (4.13) est numériquement plus abordable que l'estimation de la matrice $\boldsymbol{R}$ par la méthode décrite en sous-section 4.2.1. Cependant, elle peut s'avérer bruitée et suggère que la fonction de corrélation estimée doit être deux fois différentiable. Nous verrons pour les données exposées dans la sous-section 4.2.3 que ce n'est pas forcément le cas. Notre méthode de référence sera donc l'ajustement de modèle présentée ci-après.

4.2.3 Ajustement de modèle

Nous présentons maintenant une méthode d'estimation des paramètres de corrélation m et l dans le cas homogène, ne faisant aucune hypothèse sur la régularité des fonctions de corrélation estimées.

Ajustement du modèle de Matérn à amplitude unitaire :

Pour commencer, on utilise le diagnostic de Desroziers et al. [2005] pour estimer $\boldsymbol{R}$ à partir d'un ensemble de taille réduite. Pour s'affranchir du bruit d'échantillonnage, on choisit de calculer un profil de corrélation moyen en effectuant une moyenne sur les colonnes de $\boldsymbol{R}$ [Bormann et al., 2010, Bormann and Bauer, 2010, Waller et al., 2016a, Michel, 2018]. Si besoin, on peut également avoir recours à une moyenne temporelle (donc sur plusieurs jours successifs). Le profil ainsi obtenu est représenté sur la figure 4.5. Le modèle de corrélation de Matérn est ensuite ajusté pour réduire l'écart aux données. On résout

$$\min_{(\sigma),m,l} \sum_{k=1}^{p} \|\sigma c_{m,l}(r_k) - c_k\|_2, \tag{4.14}$$

où σ désigne l'écart-type, (c'est-à-dire l'amplitude de la fonction ajustée), $c_{m,l}$ la fonction de Matérn de paramètres m et l et c_k la valeur de la corrélation en r_k estimée à partir de l'ensemble.

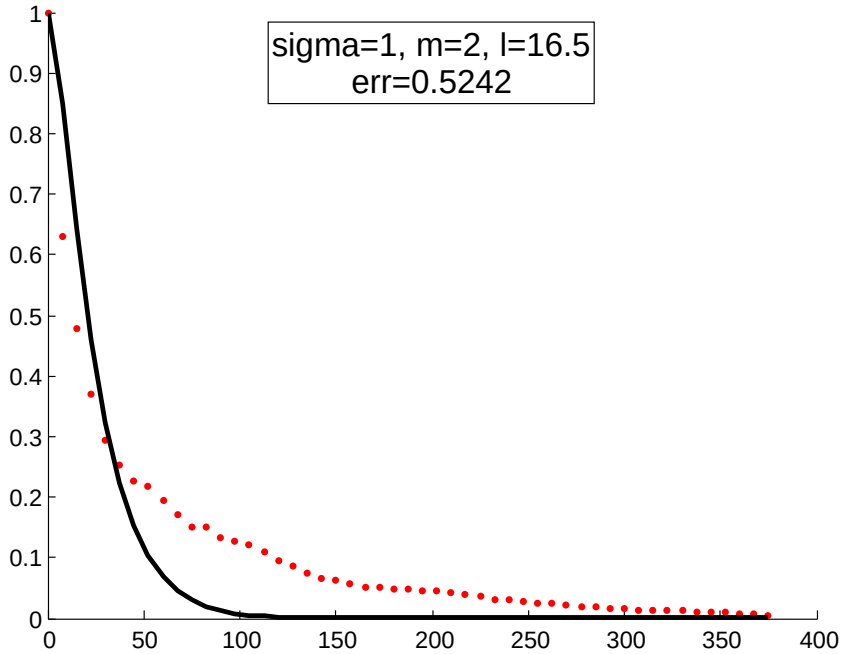


FIGURE 4.5 – Fonction de Matérn (en noir) ajustée aux données (en rouge) avec $\sigma = 1$ imposé. Le minimum de (4.14) est atteint en $m = 2$ et $l = 16.5$ km.

Ajustement du modèle de Matérn à amplitude variable :

On remarque que le modèle de Matérn ne parvient pas à représenter les données de manière satisfaisante. En particulier, il ne permet pas de représenter les corrélations à longue portée, qui restent non-négligeables bien après que la fonction ajustée ait atteint une valeur presque nulle. On se propose donc d'ajuster le modèle en autorisant également l'amplitude à varier, c'est-à-dire l'écart-type σ. On obtient alors le profil de la figure 4.6.

Le gain sur l'erreur ne semble pas être conséquent (de l'ordre de 15%). Par contre, la diminution du paramètre d'amplitude semble permettre une augmentation du paramètre de portée, ayant pour conséquence une meilleure représentation des corrélation à longue portée.

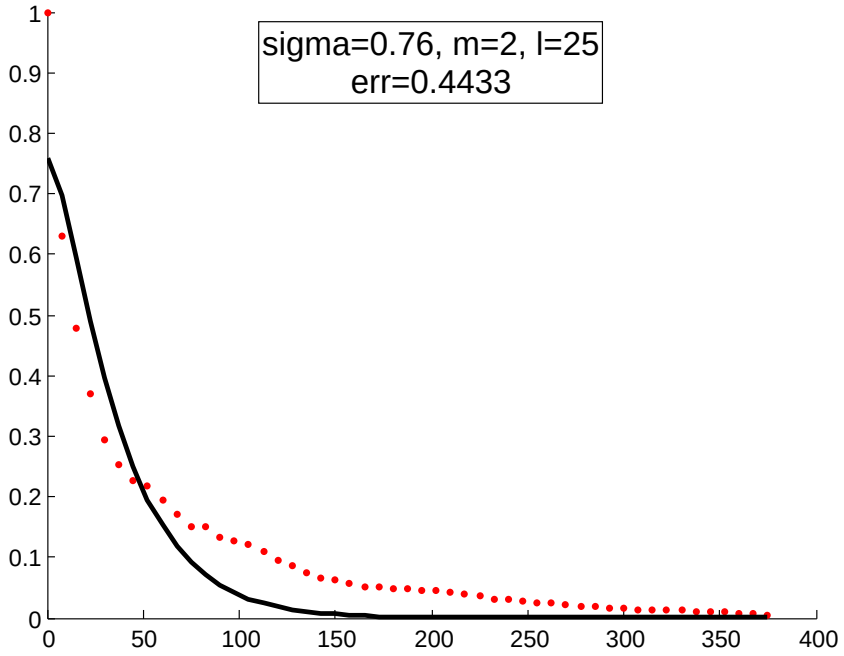


FIGURE 4.6 – Fonction de Matérn ajustées aux données avec σ variable. Le minimum de (4.14) est atteint en $\sigma = 0.76$, $m = 2$ et $l = 25$ km.

Ajustement du modèle de Matérn régularisé :

Suivant les recommandations de Waller et al. [2016a], il est possible de représenter la matrice de corrélation $\boldsymbol{C}$ comme

$$\boldsymbol{C} = \alpha \boldsymbol{I} + (1 - \alpha)\tilde{\boldsymbol{C}}, \tag{4.15}$$

où $\tilde{\boldsymbol{C}}$ est une matrice de corrélation et $\alpha \in [0, 1]$. Le paramètre d'ajustement α est relié au facteur d'amplitude σ au travers de la relation

$$\sigma = 1 - \alpha. \tag{4.16}$$

Cette approche permet d'obtenir l'ajustement de la figure 4.7, dont l'erreur a été réduite de moitié par rapport au cas initial de la figure 4.5. Néanmoins, l'obtention de l'inverse de $\boldsymbol{C}$ reste une difficulté. Suivant la façon dont $\boldsymbol{C}$ est modélisée, plusieurs stratégies seront possibles, notamment par l'utilisation de la formule de Woodburry.

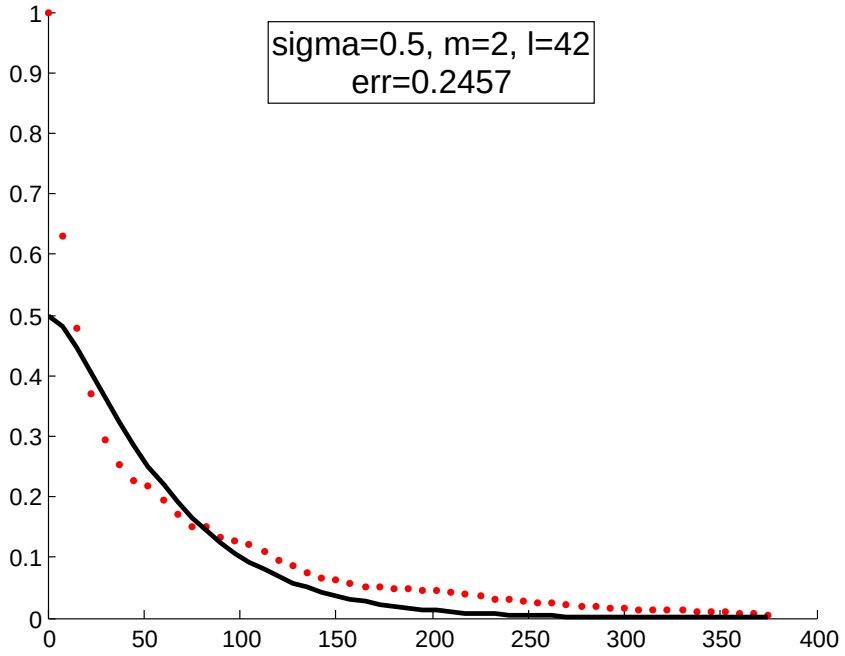


FIGURE 4.7 – Fonction de Matérn ajustées aux données avec σ variable. On ne prend pas en compte la valeur en 0. Le minimum de (4.14) est atteint en $\sigma = 0.5$ $m = 2$ et $l = 42$ km.

Ajustement de deux modèles de Matérn superposés :

Enfin, pour obtenir une correspondance parfaite, il est nécessaire de modéliser $\boldsymbol{C}$ comme

$$\boldsymbol{C} = \alpha \boldsymbol{C}_1 + (1-\alpha)\boldsymbol{C}_2, \tag{4.17}$$

où $\boldsymbol{C}_1$ et $\boldsymbol{C}_2$ sont deux matrices de corrélation. On obtient alors l'ajustement de la figure 4.8.

Bien que cette formule soit attractive, elle rend l'inversion de $\boldsymbol{C}$ difficile. En effet, posons

$$\sigma_1 = \alpha \quad \text{et} \quad \sigma_2 = 1-\alpha. \tag{4.18}$$

En vertu de la formule de Woodburry,

$$\begin{aligned} \boldsymbol{C}^{-1} &= [\sigma_1 \boldsymbol{C}_1 + \sigma_2 \boldsymbol{C}_2]^{-1} \\ &= \frac{1}{\sigma_2}\boldsymbol{C}_2^{-1} \times \left[\frac{1}{\sigma_1}\boldsymbol{C}_1^{-1} + \frac{1}{\sigma_2}\boldsymbol{C}_2^{-1}\right]^{-1} \times \frac{1}{\sigma_1}\boldsymbol{C}_1^{-1}. \end{aligned} \tag{4.19}$$

Pour inverser $\boldsymbol{C}$, il faut donc pouvoir résoudre le système linéaire

$$\left(\frac{1}{\sigma_1}\boldsymbol{C}_1^{-1} + \frac{1}{\sigma_2}\boldsymbol{C}_2^{-1}\right)\boldsymbol{x} = \boldsymbol{b}. \tag{4.20}$$

Suivant la taille de $\boldsymbol{C}_1^{-1}$ et $\boldsymbol{C}_2^{-1}$, résoudre le système (4.20) peut requérir l'usage d'une méthode itérative. Or, il n'existe pas de formule simple pour une racine carrée de $\left(\frac{1}{\sigma_1}\boldsymbol{C}_1^{-1} + \frac{1}{\sigma_2}\boldsymbol{C}_2^{-1}\right)$. Il n'est donc pas possible d'utiliser un solveur linéaire comme les itérations de Chebychev ou les méthodes multigrilles, dont on peut tronquer le nombre d'itérations. Il faut donc faire appel à une méthode non-linéaire de type gradient conjugué qui n'assure pas la symétrie de $\boldsymbol{C}^{-1}$.

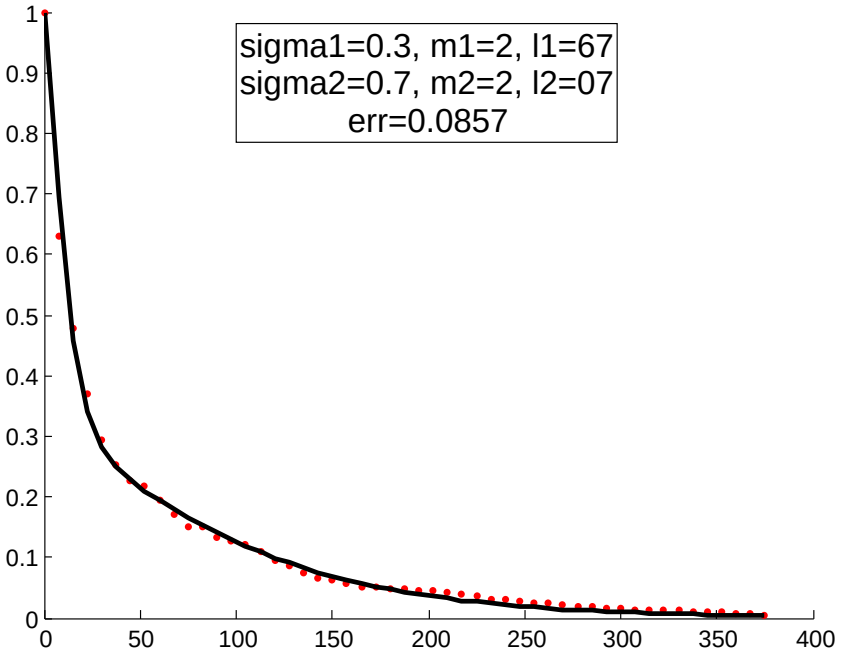


FIGURE 4.8 – Ajustement aux données par la somme de deux fonctions de Matérn. Le minimum de (4.14) est atteint en $\sigma_1 = 0.3$, $m = 2$ et $l = 67$ km pour la première fonction et $\sigma_2 = 0.7$, $m = 2$ et $l = 7$ km pour la deuxième fonction.

Quel choix de paramètres pour nos expériences ?

Au travers les quatre paramétrisations précédentes, nous avons montré que les erreurs d'observation de l'imageur SEVIRI contenaient plusieurs échelles. La plus petite est de l'ordre de 10-20km, alors que la plus grand

avoisinne 60-70km. Dans nos expériences, nous utiliserons une valeur intermédiaire. Dans leur étude, Waller et al. [2016a] montrent que les corrélations spatiale de SEVIRI atteignent la valeur de 20% à 80km, ce qui correspond à la valeur $l = 32.5$km. On décide de réutiliser la même valeur. Le graphe correspondant est visible sur la figure 4.9.

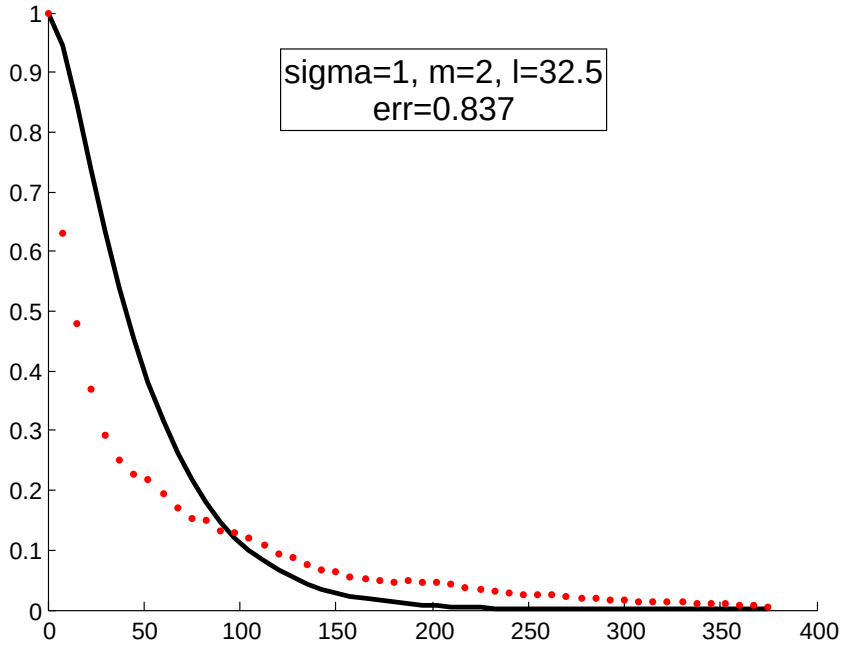


FIGURE 4.9 – Fonction de Matérn avec $m = 2$ et $l = 32.5$km.

4.3 Géométrie du maillage des observations

Au delà des aspects purements mathématiques liés à la paramétrisation de l'équation de diffusion, la modélisation des corrélations spatiales d'erreurs d'observation dépend en pratique de la répartition des données dans l'espace [4]. En effet, de cette dernière dépend la géométrie du maillage servant de support à l'assemblage des matrices d'éléments fini. On présente donc dans un premier temps la génération de maillage et les étapes-clefs de cette procédure. Dans un second temps, on discute de la qualité du maillage *a priori*. Des critères sont présentés, qui serviront à l'évaluation des résultats dans la section 4.4.

4. Dans un deuxième temps. Elle dépend bien entendu de l'instrument et de ses mesures, du modèle et de l'opérateur d'observation.

4.3.1 Génération de maillage

Contrairement à l'étude sur des maillages structurés, la modélisation des opérateurs de corrélation sur des maillages non-structurés requiert l'introduction d'une étape supplémentaire dans l'assimilation de données : la génération de maillage. Même si cette étape peut être automatisée et ne comporte pas de lien direct avec la modélisation des covariances, on en explique les grands principes afin d'éclairer le lecteur sur les ajustements possibles de la méthode. C'est aussi l'occasion d'illustrer la provenance de certaines dégénérescences géométriques qui auront des conséquences directes sur la qualité de la solution de l'équation de diffusion.

Soit

$$\mathcal{S} = \{\boldsymbol{z}_i \in \mathbb{R}^2, i \in [\![1, p]\!]\} \tag{4.21}$$

un ensemble de points du plan, communément appelés « sites ». Pour tout $i \in [\![1, p]\!]$, on définit la cellule de Voronoï $\mathcal{V}(\boldsymbol{z}_i)$ comme l'ensemble des points de $\mathbb{R}^2$ qui sont plus proches de $\boldsymbol{z}_i$ que de toute autre $\boldsymbol{z}_j, j \in [\![1, p]\!]$:

$$\mathcal{V}(\boldsymbol{z}_i) = \{\boldsymbol{z} \in \mathbb{R}^2, \|\boldsymbol{z} - \boldsymbol{z}_i\|_2 \leqslant \|\boldsymbol{z} - \boldsymbol{z}_j\|_2, \forall j \neq i\}. \tag{4.22}$$

La définition (4.22) n'exclut pas qu'un point du plan appartienne à plusieurs cellules de Voronoï à la fois. L'ensemble des points appartenant à plusieurs cellules est appelé le diagramme de Voronoï et on le note $\mathcal{V}(\mathcal{S})$ [Voronoi, 1908, Du et al., 1999].

Pour se donner une idée de la géométrie du diagramme de Voronoï, considérons un triplet de points non-colinéaires du plan $(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3) \in (\mathbb{R}^2)^3$. Le diagramme de Voronoï est construit à partir des médiatrices du triangle $(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3)$. Elles se coupent en un unique point $\boldsymbol{z}_c$ qui est le centre du cercle circonscrit au triangle (figure 4.10).

Le diagramme de Voronoï est un outil important en mathématiques en raison de ses propriétés structurelles. En effet, soit un nuage de points du plan défini comme (4.21). Etant donnée une mesure de la distance, le diagramme de Voronoï constitue un moyen d'attribuer à chaque point $\boldsymbol{z}_i$ un voisinage géométriquement pertinent, représenté par la cellule $\mathcal{V}(\boldsymbol{z}_i)$. Autrement dit, $\boldsymbol{z}_i$ « représente » les points de $\mathcal{V}(\boldsymbol{z}_i)$. La figure 4.11 représente un nuage de points et les cellules associées.

De manière générale, la donnée d'un nuage de points ne définit pas une unique triangulation, au sens de la définition donnée en sous-section 2.3.1. Il

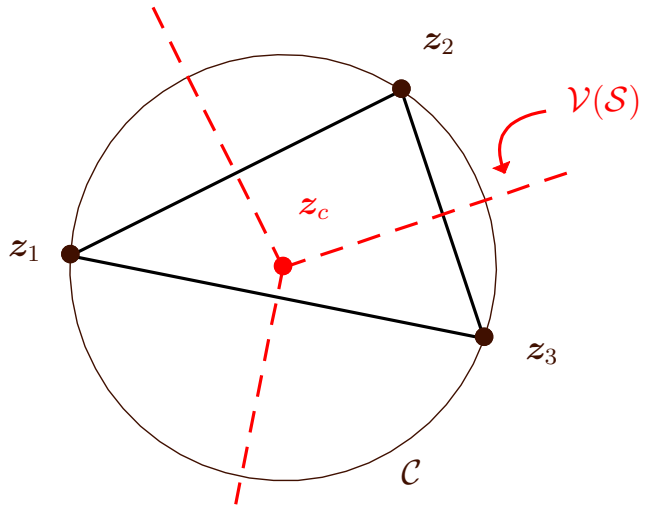


FIGURE 4.10 – Diagramme de Voronoï (en rouge) associé au triangle (z_1, z_2, z_3) (en noir) et son cercle circonscrit (en noir).

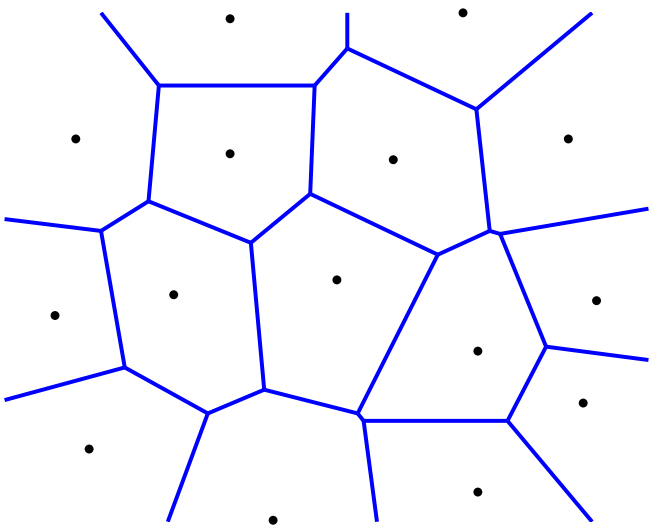


FIGURE 4.11 – Diagramme de Voronoï d'un nuage de points.

faut donc pouvoir donner un critère pour en choisir une en particulier. Pour ce faire, il est commun d'adopter la triangulation obtenue en considérant le maillage dual du diagramme de Voronoï [de Berg et al., 2008].

Pour construire la triangulation de Delaunay comme le dual du diagramme de Voronoï, on définit la matrice d'adjacence $\boldsymbol{A} = (A_{\mathrm{ij}}) \in \mathbb{N}^{p\times p}$ définie par :

$$\begin{cases} A_{\mathrm{ii}} = 0 & \\ A_{\mathrm{ij}} = 0 & \text{si} \quad \mathcal{V}(\boldsymbol{z}_i) \cap \mathcal{V}(\boldsymbol{z}_j) = \varnothing \\ A_{\mathrm{ij}} = 1 & \text{si} \quad \mathcal{V}(\boldsymbol{z}_i) \cap \mathcal{V}(\boldsymbol{z}_j) \neq \varnothing \end{cases} . \tag{4.23}$$

La triangulation de Delaunay est alors représentée par le graphe G^D dont les noeuds sont les $(\boldsymbol{z}_i)_{i\in[\![1,p]\!]}$ et les arêtes sont les segments reliants entre eux les points $(\boldsymbol{z}_i, \boldsymbol{z}_j)$ vérifiant $A_{\mathrm{ij}} \neq 0$. Cette définition assure que les arêtes de G^D ne se croisent jamais et que les faces définies par l'espace entre les arêtes sont des triangles.

Dans la suite, on confondra la triangulation de Delaunay, notée $\mathcal{T}^D$, avec la définition de son graphe G^D, car ils définissent des objets géométriques équivalents. En l'absence de triangulation alternative, cette notation sera abréviée en $\mathcal{T}$. Un exemple de triangulation de Delaunay est donné par la figure 4.12.

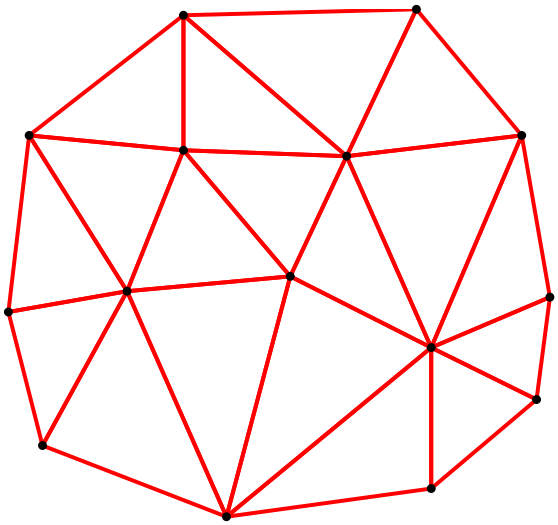


FIGURE 4.12 – Triangulation d'un nuage de points.

La triangulation de Delaunay est construite de sorte à maximiser le plus petit angle de tous ses triangles [Edelsbrunner et al., 1992]. En cela, elle est optimale pour la discrétisation d'équations aux dérivées partielles, qui souffrent de la présence d'éléments étirés ou aplatis (voir sous-section 2.4.2).

D'autre part, une propriété suffisante pour qu'une triangulation soit dite de Delaunay est que chaque cercle circonscrit à un triangle ne contienne aucun autre site. Dans le cas limite où quatre sites sont concentriques, il n'est pas possible de définir la triangulation de Delaunay de manière unique. Néanmoins, il suffit en pratique de choisir une des constructions possibles, car les autres choix conduisent à des résultats numériques semblables (avec la méthode des éléments finis). Un exemple trivial est celui ou les points de $\mathcal{S}$ sont sur un grille cartésienne.

Pour générer une triangulation de Delaunay à partir d'un nuage de points, les algorithmes populaires exploitent la technique de la bascule (*edge flipping* en anglais [Edelsbrunner et al., 1992]). Une triangulation est tout d'abord générée, éventuellement aléatoirement, puis ses arêtes sont « basculées » jusqu'à ce que le critère de Delaunay (pas de site à l'intérieur d'un cercle circonscrit) soit vérifié. Ce procédé est représenté sur la figure 4.13.

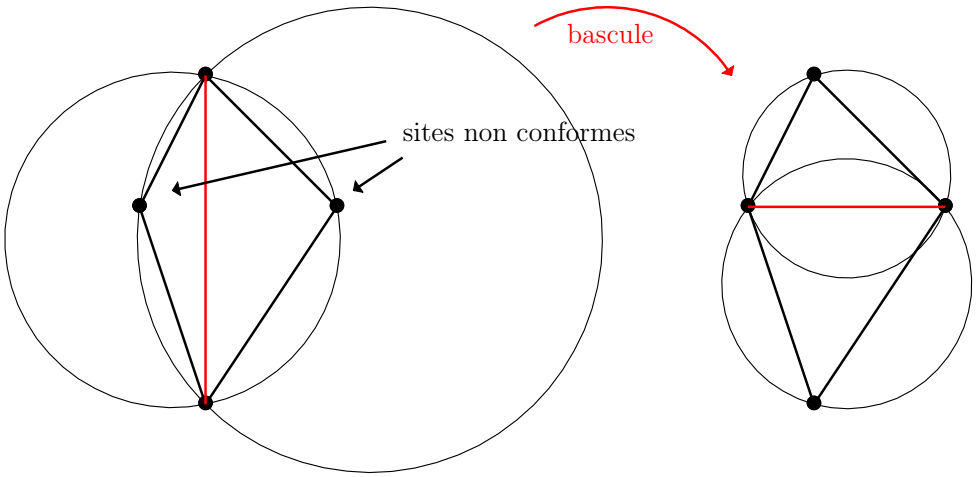


Figure 4.13 – Basculement de l'arête centrale afin de réduire les rayons des cercles circonscrits aux triangles. A gauche, la triangulation est non conforme car les cercles circonscrits contiennent des noeuds n'appartenant pas au triangle. A gauche, le basculement a rétabli la conformité.

4.3.2 Qualité du maillage

La méthode des éléments finis $\mathbb{P}_1$ discrétisée sur une triangulation $\mathcal{T}$ contient autant de degrés de liberté que de noeuds du maillage. Ainsi, cette méthode permet de construire une matrice de corrélation (3.2) inversible, dont le rang est égal au nombre d'observations. Dans cette sous-section, on évalue la qualité du maillage construit à partir de données SEVIRI.

La figure 4.14 représente le maillage associé à deux niveaux de *thinning* sur une même situation météorologique. On constate que les déserts de données occupent la même place dans les deux cas, puisque le *screening* est le même. En revanche, la différence de taille entre le plus grand et le plus petit élément du maillage est accentuée lorsque le nombre de données est plus grand, c'est-à-dire dans le cas du *thinning* moins agressif. La conséquence est une plus grande disparité dans la forme et la taille des éléments, ce qui est susceptible de dégrader la qualité de la méthode des éléments finis.

Nous sommes intéressés par le jeu de données le plus dense. En effet, la prise en compte des corrélations d'erreurs d'observation est un argument pour la réduction du *thinning*. Une conséquence immédiate est l'augmentation du nombre des données. Les méthodes développées dans cette étude s'adresse à ce cas d'utilisation future et se focalise donc sur l'exploitation de données à haute densité spatiale.

Pour quantifier la qualité du maillage, il convient de se donner un indicateur pertinent permettant d'indiquer *a priori* où seront situées les erreurs numériques de notre méthode. D'après la sous-section 2.4.2, plusieurs choix sont possibles, comme le rapport d'aspect des éléments, le diamètre de leur cercle circonscrit ou encore leur angle maximal. On attire l'attention du lecteur sur le fait que ces variables sont associées à des éléments du maillages (*i.e.* des triangles) et non à des noeuds. Or, l'analyse comparée avec le modèle théorique de Matérn sera menée noeud par noeud dans la section 4.4. On adopte donc l'approche suivante : soit $\boldsymbol{z}_i$ un noeud du maillage. On note $\mathfrak{E}(\boldsymbol{z}_i)$ l'ensemble des triangles ayant $\boldsymbol{z}_i$ pour sommet (figure 4.15). On calcule l'indice de qualité $\iota(\tau)$ de chaque élément $\tau \in \mathfrak{E}(\boldsymbol{z}_i)$. Plus la valeur $\iota(\tau)$ est grande (*cf.* sous-section 2.4.2), plus τ est de mauvaise qualité. On définit l'indicateur de qualité $\iota(\boldsymbol{z}_i)$ comme la plus grande valeur de $\iota(\tau)$, c'est-à-dire :

$$\iota(\boldsymbol{z}_i) = \max_{\tau \in \mathfrak{E}(\boldsymbol{z}_i)} \iota(\tau). \tag{4.24}$$

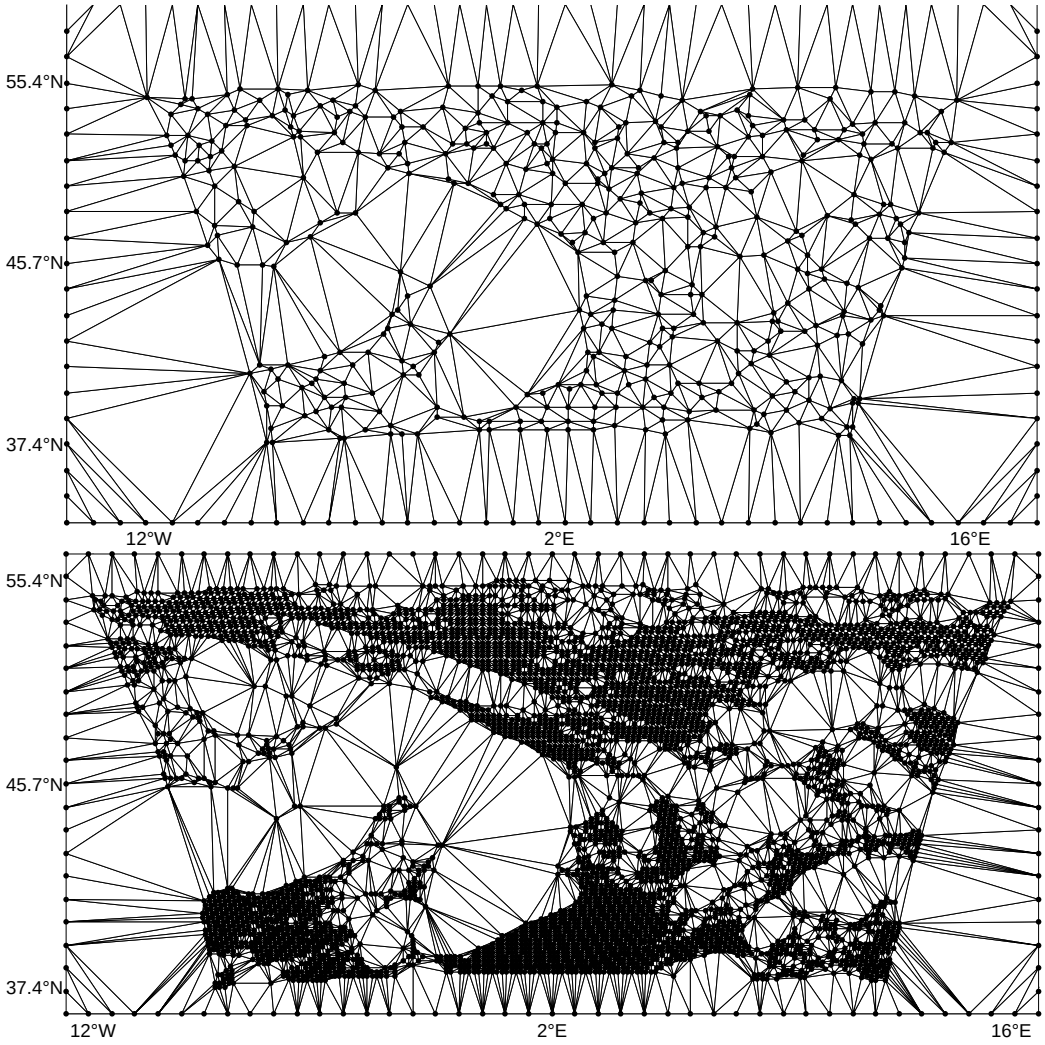


Figure 4.14 – Maillages associés à la même situation météorologique. En haut : densité d'observation opérationnelle du modèle Arome. En bas : densité plus élevée correspondant à un *thinning* moins agressif.

La procédure décrite ci-dessus est appliquée à la situation météorologique du 12/02/2016 sur le domaine Arome-France dans le canal 2 de Seviri. La figure 4.16 est une carte de rapports d'aspect, la figure 4.17 est une carte d'angles maximaux et la figure 4.18 est une carte de diamètres de cercles circonscrits.

On remarque que les cartes d'indices ne mettent pas exactement les

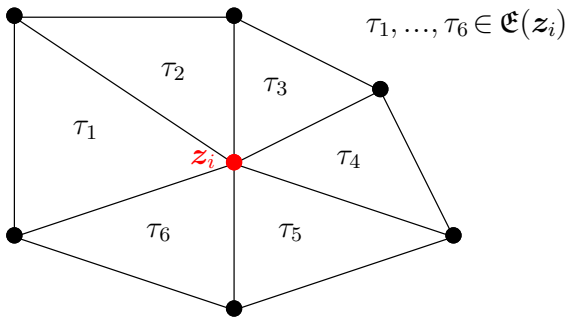


FIGURE 4.15 – Voisinage (en noir) du noeud z_i (en rouge).

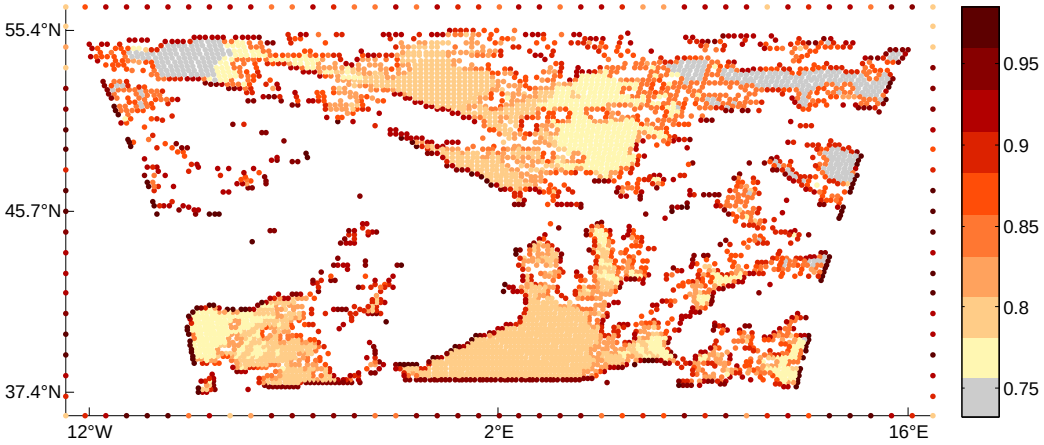


FIGURE 4.16 – Rapports d'aspect.

mêmes régions en valeur. Globalement, les valeurs d'indice élevées sont situées proche des déserts de données. Nous verrons toutefois que ces différences subtiles auront de l'importance dans la prédiction des erreurs dans la section 4.4. On peut faire les constats suivants : le rapport d'aspect semble faire ressortir tous les bords, intérieurs (bords de désert) comme extérieurs (bords du domaine), avec une intensité comparable. L'angle maximal permet de concentrer l'analyse sur certaines zones, mais il est plus bruité que l'indice basé sur le diamètre du cercle circonscrit, qui semble cibler sensiblement les mêmes zones. Ce dernier fait ressortir une zone sur l'est du domaine qui sera pertinente lors de l'analyse d'erreur.

On précise que le développement d'un indice de qualité adapté à la représentation des corrélations sur des maillages non structurés est une nouveauté.

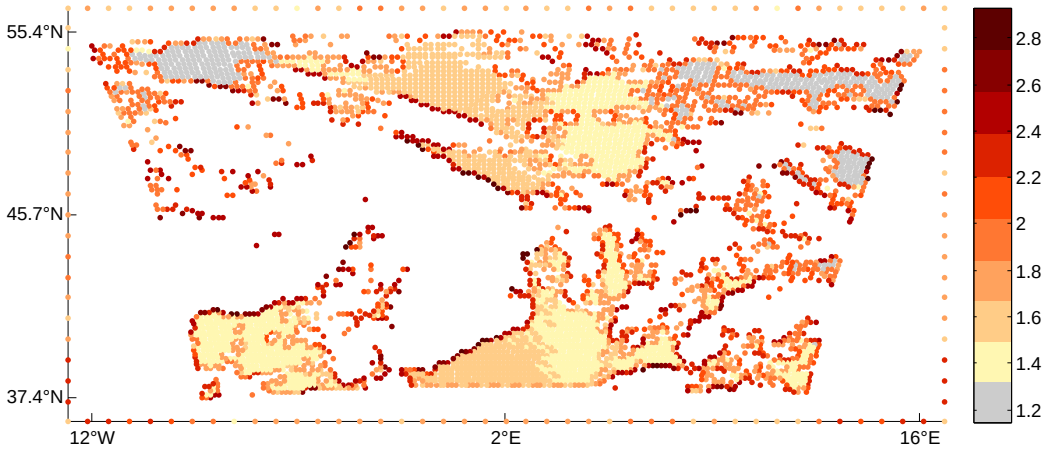


Figure 4.17 – Angles maximaux.

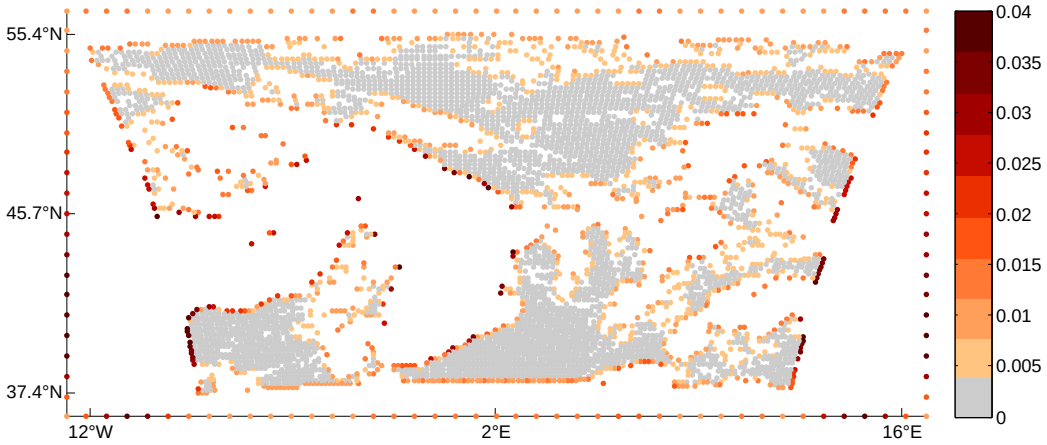


Figure 4.18 – Diamètres de cercles circonscrits.

En effet, les données satellitaires ont la particularité d'être très non structurées : leur distribution spatiale est très dense par endroits, alors que certaines zones sont vides d'observations. C'est la raison pour laquelle les indicateurs traditionnels ne sont pas adaptés à ce cas d'étude.

4.3.3 Noeuds de frontière

On donne quelques précisions importantes concernant le domaine d'étude et les noeuds de frontière.

Nous souhaitons résoudre une équation aux dérivées partielles pour simuler l'action d'un opérateur de corrélation. Cette équation aux dérivées

partielles doit être résolue sur un domaine. En pratique, le domaine est défini par sa frontière et cette dernière par les noeuds qui l'échantillonnent. Sans les noeuds de frontière, le domaine de résolution n'est pas clairement défini et les conditions aux limites sont imposées directement sur l'enveloppe convexe des observations.

On distingue ainsi trois types de noeuds dans le maillage :

- Les noeuds correspondant aux observations ;
- Les noeuds « supplémentaires », qui interviendront dès le chapitre 5 dans la définition d'un maillage auxiliaire destiné à la stabilisation de la résolution de l'équation de diffusion ;
- Les noeuds de frontière, qui sont également des noeuds supplémentaires. Cependant, on les distingue du deuxième type car il n'est pas possible de s'affranchir de leur introduction. Ce sont donc des points qu'on ajoute obligatoirement pour résoudre l'équation de diffusion.

Intuitivement, les noeuds de frontière posent deux genres de problèmes :

- En plaçant les frontières suffisamment proche des observations, on peut faire en sorte que les éléments de bords soient de bonne qualité (petit, pas étirés). Toutefois, les conditions aux limites se font alors ressentir et on observe une déformation des fonctions de corrélation proche de la frontière. A l'inverse, on peut facilement nullifié l'effet des conditions aux limites en écartant les frontières des observations. Ce faisant, on introduit cependant des éléments de mauvaise qualité (des triangles longs et fins) proches des bords. En pratique, il faut donc trouver un compromis entre l'éloignement des frontières et la qualité du maillage.
- L'ajout de noeuds à la frontière correspond à l'introduction d'un opérateur d'extension $\boldsymbol{S}$ dont le nombre de lignes est supérieur au nombre de colonnes. A la fin de la procédure, il est nécessaire de supprimer les noeuds de frontière, qui ne représentent pas des variables d'intérêt. Cette opération recquiert l'application de l'opérateur $\boldsymbol{S}^{\mathrm{T}}$. Désignons par $\boldsymbol{C}$ l'opérateur de corrélation agissant dans l'espace des observations et par $\breve{\boldsymbol{C}}$ l'opérateur de corrélation agissant dans l'espace du maillage (observations + noeuds de frontière). On a

$$\boldsymbol{C} = \boldsymbol{S}^{\mathrm{T}} \breve{\boldsymbol{C}} \boldsymbol{S}. \tag{4.25}$$

  Dans nos expériences, les noeuds de frontière n'ont pas d'influence sur le calcul de l'inverse. Mathématiquement, on vérifie que

$$\boldsymbol{C}^{-1} = (\boldsymbol{S}^{\mathrm{T}} \breve{\boldsymbol{C}} \boldsymbol{S})^{-1} \simeq \boldsymbol{S}^{\mathrm{T}} \breve{\boldsymbol{C}}^{-1} \boldsymbol{S}. \tag{4.26}$$

Pour ce faire, on calcule l'erreur d'approximation

$$\frac{\|\boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{w} - \boldsymbol{w}\|_2}{\|\boldsymbol{w}\|_2} < \varepsilon \simeq 10^{-16}, \tag{4.27}$$

où $\boldsymbol{w}$ est un vecteur aléatoire gaussien (bruit blanc).

Pour cette raison, on négligera l'effet des noeuds additionnels aux bords du domaine et on ne représentera pas les opérateurs $\boldsymbol{S}$ et $\boldsymbol{S}^{\mathrm{T}}$ dans nos équations.

4.4 Validation des opérateurs de corrélation

Cette section est dédiée à la validation numérique de la représentation des opérateurs de corrélation par la méthode diffusion-éléments-finis. Dans un premier temps, on étudie la réponse impulsionnelle de l'opérateur décrit dans le chapitre 2. Cela permet de visualiser les fonctions de corrélation contenues dans la matrice $\boldsymbol{R}$. Ce faisant, plusieurs types d'erreurs numériques sont diagnostiquées et mises en relation avec les indicateurs de qualité introduits dans la sous-section 4.3.2. Ensuite, on présente un test d'adjonction permettant de montrer que les propriétés mathématiques (1.36) et (1.37) sont vérifiées du point de vue discret.

Les expériences sont menées en utilisant les paramètres de corrélation diagnostiqués dans la section 4.2 pour le canal 2 de Seviri. En particulier, on choisit de fixer les paramètres de Matérn m et l aux valeurs suivantes :

$$m = 2 \quad \text{et} \quad l = 32.5\,\mathrm{km}\,. \tag{4.28}$$

4.4.1 Fonctions de corrélation

On souhaite appliquer notre modèle de corrélation à des maillages non-structurés pour visualiser les fonctions de corrélations modélisées et les comparer au modèle théorique de Matérn (voir sous-section 1.1.3). On a vu dans le chapitre 3 que la méthode des éléments finis sur des maillages cartésiens permettait de reproduire fidèlement le modèle théorique. Bien que les conditions aux limites soient responsables d'une sur-estimation de l'amplitude aux bords du domaine, on considère ce phénomène comme partie intégrante de notre modèle de corrélation, indépendante des notions de maillages (non-)structurés. A l'inverse, on s'intéresse aux éventuelles erreurs numériques

apparaissant dans l'intérieur du domaine et on cherche à établir un lien avec la qualité du maillage.

La première expérience est celle de la réponse impulsionnelle. Etant donnée une image satellite, on choisit l'une des observations de façon arbitraire. On introduit une impulsion de Dirac au noeud $\boldsymbol{z}_i$ correspondant dans le maillage et on applique l'opérateur de corrélation (1.44). La réponse obtenue est la fonction de corrélation centrée en $\boldsymbol{z}_i$, en vertu de la relation

$$\mathcal{C}(\delta_{\boldsymbol{z}_i})(\boldsymbol{z}) = c(\boldsymbol{z}, \boldsymbol{z}_i). \tag{4.29}$$

Cela est possible car la distribution de Dirac (3.3) est définie dans $H^{-1}(\mathcal{D})$ [Mitrovic and Zubrinic, 1997] et que les fonctions de Matérn sont au moins dans H^1 [Mirouze and Weaver, 2010]. Il est ensuite facile de comparer $c(\boldsymbol{z}, \boldsymbol{z}_i)$ à la fonction de Matérn $c_{m,l}(\boldsymbol{z}, \boldsymbol{z}_i)$ définie par l'expression (1.21).

En pratique, on approxime la distribution de Dirac $\delta_{\boldsymbol{z}_i}$ par un vecteur dont les éléments sont nuls, à l'exception du i-ème élément, qui est égal à 1. Ce vecteur est noté $\boldsymbol{\delta}_i$. Pour réaliser le test, on multiplie $\boldsymbol{\delta}_i$ à gauche par la matrice de corrélation (2.35). Le résultat est un vecteur égal à la i-ème colonne de $\boldsymbol{C}$. Pour s'en convaincre, il suffit d'effectuer la multiplication symbolique comme dans l'équation (4.30).

$$\begin{pmatrix} \circ & \bullet & \circ & & \circ \\ \circ & \bullet & \circ & \cdots & \circ \\ \circ & \bullet & \circ & & \circ \\ & \vdots & & \ddots & \\ \circ & \bullet & \circ & & \circ \end{pmatrix} \times \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \bullet \\ \bullet \\ \bullet \\ \vdots \\ \bullet \end{pmatrix}. \tag{4.30}$$

L'expérience décrite ci-dessus est réalisée à partir des données SEVIRI. Le choix des observations a pour but d'illustrer différents cas pratiques, par exemple lorsque $\boldsymbol{z}_i$ se trouve à proximité d'un désert d'observation. Par soucis de compacité, plusieurs instances de cette expérience sont représentées sur la figure 4.19.

Six impulsions de Dirac ont été introduites en six emplacements différents. La figure 4.20 donne les écarts à la solution analytique correspondants. Pour tout $i \in [\![1, p]\!]$, on définit l'écart ε_i comme la différence entre la fonction de corrélation modélisée par la diffusion et la fonction théorique au point $\boldsymbol{z}_i$, c'est-à-dire :

$$\varepsilon_i \quad : \quad \boldsymbol{z} \mapsto \varepsilon_i(\boldsymbol{z}) = c(\boldsymbol{z}, \boldsymbol{z}_i) - c_{m,l}(\boldsymbol{z}, \boldsymbol{z}_i). \tag{4.31}$$

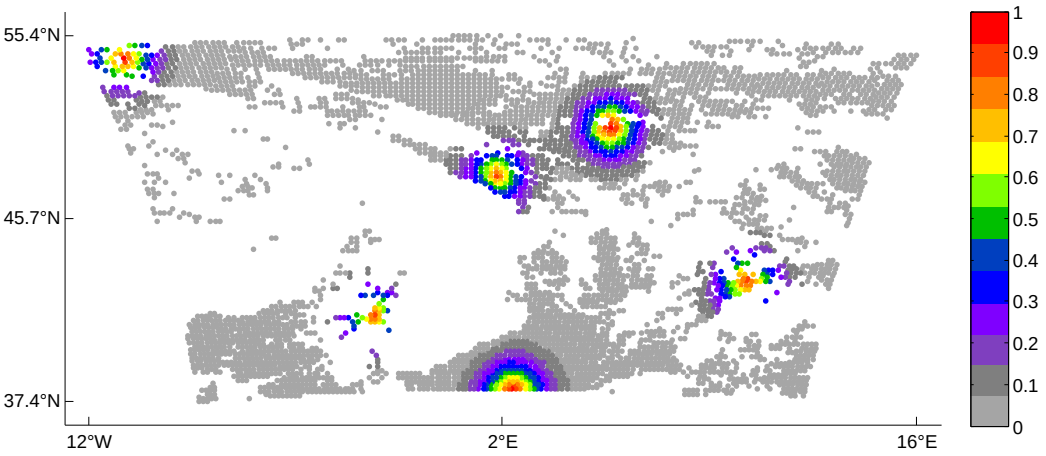


FIGURE 4.19 – Six réponses impulsionnelles en six emplacements choisis pour illustrer différents cas de la méthode.

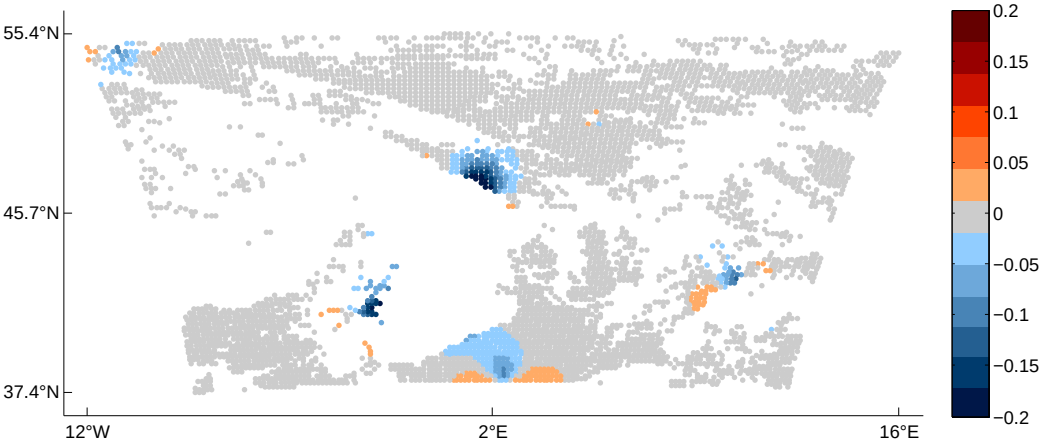


FIGURE 4.20 – Ecarts à la solution analytique correspondant aux expériences de la figure 4.19.

Cet écart peut se reformuler à partir des coefficients de la matrice de corrélation $\boldsymbol{C}$. Soit $j \in [\![1, p]\!]$,

$$\varepsilon_i(\boldsymbol{z}_j) = c(\boldsymbol{z}_j, \boldsymbol{z}_i) - c_{m,l}(\boldsymbol{z}_j, \boldsymbol{z}_i) = \boldsymbol{C}_{\mathrm{ij}} - (\boldsymbol{C}_{m,l})_{\mathrm{ij}}. \tag{4.32}$$

On attire le lecteur sur le fait que $(\boldsymbol{C}_{m,l})_{\mathrm{ii}} = 1$ (par définition). Par conséquent,

$$\varepsilon_i(\boldsymbol{z}_i) = \boldsymbol{C}_{\mathrm{ii}} - 1. \tag{4.33}$$

La valeur de (4.33) est variable dans l'espace. En particulier, nos expériences montrent que l'erreur peut être importante près des déserts d'observation ou près des bords du domaine. En revanche, les éléments finis permettent

une modélisation relativement robuste en présence d'un réseau d'observation non-structuré.

En observant la figure 4.20, on parvient à distinguer deux types d'erreurs. La première concerne la différence (4.33), c'est-à-dire la différence d'amplitude entre la fonction de corrélation modélisée par la diffusion et celle de Matérn :

$$\varepsilon_i^{\text{amp}} = \boldsymbol{C}_{\text{ii}} - 1. \tag{4.34}$$

Cette valeur doit être égale à 0 pour que $\boldsymbol{C}$ définisse une matrice de corrélation. Cette condition est nécessaire puisqu'elle signifie que les éléments diagonaux de $\boldsymbol{C}$ sont égaux à 1. En revanche, elle n'est pas suffisante, en raison des autres propriétés que $\boldsymbol{C}$ doit vérifier (symétrie, positivité, rang plein, valeurs des éléments non-diagonaux...). La quantité $\varepsilon_i^{\text{amp}}$ est calculée en tous les points d'observation $\boldsymbol{z}_i$ et représentée sur la figure 4.21.

Le deuxième type d'erreur concerne la forme des fonctions de corrélations contenues dans $\boldsymbol{C}$. Pour la visualiser, on calcule la quantité

$$\varepsilon_i^{\text{shape}} = \frac{\left(\sum_{j=1}^{p} |\overline{\boldsymbol{C}_{\text{ij}}} - (\boldsymbol{C}_{m,l})_{\text{ij}}|^2\right)^{1/2}}{\left(\sum_{j=1}^{p} |(\boldsymbol{C}_{m,l})_{\text{ij}}|^2\right)^{1/2}}, \tag{4.35}$$

où

$$\overline{\boldsymbol{C}_{\text{ij}}} = \frac{\boldsymbol{C}_{\text{ij}}}{\sqrt{\boldsymbol{C}_{\text{ii}}}\sqrt{\boldsymbol{C}_{\text{jj}}}} \tag{4.36}$$

pour tout $i \in [\![1, p]\!]$ et $j \in [\![1, p]\!]$. Le dénominateur dans (4.35) permet d'exprimer l'erreur de forme $\varepsilon_i^{\text{shape}}$ comme un pourcentage alors que le dénominateur dans (4.36) permet que s'assurer que la valeur de $\varepsilon_i^{\text{shape}}$ est indépendante de $\varepsilon_i^{\text{amp}}$. Ce deuxième type d'erreur est représenté sur la figure 4.22.

L'erreur d'amplitude comme l'erreur de forme est accentuée sur les bords intérieurs du domaine, c'est-à-dire à proximité des zones où le prétraitement des observations a supprimé beaucoup de données. Concernant l'erreur d'amplitude, elle est visuellement corrélée avec l'indicateur de qualité de maillage de la figure 4.18. En effet, les zones colorées, correspondants aux zones d'erreur / d'indicateur élevés, sont situées aux mêmes endroits. Plus précisément, le critère basé sur les rayons des cercles circonscrits est plus apte à repérer ces

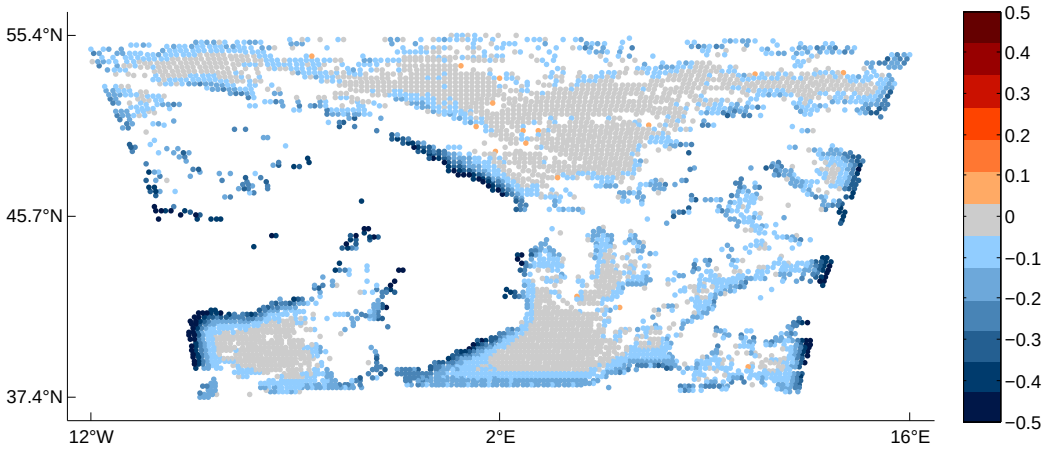


FIGURE 4.21 – Carte des erreurs d'amplitude.

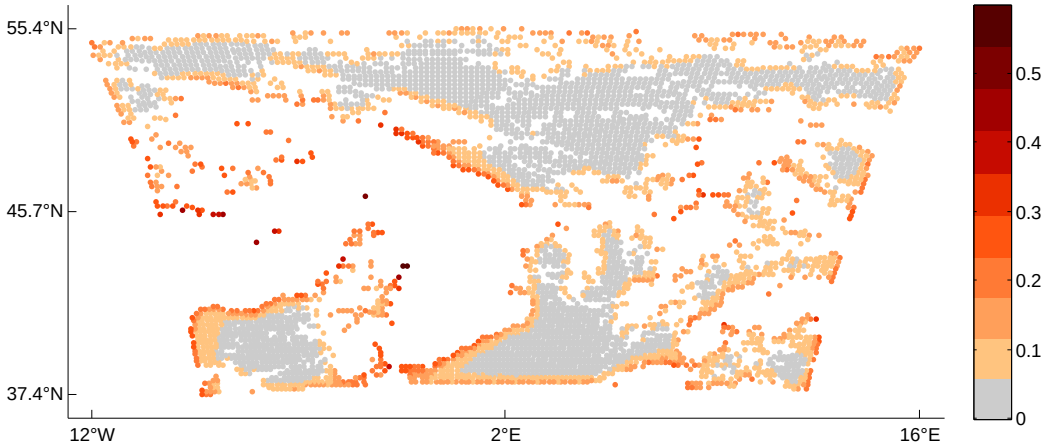


FIGURE 4.22 – Carte des erreurs de forme.

zones, contrairement au rapport d'aspect et à l'angle maximal. Il est attendu que la présence d'éléments malformés dans le maillage dégrade localement la qualité de la modélisation. Dans la sous-section 4.5.3, ce constat sera mis à profit pour suggérer une correction de l'erreur d'amplitude basée sur l'indicateur (4.24), où $\iota(\tau)$ est le diamètre du cercle circonscrit à l'élément τ.

Les figures 4.23 et 4.24 représentent respectivement $\varepsilon_i^{\text{amp}}$ et $\varepsilon_i^{\text{shape}}$ dans le cas où la matrice de masse $\boldsymbol{M}$ est approximée par un schéma de condensation de masse (voir section 3.2).

La condensation de masse a pour effet d'augmenter l'erreur de modélisation (amplitude et forme), sans changer sa dépendance géographique au

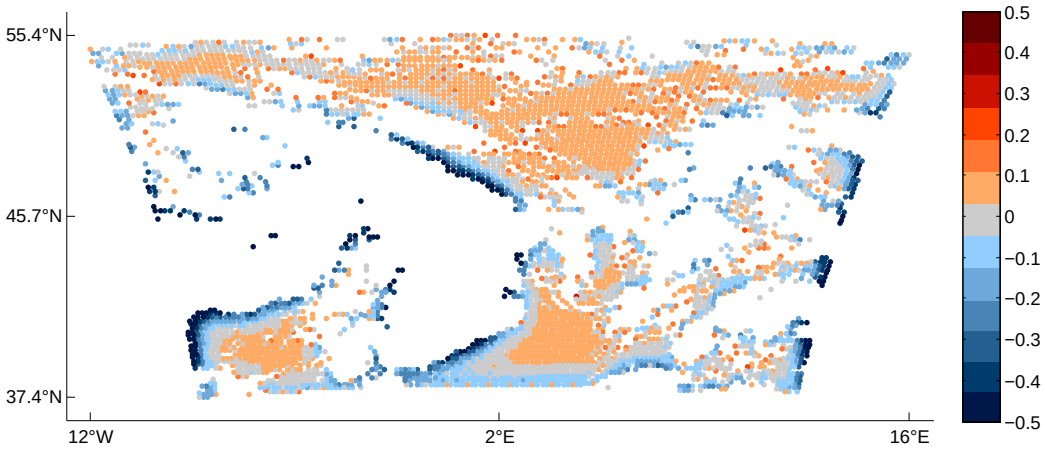


FIGURE 4.23 – Carte des erreurs d'amplitude en utilisant la condensation de masse.

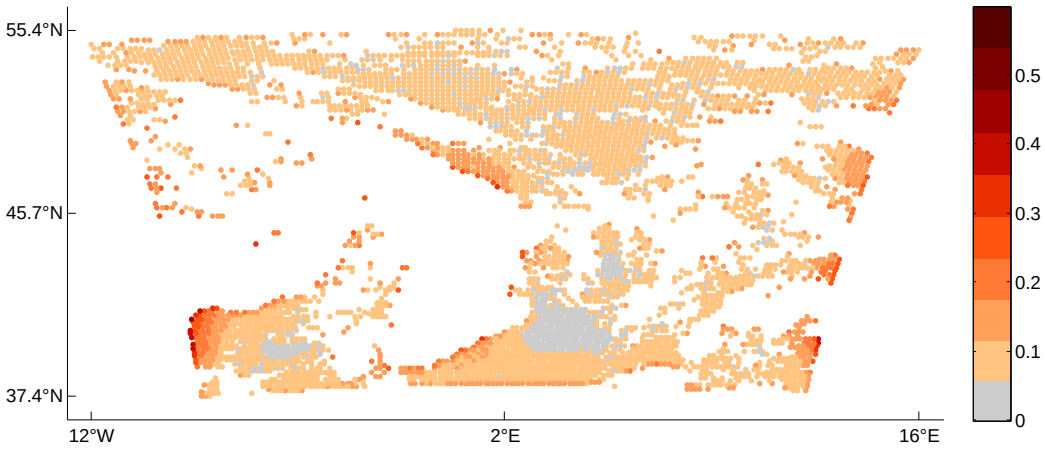


FIGURE 4.24 – Carte des erreurs de forme en utilisant la condensation de masse.

maillage. Toutefois, on constate que l'erreur d'amplitude dépasse 15% presque partout en présence de condensation de masse, alors qu'elle se situe sous le seuil des 10% avec le schéma standard. Il est donc nécessaire et très fortement recommandé d'adopter une méthode de correction de l'amplitude en cas d'utilisation de la condensation de masse. Ces méthodes seront présentées dans la section 4.5.

4.4.2 Test de l'adjoint

L'opérateur de corrélation inverse $\mathcal{C}^{-1} : H^{2m} \to L^2$ est symétrique, c'est-à-dire que pour tout $f \in H^{2m}$ et tout $g \in H^{2m}$,

$$\langle f, \mathcal{C}^{-1}(g)\rangle_{L^2} = \langle \mathcal{C}^{-1}(f), g\rangle_{H^{-2m},H^{2m}}. \tag{4.37}$$

La démonstration de (4.37) découle directement de la propriété (1.36). De même, l'opérateur de corrélation $\mathcal{C} : L^2 \to H^{2m}$ est lui-même symétrique. En effet, on a :

$$\begin{aligned} \langle f, \mathcal{C}g\rangle_{H^{-2m},H^{2m}} &= \langle \mathcal{C}^{-1}\mathcal{C}f, \mathcal{C}g\rangle_{H^{-2m},H^{2m}} \\ &= \langle \mathcal{C}f, \mathcal{C}^{-1}\mathcal{C}g\rangle_{L^2} \\ &= \langle \mathcal{C}f, g\rangle_{L^2}, \end{aligned} \tag{4.38}$$

où les parenthèses ont été omises par souci de clarté, $f \in L^2$ et $g \in L^2$.

Lorsque $\mathcal{C}^{-1} : H^m \to H^{-m}$, alors on utilise utilise la symétrie de $\mathcal{C}^{-1/2} : H^m \to L^2$ pour montrer la symétrie de $\mathcal{C}^{1/2} : L^2 \to H^m$. Les espaces de départ et d'arrivée de ces applications sont représentés sur le diagramme de la figure 4.25.

On exploite alors la relation (1.41) pour prouver la symétrie de $\mathcal{C}$:

$$\begin{aligned} \langle f, \mathcal{C}g\rangle_{H^{-m},H^m} &= \langle f, (\mathcal{C}^{1/2})(\mathcal{C}^{1/2})^{\mathrm{T}} g\rangle_{H^{-m},H^m} \\ &= \langle (\mathcal{C}^{1/2})^{\mathrm{T}} f, (\mathcal{C}^{1/2})^{\mathrm{T}} g\rangle_{L^2} \\ &= \langle (\mathcal{C}^{1/2})(\mathcal{C}^{1/2})^{\mathrm{T}} f, g\rangle_{H^m,H^{-m}} \\ &= \langle \mathcal{C}f, g\rangle_{H^m,H^{-m}}, \end{aligned} \tag{4.39}$$

où $\langle \cdot, \cdot \rangle_{H^m,H^{-m}}$ désigné le crochet de dualité à droite défini pour tout $u \in H^m$ et $v \in H^{-m}$ par

$$\langle u, v\rangle_{H^m,H^{-m}} = v(u). \tag{4.40}$$

Cette propriété de symétrie est essentielle en pratique, car son absence est la cause de nombreuses erreurs numériques parfois difficiles à diagnostiquer. On s'efforce donc de toujours vérifier que les opérateurs de corrélation et de covariance sont symétriques. On retrouve cette propriété dans sa version discrète. On a

$$\boldsymbol{w}_1^{\mathrm{T}} \boldsymbol{C} \boldsymbol{w}_2 = \boldsymbol{w}_1^{\mathrm{T}} \boldsymbol{C}^{\mathrm{T}} \boldsymbol{w}_2 \tag{4.41}$$

où $\boldsymbol{w}_1$ et $\boldsymbol{w}_2$ sont des vecteurs de $\mathbb{R}^p$. On en déduit que $\boldsymbol{C} = \boldsymbol{C}^{\mathrm{T}}$.

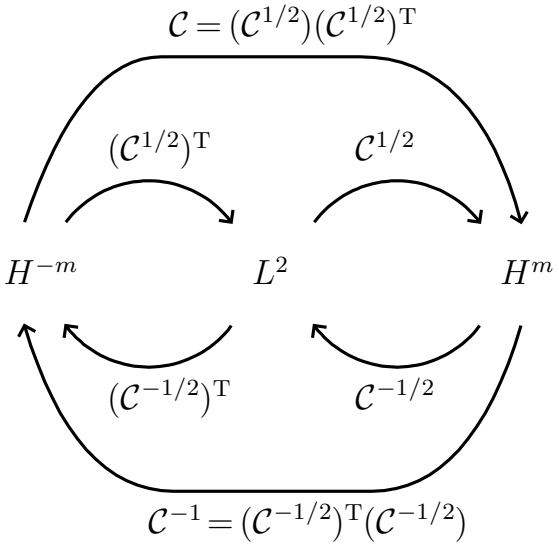


FIGURE 4.25 – Diagramme des espaces d'arrivée et de départ des applications $\mathcal{C}$, $\mathcal{C}^{1/2}$, $\mathcal{C}^{-1/2}$ et leurs transposées.

Pour vérifier (4.41), on considère deux vecteurs aléatoires indépendants de $\mathbb{R}^p$ tels que

$$\boldsymbol{w}_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \quad \text{et} \quad \boldsymbol{w}_2 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{4.42}$$

Ces vecteurs sont des bruits blancs gaussiens et font office de représentants discrets des fonctions non-régulières de H^{-m}. Comme $\boldsymbol{w}_1$ et $\boldsymbol{w}_2$ sont choisis aléatoirement, on suppose que vérifier (4.41) pour un seul couple de vecteurs est suffisant pour conclure que $\boldsymbol{C}$ est symétrique.

Le test de l'adjoint consiste à calculer $\boldsymbol{C}\boldsymbol{w}_1$ et $\boldsymbol{C}\boldsymbol{w}_2$ en étudiant le quotient

$$\frac{(\boldsymbol{C}\boldsymbol{w}_1)^{\mathrm{T}}\boldsymbol{w}_2 - \boldsymbol{w}_1^{\mathrm{T}}(\boldsymbol{C}\boldsymbol{w}_2)}{(\boldsymbol{C}\boldsymbol{w}_1)^{\mathrm{T}}\boldsymbol{w}_2}. \tag{4.43}$$

Lorsque (4.43) est égal à 0 ou a une valeur très proche de la précision machine, le test de l'adjoint est passé avec succès. Dès que cette valeur n'est pas égale à zéro, le test n'est pas passé et l'utilisation de $\boldsymbol{C}$ comme matrice de corrélation est susceptible d'introduire des erreurs numériques.

Un test analogue existe pour la validation de la racine $\boldsymbol{C}^{1/2}$. Il suffit de considérer

$$\boldsymbol{w}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \quad \text{et} \quad \boldsymbol{w}_2 = \boldsymbol{C}^{1/2}\boldsymbol{w}_1 \tag{4.44}$$

et d'étudier le rapport

$$\frac{\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{w}_2 - \boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{C}^{1/2}\boldsymbol{w}_2}{\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{w}_2} \tag{4.45}$$

en remarquant que $\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{w}_2 = (\boldsymbol{C}^{1/2}\boldsymbol{w}_1)^{\mathrm{T}}\boldsymbol{w}_2$. Encore une fois, le résultat de ce test doit être aussi proche que possible de la précision machine.

4.5 Méthodes de correction de l'amplitude

Dans la section 3.3 sont présentées des techniques de correction analytique de l'amplitude des fonctions de corrélations modélisées par l'équation de diffusion. Ces techniques sont adaptées au cas de la diffusion homogène sur un maillage régulier. Elles permettent notamment de corriger l'amplitude des corrélations près des frontières du domaine de discrétisation, où les conditions aux limites de Neumann sont susceptibles d'introduire des effets de réflection.

Toutefois, les erreurs provenant de la discrétisation et de la structure du maillage ne peuvent pas être corrigées par un coefficient analytique qui, par définition, n'est défini qu'à partir de critères continus. Il convient donc d'établir d'autres stratégies de normalisation plus générales, qui s'adaptent au cas de la diffusion hétérogène sur des maillages non-réguliers. La présentation de telles stratégies fait l'objet de cette section.

4.5.1 Normalisation exacte et randomisation

Les éléments de la matrice de corrélation $\boldsymbol{C}$ de la formule (3.2) sont égaux à 1 lorsque le maillage est régulier. En revanche, comme le montrent les expériences de la section 4.4, ce n'est pas le cas lorsque le maillage des observations est non-structuré. Une manière de renormaliser la matrice $\boldsymbol{C}$ est d'utiliser ses éléments diagonaux $\boldsymbol{C}_{\mathrm{ii}}$, lesquels correspondent aux amplitudes des fonctions de corrélation aux points $\boldsymbol{z}_i$ correspondants.

Soit $\boldsymbol{D}$ la matrice contenant les éléments diagonaux de $\boldsymbol{C}$,

$$\begin{cases} \boldsymbol{D}_{\mathrm{ii}} = \boldsymbol{C}_{\mathrm{ii}} \\ \boldsymbol{D}_{\mathrm{ij}} = 0 \quad \text{lorsque} \quad i \neq j. \end{cases} \tag{4.46}$$

La matrice renormalisée $\hat{\boldsymbol{C}}$ s'obtient en recourant à la matrice $\boldsymbol{D}^{-1/2}$ (qui existe, puisque les $\boldsymbol{C}_{\mathrm{ii}}$ sont tous non-nuls et positifs). On a ainsi

$$\hat{\boldsymbol{C}} = \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{C} \boldsymbol{D}^{-\frac{1}{2}}, \tag{4.47}$$

qui est encore symétrique.

Cette méthode est exacte, c'est-à-dire que les éléments diagonaux de $\hat{\boldsymbol{C}}$ sont exactement égaux à 1. En revanche, dans le cas où $\boldsymbol{C}$ est construite à partir d'une équation de diffusion, ses coefficients ne sont jamais calculés explicitement. On n'a donc pas d'accès direct à $\boldsymbol{C}_{\mathrm{ii}}$. Pour les calculer, il faut effectuer le calcul (4.30) pour tous les $i \in [\![1, p]\!]$, ce qui rend la méthode très coûteuse. Il est donc exclu d'utiliser la normalisation exacte, qu'on qualifiera de « force brute », en pratique.

L'estimation statistique par échantillonnage, renommée « randomisation » par le biais d'un angliscisme communément admis, consiste à estimer la matrice $\boldsymbol{D}$ à partir d'un ensemble de tirages aléatoires. Pour ce faire, on invoque le caractère symétrique défini positif de $\boldsymbol{C}$ pour écrire sa factorisation de Cholesky sous la forme

$$\boldsymbol{C} = \boldsymbol{U}\boldsymbol{U}^T, \tag{4.48}$$

et on constate que si $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ et $\boldsymbol{\zeta} = \boldsymbol{U}\boldsymbol{\eta}$, alors $\mathrm{Cov}(\boldsymbol{\zeta}) = \boldsymbol{C}$. Pour estimer $\boldsymbol{D}$, il suffit donc de calculer la somme

$$\frac{1}{p' - 1} \sum_{i \in [\![1, p']\!]} (\boldsymbol{\zeta}_i - \bar{\boldsymbol{\zeta}}) \circ (\boldsymbol{\zeta}_i - \bar{\boldsymbol{\zeta}}) \tag{4.49}$$

où $\circ$ désigne le produit de Schur [5], $(\boldsymbol{\zeta}_i)_{i \in [\![1, p']\!]}$ sont p' réalisations indépendantes de $\mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$ et $\bar{\boldsymbol{\zeta}}$ est la moyenne des réalisations, d'expression

$$\frac{1}{p'} \sum_{i \in [\![1, p']\!]} \boldsymbol{\zeta}_i. \tag{4.50}$$

La matrice $\boldsymbol{D}$ s'obtient en prenant pour éléments diagonaux les éléments de (4.49) et en imposant $\boldsymbol{D}_{\mathrm{ij}} = 0$ dès que $i \neq j$.

Bien que cette méthode soit moins coûteuse que la méthode par force brute, elle nécessite de résoudre p' fois le système linéaire complet. Dans la

5. Le produit de Schur de deux matrices $\boldsymbol{A} = (\boldsymbol{A}_{\mathrm{ij}})$ et $\boldsymbol{B} = (\boldsymbol{B}_{\mathrm{ij}})$ de même dimension correspond au produit terme à terme de ces deux matrices. Ainsi, le produit de Schur $\boldsymbol{A} \circ \boldsymbol{B}$ est la matrice de même dimension que $\boldsymbol{A}$ et $\boldsymbol{B}$ de terme général $(\boldsymbol{A} \circ \boldsymbol{B})_{\mathrm{ij}} = \boldsymbol{A}_{\mathrm{ij}} \times \boldsymbol{B}_{\mathrm{ij}}$.

pratique, on choisit $p' \ll p$, mais la convergence de (4.49) est lente [Wishart, 1928, Ménétrier et al., 2015a] et il peut être nécessaire d'effectuer $p' = 1000$ ou 10000 réalisations avant d'avoir un résultat satisfaisant.

4.5.2 Méthodes avancées

Dans Raynaud et al. [2009] (voir aussi Ménétrier et al. [2015a,b]), une méthode de filtrage est proposée, visant à réduire le nombre d'échantillons nécessaires à la randomisation. La réduction est efficace quand le tenseur de corrélation varie peu par rapport à sa moyenne et par rapport à la longueur de portée moyenne. Cependant, les erreurs numériques dues à la distribution hétorogène des données ne rentrent pas dans ce cadre, car elles introduisent des changements brutaux dans l'amplitude des corrélations, et donc dans la variance du signal. Il en va de même pour les techniques de régression, qui estiment les coefficients de $\boldsymbol{D}$ en certains points du domaine et interpolent les coefficients correspondant aux autres emplacements.

En revanche, la réponse impulsionnelle de l'opérateur de corrélation étant une fonction quasi-gaussienne, elle prend des valeurs quasi-nulles partout à l'exception d'un sous-domaine de $\mathcal{D}$, centré autour de l'origine de l'impulsion. On peut donc améliorer la méthode de force brute en ne résolvant à chaque fois l'équation que sur un sous-domaine de petites dimensions. Ce procédé permet de grandement accélérer l'estimation de $\boldsymbol{D}$.

4.5.3 Régression à partir de la qualité du maillage

La comparaison des figures 4.21 et 4.18 permet de constater que la magnitude de l'erreur d'amplitude sur les bords intérieurs du domaine est proportionnelle à la valeur de l'indice de qualité aux mêmes emplacements. Ce constat se retrouve sur la figure 4.26, qui illustre cette correspondance dans le coin sud-ouest du domaine.

Ce phénomène indique qu'il est possible d'utiliser l'information sur le maillage pour corriger l'erreur d'amplitude contenue dans l'opérateur de corrélation. Toutefois, la valeur de l'indice de qualité décroît instantannément dès qu'on s'écarte du bord intérieur, alors que l'erreur d'amplitude décroît progressivement pour atteindre sa valeur minimale dans la zone structurée. Ce phénomène est inhérent à l'équation de diffusion, dont les solutions sont naturellement régulières. Qualitativement, la situation se représente en dimension 1 comme sur la figure 4.27.

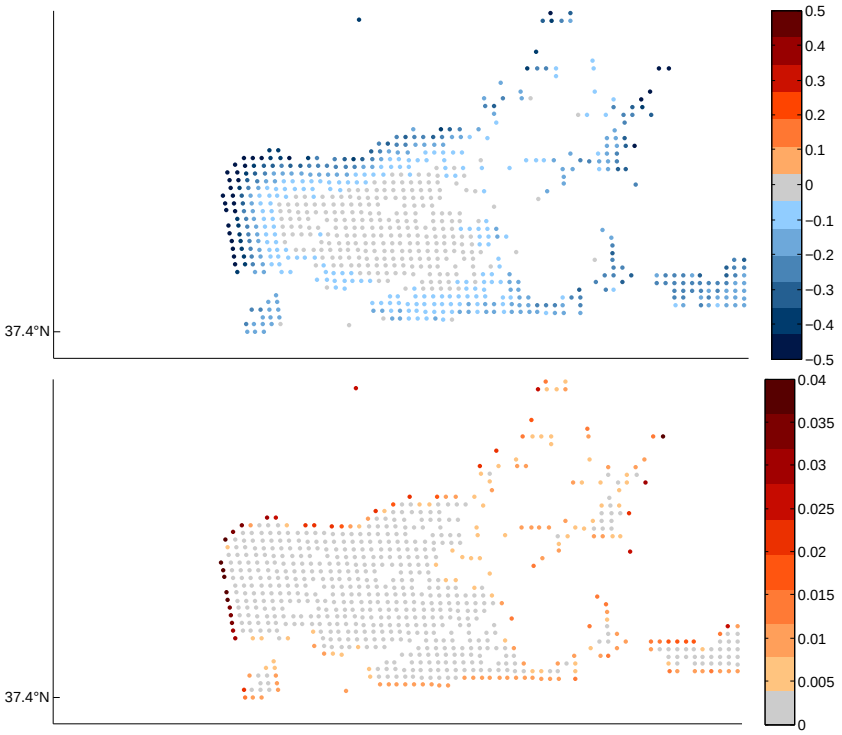


FIGURE 4.26 – Comparaison entre l'erreur d'amplitude et l'indice de qualité sur une portion du domaine. On remarque visuellement que les couleurs correspondent à proximité des désert de données, ce qui suggère l'existence d'une relation de dépendance linéaire entre les deux quantités (au moins sur les bords intérieurs du domaine.

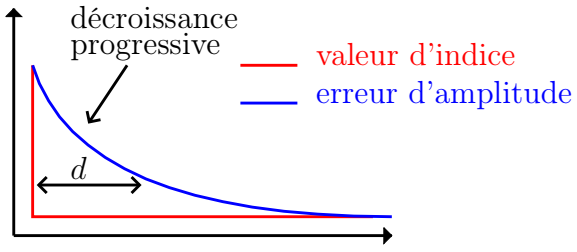


FIGURE 4.27 – Schématisation de la correspondance entre l'erreur et l'indice de qualité en dimension 1.

Pour exploiter la correspondance entre l'erreur et l'indice de qualité, on propose de calculer une régression linéaire entre les deux, à partir des données

expérimentales issues de SEVIRI (figure 4.28).

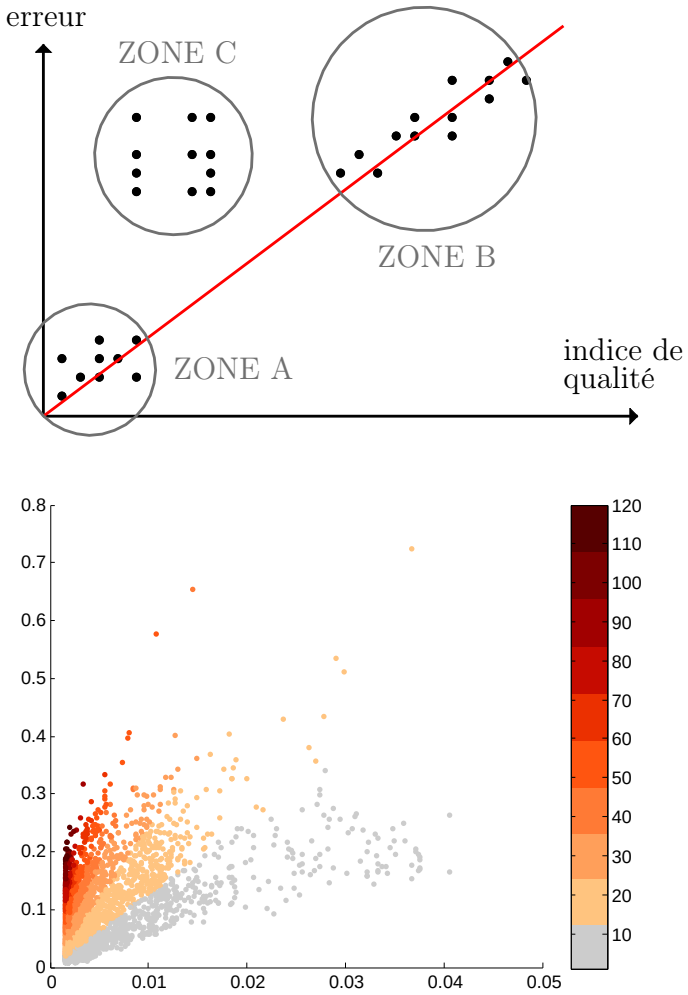


FIGURE 4.28 – En haut : allure symbolique de la régression entre l'erreur et l'indice de qualité du maillage. En bas : erreur en fonction de l'indice calculée à partir des données de SEVIRI. Les couleurs représentent le rapport entre les deux valeurs. Ces mêmes couleurs sont reportées sur la figure 4.29.

Trois zones sont à distinguer. Dans la zone A, les normes des erreurs sont faibles et correspondent à un indice de qualité proche de zéro. Elle caractérise donc les zones denses en observations qui se situent loin des déserts de données. La zone B correspond aux bords intérieurs du domaine, où la

valeur d'indice et l'erreur sont forts. Dans les zones A et B, on s'attend à retrouver une relation de proportionalité entre abscisses et ordonnées. En revanche, la zone C représente les noeuds situés sur des bons triangles, où l'erreur d'amplitude est due à la proximité d'éléments de mauvaise qualité. Ces noeuds ne se situent pas sur la droite de régression. L'estimation d'un facteur de normalisation pour les noeuds de la zone C n'est donc pas triviale. **Ce sujet doit faire l'objet d'investigations plus poussées**.

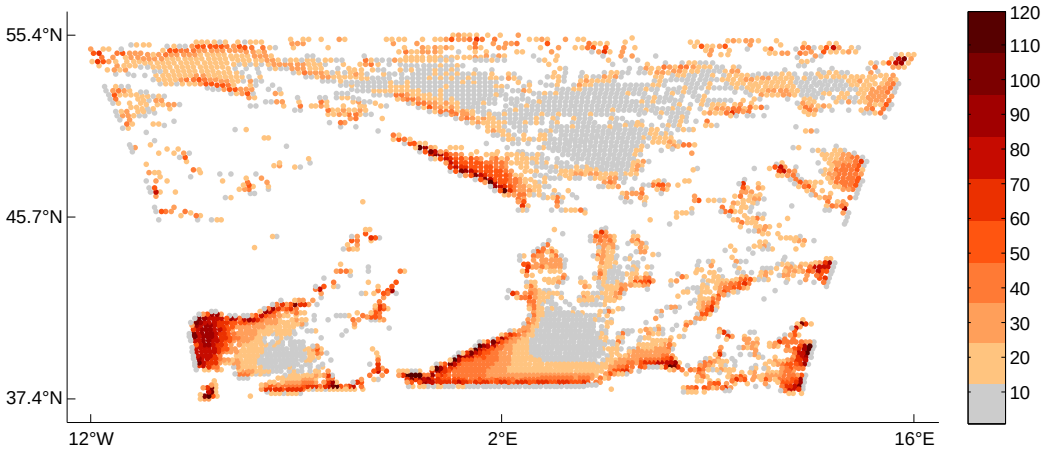


FIGURE 4.29 – Rapport entre l'erreur et l'indice de qualité du maillage en fonction de l'emplacement dans le domaine. Les bords intérieurs du domaine sont gris, comme les zones denses en observations. On confirme donc qu'on peut corriger les zones A et B en exploitant la même dépendance linéaire. Mais rien d'indique comment traiter la zone C.

4.6 L'essentiel du chapitre

La méthode des éléments finis est naturellement adaptée à la résolution des équations aux dérivées partielles sur des maillages non structurés. Toutefois, la distribution spatiale très hétérogène des observations satellitaires pousse la méthode jusqu'à ses limites. Un diagnostic a posteriori permet de mettre en évidence deux erreurs de modélisation d'origine numérique. La première est l'erreur sur l'amplitude des fonctions de corrélation contenues dans les matrices modélisées. La deuxième est l'erreur sur la forme de ces fonctions de corrélation. La compréhension du processus de génération de maillage et de ses propriétés géométrique a permis de mettre au point un indicateur de qualité a priori, qui offre un moyen de prédire l'apparition d'erreurs de

modélisation. Des méthodes de correction d'amplitude sont présentées pour remédier à ce problème.

Les contributions de ce chapitre sont donc multiples. Tout d'abord, un outil de maillage a été développé. Les détails de son implémentation ont permis de développer un indice de qualité pour notre application. L'application de la méthode des éléments finis à la modélisation des opérateurs de corrélation est inédite aux yeux de la communauté de l'assimilation de données. Enfin, une étude rapide a permis d'estimer un jeu de paramètres de corrélation réaliste à partir de données satellitaires opérationnelles.

Ce qu'il faut retenir :

- La distribution spatiale des observations satellitaires est très hétérogène à cause de l'étape de *screening*.
- Le maillage s'appuyant sur ces données est obtenu par une triangulation de Delaunay.
- La méthode des éléments finis est adaptée au maillages non structurés.
- Un grand nombre de données est jeté à cause du *thinning* et du *superobing*.
- Dans le futur, prendre en compte les corrélations spatiales d'erreurs d'observation en assimilation de données permettra de jeter moins de données.
- Pour valider un opérateur de corrélation, on visualise ses colonnes et on vérifie son caractère symétrique défini positif.
- Les expériences sont réalisées à partir d'observations réalistes.
- Deux types d'erreurs sont diagnostiquées : erreur d'amplitude et erreur de forme. Elles sont dues à la dégénérescence locale du maillage s'appuyant sur les observations.
- Il existe des moyens de corriger l'erreur d'amplitude.
- D'ailleurs cette erreur semble liée aux propriétés géométriques du maillage.
- Un indicateur de qualité fiable se base sur la présence de triangles dont les cercles circonscrit ont des rayons importants.
- Les paramètres de corrélation sont également réalistes puisqu'ils sont estimés directement à partir des données.
- Toutefois, le modèle de Matérn ne parvient pas parfaitement à décrire les données.
- Pour mieux les décrire, il faudrait ajouter un terme de régularisation ou, dans l'idéal, additionner deux modèles de Matérn.

Troisième partie

Contrôle de la précision des opérateurs

Chapitre 5

Stratégies de raffinement de maillage

Nous avons décrit une méthode basée sur les éléments finis pour modéliser les corrélations spatiales d'erreur d'observation. L'application de cette méthode à des images satellitaires opérationnelles a fait naître la question de sa dépendance aux données et au maillage construit à partir de ces données. En particulier, on a montré que la méthode permettait de représenter des fonctions de corrélation avec précision sur la quasi-totalité du domaine. Cependant, aux bords des déserts d'observation créés par le prétraitement des données, on a remarqué que les fonctions de corrélation modélisées n'atteignaient pas la bonne amplitude. Dans le chapitre 4, plusieurs solutions sont proposées, mettant toutes en avant une renormalisation *a posteriori* de la matrice de corrélation.

Dans ce chapitre, on s'intéresse aux stratégies permettant de contrôler *a priori* la précision de la modélisation des opérateurs de corrélation. L'idée générale est de mettre au point une stratégie de modélisation permettant d'éviter les écueils évoqués précédemment, afin de s'affranchir de la dépendance aux données.

Dans la section 5.1, on sensiblise le lecteur au raffinement de maillage et on discute des stratégies envisageables dans le cadre de cette étude. On explique ensuite dans la section 5.2 comment construire les opérateurs de transfert entre le maillage des observations et le maillage raffiné. Pour ce faire, on fait appel à des notions proches des méthodes multigrille. Dans la section 5.3, la formule de l'opérateur de corrélation avec raffinement de maillage est présentée et sa performance est évaluée à partir des données de

Seviri. La section 5.4 discute de la possibilité d'appliquer l'approche susmentionnée à la modélisation de l'inverse de l'opérateur de corrélation. Enfin, la section 5.5 traite de l'effet de la condensation de masse sur les performances du raffinement de maillage.

Les contributions de ce chapitre incluent l'implémentation de plusieurs stratégies de raffinement de maillage, la mise en place d'expériences comparant les avantages et les inconvénients de chacune et la mise au point d'un nouveau modèle de corrélation. Ce modèle a pour but de corriger les erreurs diagnostiquées dans le chapitre 4. Le cadre de l'analyse fonctionnelle révèle tout son intérêt quand il s'agit de travailler avec deux maillages à la fois. Encore une fois, le diagramme de dualité offre une vision globale des espaces et des opérateurs manipulés.

5.1 Aspects pratiques du raffinement de maillage

La motivation principale du raffinement de maillage est l'introduction de points de calcul additionnels dans le maillage afin d'obtenir un maillage de meilleure qualité, et ainsi d'améliorer la précision des calculs. L'objectif est donc de considérer le maillage construit à partir des données de Seviri, de cibler les zones responsables d'erreurs numériques dans la résolution de l'équation de diffusion, et de raffiner ces zones en introduisant des points supplémentaires.

On distingue ainsi deux maillages (ou « grilles » dans le vocabulaire multigrille, voir Trottenberg et al. [2001]). Le maillage des observations est qualifié de « grossier ». Il s'oppose au maillage « fin », qui comprend les observations ainsi que les points supplémentaires. La sous-section 5.1.1 présente différentes méthodes de construction du maillage fin, tandis que la sous-section 5.1.2 discute des aspects algorithmiques sous-jacents.

Remarque : les différentes techniques de raffinement de maillage évoquées dans cette section sont implémentées dans Matlab. Cet environnement offre de nombreuses facilitées de développement, particulièrement adaptées au calcul numérique. Néanmoins, les techniques présentées dans ce manuscrit peuvent être implémentées dans n'importe quel langage de programmation. En l'absence de logiciel de maillage facile à interfacer avec Matlab, les algorithmes sont codés à la main, incorporant un minimum d'optimisations.

Le code comprend néanmoins toutes les fonctions standards d'un mailleur classique, en plus des fonctions spécifiques à la modélisation des corrélations.

5.1.1 Types de raffinement

Il n'existe pas de méthode optimale pour le raffinement de maillage. A chaque application correspond une ou plusieurs stratégies optimales. On présente les stratégies logiquement applicables à la modélisation des opérateurs de corrélation sur des maillages irréguliers. Les avantages et les inconvénients de chacune sont présentés. L'objectif est d'introduire le raffinement de manière intuitive, en justifiant pourquoi certaines méthodes ne sont pas abordées dans la suite de cette étude. Enfin, cette sous-section offre l'opportunité de discuter d'approches alternatives telles que le déraffinement et les maillages hybrides, qui sont pertinentes dans le cadre de notre étude.

Pour expliquer les différences entres les diverses approches, considérons le maillage de référence de la figure 5.1. Il comprend deux triangles qui partagent une arête commune. L'un des triangles comporte un angle très large. On souhaite donc utiliser le raffinement de maillage.

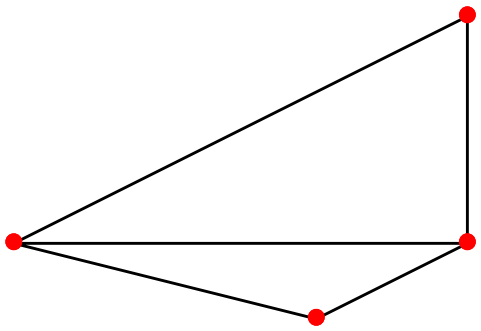


Figure 5.1 – Maillage de référence. Les noeuds de référence sont en rouge et les arêtes de référence en noir.

Approche 1 : h-raffinement non hiérarchique

Le premier type de raffinement est le raffinement en espace. Qualifié de « h-raffinement » car il joue sur la taille h des éléments (dénomination provenant des méthodes multigrilles), il est obtenu en rajoutant des noeuds dans

le maillage. La figure 5.2 donne l'exemple d'un h-raffinement non hiérarchique, dont les noeuds supplémentaires sont ajoutés sans préoccupation de la position des arêtes de référence.

Les points supplémentaires sont ajoutés dans les zones où les données manquent dans le maillage de référence. Ensuite, la position des arêtes est recalculée pour prendre en compte l'introduction des nouveaux noeuds. De ce procédé résulte un maillage souvent très homogène. Toutefois, les arêtes du maillage de référence ne se retrouvent pas systématiquement dans le maillage final, d'où l'appellation « non hiérarchique ». Dans la section 5.2, on montre que cela peut nuire à la qualité des opérateurs de transfert, qui permettent d'interpoler l'information entre le maillage de référence (« grossier ») et le maillage raffiné (« fin »).

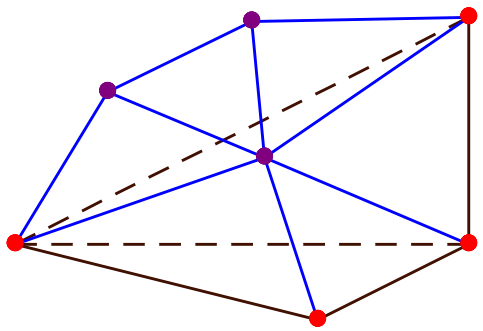


Figure 5.2 – h-raffinement non hiérarchique. Les noeuds de référence sont en rouge et les arêtes de référence en noir. Les arêtes supprimées lors du raffinement sont représentées par des tirets. Les nouveaux noeuds sont en violet et les nouvelles arêtes en bleu.

Approche 2 : h-raffinement hiérarchique

On présente maintenant le h-raffinement hiérarchique. Cette fois-ci, les points supplémentaires sont introduits sur les arêtes de référence, de telle sorte que le maillage grossier soit inclus dans le maillage fin.

Cependant, il n'existe pas une unique façon d'ajouter des points dans le maillage. On présente ici deux méthodes connues qui sont souvent utilisées en complément l'une de l'autre. La première méthode est la bisection (figure 5.3, gauche). Dans cette méthode, le triangle à raffiner est scindé en deux

triangles plus petits en introduisant un noeud additionnel au milieu de son côté le plus grand. Ensuite, le point nouveau point est raccordé à ses voisins afin qu'il ne reste pas au milieu d'une arête.

Facile à coder, la bisection est la méthode qui introduit le moins de noeuds supplémentaires en général. Toutefois, il n'est pas garanti que les éléments du maillage fin soient de meilleure qualité que les éléments du maillage grossier, cela même si le maillage fin contient globalement plus de noeuds de calcul que le maillage grossier. Il peut donc être nécessaire d'effectuer le raffinement en plusieurs étapes (figure 5.5) et d'ajouter un basculement d'arête entre chaque étape (sous-section 4.3.1).

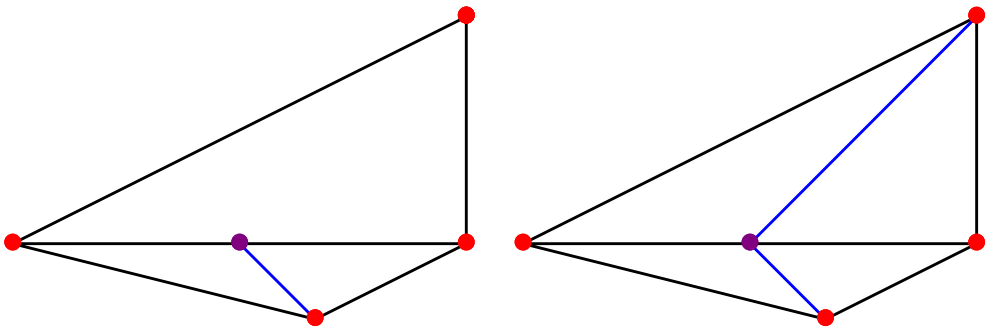


Figure 5.3 – h-raffinement hiérarchique (bisection). A gauche : première bisection. A droite : bisection nécessaire pour que le maillage reste conforme (pas de noeud isolé au milieu d'un arête). Les noeuds de référence sont en rouge et les arêtes de référence en noir. Les nouveaux noeuds sont en violet et les nouvelles arêtes en bleu.

La deuxième méthode est celle de la réduction (figure 5.4). Le triangle raffiné est scindé en quatre triangles plus petits en rejoignant les milieux de ses côtés par des arrêtes. Cette méthode a l'avantage de ne pas dégrader le rapport d'aspect des triangles lors du raffinement. Comme pour la bisection, il est toutefois nécessaire de rajouter une étape de bisection pour rendre le maillage conforme. A noter que cette deuxième étape est presque systématiquement une bisection, car elle permet d'arrêter la propagation des noeuds non conformes.

En général, la réduction introduit plus de noeuds supplémentaires que la bisection. C'est pourquoi les deux méthodes sont souvent utilisées conjoin-

tement pour obtenir le meilleur rendement [Bank et al., 1988, Zhao et al., 2015, Krysl et al., 2004]. Le principe de la réduction est illustré sur la figure 5.5.

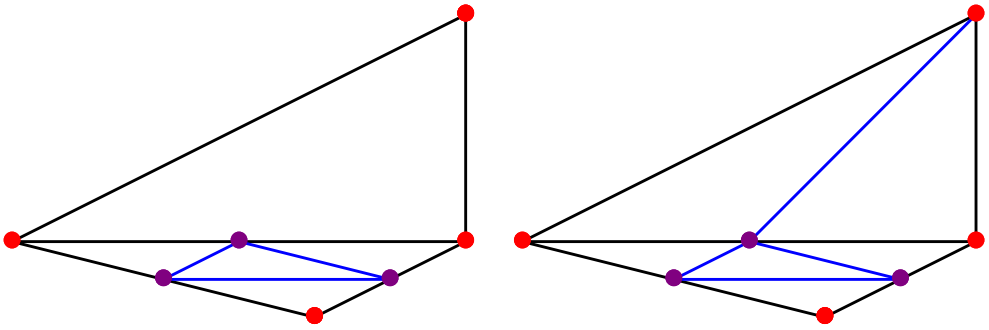


Figure 5.4 – h-raffinement hiérarchique (Réduction). A gauche : première réduction. A droite : bisection nécessaire pour que le maillage reste conforme (pas de noeud isolé au milieu d'un arête). Les noeuds de référence sont en rouge et les arêtes de référence en noir. Les nouveaux noeuds sont en violet et les nouvelles arêtes en bleu.

Ainsi, plusieurs stratégies de raffinement de maillage sont viables en pratique. Pour nos expériences, on a développé un mailleur hybride combinant plusieurs étapes de réduction et de bisection, basé sur le code en libre accès de Persson and Strang [2004]. Le nombre d'étapes est adaptatif et peut être sensiblement réglé par l'utilisateur. Le critère de raffinement combine plusieurs des indicateurs présentés en sous-section 4.3.2 pour un meilleur rendement. Les perspectives d'amélioration comportent l'introduction d'une étape de basculement d'arête entre les différents raffinements. Toutefois, il faut veiller à ce que le basculement ne rompt pas la hiérarchie entre le maillage grossier et le maillage fin.

L'avantage du h-raffinement hiérarchique est qu'il permet de définir des opérateurs de transfert de qualité. En effet, l'espace d'approximation V_c généré par les fonctions de forme sur le maillage grossier est alors inclus dans l'espace d'approximation V_f généré par les fonctions de forme sur le maillage fin (on a $V_c \subset V_f$). Ce n'est pas le cas lorsque le raffinement n'est pas hiérarchique. Plus de détails sont donnés dans la section 5.2.

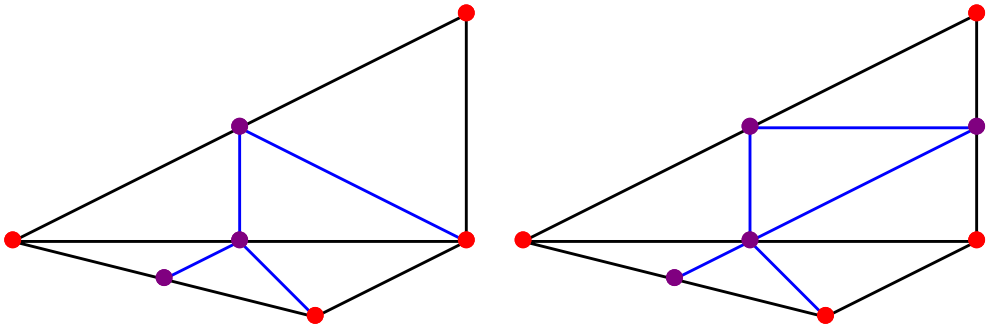


Figure 5.5 – h-raffinement hiérarchique (Multiples raffinements). A gauche : raffinement par multiples bisections. A droite : raffinement combinant bisections et réduction. Les noeuds de référence sont en rouge et les arêtes de référence en noir. Les nouveaux noeuds sont en violet et les nouvelles arêtes en bleu.

Approche 3 : p-raffinement

Le deuxième type de raffinement est le raffinement en ordre [Ern and Guermond, 2010, Canuto et al., 1987], également nommé « p-raffinement » car le « p » fait habituellement référence à l'ordre des fonctions de base d'éléments finis. De manière analogue au h-raffinement, le p-raffinement a pour effet d'augmenter la densité spatiale des degrés de liberté dans le maillage. Cette fois, néanmoins, les degrés de liberté supplémentaires ne résultent pas de l'introduction de noeuds supplémentaires. Ils proviennent de l'augmentation du degré polynomial des éléments de la base d'éléments finis choisi pour la discrétisation de l'équation de diffusion (ou, de manière équivalente, pour la définition de l'espace d'approximation). Schématiquement, les degrés de liberté du maillage fin sont situés comme sur la figure 5.6.

En raison de l'emplacement des degrés de liberté, l'espace d'approximation $\mathcal{P}_1$ est automatiquement inclus dans l'espace d'approximation $\mathcal{P}_k, k \geqslant 1$. On a donc une relation de hiérarchie entre le maillage grossier et le maillage fin. Toutefois, l'évaluation des opérateurs de transfert fait intervenir une formule de quadrature qui n'est pas forcément triviale à calculer [Ern and Guermond, 2010].

L'utilisation du p-raffinement n'est pas exploité de le cadre de cette étude. Nous tenons toutefois à garder sa mention pour de futures investigations.

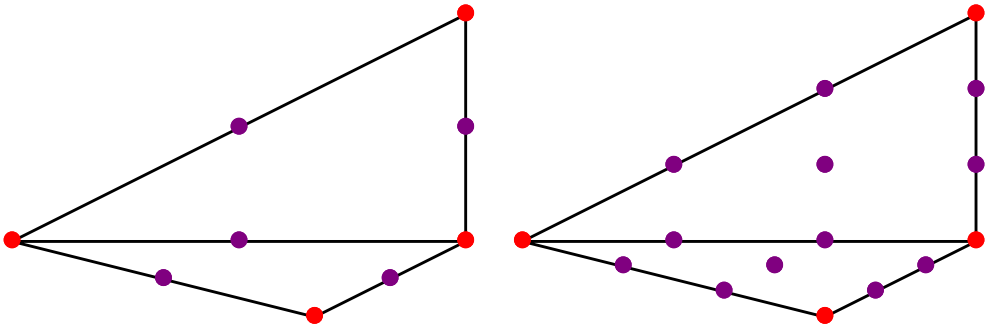


Figure 5.6 – p-raffinement. A gauche : degrés de liberté de la méthode $\mathbb{P}_2$. A droite : degrés de liberté de la méthode $\mathbb{P}_3$. Les noeuds de référence sont en rouge et les arêtes de référence en noir. Les nouveaux degrés de liberté sont en violet.

5.1.2 Types de déraffinement

Si le raffinement de maillage paraît naturel pour obtenir un maillage de qualité supérieure, son antagoniste présente également un intérêt théorique et pratique. Le déraffinement consiste à supprimer des noeuds dans le maillage grossier jusqu'à obtention d'un maillage suffisamment bon (au regard d'un nouveau critère à définir). L'idée est de repérer les noeuds qui sont responsables de l'existence d'éléments mal formés et de le supprimer pour « réparer » le maillage. Un exemple est donné sur la figure 5.7.

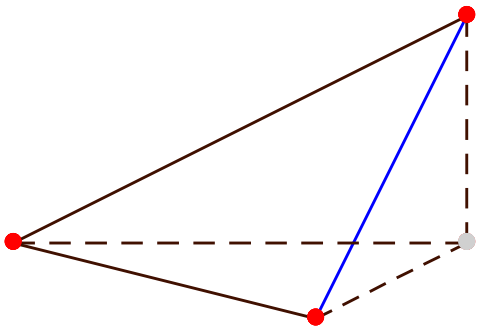


Figure 5.7 – Déraffinement. Les noeuds de référence sont en rouge et les arêtes de référence en noir. Les points supprimés lors du raffinement sont en gris et les arêtes supprimées sont représentées par des tirets. Les nouvelles arêtes sont en bleu.

Le danger d'un tel procédé réside dans la propagation potentielle du front du suppression de noeud dans une implémentation itérative. En effet, supprimer un noeud peut être responsable de l'apparition de nouveaux éléments mal formés dans son voisinage. A l'itération suivante, l'algorithme va donc tenter de supprimer de nouveaux noeuds pour réparer le voisinage en question. Cette réaction en chaîne pose la question du critère d'arrêt dans l'utilisation du déraffinement itératif. Nous ne donnons pas de réponse à cette question dans cette étude.

En revanche, on mentionne une classe de méthodes de raffinement non local, basées sur l'évaluation conjointe de plusieurs noeuds simultanés. Parmis ces méthodes, on choisit d'évoquer la recherche de sous-ensembles maximaux indépendants (MIS, pour « *Maximal Independent Subset* »), l'effrondrement d'arêtes et l'inflation de sphères.

Approche 1 : sous-ensembles maximaux indépendants

Notons $\mathcal{S}$ l'ensemble des noeuds de la triangulation de référence (définition 4.21) et $\mathcal{E}$ l'ensemble de ses arêtes. Chaque arête est définie comme un couple de noeuds $(\boldsymbol{z}_i, \boldsymbol{z}_j)$ vérifiant $i \neq j$. On appelle sous-ensemble maximal indépendant de $\mathcal{S}$ tout sous-ensemble $\mathcal{S}' \subset \mathcal{S}$ vérifiant :

- L'**indépendance** : pour tout $\boldsymbol{z}'_1 \in \mathcal{S}'$ et tout $\boldsymbol{z}'_2 \in \mathcal{S}'$, $(\boldsymbol{z}'_1, \boldsymbol{z}'_2) \notin \mathcal{E}$;
- La **maximalité** : pour tout $\boldsymbol{z} \in \mathcal{S} \setminus \mathcal{S}'$, il existe $\boldsymbol{z}' \in \mathcal{S}'$ tel que $(\boldsymbol{z}, \boldsymbol{z}') \in \mathcal{E}$. C'est-à-dire qu'aucun noeud $\boldsymbol{z} \in \mathcal{S}$ ne peut être ajouté à l'ensemble $\mathcal{S}'$ sans enfreindre son indépendance.

Pour construire le nouveau maillage, il suffit ensuite de construire une triangulation à partir du sous-ensemble $\mathcal{S}'$. Cette étape peut se faire au fur et et mesure de la construction de $\mathcal{S}'$, ou bien dans un second temps. Toutefois, rien ne garantit que le maillage obtenu est de bonne qualité. Habituellement, cette méthode est utilisée pour déraffiner des maillages réguliers (dans le cadre des techniques multigrille).

Voici une heuristique permettant de générer facilement un MIS de $\mathcal{S}$ (figure 5.8) :

1. Ordonner la liste des éléments de $\mathcal{S}$ de telle sorte que les noeuds sur la frontière de $\mathcal{D}$ apparaissent en premier, puis parcourir la liste en commençant par le premier noeud ;

2. Si le noeud actif est déjà banni de $\mathcal{S}'$, passer au noeud suivant ;
3. Sinon, ajouter le noeud actif à $\mathcal{S}'$ et bannir tous ses voisins ;
4. Passer au noeud suivant.

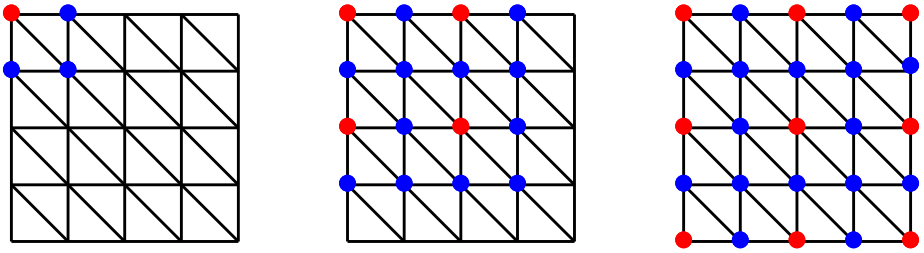


Figure 5.8 – De gauche à droite : détermination d'un sous-ensemble maximal indépendant à partir d'une triangulation cartésienne. Les noeuds sélectionnés dans $\mathcal{S}'$ sont en rouge. Les noeuds supprimés sont en bleu. On remarque que dans le cas cartésien, la procédure équivaut à sélection un noeud sur deux dans chaque dimension de l'espace.

Un exposé des techniques standards, accompagné de nouveaux algorithmes de détermination des MIS, est disponible dans l'article de Luby [1986].

Approche 2 : effrondrement d'arêtes

L'algorithme de déraffinement le plus connu est probablement l'effondrement d'arêtes (*edge collapsing* en anglais, voir par exemple Hoppe [1996]). A chaque étape, deux noeuds voisins proches sont fusionnés en supprimant l'arête du maillage qui les relie. Le maillage est ensuite réarrangé pour assurer sa conformité. Cette technique se rapproche du *superobing* utilisé en traitement des données satellitaires.

Pour commencer, une inspection du maillage permet de marquer les arêtes devant être effondrées. Ensuite, l'algorithme supprime les arêtes marquées en fusionant ses noeuds. Pour se faire, une des extrémités de l'arête est conservée, tandis que l'autre est supprimée. Les éléments dont le noeud a été déplacé voient leur volume augmenter (figure 5.9).

Toutefois, lorsque beaucoup d'arêtes sont marquées lors de l'inspection, la procédure tend à supprimer un grand nombre de noeuds. Il est donc nécessaire

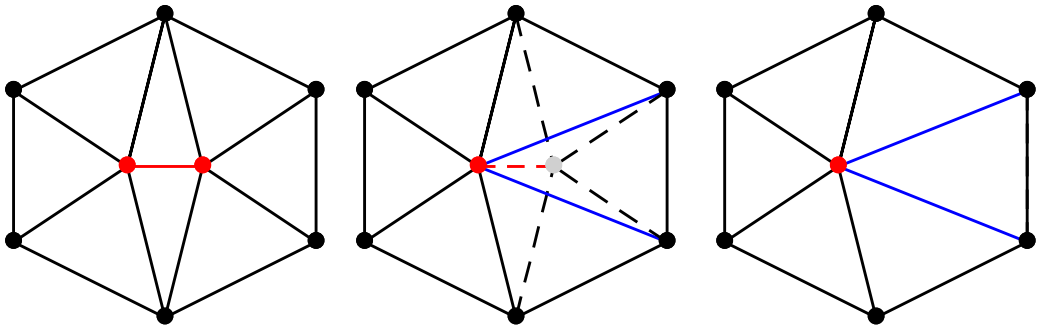


Figure 5.9 – De gauche à droite : effondrement de l'arête centrale marquée en rouge. Les arêtes supprimées sont en lignes tiretées. Les nouvelles arêtes sont en bleu. Le noeud supprimé est en gris.

de contrôler la distribution des noeuds en fin de procédure. Pour ce faire, on détermine un MIS du maillage, contenant des noeuds qui ne peuvent être supprimés.

La suppression des arêtes se fait en trois étapes. Tout d'abord, la qualité des éléments est évaluée conformément aux critères exposés dans la sous-section 4.3.2 (première étape de l'inspection). Puis, pour chaque élément marqué, on décide de supprimer son arête la plus courte (deuxième étape de l'inspection). Enfin, le noeud qui ne fait pas partie du MIS est effondré sur l'autre. Si les deux noeuds sont en dehors du MIS, alors le noeud ayant le plus de voisins est supprimé.

Enfin, l'effondrement d'arêtes est susceptible de produire des structures interdites (figure 5.10), comme des éléments de volume négatif. Une inspection de vérification permet de réparer la structure du maillage final.

Approche 3 : inflation de sphères et fonctions d'espacement

L'inflation de sphère est la dernière méthode de déraffinement présentée dans ce manuscrit. Elle est conçue spécialement pour obtenir des maillages de bonne qualité. Le principe est le suivant : on attribue une sphère à chaque noeud du maillage. Au début, chaque sphère a un rayon nul. Puis, on « gonfle » les sphères jusqu'à ce qu'elles entrent en contact ou qu'elles se recoupent. On détermine ensuite quels noeuds doivent être supprimés pour que les sphères restantes ne soient de nouveau plus en contact. Le maillage final

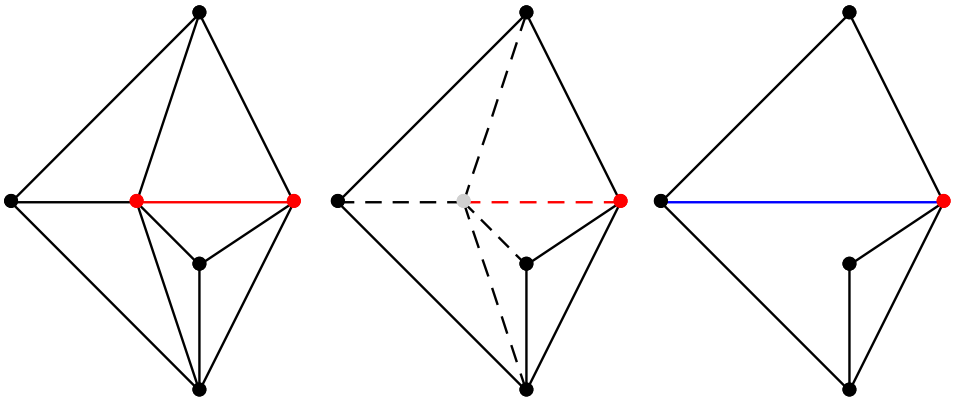


Figure 5.10 – Exemple d'effondrement d'arête illégal.

est celui qui conserve les noeuds « gagnants » (figure 5.11).

Bien sûr, la subtilité de la méthode réside en la façon de gonfler les sphères et la manière de sélectionner et supprimer les noeuds. Une variante consiste à fixer initialement le rayon de chaque sphère à celui de la plus petite boule contenant au moins deux noeuds dans le voisinage. Ce choix n'est pas unique. La fonction régissant le rayon de l'ensemble des sphères est appelée « fonction d'espacement » (*spacing function* en anglais). Le procédé de gonflage des sphères est quant à lui nommé « inflation ».

Enfin, la sélection des noeuds peut se faire en déterminant un MIS du maillage, en incorporant la définition des fonctions d'espacement dans sa construction.

Approche 4 : éléments hybrides

Enfin, tous les maillages ne sont pas constitués de triangles. Nombre de méthodes s'appuient sur des maillages constitués de quadrangles, de polygones réguliers ou encore sur des maillages hybrides mêlant différents types d'éléments. C'est le cas des éléments finis mixtes [Zienkiewicz et al., 2005], des éléments finis spectraux (qui couplent le h- et le p- raffinement, voir Canuto et al. [1987]) et des éléments virtuels (Annexe B). Certaines approches se raccrochent également à l'étude générale des laplaciens sur des graphes (Annexe C).

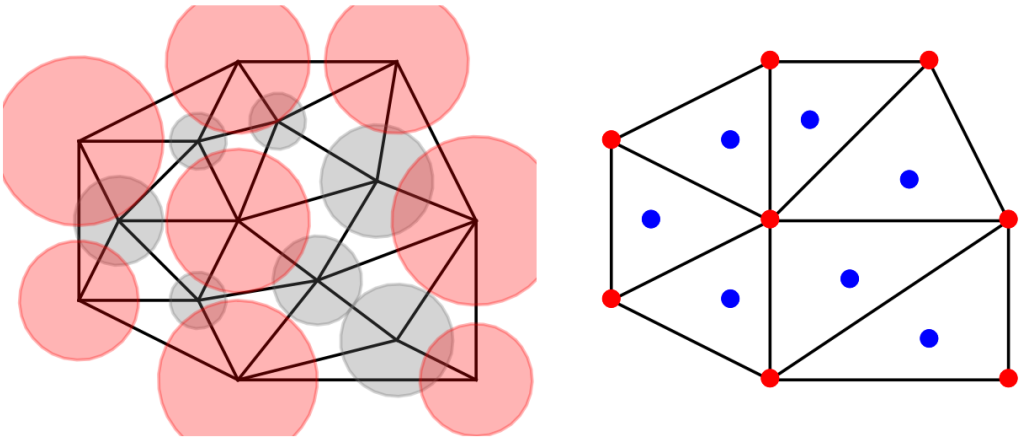


Figure 5.11 – Illustration de l'inflation de sphère (à la main). Le maillage final est composé des noeuds rouges qui avaient des sphères rouges. Les noeuds rejetés sont en bleu et leurs sphères en gris.

A titre d'exemple, on illustre le cas du déraffinement par suppression d'arête (figure 5.12). Deux triangles adjacents peuvent être fusionnés pour donner un quadrangle. Le maillage résultant est un maillage hybride comportant différents types de polygones. La méthode des éléments finis s'appuyant sur un maillage de triangles et de quadrangles est qualifiée de « $\mathbb{P}_1/\mathbb{Q}_1$ », le « $\mathbb{Q}$ » faisant référence aux éléments quadrangulaires. Les formules d'intégration $\mathbb{Q}_1$ sont obtenues similairement au cas $\mathbb{P}_1$ en ramenant les calculs à un élément de référence.

Notre cadre d'application ne profitant pas particulièrement de l'utilisation d'un maillage hybride, il n'est pas pertinent de s'orienter vers cette approche au sein de cette étude. Néanmoins, on mentionne cette approche en raison de sa généralité et des liens forts avec d'autres méthodes que les éléments finis.

Le déraffinement de maillage peut fournir une alternative pertinente au *thinning* actuellement utilisé dans le traitement en amont des données d'observation satellites. En effet, plutôt que de supprimer des observations selon des critères grossiers ou arbitraires, cette étape peut être adaptée de telle sorte de faciliter la génération de maillage et les calculs en éléments finis dans la modélisation des opérateurs de corrélation. L'impact de l'utilisation d'un tel procédé est hors du cadre de cette étude mais fait partie des points à surveiller prioritairement dans l'avenir de la représentation des corrélations d'erreur d'observation.

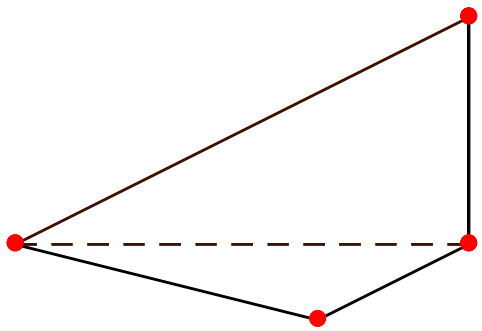


Figure 5.12 – Fusion d'éléments. Les noeuds de référence sont en rouge et les arêtes de référence en noir. Les arêtes supprimées lors du raffinement sont représentées par des tirets.

5.1.3 Maillage sur Seviri

Dans cette sous-section, on illustre les principes évoqués dans la sous-section 5.1.1 pour le raffinement de maillage à partir des données de Seviri. Le maillage initial et la valeur de l'indice de qualité sont rappelés sur les figures 5.13 et 5.14.

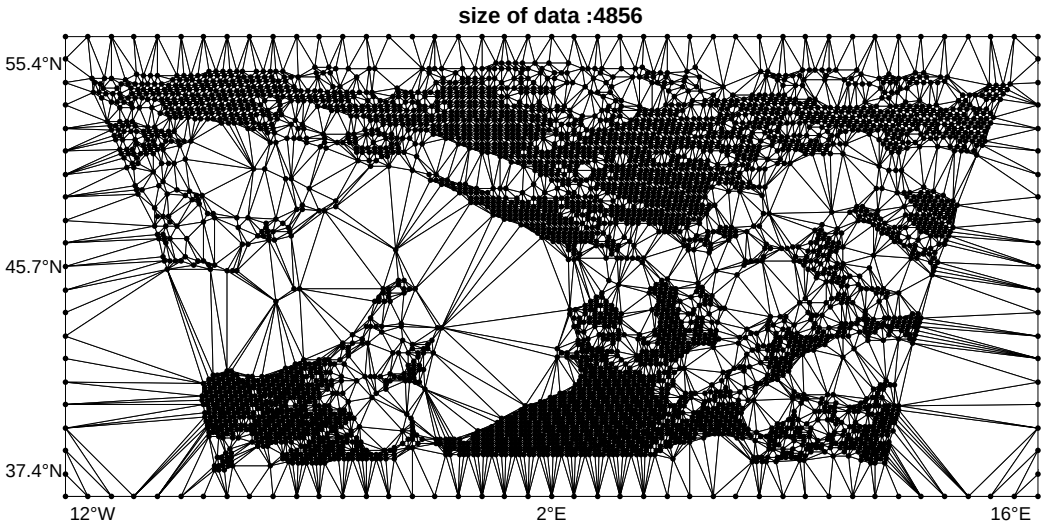


Figure 5.13 – Maillage initial des données Seviri, contenant 4856 noeuds. Les éléments sont étirés ou aplatis dans les et à proximité des déserts de données.

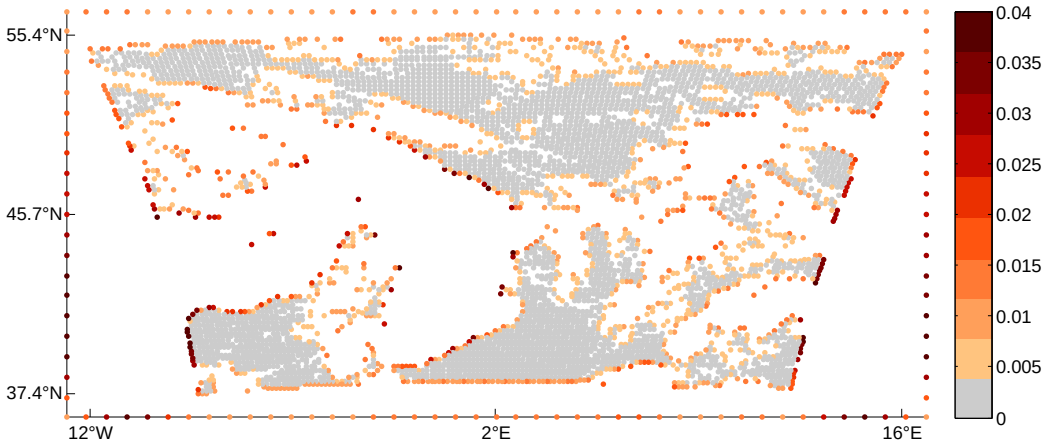


Figure 5.14 – Carte de l'indice de qualité basée sur le rayon du cercle circonscrit, avant raffinement. On remarque la présence d'indice élevé symbolisé par les couleurs foncées.

h-raffinement non hiérarchique

Le *h*-raffinement non hiérarchique est réalisé par superposition d'un maillage régulier, de finesse h, au mail-lage des observations. L'objectif est d'homogénéiser l'espacement maximal entre deux points de calcul dans le maillage final. A l'issue de cette première étape, seuls les noeuds sont conservés. En deuxième étape, les noeuds trop proches des observations sont supprimés et la position des arêtes est calculée par l'algorithme de Delaunay. La troisième étape est itérative. Le centre de masse de chaque élément est calculé. Puis, une force de répulsion est associée à chaque centre, de telle sorte que chaque noeud cherche à se situer le plus loin possible de ses voisins. Un bilan des forces permet de réajuster la position des éléments. Les centres de masse sont recalculés et la procédure est répétée jusqu'à l'équilibre. Cette troisième étape est inspirée par le travail de Persson and Strang [2004], dans lequel une procédure similaire est détaillée en MATLAB. Le lecteur intéressé peut se référer à leur travail pour les détails de l'implémentation. Illustration sur la figure 5.15.

Comme on peut le constater sur la figure 5.16, l'utilisation du *h*-raffinement non hiérarchique permet de réduire l'indice de qualité dans les zones critiques. Le nombre de noeud augmente de 12% lors du raffinement. Bien sûr, ce chiffre pourrait être amélioré en utilisant un mailleur plus performant. Une perspective d'étude consiste à raffiner plus finement à proximité des observations, quitte à garder un maillage plus grossier à distance des noeuds initiaux. La

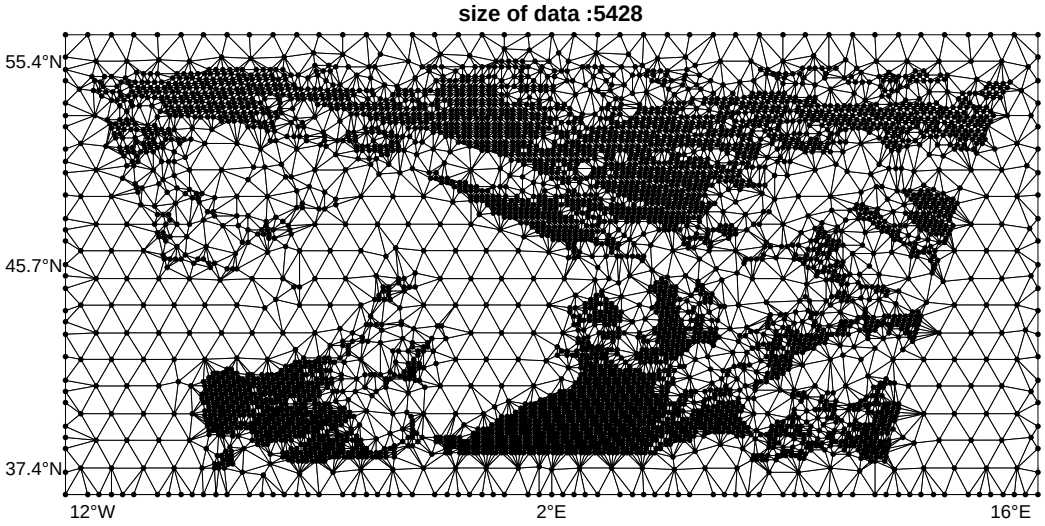


Figure 5.15 – h-raffinement non hiérarchique. La présence de triangles équilatéraux permet de deviner l'allure du maillage régulier superposé aux observations. On remarque visuellement que la qualité des triangles est améliorée près des frontières et dans les déserts de données. Ce maillage comporte 5428 noeuds, soit une augmentation légèrement inférieure à 12% par rapport au maillage initial.

réalisation de ce travail sort du cadre qualitatif de cette étude.

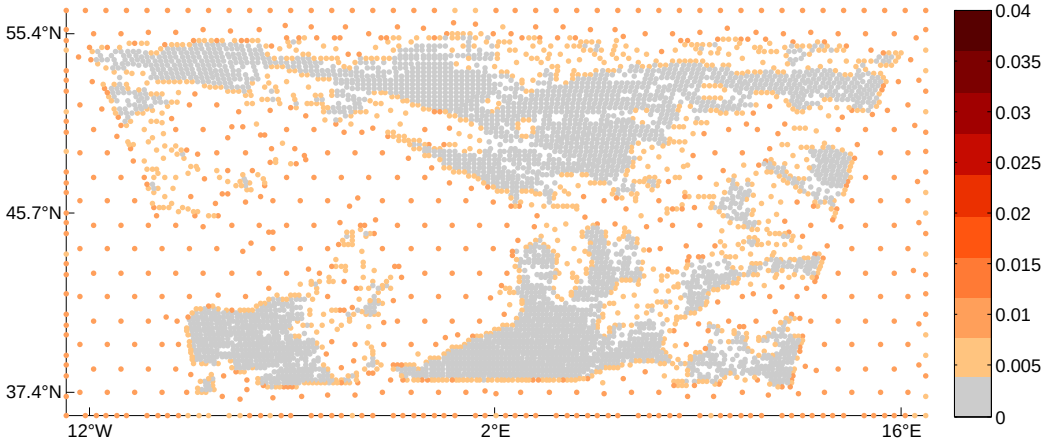


Figure 5.16 – Carte de l'indice de qualité après h-raffinement non hiérarchique. La valeur d'indice passe sous la barre des 0.01 sur la quasi-totalité du maillage. L'absence de couleurs foncées indique que les éléments malformés ont été « réparés » lors de la procédure.

h-raffinement hiérarchique : réduction

La deuxième stratégie de raffinement qu'on choisit de présenter est la réduction. Il s'agit d'un h-raffinement hiérarchique, qui a la particularité de ne requérir aucune itération. Cependant, ce raffinement augmente drastiquement le nombre de noeuds dans le maillage, comme l'atteste la figure 5.17. L'indice de qualité correspondant est observable sur la figure 5.18.

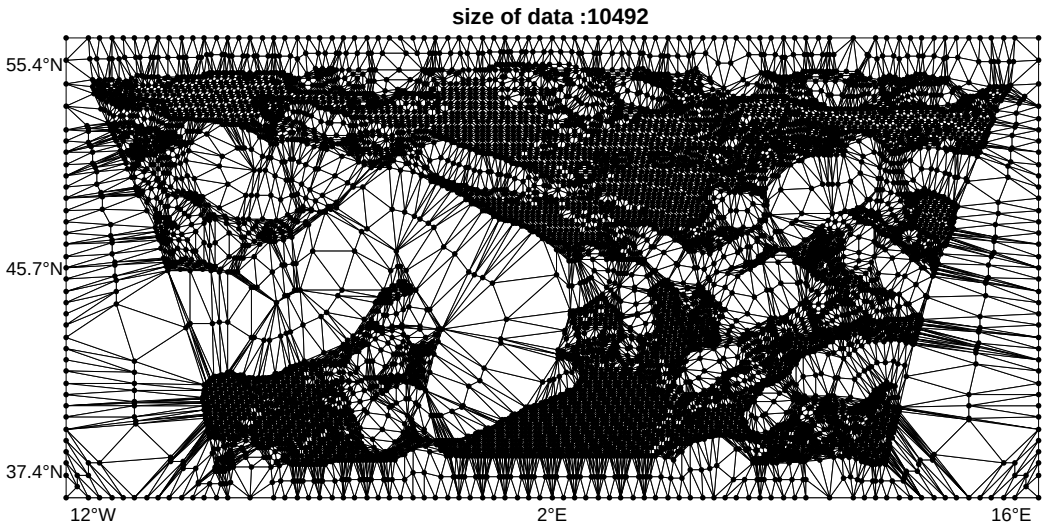


Figure 5.17 – h-raffinement hiérarchique par réduction. Le maillage final contient 10492 noeuds, soit 116% de plus que le maillage initial.

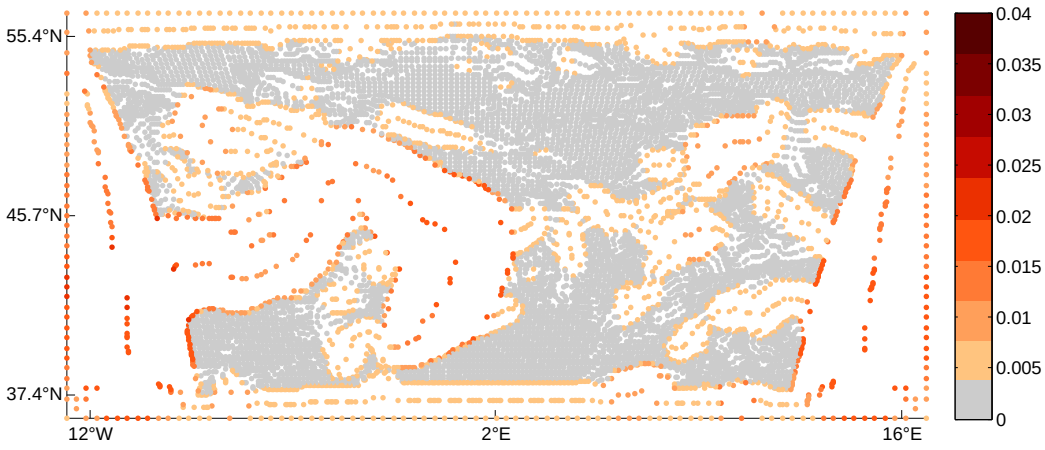


Figure 5.18 – Carte de l'indice de qualité après réduction. Seules quelques zones sont encore foncées. L'emplacement de ces zones est imprévisible et s'observe au cas par cas. Globalement, l'indice de qualité est amélioré sur tout le maillage.

h-raffinement (quasi-)hiérarchique : bisection itérée

Enfin, on présente la bisection itérée, procédé de h-raffinement hiérarchique consistant à exécuter au moins deux étapes de bisection. Comme l'attestent les figures 5.19 et 5.20, la procédure seule ne permet pas d'améliorer la qualité du maillage. Toutefois, une étape de basculement d'arête permet de pallier le problème simplement, quitte à compromettre la relation de hiérarchie par endroits. On parle alors de raffinement quasi-hiérarchique. Le résultat est visualisable sur les figures 5.21 et 5.22. Dans les deux cas, résulte une augmentation de 26% du nombre de noeuds. Une perspective d'étude est d'arriver à définir un basculement d'arête qui ne compromette pas la relation de hiérarchie.

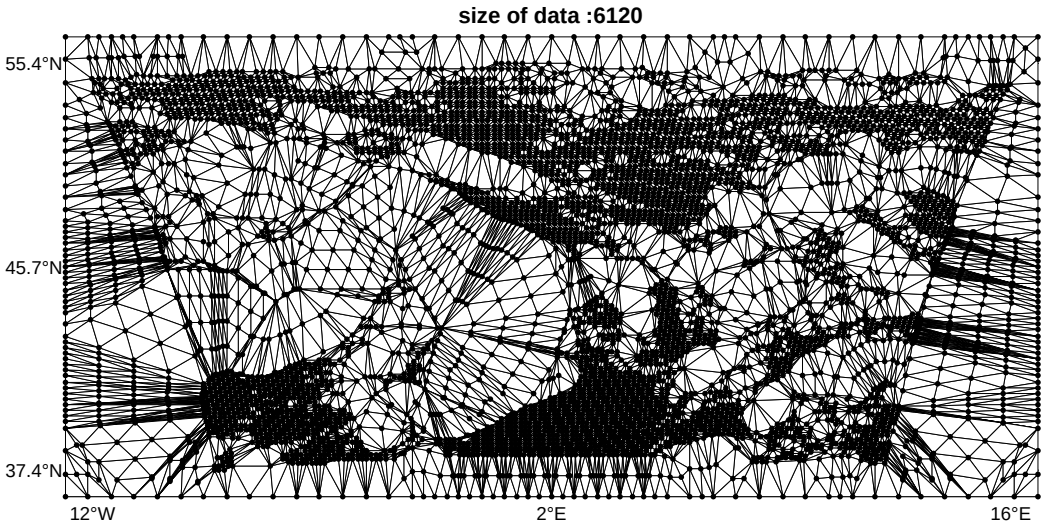


Figure 5.19 – h-raffinement hiérarchique par bisection itérée (2 itérations). Le maillage fin comprend 6120 noeuds, soit 26% de plus que le maillage initial.

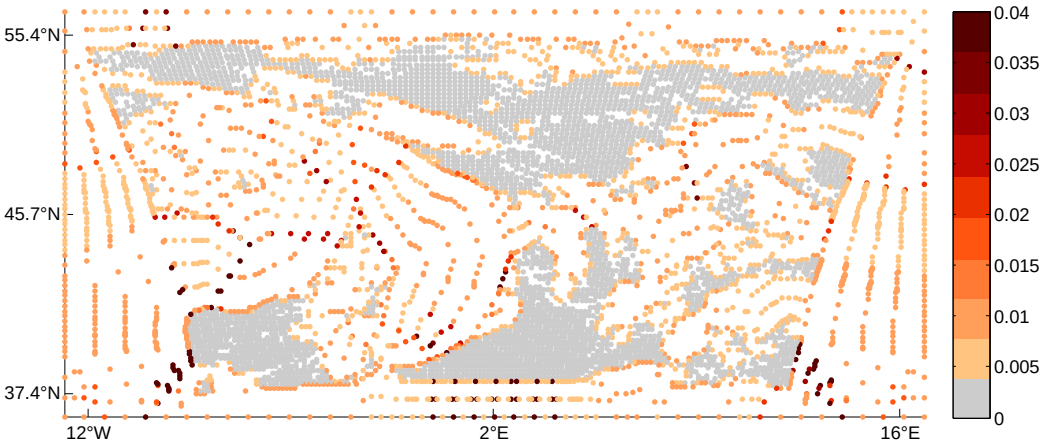


Figure 5.20 – Carte de l'indice de qualité après bisection itérée. La procédure contient des instabilités qui font apparaître des éléments de mauvaise qualité.

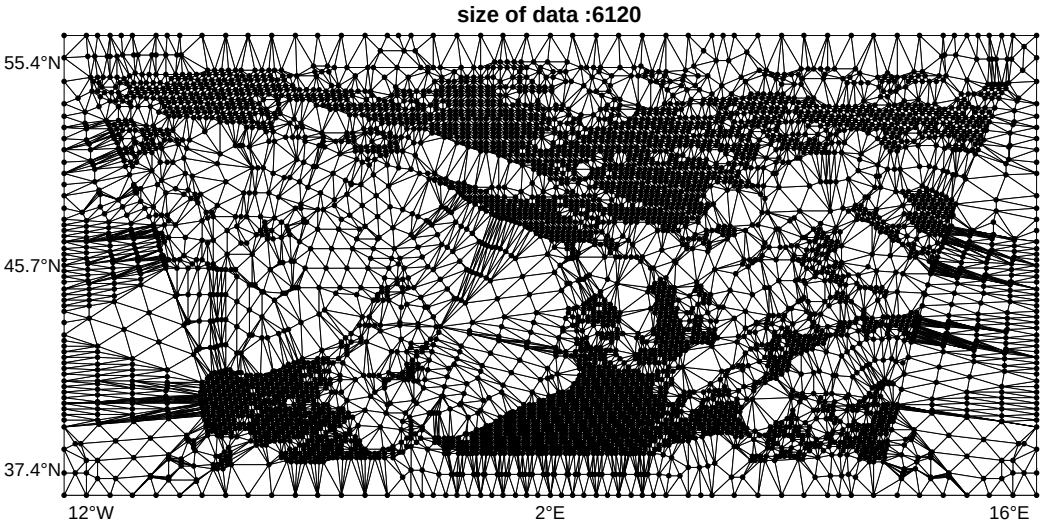


Figure 5.21 – Basculement après bisection itérée. Le nombre et l'emplacement des noeuds ne change pas.

5.2 Opérateurs de transfert

Le maillage construit à partir des observations satellites n'est pas totalement adapté à la résolution de l'équation de diffusion (résultats de la section 4.4). Dans la section 5.1, on a montré comment construire un maillage auxiliaire sur lequel les calculs peuvent être menés tout en réduisant l'apparition d'erreurs numériques. Avant de dériver une méthode à deux niveaux pour la représentation des opérateurs de corrélation, il convient de montrer com-

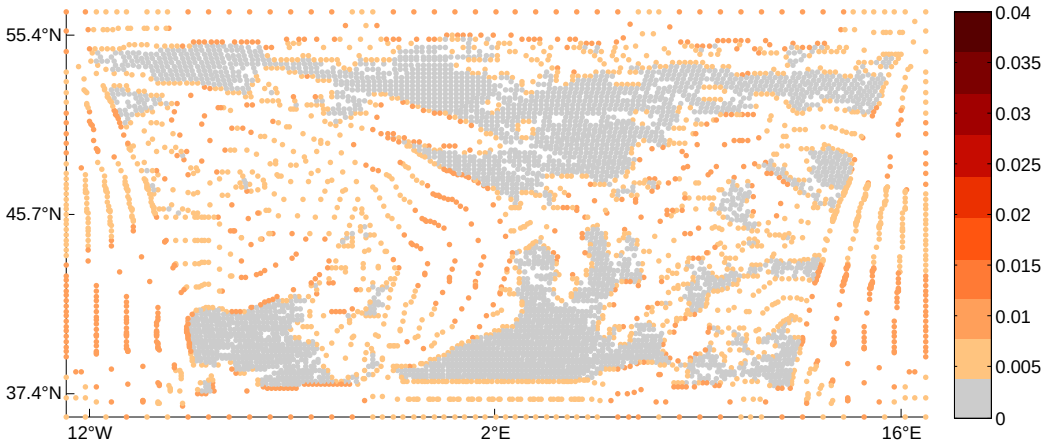


Figure 5.22 – Réparation du maillage bisecté par basculement d'arête. L'indice de qualité a des valeurs comparables au h-raffinement non hiérarchique.

ment l'information utilisée en assimilation de données peut être transférée d'un maillage à l'autre.

En effet, le modèle de corrélation présenté dans le chapitre 2 s'appuie sur une représentation continue, linéaire par morceaux, du champ d'observation. La qualité de la méthode des éléments finis dépend de la correspondance entre cette représentation continue et le maillage construit à partir des observations. Or, changer le maillage revient à changer la représentation des champs. Il faut donc exprimer un ou plusieurs critères objectifs permettant d'adapter la méthode des éléments finis au changement de maillage, et au changement d'espaces fonctionnels associé.

Bien que cette approche soit très similaire au paradigme du multigrille, sa présentation à partir du cadre fonctionnel introduit dans la partie I n'est pas tout à fait classique. Notamment, cette étude accorde beaucoup d'importance aux relations d'adjonction et de pseudo-inversion des opérateurs.

5.2.1 Transfert par interpolation linéaire

Supposons qu'on dispose d'une modélisation d'un opérateur de corrélation qui s'appuie sur une base de fonctions de forme $(\varphi_i)_{i\in[\![1,p]\!]}$ de l'espace d'approximation V_c. Ces fonctions de base sont attachées au maillage des observations. Raffiner le maillage introduit une nouvelle base de fonctions $(\psi_j)_{j\in[\![1,q]\!]}$, qui génère un espace d'approximation V_f plus grand que V_c. Cette

propriété vient du fait que

$$\dim(V_f) = q > p = \dim(V_c), \tag{5.1}$$

ce qui n'entraîne pas nécessairement l'inclusion $V_c \subset V_f$. Le rôle de cette inclusion sera discuté dans la suite.

Soit $f_c = \sum_{i=1}^{p} \alpha^i \varphi_i$ une fonction de V_c. On note $\boldsymbol{\alpha} = (\alpha_i)_{i\in[\![1,p]\!]}$ le vecteur de ses coordonnées. La projection de la fonction f_c sur la base $(\psi_j)_{j\in[\![1,q]\!]}$ de V_f s'écrit $f_f = \sum_{j=1}^{q} \beta^j \psi_j$. De même, on note $\boldsymbol{\beta} = (\beta_j)_{j\in[\![1,q]\!]}$ le vecteur des coordonnées de f_f. Pour déterminer f_f à partir de f_c, on résout le problème d'optimisation

$$\min_{\boldsymbol{\beta}} \mathcal{J}(\boldsymbol{\beta}) \quad = \quad \min_{\boldsymbol{\beta}} \left\| \sum_{i=1}^{p} \alpha^i \varphi_i - \sum_{j=1}^{q} \beta^j \psi_j \right\|_{L^2}. \tag{5.2}$$

Autrement dit, on recherche la fonction de V_f la plus proche possible de f_c en norme L^2. L'utilisation de la norme L^2 est justifiée par le rôle d'espace pivot de L^2 dans le triplet de Gelfand (1.10).

Le développement de la fonction-coût s'écrit

$$\begin{aligned}
\mathcal{J}(\boldsymbol{\beta}) &= \left\| \sum_{i=1}^{p} \alpha^i \varphi_i - \sum_{j=1}^{q} \beta^j \psi_j \right\|_{L^2}^2 \\
&= \left\langle \left(\sum_{i=1}^{p} \alpha^i \varphi_i - \sum_{j=1}^{q} \beta^j \psi_j \right), \left(\sum_{i=1}^{p} \alpha^i \varphi_i - \sum_{j=1}^{q} \beta^j \psi_j \right) \right\rangle_{L^2} \\
&= \sum_{i=1}^{p}\sum_{i=1}^{p} \alpha^i \alpha^i \langle \varphi_i, \varphi_i \rangle_{L^2} - 2\sum_{i=1}^{p}\sum_{j=1}^{q} \alpha^i \beta^j \langle \varphi_i, \psi_j \rangle_{L^2} + \sum_{j=1}^{q}\sum_{j=1}^{q} \beta^j \beta^j \langle \psi_j, \psi_j \rangle_{L^2} \\
&= \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{M}_c \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{P}_c^f \boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{M}_f \boldsymbol{\beta}.
\end{aligned} \tag{5.3}$$

Dans l'expression (5.3), $\boldsymbol{M}_c$ et $\boldsymbol{M}_f$ désignent respectivement la matrice de masse sur le maillage grossier et la matrice de masse sur le maillage fin. La matrice rectangulaire $\boldsymbol{P}_c^f$ est quant à elle la matrice de passage de la base $(\psi_j)_{j\in[\![1,q]\!]}$ vers la base $(\varphi_i)_{i\in[\![1,p]\!]}$. Son terme général est

$$(\boldsymbol{P}_c^f)_{\mathrm{ij}} = \int_{\mathcal{D}} \varphi_i(\boldsymbol{z}') \psi_j(\boldsymbol{z}') \mathrm{d}\boldsymbol{z}'. \tag{5.4}$$

Le gradient de la fonction-coût par rapport à la variable $\boldsymbol{\beta}$ est

$$\nabla \mathcal{J}(\boldsymbol{\beta}) = 2\boldsymbol{M}_f \boldsymbol{\beta} - 2(\boldsymbol{P}_c^f)^{\mathrm{T}} \boldsymbol{\alpha}. \tag{5.5}$$

En posant $\nabla \mathcal{J}(\boldsymbol{\beta}) = 0$, on obtient finalement

$$\boldsymbol{\beta} = \boldsymbol{M}_f^{-1} (\boldsymbol{P}_c^f)^{\mathrm{T}} \boldsymbol{\alpha}, \tag{5.6}$$

qui est similaire à la relation de projection établie par Trottenberg et al. [2001] dans le cadre des méthodes multigrille algébriques (AMG, *Algebraic MultiGrid* en anglais).

De même, pour revenir du maillage fin au maillage grossier, on minimise la fonctionnelle par rapport au vecteur de coordonnées $\boldsymbol{\alpha}$:

$$\min_{\boldsymbol{\alpha}} \mathcal{J}(\boldsymbol{\alpha}) \quad = \quad \min_{\boldsymbol{\alpha}} \left\| \sum_{i=1}^{p} \alpha^i \varphi_i - \sum_{j=1}^{q} \beta^j \psi_j \right\|_{L^2}. \tag{5.7}$$

On trouve alors

$$\boldsymbol{\alpha} = \boldsymbol{M}_c^{-1} \boldsymbol{P}_c^f \boldsymbol{\beta}. \tag{5.8}$$

On remarque que les opérateurs de transfert $\boldsymbol{M}_f^{-1}(\boldsymbol{P}_c^f)^{\mathrm{T}}$ et $\boldsymbol{M}_c^{-1}\boldsymbol{P}_c^f$ sont pseudo-inverse l'un de l'autre. En effet, supposons qu'il existe une relation de hiérarchie entre le maillage grossier et le maillage fin, et que la base $(\varphi_i)_{i\in[\![1,p]\!]}$ soit inclue dans la base $(\psi_j)_{j\in[\![1,q]\!]}$. Quitte à réordonner les noeuds du maillage, on peut supposer que les p premiers noeuds correspondent aux observations et que les $q-p$ suivants sont des points additionnels. On a également

$$\forall j \leqslant p \quad , \quad \psi_j = \varphi_j, \tag{5.9}$$

ce qui implique

$$\begin{aligned} &\forall j \leqslant p \quad , \quad \int_{\mathcal{D}} \varphi_i(\boldsymbol{z}')\psi_j(\boldsymbol{z}')\mathrm{d}\boldsymbol{z}' = \int_{\mathcal{D}} \varphi_i(\boldsymbol{z}')\varphi_j(\boldsymbol{z}')\mathrm{d}\boldsymbol{z}' \\ &\forall j \leqslant p \quad , \quad (\boldsymbol{P}_c^f)_{\mathrm{ij}} = (\boldsymbol{M}_c)_{\mathrm{ij}}. \end{aligned} \tag{5.10}$$

Lorsque $j > p$, on note $\boldsymbol{T}_c^f$ la matrice de taille $p \times (q-p)$ de terme général

$$(\boldsymbol{T}_c^f)_{\mathrm{ij}} = (\boldsymbol{P}_c^f)_{\mathrm{ij}}. \tag{5.11}$$

Dès lors, on peut décomposer les opérateurs de transfert comme suit :

$$\boldsymbol{M}_f^{-1}(\boldsymbol{P}_c^f)^{\mathrm{T}} = \boxed{\boldsymbol{M}_f^{-1}}\,\boxed{(\boldsymbol{P}_c^f)^{\mathrm{T}}} = \left(\begin{array}{|c|c|}\hline \boldsymbol{M}_c & \boldsymbol{T}_c^f \\ \hline (\boldsymbol{T}_c^f)^{\mathrm{T}} & \cdots \\ \hline \end{array}\right)^{-1} \begin{array}{|c|}\hline \boldsymbol{M}_c \\ \hline (\boldsymbol{T}_c^f)^{\mathrm{T}} \\ \hline \end{array} = \begin{array}{|c|}\hline \boldsymbol{I}_c \\ \hline \boldsymbol{0} \\ \hline \end{array} \tag{5.12}$$

et

$$\boldsymbol{M}_c^{-1}\boldsymbol{P}_c^f = \boxed{\boldsymbol{M}_c^{-1}}\,\boxed{\boldsymbol{P}_c^f} = \boxed{\boldsymbol{M}_c^{-1}}\,\boxed{\boldsymbol{M}_c}\,\boxed{\boldsymbol{T}_c^f} = \boxed{\boldsymbol{I}_c}\,\boxed{\boldsymbol{M}_c^{-1}\boldsymbol{T}_c^f}, \tag{5.13}$$

où $\boldsymbol{I}_c$ désigne la matrice identité sur le maillage grossier, de taille $p \times p$. La relation de pseudo-inverse se déduit immédiatement de (5.12) et (5.13) :

$$\boldsymbol{M}_c^{-1}\boldsymbol{P}_c^f \times \boldsymbol{M}_f^{-1}(\boldsymbol{P}_c^f)^{\mathrm{T}} = \boxed{\boldsymbol{I}_c}\,\boxed{\boldsymbol{M}_c^{-1}\boldsymbol{T}_c^f} \times \begin{array}{|c|}\hline \boldsymbol{I}_c \\ \hline \boldsymbol{0} \\ \hline \end{array} = \boldsymbol{I}_c. \tag{5.14}$$

Qu'advient-il lorsque $V_c \subset V_f$, mais qu'on dispose de deux bases $(\varphi_i)_{i\in[\![1,p]\!]}$ et $(\chi_j)_{j\in[\![1,q]\!]}$ qui ne sont pas inclues l'une dans l'autre ? On peut toujours trouver une base $(\psi_j)_{j\in[\![1,q]\!]}$ telle que (5.9) est vérifiée.

Notons respectivement $\boldsymbol{M}_\chi$, $\boldsymbol{M}_\psi$, $\boldsymbol{P}_\chi$ et $\boldsymbol{P}_\psi$ les matrices de masse et les matrices de passage associées aux bases $(\chi_j)_{j\in[\![1,q]\!]}$ et $(\psi_j)_{j\in[\![1,q]\!]}$. Soit $\boldsymbol{T}$ la matrice transformant $(\psi_j)_{j\in[\![1,q]\!]}$ en $(\chi_j)_{j\in[\![1,q]\!]}$. On a

$$\boldsymbol{M}_\chi = \boldsymbol{T}\boldsymbol{M}_\psi\boldsymbol{T}^{\mathrm{T}} \quad \text{et} \quad \boldsymbol{P}_\chi = \boldsymbol{P}_\psi\boldsymbol{T}^{\mathrm{T}}, \tag{5.15}$$

puis en notant $\boldsymbol{T}^{-\mathrm{T}}$ l'inverse de la transposée de $\boldsymbol{T}$,

$$\begin{aligned} \boldsymbol{M}_c^{-1}\boldsymbol{P}_\chi \times \boldsymbol{M}_\chi^{-1}\boldsymbol{P}_\chi^{\mathrm{T}} &= \boldsymbol{M}_c^{-1}(\boldsymbol{P}_\psi\boldsymbol{T}^{\mathrm{T}})(\boldsymbol{T}\boldsymbol{M}_\psi\boldsymbol{T}^{\mathrm{T}})^{-1}(\boldsymbol{P}_\psi\boldsymbol{T}^{\mathrm{T}})^{\mathrm{T}} \\ &= \boldsymbol{M}_c^{-1}\boldsymbol{P}_\psi\boldsymbol{T}^{\mathrm{T}}\boldsymbol{T}^{-\mathrm{T}}\boldsymbol{M}_\psi^{-1}\boldsymbol{T}^{-1}\boldsymbol{T}\boldsymbol{P}_\psi^{\mathrm{T}} \\ &= \boldsymbol{M}_c^{-1}\boldsymbol{P}_\psi\boldsymbol{M}_\psi^{-1}\boldsymbol{P}_\psi^{\mathrm{T}} \\ &= \boldsymbol{I}_c, \end{aligned} \tag{5.16}$$

d'après (5.14). La relation de pseudo-inversion est donc toujours vérifiée, dès lors que $V_c \subset V_f$. La relation de hiérarchie entre les bases d'éléments finis permet ainsi de construire des opérateurs de transfert qui permettent d'aller et venir entre les espace sans perdre d'information dans les hautes fréquences. En effet, en dépit de la relation (5.14), il n'est pas possible de reproduire sur le maillage grossier tous les signaux représentés sur le maillage fin.

Dans la suite, quand on parle de hiérarchie entre les maillages, on sous-entend également que les bases d'éléments finis $(\varphi_i)_{i\in[\![1,p]\!]}$ et $(\psi_j)_{j\in[\![1,q]\!]}$ sont choisies de telle sorte qu'on ait directement (5.14) et qu'il n'est pas besoin

d'introduire les transformations (5.15). A noter que dans la pratique, le choix des bases a de l'importance, puisqu'il affecte le conditionnement des matrices de masse de raideur. En particulier, soit h un paramètre caractérisant la finesse du maillage. En éléments finis de type $\mathbb{P}_1$, la matrice de masse a un conditionnement de l'ordre de $\mathcal{O}(h^{-2})$. Si on utilise des bases de Schauder, ce conditionnement devient de l'ordre de $\mathcal{O}(\log h^{-2})$ [Christon and Roach, 2000].

Pour finir, détaillons le calcul des opérateurs de transfert. Dans la sous-section 2.3.2, on détaille l'assemblage de la matrice de masse. Le même procédé est utilisé pour assembler les matrices $\boldsymbol{M}_c$ et $\boldsymbol{M}_f$. La difficulté réside dans le calcul de la matrice rectangulaire $\boldsymbol{P}_c^f$. L'intégrale (5.4) doit être calculée pour tout $i \in [\![1, p]\!]$ et pour tout $j \in [\![1, q]\!]$. Pour ce faire, on constate que l'intégrale est nulle en dehors de l'intersection des supports de φ_i et ψ_j. On a donc

$$\int_{\mathcal{D}} \varphi_i(\boldsymbol{z}')\psi_j(\boldsymbol{z}')\mathrm{d}\boldsymbol{z}' = \int_{\mathrm{supp}(\varphi_i)\bigcap\mathrm{supp}(\psi_j)} \varphi_i(\boldsymbol{z}')\psi_j(\boldsymbol{z}')\mathrm{d}\boldsymbol{z}'. \tag{5.17}$$

Or, cette intersection peut se révéler compliquée lorsque le maillage fin n'est pas obtenu par un raffinement hiérarchique. En effet, le domaine d'intégration $\mathrm{supp}(\varphi_i)\bigcap\mathrm{supp}(\psi_j)$ est dans ce cas un polygone général. En revanche, lorsque le raffinement est hiérarchique, $\mathrm{supp}(\psi_j) \subset \mathrm{supp}(\varphi_i)$ ou bien $\mathrm{supp}(\varphi_i)\bigcap\mathrm{supp}(\psi_j) = \varnothing$. Il suffit donc d'intégrer sur le maillage fin.

Dans tous les cas, on souhaite éviter le calcul de (5.17), d'autant plus qu'en l'absence de condensation de masse, il faut ensuite résoudre le système linéaire (5.6). Heureusement, l'opérateur de transfert peut être traité comme un seul bloc. Posons

$$\boldsymbol{F} = \boldsymbol{M}_f^{-1}(\boldsymbol{P}_c^f)^{\mathrm{T}}. \tag{5.18}$$

L'opérateur $\boldsymbol{F}$ est une interpolation linéaire, ce qui se devine à partir de (5.12). On remplace donc en pratique le calcul des coefficients de $\boldsymbol{F}$ par une interpolation linéaire de type *inverse distance weighting* (les poids sont proportionnels à l'inverse de la distance entre les points).

Le transfert du maillage fin vers le grossier se code ensuite comme l'adjoint du transfert du maillage grossier vers le maillage fin. On a

$$\boldsymbol{F}^\star = \boldsymbol{M}_c^{-1}\boldsymbol{P}_c^f = \boldsymbol{M}_c^{-1}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{M}_f), \tag{5.19}$$

car $\boldsymbol{M}_f$ est symétrique.

Lorsque le raffinement n'est pas hiérarchique, l'utilisation de l'interpolation linéaire n'est pas recommandée. La première solution consiste à imposer (5.14) en remplaçant la matrice de masse par

$$\boldsymbol{M}_c = \boldsymbol{F}^{\mathrm{T}} \boldsymbol{M}_f \boldsymbol{F}, \tag{5.20}$$

ce qu'on appelle l'« interpolation de la matrice de masse ». La deuxième solution consiste à utiliser le transfert par injection défini en sous-section 5.2.2.

Remarque : dans le cadre des méthodes multigrille, il est de coûtume de rajouter un coefficient multiplicateur $\sigma \geqslant 1$ devant $\boldsymbol{F}^\star$ afin de compenser d'éventuelles réductions d'amplitude dûes à l'effet de lissage des opérateurs de transfert. Ainsi, on utilise le couple $(\boldsymbol{F}, \sigma\boldsymbol{F}^\star)$ en place de $(\boldsymbol{F}, \boldsymbol{F}^\star)$. Quand le maillage est cartésien et que la maille grossière est exactement deux fois plus grande que la maille fine, Hackbusch [1985] montre que le choix $\sigma = 1$ est optimal. Lorsque le raffinement n'est pas homogène en espace, la détermination d'une valeur de σ pertinente devient compliquée.

5.2.2 Transfert par injection

On a vu dans la sous-section 5.2.1 que le signal sur le maillage fin pouvait être déduit du signal sur le maillage grossier en résolvant le problème de minimisation (5.2). On a montré que cette définition des opérateurs de transfert équivalait à effectuer une interpolation linéaire entre les deux maillages. De plus, si le raffinement est hiérarchique, on s'assure de la relation de pseudo-inverse (5.14).

On présente ici une alternative à la définition des opérateurs de transfert. Au lieu de résoudre (5.2), on choisit de définir le transfert du maillage fin vers le maillage grossier par une égalité point-à-point. Avec les notations de la sous-section 5.2.1, cherchons à vérifier l'égalité $f_c = f_f$ en tout point du maillage grossier. Cette relation s'écrit

$$\begin{aligned} &\forall k \in [\![1, p]\!] \quad , \quad \left(\sum_{i=1}^{p} \alpha^i \varphi_i \right) (\boldsymbol{z}_k) = \left(\sum_{j=1}^{q} \beta^j \psi_j \right) (\boldsymbol{z}_k) \\ &\forall k \in [\![1, p]\!] \quad , \quad \alpha^k = \beta^k. \end{aligned} \tag{5.21}$$

Pour assurer l'égalité en tout point du maillage grossier, il suffit donc de sélectionner les éléments de $\boldsymbol{\beta}$ correspondant aux noeuds des observations.

L'opérateur de transfert correspondant, qu'on note $\boldsymbol{F}^\star$ pour garder les notations de la sous-section 5.2.1, est égal à

$$\boldsymbol{F}^\star = \left[\begin{array}{c|c} \boldsymbol{I}_c & \boldsymbol{0} \end{array}\right]. \tag{5.22}$$

Son adjoint est défini comme

$$\boldsymbol{F} = \boldsymbol{M}_f^{-1}(\boldsymbol{F}^\star)^{\mathrm{T}}\boldsymbol{M}_c. \tag{5.23}$$

Ce choix d'opérateurs de transfert est naturel quand on considère le cadre matriciel de l'assimilation de données. En effet, si une matrice de corrélation est construite à partir du maillage fin, il suffit de sélectionner les lignes et les colonnes correspondant aux observations pour construire la matrice de corrélation sur le maillage grossier. Cette sélection équivaut à l'utilisation des opérateurs (5.22) et (5.23). De plus, puisque l'égalité (5.21) est définie point-à-point, il est tout à fait possible d'obtenir le maillage fin par raffinement non hiérarchique sans risquer de compromettre la précision de l'interpolation ou le temps en calcul. Il est donc avantageux d'utiliser ce critère de transfert conjointement au raffinement non hiérarchique, qui a l'avantage de produire d'excellents maillages.

Attention cependant : en élements finis $\mathbb{P}_1$, on perd *a priori* la relation de pseudo-inverse entre $\boldsymbol{F}$ et $\boldsymbol{F}^\star$. On peut toutefois retrouver la première relation de Moore-Penrose en remplaçant $\boldsymbol{M}_c^{-1}$ par

$$\boldsymbol{M}_c^{-1} = \boldsymbol{F}^\star\boldsymbol{M}_f^{-1}(\boldsymbol{F}^\star)^{\mathrm{T}}, \tag{5.24}$$

mais cette formule n'a malheureusement pas de sens physique lorsque $\boldsymbol{F}^\star$ désigne le transfert par injection.

5.3 Opérateurs de corrélations augmentés

Après le raffinement de maillage et les opérateurs de transfert, il est temps de définir les opérateurs de corrélation multi-niveaux en utilisant les définitions précédentes. L'objectif de cette partie de l'étude est de montrer qu'on peut modéliser des opérateurs de corrélation sur des maillages non structurés, sans que la précision et la stabilité de cette modélisation dépendent des données d'entrée et du maillage qui les supporte. Il s'agit donc d'une recherche de robustesse et d'adaptativité.

Les résultats clefs du chapitre sont contenus dans cette section courte. Pourtant, pour les établir, il est nécessaire d'exploiter les développements des chapitres précédents. Cela montre comment l'analyse menée dans les parties I et II peut être mise à profit pour développer de nouvelles méthodes.

5.3.1 Etablissement de la formule générale

On suppose qu'on est en présence d'un maillage grossier construit à partir des observations et d'un maillage fin obtenu par raffinement du premier (section 5.1). Soit $\boldsymbol{C}_f$ un opérateur de corrélation défini sur le maillage fin, d'expression

$$\boldsymbol{C}_f = \boldsymbol{D}_f \boldsymbol{M}_f^{-1}, \tag{5.25}$$

où $\boldsymbol{D}_f$ désigne l'opérateur de diffusion (section 2.2)

$$\boldsymbol{D}_f = [(\boldsymbol{M}_f + \boldsymbol{K}_f)^{-1} \boldsymbol{M}_f]^m. \tag{5.26}$$

L'expression de l'opérateur de corrélation $\boldsymbol{C}_c$ défini sur le maillage grossier est donnée par la relation

$$\boldsymbol{C}_c = \boldsymbol{F}^\star \boldsymbol{D}_f \boldsymbol{F} \boldsymbol{M}_c^{-1}, \tag{5.27}$$

où $\boldsymbol{F}$ et $\boldsymbol{F}^\star$ désignent respectivement l'opérateur de transfert du maillage grossier vers le maillage fin et inversement (section 5.2). On peut vérifier que la formule (5.27) est symétrique, qu'on utilise l'interpolation linéaire ou bien le transfert par injection. On vérifie numériquement qu'elle est également définie positive, et qu'elle vérifie le test de l'adjoint (sous-section 4.4.2).

Pour visualiser la formule (5.27), il est utile de se référer au diagramme de dualité de la figure 5.23. L'application de l'opérateur de corrélation raffiné (ou « augmenté », car son expression contient plus de termes) correspond ainsi à un chemin dans le diagramme de dualité. Dans la suite, l'exploitation d'autres chemins au travers du diagramme permettra de construire d'autres opérateurs de corrélation, des approximations et leurs inverses. On invite donc le lecteur à se familiariser avec cet outil.

La formulation (5.27) ne va pas sans rappeler celle de Michel [2018]. Dans son approche, les observations sont interpolées sur un maillage cartésien dans un premier temps. Ensuite, un opérateur de corrélation est appliqué sur ce maillage cartésien. Le champ obtenu est enfin ré-interpolé dans l'espace des observations. Toutefois, cette approche se heurte à certains écueils théoriques et pratiques. En effet, le maillage « fin » de son approche n'est pas nécessairement capable de représenter les mêmes échelles que le maillage des observations. Pour que ce soit le cas, il faut qu'il soit suffisamment raffiné pour que sa maille soit plus fine que la plus petite distance entre deux observations. Lorsque cette condition n'est pas vérifiée, le rang plein de l'opérateur de corrélation global n'est pas garanti théoriquement. Enfin, l'opérateur total est compliqué à inverser. Pour ce faire, Michel [2018] utilise un algorithme de

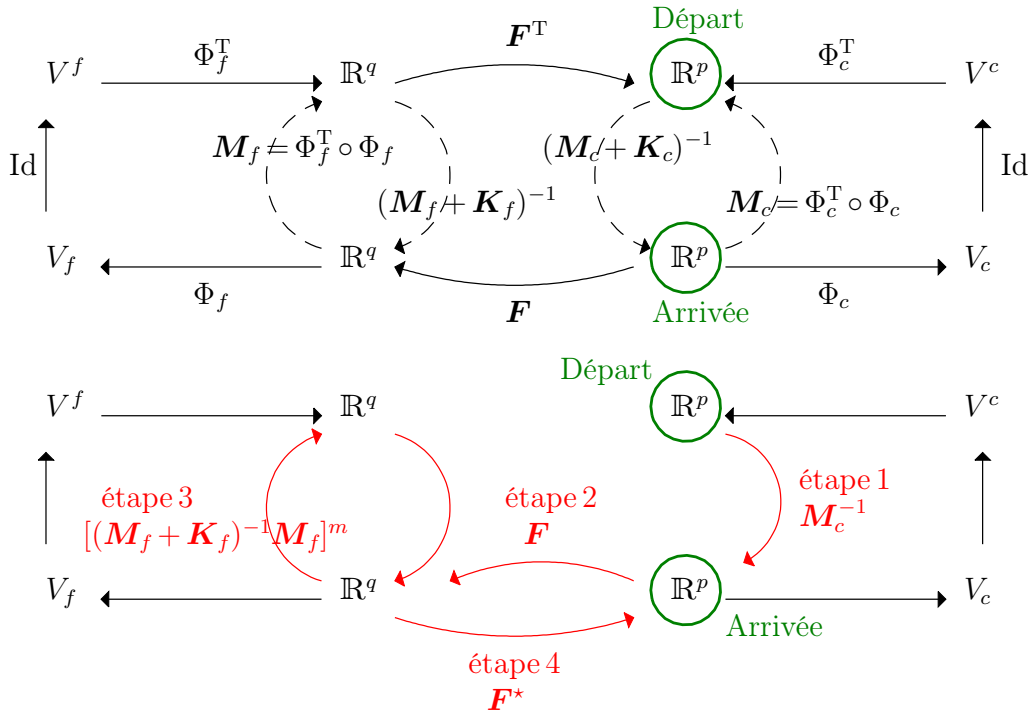


Figure 5.23 – Diagrammes de dualité. Le premier diagramme représente les applications, leurs espaces de départ et d'arrivée. Le deuxième donne le sens de lecture (en rouge) auquel correspond l'opérateur de corrélation à deux niveaux (5.27). Lire le premier diagramme : chaque application est associée à une flèche pointant de l'espace de départ vers l'espace d'arrivée. Les espaces primaux se trouvent en bas du diagramme, tandis que leurs espaces duaux respectifs se trouvent en haut. Les espaces de dimension p correspondant au maillage grossier se trouvent à droite, tandis que les espaces de dimension q correspondant au maillage fin se trouvent à gauche. Au milieu, on retrouve les opérateurs de transfert $\boldsymbol{F}$ et $\boldsymbol{F}^{\mathrm{T}}$. Les flèches tiretées correspondent aux applications inversibles (et dont on utilise l'inverse dans l'étude).

Lanczos pour en obtenir une représentation de rang réduit. Cette approximation est ajoutée à un terme de régularisation et le tout est inversé de manière explicite. Cependant, le nombre de paires de Ritz requises pour maîtriser l'erreur d'approximation atteint plusieurs centaines, ce qui rend le stockage de cette information difficile, voire infaisable en pratique.

Les techniques d'inversion de (5.27) font l'objet du chapitre 6.

5.3.2 Application aux données SEVIRI

Comme dans la section 4.4, on souhaite valider la formule (5.27) à partir des données de SEVIRI. On envisage deux cas-tests. Les paramètres de corrélation sont les mêmes qu'en section 4.4.

Cas-test 1 : interpolation linéaire

Dans un premier temps, le maillage fin est obtenu par un raffinement hiérarchique. L'opérateur de transfert $\boldsymbol{F}$ est construit par interpolation linéaire. Développons la formule (5.27) :

$$\begin{aligned} \boldsymbol{C}_c &= \boldsymbol{F}^\star \boldsymbol{D}_f \boldsymbol{F} \boldsymbol{M}_c^{-1} \\ &= \boldsymbol{M}_c^{-1} \boldsymbol{F}^{\mathrm{T}} \boldsymbol{M}_f [(\boldsymbol{M}_f + \boldsymbol{K}_f)^{-1} \boldsymbol{M}_f]^m \boldsymbol{F} \boldsymbol{M}_c^{-1}. \end{aligned} \tag{5.28}$$

Comme on peut le voir, l'inverse de la matrice de masse $\boldsymbol{M}_c$ intervient à l'extérieur des opérateurs de transfert. Cette matrice est construite à partir du maillage des observations. C'est donc la seule vraie source d'erreur possible dans cette formule. Pour la même raison, l'utilisation de la condensation de masse est réservée à la modélisation de $\boldsymbol{M}_f$. Toutefois, on n'en fait pas usage dans cette expérience.

La carte d'erreur d'amplitude est donnée dans la figure (5.24). On constate qu'il n'y a aucune amélioration par rapport au cas sans raffinement. Cela vient de la mauvaise représentation de la matrice $\boldsymbol{M}_c$. Toutefois, l'interpolation linéaire ne doit pas être rejetée. Le chapitre 7 montre comment l'utiliser pour définir $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$ de manière alternative. Cette alternative donne lieu à un nouveau schéma de construction de $\boldsymbol{C}_c$ qui présente de grands avantages sur la formulation classique ou sans raffinement.

Cas-test 2 : transfert par injection

Cette fois-ci, on construit le maillage fin par raffinement non hiérarchique. L'opérateur de transfert $\boldsymbol{F}^\star$ est quant à lui défini comme une injection. La formule (5.27) se réécrit

$$\begin{aligned} \boldsymbol{C}_c &= \boldsymbol{F}^\star \boldsymbol{D}_f \boldsymbol{F} \boldsymbol{M}_c^{-1} \\ &= \boldsymbol{F}^\star [(\boldsymbol{M}_f + \boldsymbol{K}_f)^{-1} \boldsymbol{M}_f]^m \boldsymbol{M}_f^{-1} (\boldsymbol{F}^\star)^{\mathrm{T}} \boldsymbol{M}_c \boldsymbol{M}_c^{-1} \\ &= \boldsymbol{F}^\star [(\boldsymbol{M}_f + \boldsymbol{K}_f)^{-1} \boldsymbol{M}_f]^m \boldsymbol{M}_f^{-1} (\boldsymbol{F}^\star)^{\mathrm{T}} \end{aligned} \tag{5.29}$$

La formule (5.29) a la particularité de ne pas dépendre de la matrice de masse $\boldsymbol{M}_c$. En effet, le dernier facteur $\boldsymbol{M}_c^{-1}$ se simplifie quand il est apposé à l'opérateur $\boldsymbol{M}_f^{-1}(\boldsymbol{F}^\star)^{\mathrm{T}} \boldsymbol{M}_c$. Si on définit l'opérateur de corrélation sur le maillage fin par l'expression

$$\boldsymbol{C}_f = [(\boldsymbol{M}_f + \boldsymbol{K}_f)^{-1} \boldsymbol{M}_f]^m \boldsymbol{M}_f^{-1}, \tag{5.30}$$

et en vertu de la relation (5.22), alors on peut conclure que $\boldsymbol{C}_c$ est obtenu par sélection des lignes et des colonnes dans $\boldsymbol{C}_f$.

Comme on peut le voir sur la figure 5.25, la formule (5.29) permet de corriger la quasi-totalité des erreurs d'amplitude, en particulier les erreurs dues à la présence des bords intérieurs au domaine. C'est en outre lié à l'indépendance de (5.29) en $\boldsymbol{M}_c$. De plus, la propriété de sélection des lignes et des colonnes indique qu'un raffinement plus intense mène naturellement à une approximation de meilleure qualité.

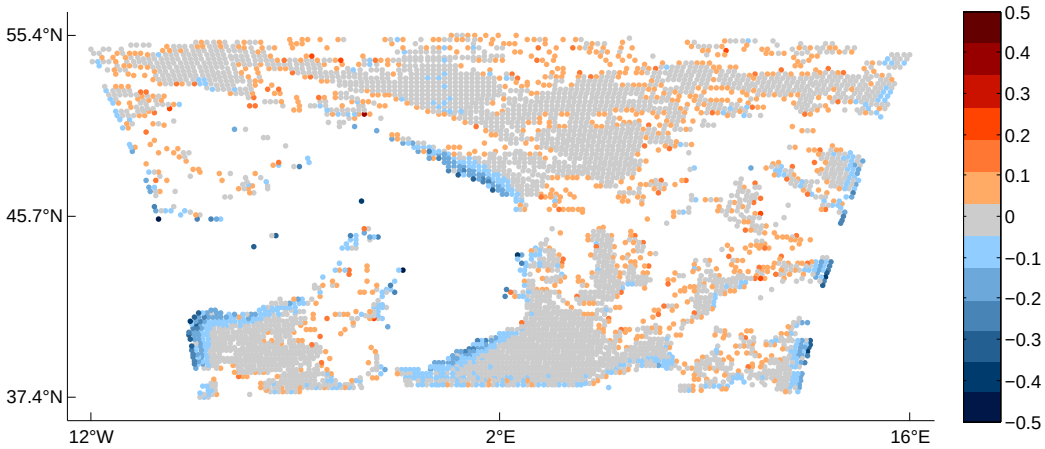


FIGURE 5.24 – Erreur d'amplitude associée à l'opérateur (5.28). Dans ce cas, il n'y a pas d'amélioration par rapport à l'erreur initiale.

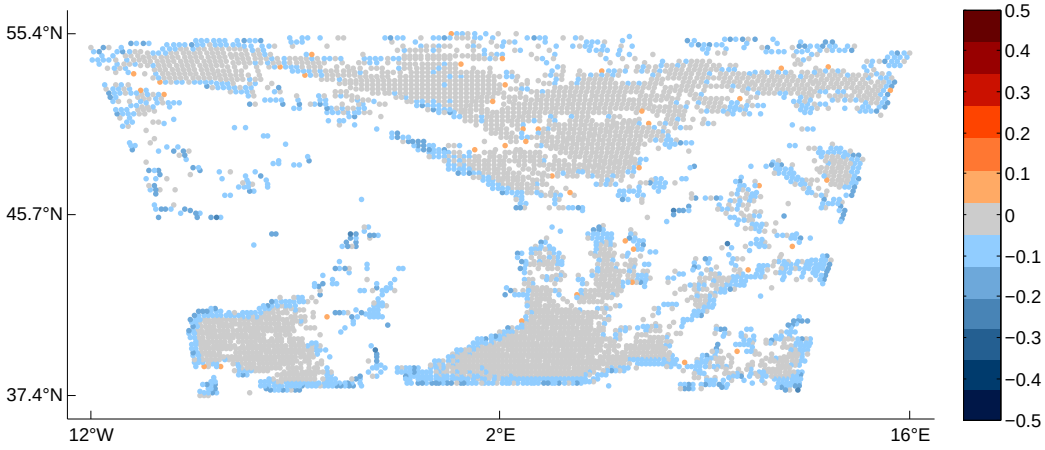


FIGURE 5.25 – Erreur d'amplitude associée à l'opérateur (5.29). La réduction de l'erreur est nette en dehors de certaines frontières. C'est du à la difficulté de normaliser l'opérateur de corrélation près des bords (section 3.3).

5.4 Interpolation de l'opérateur de corrélation inverse

Pour l'instant, le raffinement de maillage a été utilisé seulement pour construire un meilleur opérateur de corrélation. Dans cette section, on applique les mêmes principes que dans la section 5.3 à la représentation de l'opérateur de précision $\boldsymbol{C}_c^{-1}$. Cette approche est suggérée par le diagramme de dualité (5.23), en échangeant l'espace de départ et l'espace d'arrivée. Le nouveau chemin dans le diagramme est représenté sur la figure 5.26.

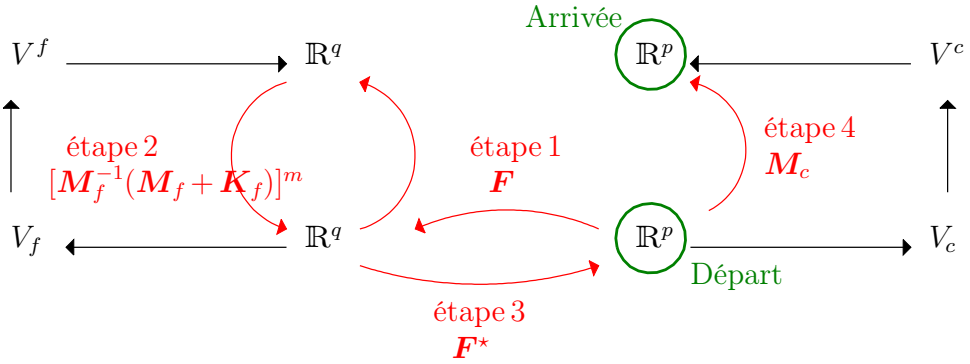


Figure 5.26 – Diagramme de dualité. Chemin associé à la définition (5.31). Les espaces de départ et d'arrivée ont été échangés puisque l'inverse de l'opérateur de corrélation va « du primal vers le dual ».

Comme on le voit, il suffit de considérer l'inverse (ou la pseudo-inverse) de chacun des termes de l'équation (5.27). En se souvenant que $\boldsymbol{F}^\star$ est la pseudo inverse de $\boldsymbol{F}$, on obtient

$$\mathbf{Q}_c = \boldsymbol{M}_c \boldsymbol{F}^\star \boldsymbol{D}_f^{-1} \boldsymbol{F}, \tag{5.31}$$

où $\boldsymbol{Q}_c$ désigne l'opérateur de précision recherché. Attention, il est utile de remarquer que

$$\boldsymbol{Q}_c \neq \boldsymbol{C}_c^{-1}, \tag{5.32}$$

car $\boldsymbol{C}_c$ est un produit de matrices non carrées, et donc n'est pas inversible terme-à-terme. La question de

$$\boldsymbol{Q}_c \simeq \boldsymbol{C}_c^{-1} \tag{5.33}$$

reste ouverte. Le préconditionnement du système linéaire $\boldsymbol{C}_c \boldsymbol{z} = \boldsymbol{b}$ est inspecté dans le chapitre 6.

La formule (5.31) n'est pas sans défaut. En effet, $\boldsymbol{D}_f^{-1}$ introduit des hautes fréquences dans le signal représenté sur le maillage fin. Ces hautes fréquences ne peuvent pas nécessairement être représentées sur le maillage grossier, en vertu du théorème de Nyquist-Shannon. L'introduction des opérateurs de transfert avant et après l'opérateur de diffusion sur le maillage fin va donc à l'encontre de l'effet recherché (*i.e.* de l'introduction de hautes fréquences).

La construction de $\boldsymbol{Q}_c$ à l'aide d'un raffinement de maillage se rapproche du travail de Lindgren et al. [2011]. Dans leur étude, ils introduisent des points dans tout le domaine, de façon à pouvoir construire une triangulation quasi-régulière sur l'intégralité du domaine. L'opérateur inverse de corrélation est ensuite modélisé par éléments finis sur le maillage fin. Toutefois, il n'est pas question de supprimer les points additionnels, ni ne chercher l'opérateur de corrélation correspondant. En effet, leur application dans les statistiques multidimensionnelles ne requiert pas les mêmes opérateurs que l'assimilation de données variationnelle.

5.5 Effets de la condensation de masse

Dans cette dernière section du chapitre, on discute des effets de la condensation de masse sur la définition des opérateurs de transfert et de corrélation. Remarque : cette section jette un regard atypique sur la condensation de masse et les opérateurs de transfert en étudiant leur lien avec l'interpolation par voisins naturels. Puis, la présentation se recentre sur l'étude des corrélations en montrant la cartes d'erreurs associées à l'utilisation de la condensation de masse.

Dans la formule de l'interpolation linéaire (5.18), on voit que la définition de $\boldsymbol{F}$ dépend de la matrice de masse $\boldsymbol{M}_c$. Qu'advient-il lorsqu'on remplace $\boldsymbol{M}_f$ par la matrice de masse condensée $\boldsymbol{M}_f^{\#}$? En notant $\boldsymbol{F}^{\#}$ l'opérateur de transfert résultant, on a

$$\boldsymbol{F}^{\#} = (\boldsymbol{M}_f^{\#})^{-1}(\boldsymbol{P}_c^f)^{\mathrm{T}}, \tag{5.34}$$

et son adjoint s'écrit

$$(\boldsymbol{F}^{\#})^{\star} = \boldsymbol{M}_c^{-1}(\boldsymbol{F}^{\#})^{\mathrm{T}}\boldsymbol{M}_f^{\#}. \tag{5.35}$$

La matrice $\boldsymbol{M}_f^{\#}$ est diagonale et, en vertu de (3.5), ses éléments diagonaux

ont pour expression

$$\forall j \in [\![1, q]\!] \quad , \quad (\boldsymbol{M}_f^{\#})_{\mathrm{jj}} = \int_{\mathcal{D}} \psi_j(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}. \tag{5.36}$$

En éléments finis $\mathbb{P}_1$, cette intégrale est égale à l'aire de la cellule de Voronoï modifiée associée au noeud $\boldsymbol{z}_j$ dans le maillage fin (sous-section 3.2.1). D'après (5.4), le terme général de $\boldsymbol{F}^{\#}$ est donc égal au rapport

$$(\boldsymbol{F}^{\#})_{\mathrm{ij}} = \frac{\int_{\mathcal{D}} \varphi_i(\boldsymbol{z}) \psi_j(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}}{\int_{\mathcal{D}} \psi_j(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}} = \frac{N}{D}, \tag{5.37}$$

où N désigne le numérateur et D le dénominateur.

Soit $j \in [\![1, q]\!]$. Supposons qu'on dispose d'un maillage qui ne contienne pas $\boldsymbol{z}_j$ initialement. A chaque point $\boldsymbol{z}_i, i \in [\![1, q]\!]$ correspond l'aire de sa cellule de Voronoï modifiée. Supposons maintenant qu'on introduise $\boldsymbol{z}_j$ dans le maillage. En l'insérant dans le maillage, on lui attribue une cellule de Voronoï modifiée. L'aire de cette cellule est obtenue en réduisant l'aire des cellules voisines. Ainsi, dans l'expression (5.37), on peut interpréter le numérateur N comme l'aire de $\boldsymbol{z}_i$ « volée » par $\boldsymbol{z}_j$ lors de son introduction. En réécrivant

$$D = \int_{\mathcal{D}} \psi_j(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \sum_{i=1}^{p} \int_{\mathcal{D}} \varphi_i(\boldsymbol{z}) \psi_j(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}, \tag{5.38}$$

on voit que le dénominateur s'interprête comme un terme de normalisation, assurant que le rapport (5.37) se situe dans l'intervalle $[0, 1]$.

On retrouve la quantité (5.37) dans la définition de l'interpolation par voisins naturels [Sibson, 1981]. Toutefois, cette méthode se base sur le diagramme de Voronoï classique, et non sur le diagramme de Voronoï modifié. Autrement dit, pour que la formule (5.37) décrive une interpolation par voisins naturels, il faudrait que l'aire volée par $\boldsymbol{z}_j$ à $\boldsymbol{z}_i$ soit calculée à partir des cellules de Voronoï, ce qui n'est pas le cas.

Néanmoins, quand les triangles sont équilatéraux, les centres de masse des triangles et les centres de leurs cercles circonscrits sont confondus. L'utilisation de la condensation de masse dans la définition de $\boldsymbol{F}$ revient donc à remplacer cet opérateur de transfert par une interpolation par voisins naturels. Quand les triangles ne sont pas équilatéraux, la quantité (5.37) est une approximation du rapport trouvé dans l'interpolation par voisins naturels. Par ailleurs, plus les éléments sont pointus ou obtus, plus cette approximation est grossière.

Le lien évoqué ci-dessus permet de remplacer le calcul des coefficients de $\boldsymbol{F}$ par l'utilisation d'une méthode d'interpolation (et donc, de réduire le coût numérique du transfert). Pour que ce remplacement soit justifié, il est nécessaire que les éléments du maillage fin soient équilatéraux ou presque. La notion de « presque » et la qualité de l'approximation lorsque les éléments ne sont pas équilatéraux n'est pas quantifiée dans cette étude. Pour notre application, on s'en tient à l'intuition géométrique et l'expérience pour juger de sa pertinence.

Quand $\boldsymbol{F}^{\#}$ désigne l'interpolation par voisins naturels, l'opérateur de corrélation augmenté s'écrit

$$\boldsymbol{C}_c^{\#} = (\boldsymbol{F}^{\#})^{\star} \boldsymbol{D}_f^{\#} \boldsymbol{F}^{\#} \boldsymbol{M}_c^{-1}, \tag{5.39}$$

où

$$\boldsymbol{D}_f^{\#} = [(\boldsymbol{M}_f^{\#} + \boldsymbol{K}_f)^{-1} \boldsymbol{M}_f^{\#}]^m. \tag{5.40}$$

La figure 5.27 représente l'erreur d'amplitude associée à (5.39).

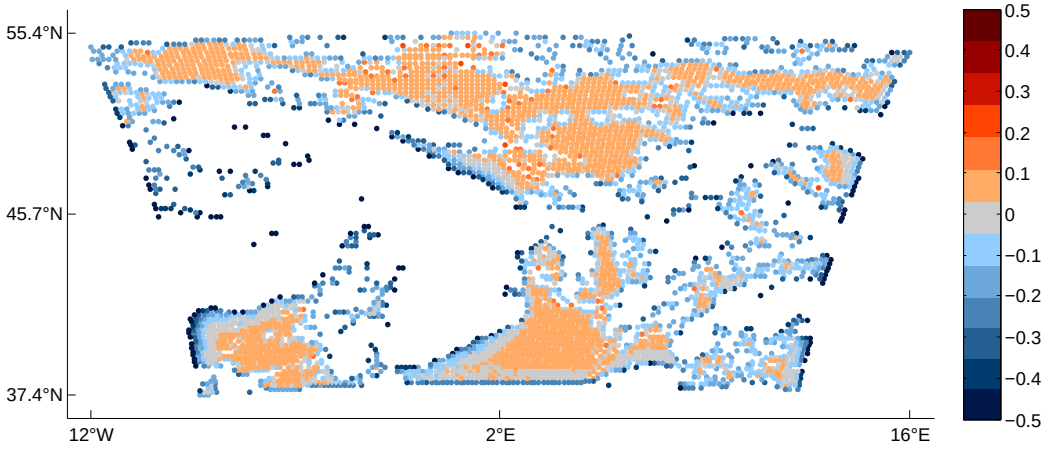


Figure 5.27 – Carte d'erreur d'amplitude. On utilise la condensation de masse pour modéliser $\boldsymbol{M}_f$ et l'interpolation de la masse pour modéliser $\boldsymbol{M}_c$. L'erreur est plus large que sans interpolation de masse.

Si l'on souhaite sélectionner les lignes et colonnes, alors, on ne change pas la définition de $\boldsymbol{F}^{\star}$ et en posant

$$\boldsymbol{F} = (\boldsymbol{M}_f^{\#})^{-1} (\boldsymbol{F}^{\star})^{\mathrm{T}} \boldsymbol{M}_c, \tag{5.41}$$

on obtient

$$\boldsymbol{C}_c^{\#} = \boldsymbol{F}^{\star}[(\boldsymbol{M}_f^{\#} + \boldsymbol{K}_f)^{-1}\boldsymbol{M}_f^{\#}]^m(\boldsymbol{M}_f^{\#})^{-1}(\boldsymbol{F}^{\star})^{\mathrm{T}}. \tag{5.42}$$

La figure 5.28 représente l'erreur d'amplitude associée à (5.42).

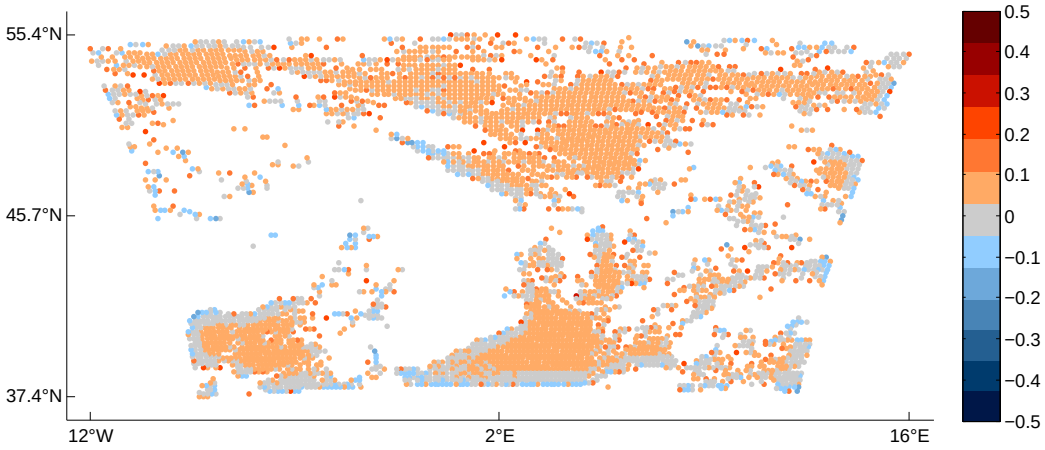


Figure 5.28 – Carte d'erreur d'amplitude. On utilise la condensation de masse pour modéliser $\boldsymbol{M}_f$. L'erreur est très légèrement supérieure que sans condensation de masse.

5.6 L'essentiel du chapitre

Deux types de raffinement en espace sont mis en exergue dans ce chapitre : le raffinement hiérarchique et le raffinement non hiérarchique. Allant de paire avec ces deux stratégies, deux modèles d'opérateurs de transfert et de corrélation voient le jour. L'approche consistant à sélectionner les lignes et colonnes d'une matrice de corrélation construite à partir du maillage fin semble la plus prometteuse. Elle permet de modéliser des fonctions de corrélation d'amplitude unitaire sur tout le domaine, faisant par la même fi de la nécessité d'employer des méthodes de renormalisation.

La possibilité d'interpoler l'opérateur de précision plutôt que l'opérateur de corrélation conduit à une formule difficile à interpréter par son manque de sens statistique. Toutefois, ce manquement sera comblé dans le chapitre 6.

Enfin, raffiner le maillage n'est pas la seule stratégie viable. On mentionne également la possibilité de déraffinement, qui s'intègre au processus de *thinning* utilisé dans le prétraitement des données satellitaires. La problématique de définir un algorithme de *thinning* intelligent, prenant en compte des critères de génération de maillage est mentionnée. Toutefois, seules des pistes de recherche sont données, car le développement et la mise à l'épreuve de ces techniques constituent une étude à part entière. Le cadre fonctionnel présenté dans ce manuscrit offre un point de départ confortable à un éventuel travail ultérieur.

Ce qu'il faut retenir :

- Ils existent plusieurs stratégies de raffinement de maillage.
- Dans ce manuscrit, on exploite essentiellement le raffinement en espace, pour sa versatilité et sa facilité d'implémentation.
- On reprend le cadre théorique des parties précédentes pour dériver un nouveau modèle de corrélation à deux niveaux

  $$\boldsymbol{C}_c = \boldsymbol{F}^\star \boldsymbol{D}_f \boldsymbol{F} \boldsymbol{M}_c^{-1}$$

  Ce modèle peut se retrouver grâce à un diagramme de dualité.
- Dans tous les cas, gagner en précision se fait au prix de rendre l'inversion de $\boldsymbol{C}_c$ difficile.
- La condensation de masse ne doit être utilisée que sur le maillage fin, qui est de bonne qualité.

Chapitre 6

Inversion dans l'espace des observations

La présente étude a pour objectif d'élaborer et d'évaluer différentes modélisations des corrélations spatiales d'erreurs d'observation en assimilation de données. La possibilité de modéliser l'inverse de l'opérateur de corrélation est un enjeu essentiel dans la réalisation de cet objectif. Dans la partie II, la méthode des élements finis est utilisée pour modéliser la matrice $\boldsymbol{C}_c$ et son inverse $\boldsymbol{C}_c^{-1}$. Toutefois, la comparaison entre le modèle théorique de Matérn et les fonctions de corrélations contenues dans $\boldsymbol{C}_c$ font ressortir la présence d'erreurs numériques dépendantes des données.

Le chapitre 5 de la partie III présente des méthodes de raffinement de maillage qui permettent d'améliorer la précision et la robustesse de la modélisation de $\boldsymbol{C}_c$. Cependant, cette modélisation ne permet plus d'exprimer $\boldsymbol{C}_c^{-1}$ aussi simplement que dans la partie II. En effet, l'opérateur de corrélation contenu dans $\boldsymbol{C}_c$ est maintenant un produit de matrices non carrées. Pour l'inverser, il convient d'être en mesure de résoudre le système linaire d'inconnue $\boldsymbol{x} \in \mathbb{R}^p$

$$\boldsymbol{C}_c \boldsymbol{x} = \boldsymbol{b}, \tag{6.1}$$

pour tout vecteur $\boldsymbol{b} \in \mathbb{R}^p$.

En raison du nombre d'observations dans nos applications, il est hors de question de calculer les coefficients de la matrice $\boldsymbol{C}_c$. Dans ce chapitre, on montre comment résoudre le système linéaire (6.1) à l'aide des méthodes de Krylov, qui font partie des méthodes de résolution dites « itératives ». En effet, ces dernières ne nécessitent pas de connaître explicitement les coefficients de l'opérateur à inverser. Elles exploitent à la place l'expression de cet

opérateur sous la forme de produits matrices-vecteurs, procédé courant en assimilation de données. Par contraste, les méthodes de résolution « directes », qui nécessitent l'accès aux coefficients, sont plutôt réservées à l'inversion des matrices creuses. Dans notre cas, cette approche concerne essentiellement l'inversion des matrices $\boldsymbol{M}_c + \boldsymbol{K}_c$ et $\boldsymbol{M}_f + \boldsymbol{K}_f$.

La description des méthodes de Krylov fait l'objet de la section 6.1. Dans les sections 6.2 et 6.3, la résolution pratique de (6.1) est abordée et les performances des différentes méthodes sont évaluées. Enfin, la section 6.4 discute de l'utilisation des préconditionneurs de deuxième ordre, partant du principe que la convergence des méthodes de Krylov dépend fortement des faibles valeurs propres de l'opérateur de corrélation.

Les innovations de ce chapitre concernent l'application des méthodes de Krylov à la résolution de (6.1).

6.1 Méthodes itératives

Dans cette section, on effectue un panorama des différentes méthodes itératives conçues pour résoudre le système linéaire $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Cette introduction commence avec les méthodes de point fixe, qui rejoignent naturellement les méthodes de Richardson. Une analyse de la solution du système linéaire permet de déduire les méthodes de Krylov, qu'on utilise dans la suite de l'étude. Le point de vue équivalent entre l'optimisation et l'utilisation des conditions de Petrov-Galerkin offre quant à lui un moyen de dériver certains grands algorithmes de manière intuitive.

Attention, les notations rencontrées dans cette section diffèrent du reste du manuscrit. Il n'est pas possible de faire une présentation, même incomplète, des méthodes itératives sans introduire de nouvelles notations qui interfèrent avec les anciennes. On choisit donc d'ignorer temporairement ces dernières au profit d'une présentation claire.

6.1.1 Méthodes de point fixe

A partir de maintenant, on cherche à résoudre l'équation

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{6.2}$$

où $\boldsymbol{x} \in \mathbb{R}^n$ est la solution du système, $\boldsymbol{b} \in \mathbb{R}^n$ est son second membre et $\boldsymbol{A}$ est une matrice inversible de taille $n \times n$.

Les méthodes de point fixe consistent à construire une suite d'itérés $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ qui converge vers la solution $\boldsymbol{x}$ lorsque $k \to +\infty$ au travers d'une relation de récurrence du type :

$$\begin{aligned} \boldsymbol{x}^{(0)} &= \mathcal{F}_0(\boldsymbol{A}, \boldsymbol{b}) \\ \boldsymbol{x}^{(k+1)} &= \mathcal{F}_{k+1}(\boldsymbol{A}, b, \{\boldsymbol{x}^{(k')}, k' < k+1\}). \end{aligned} \tag{6.3}$$

Lorsque les fonctions $(\mathcal{F}_k)_{k\in\mathbb{N}}$ sont linéaires et indépendantes de k (on parle de stationnarité), cette relation se réécrit

$$\begin{aligned} \boldsymbol{x}^{(0)} &= \boldsymbol{x}_0 \\ \boldsymbol{x}^{(k+1)} &= \boldsymbol{F}\boldsymbol{x}^{(k)} + \boldsymbol{f}, \end{aligned} \tag{6.4}$$

où $\boldsymbol{F}$ est une matrice de taille $n \times n$ et $\boldsymbol{f} \in \mathbb{R}^n$. La suite d'itérations (6.4) possède les propriétés suivantes [Quarteroni et al., 2008] :

- Si la solution $\boldsymbol{x}$ de (6.2) est un point fixe de (6.4), alors la suite est consistante. On a donc $\boldsymbol{x} = \boldsymbol{F}\boldsymbol{x} + \boldsymbol{f}$ et $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, ce qui entraîne $\boldsymbol{f} = (\boldsymbol{I} - \boldsymbol{F})\boldsymbol{A}^{-1}\boldsymbol{b}$.
- Si de plus, l'application $\boldsymbol{F}$ est contractante, alors la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ converge vers l'unique point fixe de (6.4), donc vers $\boldsymbol{x}$.

Supposons que ces propriétés soient vérifiées. Dès lors, l'itération k s'écrit

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= \boldsymbol{F}\boldsymbol{x}^{(k)} + (\boldsymbol{I} - \boldsymbol{F})\boldsymbol{A}^{-1}\boldsymbol{b} \\ &= \boldsymbol{x}^{(k)} - (\boldsymbol{I} - \boldsymbol{F})\boldsymbol{x}^{(k)} + (\boldsymbol{I} - \boldsymbol{F})\boldsymbol{A}^{-1}\boldsymbol{b} \\ &= \boldsymbol{x}^{(k)} + (\boldsymbol{I} - \boldsymbol{F})\boldsymbol{A}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k)}) \\ &= \boldsymbol{x}^{(k)} + (\boldsymbol{I} - \boldsymbol{F})\boldsymbol{A}^{-1}\boldsymbol{r}^{(k)}, \end{aligned} \tag{6.5}$$

où

$$\boldsymbol{r}^{(k)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k)} \tag{6.6}$$

est le k-ième vecteur résidu.

On cherche maintenant à préciser le choix de $\boldsymbol{F}$ pour approcher la solution de (6.2). La technique de *splitting* (« séparation » en français) consiste à prendre

$$\boldsymbol{A} = \boldsymbol{P} - \boldsymbol{N}, \tag{6.7}$$

où $\boldsymbol{P}$ est une matrice inversible. Par conséquent, on a

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} &\Leftrightarrow \boldsymbol{P}\boldsymbol{x} = \boldsymbol{N}\boldsymbol{x} + \boldsymbol{b} \\ &\Leftrightarrow \boldsymbol{x} = \boldsymbol{P}^{-1}\boldsymbol{N}\boldsymbol{x} + \boldsymbol{P}^{-1}\boldsymbol{b}, \end{aligned} \tag{6.8}$$

ce qui suggère de considérer $\boldsymbol{F} = \boldsymbol{P}^{-1}\boldsymbol{N}$ dans (6.4). De même, il en découle que

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{P}^{-1}\boldsymbol{r}^{(k)}, \tag{6.9}$$

puisque

$$\begin{aligned}(\boldsymbol{I} - \boldsymbol{F})\boldsymbol{A}^{-1} &= (\boldsymbol{I} - \boldsymbol{P}^{-1}\boldsymbol{N})(\boldsymbol{P} - \boldsymbol{N})^{-1} \\ &= \boldsymbol{P}^{-1}(\boldsymbol{P} - \boldsymbol{N})(\boldsymbol{P} - \boldsymbol{N})^{-1} \\ &= \boldsymbol{P}^{-1}. \end{aligned} \tag{6.10}$$

On remarque qu'à chaque itération (6.9), il est nécessaire d'appliquer $\boldsymbol{P}^{-1}$ et de calculer le résidu $\boldsymbol{r}^{(k)}$. La matrice $\boldsymbol{N}$ n'intervient pas dans le calcul. Il faut donc choisir une matrice $\boldsymbol{P}$ facile à inverser, mais suffisamment proche de $\boldsymbol{A}$ pour que la méthode converge vite. On se souvient également que $\boldsymbol{F} = \boldsymbol{P}^{-1}\boldsymbol{N}$ doit être contractante.

On peut généraliser la formule (6.9) en introduisant un paramètre de relaxation α_k servant à accélérer la convergence de la méthode. On parle alors de méthode de Richardson [Quarteroni et al., 2008, Saad, 2003]. Le calcul de $\boldsymbol{x}^{(k+1)}$ mène à l'expression :

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{P}^{-1}\boldsymbol{r}^{(k)}. \tag{6.11}$$

Posons $\boldsymbol{v}^{(k)} = \boldsymbol{P}^{-1}\boldsymbol{r}^{(k)}$. L'itération $k+1$ de la méthode de Richarson s'écrit simplement

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{v}^{(k)}. \tag{6.12}$$

L'algorithme correspondant est donc composé des quatre étapes :

1. Résoudre $\boldsymbol{P}\boldsymbol{v}^{(k)} = \boldsymbol{r}^{(k)}$.
2. Calculer α_k (dépend de la stratégie adoptée).
3. Calculer la solution $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{v}^{(k)}$.
4. Mettre à jour le résidu $\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - \alpha_k\boldsymbol{A}\boldsymbol{v}^{(k)}$.

6.1.2 Méthodes de Krylov

Considérons $\boldsymbol{P} = \boldsymbol{I}$. Dans l'approche de Richardson, une récurrence immédiate sur k permet d'établir que [Quarteroni et al., 2008, Saad, 2003] :

$$\boldsymbol{r}^{(k+1)} = \left[\prod_{i=0}^{k}(\boldsymbol{I} - \alpha_i \boldsymbol{A})\right] \boldsymbol{r}^{(0)} = p_{k+1}(\boldsymbol{A})\boldsymbol{r}^{(0)} \tag{6.13}$$

et

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(0)} + \sum_{i=0}^{k} \alpha_i \boldsymbol{r}^{(i)} = \boldsymbol{x}^{(0)} + q_k(\boldsymbol{A})\boldsymbol{r}^{(0)}, \tag{6.14}$$

où $p_{k+1}(\boldsymbol{A})$ est un polynôme en $\boldsymbol{A}$ d'ordre $k+1$ et $q_k(\boldsymbol{A})$ un polynôme en $\boldsymbol{A}$ d'ordre k. Ainsi, en définissant le p-ième espace de Krylov par

$$\mathcal{K}_p(\boldsymbol{A}, \boldsymbol{r}) = \text{vect}\{\boldsymbol{r}, \boldsymbol{A}\boldsymbol{r}, \boldsymbol{A}^2\boldsymbol{r}, \ldots, \boldsymbol{A}^{p-1}\boldsymbol{r}\}, \tag{6.15}$$

on trouve que pour tout $k \geqslant 1$,

$$\boldsymbol{r}^{(k)} \in \mathcal{K}_{k+1}(\boldsymbol{A}, \boldsymbol{r}^{(0)}) \quad \text{et} \quad \boldsymbol{x}^{(k)} \in \boldsymbol{x}^{(0)} + \mathcal{K}_k(\boldsymbol{A}, \boldsymbol{r}^{(0)}). \tag{6.16}$$

On peut montrer que la suite des espaces de Krylov est strictement croissante pour l'inclusion, jusqu'à un entier $p_{\max}$ (le « grade ») à partir duquel elle devient stagnante [Saad, 2003].

Définir une méthode de Krylov, c'est justement chercher à approximer $\boldsymbol{x}$ dans $\boldsymbol{x}^{(0)} + \mathcal{K}_k(\boldsymbol{A}, \boldsymbol{r}^{(0)})$ en trouvant un polynôme q_k d'ordre k tel que $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(0)} + q_k(\boldsymbol{A})\boldsymbol{r}^{(0)}$. De manière équivalente, cela revient à déterminer les coefficients $(\alpha_i)_{i \in [\![0,k]\!]}$ de la somme (6.14).

Nous allons maintenant décrire une approche pour déterminer le polynôme q_k.

On adopte le point de vue de l'optimisation. En effet, résoudre le système (6.2) équivaut à minimiser la fonctionnelle

$$\Phi(\boldsymbol{y}) = \frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}. \tag{6.17}$$

Dans le cas de la méthode de Richardson, ce principe s'applique à chaque itération en minimisant $\Phi(\boldsymbol{x}^{(k+1)})$ par rapport à α_k :

$$\Phi(\boldsymbol{x}^{(k+1)}) = \Phi(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{r}^{(k)}). \tag{6.18}$$

Après minimisation, on trouve que l'expression de α_k est donnée par

$$\alpha_k = \frac{\boldsymbol{r}^{(k)\mathrm{T}}\boldsymbol{r}^{(k)}}{\boldsymbol{r}^{(k)\mathrm{T}}\boldsymbol{A}\boldsymbol{r}^{(k)}}, \tag{6.19}$$

et on obtient la méthode du gradient (ou méthode de la plus forte pente). Cette méthode se généralise facilement au cas où $\boldsymbol{P} \neq \boldsymbol{I}$ en minimisant $\Phi(\boldsymbol{x}^{(k+1)}) = \Phi(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{P}^{-1}\boldsymbol{r}^{(k)})$. Dans ce cas, on trouve que

$$\alpha_k = \frac{\boldsymbol{r}^{(k)\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{r}^{(k)}}{\boldsymbol{r}^{(k)\mathrm{T}}\boldsymbol{P}^{-\mathrm{T}}\boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{r}^{(k)}} = \frac{\boldsymbol{r}^{(k)\mathrm{T}}\boldsymbol{v}^{(k)}}{\boldsymbol{v}^{(k)\mathrm{T}}\boldsymbol{A}\boldsymbol{v}^{(k)}}, \tag{6.20}$$

où $\boldsymbol{P}^{-\mathrm{T}} = (\boldsymbol{P}^{-1})^{\mathrm{T}}$.

Algorithme du gradient préconditionné

- Initialisation
  $\boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$
  $\boldsymbol{v}^{(0)} = \boldsymbol{r}^{(0)}$
- Calcul de l'itération numéro $j+1$
  $\boldsymbol{v}^{(j)} = \boldsymbol{P}^{-1}\boldsymbol{r}^{(j)}$
  $\alpha_j = (\boldsymbol{r}^{(j)\mathrm{T}}\boldsymbol{v}^{(j)})/(\boldsymbol{v}^{(j)\mathrm{T}}\boldsymbol{A}\boldsymbol{v}^{(j)})$
  $\boldsymbol{x}^{(j+1)} = \boldsymbol{x}^{(j)} + \alpha_j \boldsymbol{v}^{(j)}$
  $\boldsymbol{r}^{(j+1)} = \boldsymbol{r}^{(j)} - \alpha_j \boldsymbol{A}\boldsymbol{v}^{(j)}$
  Si $\|\boldsymbol{r}^{(j+1)}\|_2/\|\boldsymbol{b}\|_2 < \varepsilon$, alors fin de la procédure

Dans la méthode du gradient préconditionné, $\boldsymbol{v}^{(k)}$ s'interprête comme une direction de descente. Toutefois, la suite des directions de descente successives n'est pas optimale, car il peut arriver que la direction $\boldsymbol{v}^{(k)}$ à l'itération k soit la même que la direction $\boldsymbol{v}^{(l)}$ à l'itération l, tout en ayant $k \neq l$. L'idéal serait de ne faire appel qu'une seule fois à chaque direction.

Il convient donc de trouver un critère pour définir une méthode de Krylov dont les directions de descente $(\boldsymbol{v}^{(k)})_{k\geqslant 0}$ ne sont pas colinéaires. A cette fin,

on commence par établir que

$$\begin{aligned}\Phi(\boldsymbol{x}^{(k+1)}) &= \Phi(\boldsymbol{x}+(\boldsymbol{x}^{(k+1)}-\boldsymbol{x}))\\ &= \Phi(\boldsymbol{x})+\frac{1}{2}(\boldsymbol{x}^{(k+1)}-\boldsymbol{x})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}^{(k+1)}-\boldsymbol{x})\\ &= \Phi(\boldsymbol{x})+\frac{1}{2}\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}\|_{\boldsymbol{A}}^2. \end{aligned} \tag{6.21}$$

Ainsi, minimiser $\Phi(\boldsymbol{x}^{(k)})$ équivaut à minimiser

$$\Phi(\boldsymbol{x}^{(k)})-\Phi(\boldsymbol{x})=\frac{1}{2}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}\|_{\boldsymbol{A}}^2. \tag{6.22}$$

La solution approchée $\boldsymbol{x}^{(k)}$ est donc la projection orthogonale de $\boldsymbol{x}$ sur l'espace affine $\boldsymbol{x}^{(0)}+\mathcal{K}_k(\boldsymbol{A},\boldsymbol{r}^{(0)})$ [1]. Autrement dit, résoudre (6.2) en cherchant (6.16) revient à imposer la condition de Petrov-Galerkin [Saad, 2003] :

$$\boldsymbol{r}^{(k)}\perp\mathcal{K}_k(\boldsymbol{A},\boldsymbol{r}^{(0)}) \tag{6.23}$$

pour le produit scalaire associé à la norme euclidienne. Cette condition est la base des méthodes de Krylov orthogonales.

Soit $(\boldsymbol{v}^{(i)})_{i\in[\![1,k]\!]}$ une base de $\mathcal{K}_k(\boldsymbol{A},\boldsymbol{r}^{(0)})$ et $\boldsymbol{V}_k$ la matrice de taille $n\times k$ dont les colonnes sont les éléments de cette base. En vertu de (6.16), il est normal de chercher $\boldsymbol{x}^{(k)}$ sous la forme

$$\boldsymbol{x}^{(k)}=\boldsymbol{x}^{(0)}+\boldsymbol{V}_k\boldsymbol{z}^{(k)}, \tag{6.24}$$

où $\boldsymbol{z}^{(k)}$ est un vecteur de coordonnées.

La quantité (6.22) s'écrit

$$\begin{aligned}\frac{1}{2}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}\|_{\boldsymbol{A}}^2 &= \frac{1}{2}(\boldsymbol{x}^{(k)}-\boldsymbol{x})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}^{(k)}-\boldsymbol{x})\\ &= \frac{1}{2}(\boldsymbol{x}^{(0)}+\boldsymbol{V}_k\boldsymbol{z}^{(k)}-\boldsymbol{x})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}^{(0)}+\boldsymbol{V}_k\boldsymbol{z}^{(k)}-\boldsymbol{x})\\ &= \frac{1}{2}\boldsymbol{z}^{(k)\mathrm{T}}\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_k\boldsymbol{z}^{(k)}+(\boldsymbol{x}^{(0)}-\boldsymbol{x})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_k\boldsymbol{z}^{(k)}+\ldots \end{aligned} \tag{6.25}$$

En annulant son gradient, on se retrouve à devoir résoudre le système linéaire

$$(\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_k)\boldsymbol{z}^{(k)}=\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{r}^{(0)}. \tag{6.26}$$

Il est donc primordial de choisir la base $\boldsymbol{V}_k$ telle que le système linéaire (6.26) soit facile à résoudre. Dans la suite, on note

$$\boldsymbol{H}_k=\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_k. \tag{6.27}$$

1. Quand la solution approchée appartient à E et que le résidu est orthogonal à F, on dit que la méthode de projection est orthogonale si $E=F$. Lorsque $E\neq F$, on dit qu'elle est oblique.

6.1.3 Cas des matrices définies positives

Au premier abord, on serait tenté de considérer la base :

$$(\boldsymbol{r}^{(0)}, \boldsymbol{A}\boldsymbol{r}^{(0)}, \boldsymbol{A}^2\boldsymbol{r}^{(0)}, \ldots, \boldsymbol{A}^{p-1}\boldsymbol{r}^{(0)}). \tag{6.28}$$

Cependant, ce n'est pas une bonne idée en raison de sa dégénérescence numérique. En effet, supposons la matrice $\boldsymbol{A}$ diagonalisable et considérons la base de ses vecteur propres $(\boldsymbol{p}^{(i)})_{i\in[\![1,n]\!]}$. Notons $(\lambda_i)_{i\in[\![1,n]\!]}$ ses valeurs propres et « max » l'indice de la valeur propre maximale. On a

$$\boldsymbol{r}^{(0)} = \sum_{i=1}^{n} \alpha_i \boldsymbol{p}^{(i)} \quad \text{et} \quad \boldsymbol{A}^k\boldsymbol{r}^{(0)} = \sum_{i=1}^{n} \alpha_i \lambda_i^k \boldsymbol{p}^{(i)}. \tag{6.29}$$

La suite $(\boldsymbol{A}^k\boldsymbol{r}^{(0)})_{k\in\mathbb{N}}$ se comporte donc comme $(\alpha_{\max}\lambda_{\max}^k\boldsymbol{p}^{(\max)})_{p\in\mathbb{N}}$ lorsque $k \to +\infty$. Par conséquent, dès que le facteur $\left(\frac{\alpha_{\max}}{\alpha_i}\right)\left(\frac{\lambda_{\max}}{\lambda_i}\right)^k$ dépasse l'ordre de grandeur de la représentation des réels en machine, le terme $\alpha_i\lambda_i^k\boldsymbol{p}^{(i)}$ disparaît complètement. Les vecteurs $\boldsymbol{A}^k\boldsymbol{r}^{(0)}$ deviennent donc très rapidement colinéaires numériquement [Magoules and Roux, 2013].

Afin de résoudre ce problème de dégénérescence numérique, on peut orthonormaliser la base $\left(\boldsymbol{r}^{(0)}, \boldsymbol{A}\boldsymbol{r}^{(0)}, \boldsymbol{A}^2\boldsymbol{r}^{(0)}, \ldots, \boldsymbol{A}^{p-1}\boldsymbol{r}^{(0)}\right)$ par un procédé de Gram-Schmidt modifié, de telle sorte que

$$\boldsymbol{V}_p^{\mathrm{T}}\boldsymbol{V}_p = \boldsymbol{I}_p. \tag{6.30}$$

Il en résulte l'algorithme d'Arnoldi (tableau 6.1, à gauche). Remarque : la construction de la base d'Arnoldi reste valable dans le cas où $\boldsymbol{A}$ est simplement inversible (mais pas forcément symétrique définie positive).

Les vecteurs de la base d'Arnoldi vérifient la relation de récurrence

$$\boldsymbol{A}\boldsymbol{v}^{(j)} = \sum_{i=1}^{j+1} h_{ij}\boldsymbol{v}^{(i)}. \tag{6.31}$$

Matriciellement, cette relation s'écrit

$$\boldsymbol{A}\boldsymbol{V}_k = \boldsymbol{V}_{k+1}\boldsymbol{H}_{k+1k}, \tag{6.32}$$

Algorithme d'Arnoldi ($\boldsymbol{A}$ inversible)

— Initialisation

$\boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$

$\boldsymbol{v}^{(1)} = \frac{1}{\|\boldsymbol{r}^{(0)}\|_2}\boldsymbol{r}^{(0)}$

— Calcul de l'itération $j+1$

$\boldsymbol{w} = \boldsymbol{A}\boldsymbol{v}^{(j)}$

Pour $i = 1$ à j

$h_{ij} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}^{(i)}$

$\boldsymbol{w} = \boldsymbol{w} - h_{ij}\boldsymbol{v}^{(i)}$

Fin Pour

$h_{j+1j} = \|\boldsymbol{w}\|_2$

Si $h_{j+1j} = 0$, arrêt

$\boldsymbol{v}^{(j+1)} = \frac{1}{h_{j+1j}}\boldsymbol{w}$

Algorithme de Lanczos ($\boldsymbol{A}$ symétrique définie positive)

— Initialisation

$\boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$

$\boldsymbol{v}^{(1)} = \frac{1}{\|\boldsymbol{r}^{(0)}\|_2}\boldsymbol{r}^{(0)}$

— Calcul de l'itération 1

$\boldsymbol{w} = \boldsymbol{A}\boldsymbol{v}^{(1)}$

$h_{11} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}^{(1)}$

$\boldsymbol{w} = \boldsymbol{w} - h_{11}\boldsymbol{v}^{(1)}$

$h_{21} = \|\boldsymbol{w}\|_2$

$\boldsymbol{v}^{(2)} = \frac{1}{h_{21}}\boldsymbol{w}$

— Calcul de l'itération $j+1$

$\boldsymbol{w} = \boldsymbol{A}\boldsymbol{v}^{(j)}$

$h_{j-1j} = h_{jj}$

$\boldsymbol{w} = \boldsymbol{w} - h_{j-1j}\boldsymbol{v}^{(j-1)}$

$h_{jj} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}^{(j)}$

$\boldsymbol{w} = \boldsymbol{w} - h_{jj}\boldsymbol{v}^{(j)}$

$h_{j+1j} = \|\boldsymbol{w}\|_2$

Si $h_{j+1j} = 0$, arrêt

$\boldsymbol{v}^{(j+1)} = \frac{1}{h_{j+1j}}\boldsymbol{w}$

TABLE 6.1 – Construction de la base d'Arnoldi dans le cas non symétrique (à gauche) et dans le cas symétrique (à droite).

où

$$\boldsymbol{H}_{k+1k} = \begin{pmatrix} h_{11} & h_{12} & \cdots & \cdots & h_{1k} \\ h_{21} & h_{22} & & & \vdots \\ 0 & h_{32} & & & \vdots \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & h_{kk} \\ 0 & \cdots & \cdots & 0 & h_{k+1k} \end{pmatrix} \tag{6.33}$$

En multipliant (6.32) par $\boldsymbol{V}_k^{\mathrm{T}}$ et en utilisant (6.30), on déduit que $\boldsymbol{H}_k$ est la matrice constituée des k premières lignes de $\boldsymbol{H}_{k+1k}$. La matrice $\boldsymbol{H}_k$ est de forme Hessenberg supérieure, c'est-à-dire que ses coefficients sont nuls en dessous de la première sous-diagonale.

Quand on sait que la matrice $\boldsymbol{A}$ est symétrique, on constate que la matrice $\boldsymbol{H}_k$ l'est également. C'est alors une matrice tridiagonale, ce qui implique que le calcul des vecteurs de la base d'Arnoldi peut s'effectuer par une récurrence courte, c'est-à-dire que seuls les vecteurs $\boldsymbol{v}^{(j-1)}$ et $\boldsymbol{v}^{(j)}$ sont requis pour calculer $\boldsymbol{v}^{(j+1)}$. L'algorithme résultant est l'algorithme de Lanczos (tableau 6.1, à droite). Dans ce cas, calculer la solution approchée $\boldsymbol{x}^{(k)}$ est facile, puisque le système (6.26) est lui-même tridiagonal [Saad, 2003].

L'algorithme de Lanczos peut être utilisé pour estimer le spectre de $\boldsymbol{A}$ [Golub and Van Loan, 1996]. En effet, quand on atteint la convergence, la dernière direction $\boldsymbol{w}$ calculée est nulle. On a ainsi $\boldsymbol{A}\boldsymbol{V}_k = \boldsymbol{V}_k\boldsymbol{H}_k$. Supposons qu'il existe $(\lambda, \boldsymbol{u})$ tels que $\boldsymbol{H}_k\boldsymbol{u} = \lambda\boldsymbol{u}$. Alors $\boldsymbol{A}(\boldsymbol{V}_k\boldsymbol{u}) = \boldsymbol{V}_k\boldsymbol{H}_k\boldsymbol{u} = \lambda(\boldsymbol{V}_k\boldsymbol{u})$. On en déduit que les valeurs propres de $\boldsymbol{H}_k$ sont des valeurs propres de $\boldsymbol{A}$ avec des multiplicités éventuellement différentes. En réalité, on peut même approximer le spectre de $\boldsymbol{A}$ sans aller jusqu'à la convergence des itérations.

Si l'on tient absolument à résoudre un système diagonal, l'alternative évidente est d'orthonormaliser la base $(\boldsymbol{r}^{(0)}, \boldsymbol{A}\boldsymbol{r}^{(0)}, \boldsymbol{A}^2\boldsymbol{r}^{(0)}, \ldots, \boldsymbol{A}^{p-1}\boldsymbol{r}^{(0)})$ pour le produit scalaire associé à $\boldsymbol{A}$. Se pose alors la question du critère de convergence de la méthode. En effet, il n'est pas possible de calculer (6.22), puisque la solution exacte $\boldsymbol{x}$ est inconnue. L'astuce est de calculer à la place la norme du résidu adimensionnalisée et de la comparer à un paramètre de tolérance ε. A l'itération k, on arrête donc l'algorithme si

$$\frac{\|\boldsymbol{r}^{(k)}\|_2}{\|\boldsymbol{b}\|_2} = \frac{\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k)}\|_2}{\|\boldsymbol{b}\|_2} < \varepsilon. \tag{6.34}$$

Il faut donc de toute façon calculer les résidus successifs $(\boldsymbol{r}^{(k)})_{k\geqslant 0}$. Or, d'après (6.16) et (6.23), les résidus successifs forment une base de $\mathcal{K}_k(\boldsymbol{A}, \boldsymbol{r}^{(0)})$. Donc, quitte à orthogonaliser une base, autant orthogonaliser la base des résidus. On applique donc le procédé de Gram-Schmidt modifié à la base $(\boldsymbol{r}^{(0)}, \boldsymbol{r}^{(1)}, \boldsymbol{r}^{(2)}, \ldots, \boldsymbol{r}^{(k-1)})$ en utilisant le produit scalaire associé à $\boldsymbol{A}$. On obtient la méthode du gradient conjugué [Saad, 2003], qui garantit que (6.27) est diagonale. Les coefficients $(\alpha_k)_{k\geqslant 1}$ sont quant à eux calculés de manière analogue à la méthode du gradient préconditionné en minimisant

$$\Phi(\boldsymbol{x}^{(k+1)}) = \Phi(\boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{v}^{(k)}). \tag{6.35}$$

Attention, les $\boldsymbol{v}^{(k)}$ désignent cette fois-ci les directions de descente de la méthode du gradient conjugué. Un deuxième paramètre, β_k, est déductible des relations de $\boldsymbol{A}$-orthogonalité entre les directions de descente. On trouve que

$$\alpha_k = \frac{(\boldsymbol{r}^{(j)\mathrm{T}}\boldsymbol{r}^{(j)})}{(\boldsymbol{v}^{(j)\mathrm{T}}\boldsymbol{A}\boldsymbol{v}^{(j)})} \quad \text{et} \quad \beta_k = \frac{(\boldsymbol{r}^{(j+1)\mathrm{T}}\boldsymbol{r}^{(j+1)})}{(\boldsymbol{r}^{(j)\mathrm{T}}\boldsymbol{r}^{(j)})}. \tag{6.36}$$

Certains ouvrages privilégient les formulations

$$\alpha_k = \frac{(\boldsymbol{r}^{(j)\mathrm{T}}\boldsymbol{v}^{(j)})}{(\boldsymbol{v}^{(j)\mathrm{T}}\boldsymbol{A}\boldsymbol{v}^{(j)})} \quad \text{et} \quad \beta_k = -\frac{(\boldsymbol{r}^{(j+1)\mathrm{T}}\boldsymbol{A}\boldsymbol{v}^{(j)})}{(\boldsymbol{v}^{(j)\mathrm{T}}\boldsymbol{A}\boldsymbol{v}^{(j)})}, \tag{6.37}$$

qui sont équivalentes et mettent en évidence le lien avec la méthode du gradient.

Algorithme du gradient conjugué
($\boldsymbol{A}$ symétrique définie positive)

- Initialisation
  $\boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$
  $\boldsymbol{v}^{(0)} = \boldsymbol{r}^{(0)}$
- Calcul de l'itération numéro $j+1$
  $\alpha_j = (\boldsymbol{r}^{(j)\mathrm{T}}\boldsymbol{r}^{(j)})/(\boldsymbol{v}^{(j)\mathrm{T}}\boldsymbol{A}\boldsymbol{v}^{(j)})$
  $\boldsymbol{x}^{(j+1)} = \boldsymbol{x}^{(j)} + \alpha_j \boldsymbol{v}^{(j)}$
  $\boldsymbol{r}^{(j+1)} = \boldsymbol{r}^{(j)} - \alpha_j \boldsymbol{A}\boldsymbol{v}^{(j)}$
  Si $\|\boldsymbol{r}^{(j+1)}\|_2/\|\boldsymbol{b}\|_2 < \varepsilon$, alors fin de la procédure
  $\beta_j = (\boldsymbol{r}^{(j+1)\mathrm{T}}\boldsymbol{r}^{(j+1)})/(\boldsymbol{r}^{(j)\mathrm{T}}\boldsymbol{r}^{(j)})$
  $\boldsymbol{v}^{(j+1)} = \boldsymbol{r}^{(j+1)} + \beta_j \boldsymbol{v}^{(j)}$

Le principe du préconditionnement dans la méthode du gradient conjugué sera abordé dans la sous-section 6.2.1.

6.1.4 Cas des matrices non symétriques définies positives

Lorsque la matrice $\boldsymbol{A}$ n'est pas symétrique définie positive (*i.e.* lorsqu'une ou plusieurs conditions ne sont pas vérifiées), considérer la norme associée n'a pas de sens et minimiser (6.22) ne définit plus un critère d'optimalité pour trouver la solution de (6.2). A la place, on choisit de minimiser le carré de la norme-2 du résidu

$$\frac{1}{2}\|\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}^{(k)}\|_2^2=\frac{1}{2}\|\boldsymbol{r}^{(k)}\|_2^2, \tag{6.38}$$

ce qui équivaut à imposer la condition de Petrov-Galerkin [2]

$$\boldsymbol{r}^{(k)}\perp\boldsymbol{A}\mathcal{K}_k(\boldsymbol{A},\boldsymbol{r}^{(0)}), \tag{6.39}$$

où $\boldsymbol{A}\mathcal{K}_k(\boldsymbol{A},\boldsymbol{r}^{(0)})=\{\boldsymbol{A}\boldsymbol{v},\boldsymbol{v}\in\mathcal{K}_k(\boldsymbol{A},\boldsymbol{r}^{(0)})\}$. Comme dans l'approche de Richardson, $\boldsymbol{r}^{(k)}=0$ signifie que les suite des itérés à atteint l'unique point fixe de la méthode, donc que $\boldsymbol{x}^{(k)}=\boldsymbol{x}$.

Dès que $\boldsymbol{A}$ est inversible, $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ est symétrique définie positive. Résoudre (6.38) revient à chercher la solution du système équivalent

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}, \tag{6.40}$$

appelé « équation normale ». On est alors tenté de le résoudre par la méthode du gradient conjugué. Toutefois, c'est une mauvaise idée, puisque l'ordre de grandeur du conditionnement de $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ est égal au carré du conditionnement de $\boldsymbol{A}$. A moins que le conditionnement de $\boldsymbol{A}$ ne soit très petit, cela entraîne un ralentissement drastique de la convergence du gradient conjugué [Magoules and Roux, 2013].

La méthode du Gmres (*Generalized Minimum RESidual* en anglais) consiste à minimiser (6.38) en utilisant la base d'Arnoldi $\boldsymbol{V}_k$ [Saad and Schultz, 1986].

Similairement à (6.25), on peut développer (6.38) de la manière suivante :

$$\begin{aligned}
&\tfrac{1}{2}\|\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}^{(k)}\|_2^2 && (6.41)\\
=\;&\tfrac{1}{2}(\boldsymbol{x}^{(k)}-\boldsymbol{x})^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}^{(k)}-\boldsymbol{x})\\
=\;&\tfrac{1}{2}(\boldsymbol{x}^{(0)}+\boldsymbol{V}_k\boldsymbol{z}^{(k)}-\boldsymbol{x})^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}^{(0)}+\boldsymbol{V}_k\boldsymbol{z}^{(k)}-\boldsymbol{x})\\
=\;&\tfrac{1}{2}\boldsymbol{z}^{(k)\mathrm{T}}\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_k\boldsymbol{z}^{(k)}+(\boldsymbol{x}^{(0)}-\boldsymbol{x})^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_k\boldsymbol{z}^{(k)}+\dots && (6.42)
\end{aligned}$$

2. Dans ce cas, la condition de Petrov-Galerkin définit une projection oblique, puisque $\boldsymbol{x}^{(k)}\in\boldsymbol{x}^{(0)}+\mathcal{K}_k$ et $\boldsymbol{r}^{(k)}\perp\boldsymbol{A}\mathcal{K}_k$.

On parvient alors au système linéaire

$$(\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_k)\boldsymbol{z}^{(k)} = \boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{r}^{(0)}, \tag{6.43}$$

qui ne possède pas de structure favorable à sa résolution. Il faut donc changer de stratégie.

On remarque que le résidu de l'itération k s'exprime en fonction du résidu initial :

$$\begin{aligned}\boldsymbol{r}^{(k)} &= \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k)} \\ &= \boldsymbol{b} - \boldsymbol{A}(\boldsymbol{x}^{(0)} + \boldsymbol{V}_k\boldsymbol{z}^{(k)}) \\ &= \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)} - \boldsymbol{A}\boldsymbol{V}_k\boldsymbol{z}^{(k)} \\ &= \boldsymbol{r}^{(0)} - \boldsymbol{A}\boldsymbol{V}_k\boldsymbol{z}^{(k)}. \end{aligned} \tag{6.44}$$

En vertu de (6.32), on a également

$$\boldsymbol{r}^{(k)} = \boldsymbol{r}^{(0)} - \boldsymbol{V}_{k+1}\boldsymbol{H}_{k+1k}\boldsymbol{z}^{(k)}, \tag{6.45}$$

ce qui permet d'écrire

$$\frac{1}{2}\|\boldsymbol{r}^{(k)}\|_2^2 = \frac{1}{2}\|\boldsymbol{r}^{(0)} - \boldsymbol{V}_{k+1}\boldsymbol{H}_{k+1k}\boldsymbol{z}^{(k)}\|_2^2. \tag{6.46}$$

Comme $\boldsymbol{V}_{k+1}$ est construit par l'algorithme d'Arnoldi, sa première composante $\boldsymbol{v}^{(1)}$ est égale à $\frac{1}{\|\boldsymbol{r}^{(0)}\|_2}\boldsymbol{r}^{(0)}$. Soit $\boldsymbol{e}_1$ le premier vecteur de la base canonique de $\mathbb{R}^{k+1}$. On a

$$\begin{aligned}\boldsymbol{r}^{(0)} &= \|\boldsymbol{r}^{(0)}\|_2\boldsymbol{v}^{(1)} \\ &= \|\boldsymbol{r}^{(0)}\|_2\boldsymbol{V}_{k+1}\boldsymbol{e}_1, \end{aligned} \tag{6.47}$$

ce qui entraîne la factorisation

$$\begin{aligned}\frac{1}{2}\|\boldsymbol{r}^{(k)}\|_2^2 &= \frac{1}{2}\|\boldsymbol{V}_{k+1}(\|\boldsymbol{r}^{(0)}\|_2\boldsymbol{e}_1 - \boldsymbol{H}_{k+1k}\boldsymbol{z}^{(k)})\|_2^2 \\ &= \frac{1}{2}(\|\boldsymbol{r}^{(0)}\|_2\boldsymbol{e}_1 - \boldsymbol{H}_{k+1k}\boldsymbol{z}^{(k)})^{\mathrm{T}}\boldsymbol{V}_{k+1}^{\mathrm{T}}\boldsymbol{V}_{k+1}(\|\boldsymbol{r}^{(0)}\|_2\boldsymbol{e}_1 - \boldsymbol{H}_{k+1k}\boldsymbol{z}^{(k)}) \\ &= \frac{1}{2}(\|\boldsymbol{r}^{(0)}\|_2\boldsymbol{e}_1 - \boldsymbol{H}_{k+1k}\boldsymbol{z}^{(k)})^{\mathrm{T}}\boldsymbol{I}_{k+1}(\|\boldsymbol{r}^{(0)}\|_2\boldsymbol{e}_1 - \boldsymbol{H}_{k+1k}\boldsymbol{z}^{(k)}) \\ &= \frac{1}{2}\|\|\boldsymbol{r}^{(0)}\|_2\boldsymbol{e}_1 - \boldsymbol{H}_{k+1k}\boldsymbol{z}^{(k)}\|_2^2. \end{aligned} \tag{6.48}$$

Pour établir (6.48), on a utilisé la relation d'orthonormalité (6.30) de $\boldsymbol{V}_{k+1}$.

Cette fois, le problème fait intervenir la matrice $\boldsymbol{H}_{k+1k}$, qui est de type Hessenberg supérieure. La stratégie de résolution consiste à factoriser $\boldsymbol{H}_{k+1k}$ sous la forme $\boldsymbol{QR}$, où $\boldsymbol{Q}$ est une matrice orthogonale et $\boldsymbol{R}$ une matrice triangulaire supérieure. Il suffit ensuite de résoudre un système de la forme

$$\boldsymbol{R}\boldsymbol{z}^{(k)} = \boldsymbol{y}^{(k)}. \tag{6.49}$$

A ce stade, plusieurs factorisations sont envisageables. La première fait appel au procédé de Gram-Schmidt modifié, qui est malheureusement connu pour être instable numériquement. La méthode de Householder exploite les propriétés des matrices de réflection. Ce procédé est stable, bien que non-parallélisable. En pratique, on utilise plutôt la troisième méthode, qui fait appel aux rotations planes de Givens.

On remarque que quand $\boldsymbol{A}$ est symétrique, la base d'Arnoldi peut être calculée par la méthode de Lanczos. La méthode prend le nom de MINRES.

Algorithme du GMRES / MINRES

- Initialisation
  $\boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$
  $\boldsymbol{v}^{(1)} = \frac{1}{\|\boldsymbol{r}^{(0)}\|_2}\boldsymbol{r}^{(0)}$
- Calcul du vecteur de base numéro $j+1$
  Si $\boldsymbol{A}$ n'est pas symétrique, algorithme d'Arnoldi
  Si $\boldsymbol{A}$ est symétrique, algorithme de Lanczos
- Critère d'arrêt :
  Fin de la procédure si $h_{j+1j} = 0$
- Détermination de la solution à l'itération $j+1$
  Définir la matrice Hessenberg supérieure $\boldsymbol{H}_{j+1j}$
  Factoriser $\boldsymbol{H}_{j+1j}$ par Gram-Schmidt, Householder ou Givens
  Résoudre le système triangulaire pour déterminer $\boldsymbol{z}^{(k)}$
  Mettre à jour $\boldsymbol{x}^{(k+1)}$

6.1.5 Iterations de Chebyshev

On présente une autre méthode polynomiale qui utilise des polynômes orthogonaux, mais dont les itérés n'appartiennent pas aux espaces de Krylov. Les polynômes orthogonaux en question sont les polynômes de Chebychev. Ils sont orthogonaux pour le produit scalaire

$$\langle p, q\rangle_w = \int_{-1}^{+1} p(t)q(t)w(t)\mathrm{d}t \quad , \quad w(t) = \frac{1}{\sqrt{1-t^2}} \tag{6.50}$$

qui ne dépend pas du résidu initial $\boldsymbol{r}^{(0)}$ ou du second membre $\boldsymbol{b}$ [Fischer, 1996].

Ecrivons $p_k(\boldsymbol{A})$ sous la forme

$$p_k(\boldsymbol{A}) = \sum_{i=0}^{k} \gamma_i \boldsymbol{A}^i \tag{6.51}$$

et posons

$$p_k(t) = \sum_{i=0}^{k} \gamma_i t^i \tag{6.52}$$

pour tout $t \in \mathbb{R}$. D'après (6.13), $p_k(\boldsymbol{A})$ s'écrit nécessairement sous la forme (faire le développement pour s'en convaincre)

$$p_k(\boldsymbol{A}) = \boldsymbol{I} - s(\boldsymbol{A})\boldsymbol{A}, \tag{6.53}$$

où $s(\boldsymbol{A})$ est un polynôme en $\boldsymbol{A}$ de degré égal à $k-1$. Cette relation entraîne automatiquement que

$$p_k(1) = 1. \tag{6.54}$$

Dans (6.53), le polynôme $s(\boldsymbol{A})$ peut s'interpréter comme un préconditionneur. En effet, si $s(\boldsymbol{A}) = \boldsymbol{A}^{-1}$, alors $p_k(\boldsymbol{A}) = \boldsymbol{0}$ et le système $s(\boldsymbol{A})\boldsymbol{A}\boldsymbol{x} = s(\boldsymbol{A})\boldsymbol{b}$ converge en une seule itération. On veut donc trouver le polynôme $p_k(\boldsymbol{A})$ « le plus petit possible », dans une norme à définir.

Pour ce faire, on introduit la norme $\|\cdot\|_S$ définie pour toute matrice $\boldsymbol{Z}$ [Saad, 2003] par

$$\|\boldsymbol{Z}\|_S = \operatorname*{Sup}_{\|\boldsymbol{x}\|_2 \leqslant 1} \|\boldsymbol{Z}\boldsymbol{x}\|_2. \tag{6.55}$$

On peut montrer que, si $\mathrm{Sp}(\boldsymbol{Z})$ désigne l'ensemble des valeurs propres de $\boldsymbol{Z}$, alors

$$\|\boldsymbol{Z}\|_S = \max_{\lambda \in \mathrm{Sp}(\boldsymbol{Z})} |\lambda|. \tag{6.56}$$

Remarquons que si λ est une valeur propre de $\boldsymbol{A}$, alors $p_k(\lambda)$ est une valeur propre de $p_k(\boldsymbol{A})$. Dès lors, on cherche le polynôme $p_k(\boldsymbol{A})$ qui minimise

$$\|p_k(\boldsymbol{A})\|_S = \max_{\lambda \in \mathrm{Sp}(\boldsymbol{A})} |p_k(\lambda)|. \tag{6.57}$$

Toutefois, la minimisation de (6.57) requiert la connaissance *a priori* du spectre de $\boldsymbol{A}$, ce qui n'est pas le cas en général. On remplace donc la condition $\lambda \in \mathrm{Sp}(\boldsymbol{A})$ par une condition sur un intervalle [3] de $\mathbb{R}$ tel que $\mathrm{Sp}(\boldsymbol{A}) \subset [\alpha, \beta]$. Dès lors, la nouvelle condition s'écrit

$$\max_{\lambda \in [\alpha,\beta]} |p_k(\lambda)| \tag{6.58}$$

et on peut montrer que le polynôme optimal qui minimise (6.58) tout en vérifiant (6.54) appartient aux polynômes de Chebyshev de type 1 et d'ordre k. Une relation de récurrence d'ordre trois portant sur ces polynômes permet d'obtenir l'algorithme connu sous le nom d'« itérations de Chebyshev » [Gutknecht and Röllin, 2002].

Itérations de Chebyshev

- Initialisation

  $\boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$

  $\theta = \frac{\beta+\alpha}{2}$; $\delta = \frac{\beta-\alpha}{2}$; $\sigma_1 = \frac{\theta}{\delta}$

  $\alpha_0 = \frac{1}{\sigma_1}$; $\boldsymbol{v}^{(0)} = \frac{1}{\theta}\boldsymbol{r}^{(0)}$
- Calcul de l'itération numéro $j+1$

  $\boldsymbol{x}^{(j+1)} = \boldsymbol{x}^{(j)} + \alpha_j \boldsymbol{v}^{(j)}$

  $\boldsymbol{r}^{(j+1)} = \boldsymbol{r}^{(j)} - \boldsymbol{A}\boldsymbol{v}^{(j)}$

  $\alpha_{j+1} = (2\sigma_1 - \alpha_j)^{-1}$

  $\boldsymbol{v}^{(j+1)} = \alpha_j \boldsymbol{v}^{(j)} + \frac{2}{\delta}\boldsymbol{r}^{(j+1)}$

Remarque : même si la connaissance exacte du spectre de $\boldsymbol{A}$ n'est plus requise, il est tout de même nécessaire d'en déterminer un minorant α et un majorant β afin de pouvoir résoudre (6.58). En pratique, plus cette connaissance est précise, plus l'algorithme converge vite.

3. Un intervalle réel lorsque $\boldsymbol{A}$ est symétrique définie positive. Une ellipse du plan complexe lorsqu'elle ne l'est pas.

6.1.6 Discussion

Dans cette section, nous avons présenté de nombreux algorithmes itératifs pour la résolution des systèmes linéaires. Selon le cas d'étude, certains algorithmes sont connus pour être plus performants que les autres. Le diagramme de la figure 6.1 récapitule les choix possibles en fonction des propriétés connues de la matrice $\boldsymbol{A}$. Anticipant sur les sections suivantes, on indique également la forme du préconditionneur à utiliser dans chaque cas.

La notion de préconditionneur est d'importance capitale dans les applications récentes. En effet, les conditions de Petrov-Galerkin imposent que le résidu final soit orthogonal au plus grand espace possible. Cependant, rien n'indique que les espaces de Krylov soient un choix pertinent pour définir cette condition d'optimalité. Plus récemment, les méthodes de Krylov sont devenues importantes quand la recherche de précondionnements efficaces est elle-même devenue une priorité. Aujourd'hui, les approches par *splitting* ou la descente de gradient [4] sont principalement utilisées à des fins pédagogiques, ou à des fins de préconditionnement.

Le choix entre la méthode du gradient conjugué et MINRES est relatif à la quantité qu'on souhaite minimiser en cas de troncature des itérations (*i.e.* quand on arrête l'algorithme avant d'atteindre la convergence). Le premier minimise la norme-$\boldsymbol{A}$ de l'erreur, qui définit un critère d'optimalité pertinent du point de vue physique (on l'interprète comme une énergie), alors que le second minimise la norme-2 du résidu. C'est pourquoi, en pratique, on opte pour la méthode du gradient conjugué dès que la matrice $\boldsymbol{A}$ est symétrique définie positive. Le MINRES reste quant à lui réservé au cas des matrice non positives.

Concernant les itérations de Chebyshev, cette approche est particulièrement utile dans un contexte massivement parallèle. Contrairement à la méthode du gradient conjugué, elle ne requiert pas de produit scalaires. Cet avantage est aussi un inconvénient, puisqu'il impose à tout éventuel préconditionnement de lui-même ne pas faire appel à des produits scalaires.

On précise enfin qu'il existe une multitude de méthodes itératives qui ne sont pas abordées dans ce manuscrit, comme les méthodes avec bi-orthogona-

4. De plus, la méthode du gradient minimise en réalité $\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{A}+\boldsymbol{A}^{\mathrm{T}})\boldsymbol{x}-\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}$. Quand la matrice $\boldsymbol{A}$ n'est pas symétrique, cette quantité n'égale pas $\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}$.

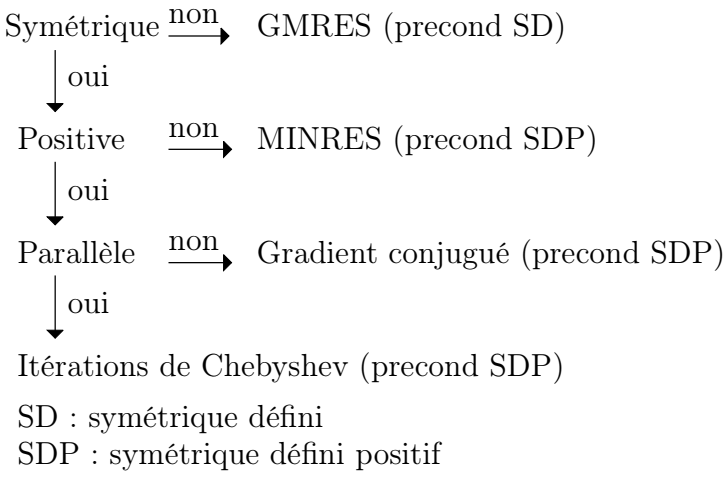


Figure 6.1 – Arbre de décision : quel algorithme itératif choisir en fonction des propriétés de la matrice du système linéaire ? On suppose la matrice inversible.

lisation, les techniques multigrille ou encore la décomposition de domaine. Ces différentes approches n'ont pas trouvé de motivation dans cette étude. Néanmoins, elles pourront être exploitées dans le contexte d'une implémentation opérationnelle, si les méthodes plus classiques viennent à faire défaut.

6.2 Préconditionnement et résolution

On considère de nouveau la problématique générale de la modélisation des opérateurs de corrélations. On cherche à appliquer l'algorithme du gradient conjugué au système linéaire (6.1). Tout d'abord, la notion de préconditionnement est détaillée. Puis, l'étude se concentre sur l'utilisation de deux préconditionnements issus des chapitres 2 et 5. Pour chacun, une étude de performance est menée à partir des données de Seviri.

Le lecteur trouvera une présentation détaillée des techniques des préconditionnement dans les ouvrages de Barrett [1994], Benzi [2002], Chan et al. [1996] et Wathen [2015].

6.2.1 Principe du préconditionnement

Soit le système linéaire modèle

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{6.59}$$

où $\boldsymbol{A}$ désigne une matrice symétrique définie positive et soit $\boldsymbol{P}$ une matrice de préconditionnement également symétrique définie positive telle que

$$\boldsymbol{P}^{-1} \simeq \boldsymbol{A}^{-1}. \tag{6.60}$$

L'idée du préconditionnement est de tirer profit de (6.60) pour accélérer la résolution de (6.59) à l'aide de méthodes itératives comme le gradient conjugué ou GMRES. On distingue trois types de préconditionnement. Remarque : quand $\boldsymbol{P}$ ne change pas au cours des itérations, on parle de préconditionneur du premier ordre.

Préconditionnement à gauche

Le préconditionnement à gauche consiste à résoudre le système équivalent

$$\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{P}^{-1}\boldsymbol{b}. \tag{6.61}$$

La symétrie de $\boldsymbol{A}$ et de $\boldsymbol{P}$ n'impliquant pas toujours celle du produit $\boldsymbol{P}^{-1}\boldsymbol{A}$, il n'est pas possible d'appliquer directement la méthode du gradient conjugué à (6.61). Pour parer cet écueil, il suffit de changer le produit scalaire canonique par le produit scalaire associé à $\boldsymbol{P}$ dans le calcul des coefficients α_k et β_k de (6.36). Autrement, il est possible de recourir à des méthodes non symétriques (figure 6.1).

On remarque que (6.61) se réécrit $\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{P}^{-1}\boldsymbol{A})\boldsymbol{x} + \boldsymbol{P}^{-1}\boldsymbol{b}$, ce qui permet d'interpréter le préconditionnement comme un schéma de *splitting* appliqué à une méthode de point fixe (sous-section 6.1.1).

Préconditionnement à droite

Le préconditionnement à droite consiste à résoudre le système équivalent

$$\boldsymbol{A}\boldsymbol{P}^{-1}\hat{\boldsymbol{x}} = \boldsymbol{b}, \tag{6.62}$$

où $\hat{\boldsymbol{x}} = \boldsymbol{P}\boldsymbol{x}$. En général, cette technique est utilisée conjointement avec des algorithmes comme le GMRES, qui ne requièrent pas que la matrice du système linéaire préconditionné soit symétrique. Appliqué à (6.62), le GMRES

minimise le même résidu que dans sa version non préconditionnée, c'est-à-dire $\|\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b}\|_2$. En revanche, appliqué à (6.61), le GMRES minimise $\|\boldsymbol{P}^{-1}(\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b})\|_2$.

Préconditionnement centré

Enfin, le préconditionnement centré (*split-preconditioning* en anglais, à ne pas confondre avec les méthodes de *splitting*) consiste à résoudre le système équivalent

$$\boldsymbol{L}^{-1}\boldsymbol{A}\boldsymbol{U}^{-1}\hat{\boldsymbol{x}} = \boldsymbol{L}^{-1}\boldsymbol{b}, \tag{6.63}$$

où $\hat{\boldsymbol{x}} = \boldsymbol{U}\boldsymbol{x}$. Les matrices $\boldsymbol{L}$ et $\boldsymbol{U}$ sont en général issues de la factorisation d'un préconditionneur valide $\boldsymbol{P}$ vérifiant

$$\boldsymbol{P} = \boldsymbol{L}\boldsymbol{U}. \tag{6.64}$$

Lorsqu'on impose la relation (6.64), la méthode du gradient conjugé appliquée au système (6.63) sans préconditionnement est équivalente à la méthode du gradient conjugué préconditionnée par (6.61) ou (6.63). Remarque : une astuce de calcul permet de s'affranchir de la définition de $\boldsymbol{L}$ et $\boldsymbol{U}$, qui n'interviennent pas explicitement dans l'écriture de l'algorithme final [Golub and Van Loan, 1996].

L'application de (6.63) à la méthode du GMRES minimise quant à elle le résidu préconditionné $\|\boldsymbol{U}^{-1}(\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b})\|_2$ [Saad and Schultz, 1986].

Dans tous les cas, les spectres des matrices $\boldsymbol{A}\boldsymbol{P}^{-1}$, $\boldsymbol{A}\boldsymbol{P}^{-1}$ et $\boldsymbol{L}^{-1}\boldsymbol{A}\boldsymbol{U}^{-1}$ sont identiques. Néanmoins, cela ne signifie pas que les algorithmes correspondants convergent de la même manière. En particulier, le critère d'arrêt est défini à partir de la norme du résidu, dont la définition diffère d'un à l'autre.

Algorithme du gradient conjugué préconditionné ($\boldsymbol{A}$ et $\boldsymbol{P}$ symétriques définies positives)

- Initialisation

  $\boldsymbol{r}^{(0)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(0)}$

  $\boldsymbol{v}^{(0)} = \boldsymbol{r}^{(0)}$

- Calcul de l'itération numéro $j+1$

  $\alpha_j = (\boldsymbol{r}^{(j)\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{r}^{(j)})/(\boldsymbol{v}^{(j)\mathrm{T}}\boldsymbol{A}\boldsymbol{v}^{(j)})$

  $\boldsymbol{x}^{(j+1)} = \boldsymbol{x}^{(j)} + \alpha_j \boldsymbol{v}^{(j)}$

  $\boldsymbol{r}^{(j+1)} = \boldsymbol{r}^{(j)} - \alpha_j \boldsymbol{A}\boldsymbol{v}^{(j)}$

  Si $\|\boldsymbol{r}^{(j+1)}\|_2/\|\boldsymbol{b}\|_2 < \varepsilon$, alors fin de la procédure

  $\beta_j = (\boldsymbol{r}^{(j+1)\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{r}^{(j+1)})/(\boldsymbol{r}^{(j)\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{r}^{(j)})$

  $\boldsymbol{v}^{(j+1)} = \boldsymbol{P}^{-1}\boldsymbol{r}^{(j+1)} + \beta_j \boldsymbol{v}^{(j)}$

6.2.2 Approximation grossière de l'inverse

Cette sous-section montre que l'opérateur de corrélation augmenté (5.27) est facilement inversible en faisant appel au bon préconditionneur. Ce résultat fort justifie à lui seul le développement et l'utilisation du raffinement de maillage dans la représentation des opérateurs de corrélation sur des maillages non structurés. De plus, ce travail sert de base à la mise au point des précontionneurs du deuxième ordre de la section 6.4.

Utilisons les notions introduites précédemment pour résoudre le système linéaire (6.1) associé à l'opérateur de corrélation augmenté (5.27). Au vu des résultats de la sous-section 5.3.2, on décide de se concentrer uniquement sur l'inversion de la formule (5.29), au sein de laquelle les opérateurs de transfert sont définis par injection. A partir de maintenant, on réadopte les notations antérieures à la section 6.1. On suppose que toute digression faisant référence aux notations de la section 6.1 est suffisamment claire pour ne pas introduire de confusion.

Notons respectivement $\boldsymbol{C}_c^{nr}$ et $\boldsymbol{C}_c^{r}$ les opérateurs de corrélation sans et avec raffinement, agissant dans l'espace des observations. On rappelle leur

expression :

$$\boldsymbol{C}_c^{nr} = [(\boldsymbol{M}_c + \boldsymbol{K}_c)^{-1}\boldsymbol{M}_c]^m \boldsymbol{M}_c^{-1}, \tag{6.65}$$

$$\boldsymbol{C}_c^{r} = \boldsymbol{F}^\star[(\boldsymbol{M}_f + \boldsymbol{K}_f)^{-1}\boldsymbol{M}_f]^m \boldsymbol{M}_f^{-1}(\boldsymbol{F}^\star)^{\mathrm{T}}. \tag{6.66}$$

Le maillage fin est construit par raffinement non-hiérarchique (sous-section 5.1.1) à partir des données du sondeur SEVIRI (5.1.3). La matrice de masse grossière $\boldsymbol{M}_c$ est construite indépendamment de la matrice de masse fine $\boldsymbol{M}_f$. Enfin, l'opérateur de transfert $\boldsymbol{F}^\star$ est défini par injection (sous-section 5.2.2).

Soit le système linéaire de taille $p \times p$ ($p = 4856$) :

$$\boldsymbol{C}_c^r \boldsymbol{x} = \boldsymbol{b}. \tag{6.67}$$

Comme l'opérateur de corrélation $\boldsymbol{C}_c^r$ est symétrique défini positif, on résout (6.67) par la méthode du gradient conjugué. Les itérations sont arrêtées arbitrairement quand la norme-2 du résidu devient inférieure à $\epsilon = 10^{-6}$. Résultat : sans préconditionnement, l'algorithme converge en 602 itérations.

L'algorithme du gradient conjugué converge lentement principalement à cause des petites valeurs propres de $\boldsymbol{C}_c^r$. La présence de ces petites valeurs propres est inhérente à l'équation de diffusion, dont le spectre est fondamentalement très étalé (sous-section 1.1.4). Toutefois, d'après Allaire [2005], seules les grandes valeurs propres de l'opérateur de corrélation discret sont des approximations correctes des valeurs propres exactes (*i.e.* du problème continu). Les plus petites valeurs propres n'ont pas de signification physique. Ceci est dû à la projection de l'équation de diffusion sur un espace de dimension finie. On est donc tenté de développer un préconditionneur à partir d'une approximation de $\boldsymbol{C}_c^r$ qui partage sensiblement les mêmes grandes valeurs propres. L'opérateur $\boldsymbol{C}_c^{nr}$ émerge naturellement parmis les candidats potentiels.

On préconditionne donc le système (6.67) par $\boldsymbol{P}^{-1} = (\boldsymbol{C}_c^{nr})^{-1}$ et on observe la convergence en 38 itérations seulement sous le seuil ϵ fixé. Soit une diminution drastique du nombre d'itérations du gradient conjugué. Ainsi, on peut affirmer que l'opérateur de corrélation sans raffinement est un excellent préconditionneur de premier ordre pour inverser le système associé à l'opérateur de corrélation augmenté.

L'apport du préconditionnement du premier ordre est résumé dans le tableau 6.2.

Méthode	Préconditionneur	Itérations
Gradient Conjugué	$\boldsymbol{I}$	602
Gradient Conjugué préconditionné	$\boldsymbol{C}_c^{nr}$	38

Figure 6.2 – Performance comparée du gradient conjugué avec et sans préconditionnement pour l'inversion de l'opérateur de corrélation augmenté. Le préconditionneur est obtenu par l'équation de diffusion discrétisée par éléments finis sans raffinement de maillage. Norme du résidu à convergence : inférieure à $\epsilon = 10^{-6}$. Le préconditionnement diminue largement le nombre d'itérations requis pour l'inversion.

6.2.3 Approximation fine de l'inverse

Le second préconditionneur qui semble adapté à la résolution de (6.67) est l'approximation de l'inverse par raffinement exposé dans la section 5.4. Or, il se trouve que c'est une mauvais choix en pratique. La convergence du gradient conjugué s'en retrouve largement ralentie et la procédure ne converge pas sous le seuil ϵ en moins de 700 itérations (tableau 6.3).

Pour valider ce résultat (négatif), l'expérience est reproduite sur un maillage cartésien de même dimension approximative (taille $69 \times 69 = 4761$). La conclusion est la même, indiquant que la non-convergence n'est pas due à l'aspect non structuré des données et du maillage.

Dans la suite, on s'abtient donc de considérer le préconditionneur construit à partir de l'opérateur de précision rafinné.

Méthode	Préconditionneur	Itérations
Gradient Conjugué	$\boldsymbol{I}$	602
Gradient Conjugué préconditionné	$\boldsymbol{Q}_c^{-1}$	>700

Figure 6.3 – Performance comparée du gradient conjugué avec et sans préconditionnement pour l'inversion de l'opérateur de corrélation augmenté. Le préconditionneur est obtenu par interpolation de l'inverse de la diffusion depuis le maillage fin. Norme du résidu à convergence : inférieure à $\epsilon = 10^{-6}$. Les chiffres indiquent que le préconditionneur ne possède pas les bonnes propriétés pour accélérer la convergence du gradient conjugué.

6.3 Reformulation comme point-selle

Cette section s'adresse toujours à l'inversion du système (6.67), en constatant cette fois-ci que la matrice $\boldsymbol{C}_c^r$ peut s'interpréter comme le complément de Schur [Benzi et al., 2005] du système point selle

$$\begin{pmatrix} \boldsymbol{C}_f^{-1} & (\boldsymbol{F}^\star)^{\mathrm{T}} \\ \boldsymbol{F}^\star & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ -\boldsymbol{b} \end{pmatrix}, \tag{6.68}$$

où $\boldsymbol{C}_f$ désigne l'opérateur de corrélation discrétisé sur le maillage fin. Cette formulation provient directement de l'interprétation du raffinement comme un problème d'optimisation sous contraintes, où les contraintes sont exprimées à l'aide des opérateurs de transfert. La résolution de (6.68) est ainsi équivalente à la recherche de la solution de

$$\begin{cases} \min\limits_{\boldsymbol{y}} \frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{C}_f^{-1}\boldsymbol{y} \\ \text{sachant} \quad \boldsymbol{F}^\star\boldsymbol{y} + \boldsymbol{b} = \boldsymbol{0} \end{cases}. \tag{6.69}$$

Dans un premier temps, on montre l'équivalence entre le système initial (6.67) et la formulation point-selle (6.68). Puis, on s'adresse à la question du préconditionnement. Enfin, on présente les performance de la méthode dans une sous-section indépendante.

6.3.1 Formulation point-selle

On commence par établir la relation d'implication suivante :

$$\begin{aligned} \begin{pmatrix} \boldsymbol{C}_f^{-1} & (\boldsymbol{F}^\star)^{\mathrm{T}} \\ \boldsymbol{F}^\star & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ -\boldsymbol{b} \end{pmatrix} \quad &\Rightarrow \quad \begin{cases} \boldsymbol{C}_f^{-1}\boldsymbol{y} + (\boldsymbol{F}^\star)^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{0} \\ \boldsymbol{F}^\star\boldsymbol{y} + \boldsymbol{b} = \boldsymbol{0} \end{cases} \\ &\Rightarrow \quad \begin{cases} \boldsymbol{y} + \boldsymbol{C}_f(\boldsymbol{F}^\star)^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{0} \\ \boldsymbol{F}^\star\boldsymbol{y} + \boldsymbol{b} = \boldsymbol{0} \end{cases} \\ &\Rightarrow \quad \begin{cases} \boldsymbol{F}^\star\boldsymbol{y} + \boldsymbol{F}^\star\boldsymbol{C}_f(\boldsymbol{F}^\star)^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{0} \\ \boldsymbol{F}^\star\boldsymbol{y} = -\boldsymbol{b} \end{cases} \\ &\Rightarrow \quad -\boldsymbol{b} + \boldsymbol{F}^\star\boldsymbol{C}_f(\boldsymbol{F}^\star)^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{0} \\ &\Rightarrow \quad \boldsymbol{x} = [\boldsymbol{F}^\star\boldsymbol{C}_f(\boldsymbol{F}^\star)^{\mathrm{T}}]^{-1}\boldsymbol{b} \\ &\Rightarrow \quad \boldsymbol{x} = [\boldsymbol{C}_c^r]^{-1}\boldsymbol{b}. \end{aligned} \tag{6.70}$$

Voilà qui justifie la formulation point-selle. Dans la suite, on note

$$\boldsymbol{\Xi} = \begin{pmatrix} \boldsymbol{C}_f^{-1} & (\boldsymbol{F}^\star)^{\mathrm{T}} \\ \boldsymbol{F}^\star & \boldsymbol{0} \end{pmatrix}. \tag{6.71}$$

Soit $\boldsymbol{S}$ une approximation du complément de Schur $\boldsymbol{F}^\star \boldsymbol{C}_f (\boldsymbol{F}^\star)^{\mathrm{T}}$, donc de $\boldsymbol{C}_c^r$. Deux types de préconditionneurs présentent un intérêt pour notre étude.

Le premier préconditionneur, noté $\boldsymbol{P}_d$, est diagonal par bloc. Les expressions de $\boldsymbol{P}_d$ et de son inverse dont données par [Benzi et al., 2005] :

$$\boldsymbol{P}_d = \begin{pmatrix} \boldsymbol{C}_f^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{S} \end{pmatrix} \quad \text{et} \quad \boldsymbol{P}_d^{-1} = \begin{pmatrix} \boldsymbol{C}_f & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{S}^{-1} \end{pmatrix}. \tag{6.72}$$

En pratique, on calcule le produit

$$\boldsymbol{P}_d^{-1} \boldsymbol{\Xi} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{C}_f^{-1} (\boldsymbol{F}^\star)^{\mathrm{T}} \\ -\boldsymbol{S}\boldsymbol{F}^\star & \boldsymbol{0} \end{pmatrix}, \tag{6.73}$$

ce qui permet d'éviter d'appliquer $\boldsymbol{C}_f$.

Le deuxième préconditionneur, noté $\boldsymbol{P}_t$, est triangulaire par bloc. Les expressions de $\boldsymbol{P}_t$ et de son inverse dont données par [Benzi et al., 2005] :

$$\boldsymbol{P}_t = \begin{pmatrix} \boldsymbol{C}_f^{-1} & (\boldsymbol{F}^\star)^{\mathrm{T}} \\ \boldsymbol{0} & \boldsymbol{S} \end{pmatrix} \tag{6.74}$$

et

$$\boldsymbol{P}_t^{-1} = \begin{pmatrix} \boldsymbol{C}_f & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & (\boldsymbol{F}^\star)^{\mathrm{T}} \\ \boldsymbol{0} & -\boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{S}^{-1} \end{pmatrix}. \tag{6.75}$$

En pratique, il n'y a pas besoin de calculer $\boldsymbol{P}_t^{-1} \boldsymbol{\Xi}$ et l'application de $\boldsymbol{P}_t$ est au final légèrement plus coûteux que l'application de $\boldsymbol{P}_d$.

Lorsque $\boldsymbol{S}$ et exactement égal au complément de Schur, le Gmres appliqué au système préconditionné par $\boldsymbol{P}_d$ converge en exactement trois itérations. Le Gmres appliqué au système préconditionné par $\boldsymbol{P}_t$ converge, lui, en deux itérations au plus [Benzi et al., 2005]. Cependant, les deux stratégies font intervenir $\boldsymbol{S}^{-1}$, qui n'est pas disponible, puisque c'est ce qu'on cherche à déterminer. Il est donc nécessaire de déterminer une approximation $\boldsymbol{S}$ facile à inverser. La sous section 6.2.2 indique qu'on peut faire appel à l'opérateur sans raffinement $\boldsymbol{C}_c^{nr}$.

6.3.2 Performances

En pratique, les expériences montrent que le préconditionnement à droite est beaucoup plus performant que le préconditionnement à gauche pour la résolution, du problème point-selle (6.68) par la méthode du GMRES. En fait, le système préconditionné à gauche ne converge pas dans la plupart de nos cas-tests.

On évalue donc l'effet du préconditionnement à droite. Le préconditionneur triangulaire (6.74) est, comme attendu, légèrement meilleur que le préconditionneur diagonal (6.72). Néanmoins, la vraie amélioration tient à la qualité de l'approximation du complément de Schur. Suivant les conclusions de la sous-section 6.2.2, on envisage les deux approximations $\boldsymbol{S} = \boldsymbol{I}$ et $\boldsymbol{S} = \boldsymbol{C}_c^{nr}$. Les résultats sont sans appel : l'utilisation de $\boldsymbol{S} = \boldsymbol{C}_c^{nr}$ permet de diminuer largement le nombre d'itérations, qu'on utilise le préconditionnement diagonal ou triangulaire. Les chiffres correspondants sont donnés dans le tableau 6.4.

Méthode	Préconditionneur	Complément de Schur	Itérations
GMRES	Identité	$\boldsymbol{I}$	>700
GMRES préconditionné	Diagonal	$\boldsymbol{I}$	544
GMRES préconditionné	Triangulaire	$\boldsymbol{I}$	>700
GMRES préconditionné	Diagonal	$\boldsymbol{C}_c^{nr}$	68
GMRES préconditionné	Triangulaire	$\boldsymbol{C}_c^{nr}$	54

FIGURE 6.4 – Performance des différents préconditionneurs pour l'inversion de l'opérateur de corrélation augmenté. Norme du résidu à convergence : inférieure à $\epsilon = 10^{-6}$.

Globalement, la performance du GMRES pour résoudre le système point-selle n'est pas suffisante (en l'état / en attente d'optimisations futures) pour justifier son utilisation. En effet, une itération du GMRES (préconditionné) appliqué à (6.68) coûte plus cher qu'une itération de la méthode du gradient conjugué (préconditionnée) appliquée à (6.67). De plus, la méthode du

GMRES recquiert plus d'itérations. On choisit donc de se concentrer sur la méthode du gradient conjugué et ses variantes (voir section 6.4).

6.4 Préconditionnement du deuxième ordre

Deux approches du préconditionnement du deuxième ordre sont présentées dans cette section. La première fait usage de la déflation. L'information collectée au cours d'une première résolution du système linéaire est mise à profit pour accélérer la convergence du même système linéaire, s'il doit être résolu une seconde fois. Néanmoins, cette approche peine à réutiliser cette information pour résoudre un système linéaire différent (en changeant le second membre, par exemple). Pour y remédier, on introduit donc les préconditionneurs à mémoire limitée, qui mettent à jour le préconditionneur au fur et à mesure des itérations. Les deux approches bénéficient grandement de la disponibilité d'un bon préconditionneur du premier ordre, comme présenté en section 6.2.

6.4.1 Déflation

La résolution de $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ par les méthodes de Krylov souffre de la présence de petites valeurs propres dans le spectre de $\boldsymbol{A}$. Considérons l'espace W généré par les vecteurs propres correspondants. L'idée de la déflation est de chercher la solution dans l'orthogonale de W pour le produit scalaire associé à $\boldsymbol{A}$. On s'affranchit ainsi de la difficulté posée par la présence des petites valeurs propres.

Soient $(\boldsymbol{w}_i)_{i\in[\![1,l]\!]}$ un ensemble de vecteurs linéairement indépendants contenus dans une matrice $\boldsymbol{W}$. On pose

$$\mathcal{K}_k(\boldsymbol{A}, \boldsymbol{W}, \boldsymbol{r}^{(0)}) = \text{vect}\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_l, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}, \tag{6.76}$$

où les vecteurs $(\boldsymbol{v}_j)_{j\in[\![1,k]\!]}$ vérifient

$$\|\boldsymbol{v}_j\|_2 = 1 \quad \text{et} \quad \boldsymbol{v}_j \perp \mathcal{K}_{j-1}(\boldsymbol{A}, \boldsymbol{W}, \boldsymbol{r}^{(0)}). \tag{6.77}$$

Utiliser une méthode de Krylov avec déflation [Saad et al., 1999, Gaul et al., 2018], c'est rechercher l'approximation $\boldsymbol{x}^{(k)}$ de la solution exacte $\boldsymbol{x}$ dans l'espace $\boldsymbol{x}^{(0)} + \mathcal{K}_k(\boldsymbol{A}, \boldsymbol{W}, \boldsymbol{r}^{(0)})$ en imposant la condition de Petrov-Galerkin

$$\boldsymbol{r}^{(k)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \perp \mathcal{K}_{k-1}(\boldsymbol{A}, \boldsymbol{W}, \boldsymbol{r}^{(0)}). \tag{6.78}$$

Notamment, cela entraîne que le résidu initial $\boldsymbol{r}^{(0)}$ doit lui-même se trouver dans l'orthogonale de $\boldsymbol{W}$. C'est possible en posant

$$\boldsymbol{x}^{(0)} = \boldsymbol{W}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{b} \tag{6.79}$$

$$\Leftrightarrow \quad \boldsymbol{r}^{(0)} = [\boldsymbol{I} - \boldsymbol{A}\boldsymbol{W}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}]\boldsymbol{b}. \tag{6.80}$$

Dans la suite, on note $\boldsymbol{Q}$ la projection oblique sur W définie par

$$\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{A}\boldsymbol{W}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}. \tag{6.81}$$

Dans notre application, on s'intéresse à l'apport de la déflation dans la méthode du gradient conjugué. On reprend donc la méthodologie de la sous-section 6.1.3 en observant que $\boldsymbol{x}^{(k)}$ s'écrit maintenant sous la forme

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(0)} + \boldsymbol{V}_k\boldsymbol{z}^{(k)} + \boldsymbol{W}\boldsymbol{\eta}^{(k)}, \tag{6.82}$$

ce qui implique immédiatement que

$$\boldsymbol{r}^{(k)} = \boldsymbol{r}^{(0)} - \boldsymbol{A}\boldsymbol{V}_k\boldsymbol{z}^{(k)} - \boldsymbol{A}\boldsymbol{W}\boldsymbol{\eta}^{(k)}. \tag{6.83}$$

Comme $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{r}^{(0)} = \boldsymbol{0}$ et $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{r}^{(k)} = \boldsymbol{0}$, la multiplication de (6.83) par $\boldsymbol{W}^{\mathrm{T}}$ conduit à la relation

$$\boldsymbol{\eta}^{(k)} = -(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V}_k\boldsymbol{z}^{(k)}. \tag{6.84}$$

Grâce à (6.81) et (6.83), on en déduit que

$$\boldsymbol{r}^{(k)} = \boldsymbol{r}^{(0)} - \boldsymbol{Q}\boldsymbol{A}\boldsymbol{V}_k\boldsymbol{z}^{(k)}, \tag{6.85}$$

ou de manière équivalente,

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(0)} + \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{V}_k\boldsymbol{z}^{(k)}, \tag{6.86}$$

En réécrivant la relation de Petrov-Galerkin (6.78) sous la forme $\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{r}^{(k)} = \boldsymbol{0}$ et en multipliant (6.85) par $\boldsymbol{V}_k^{\mathrm{T}}$, on obtient que $\boldsymbol{z}^{(k)}$ est solution du système linéaire

$$(\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{A}\boldsymbol{V}_k)\boldsymbol{z}^{(k)} = \boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{r}^{(0)}, \tag{6.87}$$

qui est similaire à l'équation (6.26) des méthodes de Krylov orthogonale. On remarque que le produit $\boldsymbol{Q}\boldsymbol{A}$ est symétrique et que les matrices $\mathbf{A}$ et $\boldsymbol{Q}$ vérifient

$$\boldsymbol{Q}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{Q}^{\mathrm{T}} = \boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^{\mathrm{T}}. \tag{6.88}$$

Ainsi, pour construire l'algorithme du gradient conjugué avec déflation [Saad et al., 1999], il suffit d'appliquer l'algorithme du gradient conjugué standard au système

$$\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{Q}\boldsymbol{b}. \tag{6.89}$$

Dans cette formulation, $\boldsymbol{Q}$ s'interprête comme un préconditionneur.

Remarque : le système (6.89) est symétrique, semi-défini positif seulement. Toutefois, la méthode du gradient conjugué parvient quand même à résoudre le système tant que celui-ci est consistant, c'est à dire tant que le second membre ne se trouve pas dans le noyau de la matrice.

En pratique, la procédure nécessite de construire la matrice

$$\boldsymbol{Q}' = \boldsymbol{W}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}. \tag{6.90}$$

Pour ce faire, on résout une première fois le système linéaire $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ par une méthode de Lanczos (en tronquant éventuellement les itérations), et on récupère les paires de Ritz $(\theta_i, \boldsymbol{w}_i)$ qui approximent les valeurs propres et les vecteurs propres de $\boldsymbol{A}$. Les vecteurs $\boldsymbol{w}_i$ sont stockés dans la matrice $\boldsymbol{W}$ qui vérifie alors

$$\boldsymbol{W}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{W} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{\Theta} = \boldsymbol{\Theta}, \tag{6.91}$$

où $\boldsymbol{\Theta}$ est la matrice diagonale des valeurs propres θ_i. L'inversion de $\boldsymbol{W}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{W}$ ne pose donc pas de difficulté technique.

Présentons maintenant l'utilité pratique de la déflation.

En l'absense de préconditionneur du premier ordre, l'intérêt de la déflation est limité (569 itérations au lieu de 602). Par contre, la déflation donne de bons résultats lorsqu'elle est couplée à l'utilisation d'une préconditionneur du premier ordre. En effet, la déflation utilisant N paires de Ritz permet de réduire le nombre d'itérations de la même valeur N en moyenne. Ce résultat est illustré dans le tableau 6.5.

Algorithme du gradient conjugué avec déflation et préconditionneur P

- Evaluation de $\boldsymbol{W}$, $\boldsymbol{Q}$ et $\boldsymbol{Q}'$ en amont
- Initialisation

  $\boldsymbol{x}^{(0)} = \boldsymbol{Q}'\boldsymbol{b}$

  $\boldsymbol{r}^{(0)} = \boldsymbol{Q}\boldsymbol{b}$

  $\boldsymbol{z}^{(0)} = \boldsymbol{P}^{-1}\boldsymbol{r}^{(0)}$

  $\boldsymbol{v}^{(0)} = \boldsymbol{z}^{(0)}$
- Calcul de l'itération numéro $j+1$

  $\alpha_j = (\boldsymbol{r}^{(j)\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{r}^{(j)})/(\boldsymbol{v}^{(j)\mathrm{T}}\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{v}^{(j)})$

  $\boldsymbol{x}^{(j+1)} = \boldsymbol{x}^{(j)} + \alpha_j\boldsymbol{v}^{(j)}$

  $\boldsymbol{r}^{(j+1)} = \boldsymbol{r}^{(j)} - \alpha_j\boldsymbol{Q}\boldsymbol{A}\boldsymbol{v}^{(j)}$

  Si $\|\boldsymbol{r}^{(j+1)}\|_2/\|\boldsymbol{b}\|_2 < \varepsilon$, alors fin de la procédure

  $\beta_j = (\boldsymbol{r}^{(j+1)\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{r}^{(j+1)})/(\boldsymbol{r}^{(j)\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{r}^{(j)})$

  $\boldsymbol{v}^{(j+1)} = \boldsymbol{P}^{-1}\boldsymbol{r}^{(j+1)} + \beta_j\boldsymbol{v}^{(j)}$
- Calcul de la solution finale (itération k)

  $\hat{\boldsymbol{x}} = \boldsymbol{Q}'\boldsymbol{b} + \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{x}^{(k)}$

Méthode	Préconditionneur de premier ordre	Nombre de paires de Ritz	Itérations
Gradient Conjugué	$\boldsymbol{I}$	0	602
Gradient Conjugué avec déflation	$\boldsymbol{I}$	30	469
Gradient Conjugué préconditionné	$\boldsymbol{C}_c^{nr}$	0	38
Gradient Conjugué préconditionné avec déflation	$\boldsymbol{C}_c^{nr}$	30	6

Figure 6.5 – Réduction du nombre d'itérations par l'utilisation de la déflation construite à partir des vecteurs de Ritz. La déflation couplée à l'utilisation d'un préconditionneur du premier ordre permet de diviser par 100 le nombre d'itérations requises pour l'inversion de l'opérateur de corrélation augmenté. Norme du résidu à convergence : inférieure à $\epsilon = 10^{-6}$.

6.4.2 Mise à jour du préconditionneur

Comme précédemment, on suppose que le système linéaire est résolu une première fois. L'information provenant de cette résolution peut être utilisée pour définir un préconditionneur du deuxième ordre. Pour ce faire, on croise l'information contenue dans les paires de Ritz avec un « bon » préconditionneur du premier ordre. Ce croisement s'inspire des techniques de quasi-Newton, habituellement utilisées en optimisation. L'analyse se concentre sur la méthode du BFGS [Nocedal and Wright, 2006].

Soit $\boldsymbol{A}$ une matrice symétrique définie positive. On rappelle tout d'abord l'expression de la fonctionnelle quadratique $\Phi(\boldsymbol{x})$ définie par :

$$\Phi(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}. \tag{6.92}$$

Tout comme la méthode du gradient conjugué, la méthode de Newton a pour objet la minimisation de (6.92) selon la variable $\boldsymbol{x}$. Elle peut s'interpréter comme une méthode de point fixe appliquée à la fonction

$$\Psi(\boldsymbol{x}) = \boldsymbol{x} - \nabla^2\Phi(\boldsymbol{x})^{-1}\nabla\Phi(\boldsymbol{x}), \tag{6.93}$$

ce qui donne lieu à la relation de récurrence

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{v}^{(k)}, \tag{6.94}$$

où

$$\boldsymbol{v}^{(k)} = -\nabla^2\Phi(\boldsymbol{x}^{(k)})^{-1}\nabla\Phi(\boldsymbol{x}^{(k)}). \tag{6.95}$$

L'identité $\nabla^2\Phi(\boldsymbol{x})^{-1} = \boldsymbol{A}^{-1}$ rend la méthode de Newton inutilisable en grande dimension. Les méthodes de quasi-Newton consistent à remplacer le terme $\nabla^2\Phi(\boldsymbol{x}^{(k)})$ par une approximation $\boldsymbol{P}_k$ dont l'inverse est facile à calculer [5]. Incidemment, on remarque que l'approximation $\boldsymbol{P}_k = \boldsymbol{I}$ conduit à la méthode de la plus forte pente. Dans la suite, on cherche $\boldsymbol{P}_k$ symétrique définie positive.

La matrice $\boldsymbol{P}_k$ doit vérifier l'équation dite « de la sécante »

$$\boldsymbol{P}_k\boldsymbol{s}^{(k)} = \boldsymbol{y}^{(k)}, \tag{6.97}$$

où

$$\begin{aligned} \boldsymbol{r}^{(k)} &= -\nabla\Phi(\boldsymbol{x}^{(k)}), & (6.98)\\ \boldsymbol{s}^{(k)} &= \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}, & (6.99)\\ \boldsymbol{y}^{(k)} &= \boldsymbol{r}^{(k-1)} - \boldsymbol{r}^{(k)}. & (6.100) \end{aligned}$$

On remarque que ces relations engendrent l'égalité $\boldsymbol{y}^{(k)} = \boldsymbol{A}\boldsymbol{s}^{(k)}$. De plus, le caractère symétrique défini positif de $\boldsymbol{P}_k$ entraîne l'inégalité de courbure

$$\boldsymbol{s}^{(k)\mathrm{T}}\boldsymbol{y}^{(k)} = \boldsymbol{s}^{(k)\mathrm{T}}\boldsymbol{A}\boldsymbol{s}^{(k)} > 0. \tag{6.101}$$

Pour que la matrice $\boldsymbol{P}_k$ soit définie de manière unique, il convient d'ajouter une dernière condition. On impose que cette matrice minimise la norme de $\boldsymbol{P}_k - \boldsymbol{P}_{k-1}$ où $\boldsymbol{P}_{k-1}$ est l'approximation de $\nabla^2\Phi(\boldsymbol{x}^{(k-1)})$ calculée à l'itération $k-1$.

La plus populaire des méthodes de quasi-Newton, nommé BFGS d'après ses auteurs (Broyden–Fletcher–Goldfarb–Shanno, voir article de Nocedal and Wright [2006]), repose sur le choix de la norme

$$\|\boldsymbol{B}\| = \|\boldsymbol{H}_k^{-1/2}\boldsymbol{B}\boldsymbol{H}_k^{-1/2}\|_F, \tag{6.102}$$

5. L'approche de quasi-Newton s'interpète donc comme une version instationnaire de la méthode de Richardson. On a en effet :

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{P}_k^{-1}\boldsymbol{r}^{(k)}, \tag{6.96}$$

où $\boldsymbol{P}_k$ est une approximation de la matrice $\boldsymbol{A}$, différente à chaque itération. En pratique, le pas α_k peut-être égal à 1, ou bien calculée par recherche linéaire.

où $\|\cdot\|_F$ désigne la norme de Frobenius et $\boldsymbol{H}$ est la valeur moyenne de la Hessienne définie par

$$\boldsymbol{H}_k = \int_0^1 \nabla^2 \Phi(\boldsymbol{x}^{(k)} + \tau \boldsymbol{v}^{(k)}) \mathrm{d}\tau. \tag{6.103}$$

Dans le cas où Φ est une fonction quadratique, $\boldsymbol{H}_k = \boldsymbol{A}$. Finalement, $\boldsymbol{P}_k$ est solution du problème d'optimisation

$$\begin{aligned} \min_{\boldsymbol{P}_k} \|\boldsymbol{A}^{-1/2}(\boldsymbol{P}_k - \boldsymbol{P}_{k-1})\boldsymbol{A}^{-1/2}\|_F \quad \text{sachant} \quad & \boldsymbol{P}_k \boldsymbol{s}^{(k)} = \boldsymbol{y}^{(k)} \\ \text{et} \quad & \boldsymbol{P}_k = \boldsymbol{P}_k^{\mathrm{T}}. \end{aligned} \tag{6.104}$$

Dès lors que (6.101) est vérifiée, on peut montrer que l'unique solution de (6.104) admet pour expression

$$\boldsymbol{P}_k = \left(\boldsymbol{I} - \frac{\boldsymbol{y}^{(k)}\boldsymbol{s}^{(k)\mathrm{T}}}{\boldsymbol{y}^{(k)\mathrm{T}}\boldsymbol{s}^{(k)}}\right)^{\mathrm{T}} \boldsymbol{P}_{k-1} \left(\boldsymbol{I} - \frac{\boldsymbol{s}^{(k)}\boldsymbol{y}^{(k)\mathrm{T}}}{\boldsymbol{y}^{(k)\mathrm{T}}\boldsymbol{s}^{(k)}}\right) + \frac{\boldsymbol{y}^{(k)}\boldsymbol{y}^{(k)\mathrm{T}}}{\boldsymbol{y}^{(k)\mathrm{T}}\boldsymbol{s}^{(k)}}, \tag{6.105}$$

La formule du préconditionneur à l'itération k s'obtient donc par mise à jour du préconditonneur à l'itération $k-1$, en faisant appel à la dernière solution et au dernier résidu calculés. On peut montrer que si $\boldsymbol{P}_0 = \boldsymbol{I}$, alors l'algorithme du BFGS produit les mêmes itérés que la méthode du gradient conjugué sans préconditionnement [Nazareth, 1979].

En inversant les rôles respectifs de $\boldsymbol{s}^{(k)}$ et $\boldsymbol{y}^{(k)}$ dans (6.97), on établit la formule de mise à jour de l'inverse :

$$\boldsymbol{P}_k^{-1} = \left(\boldsymbol{I} - \frac{\boldsymbol{s}^{(k)}\boldsymbol{y}^{(k)\mathrm{T}}}{\boldsymbol{y}^{(k)\mathrm{T}}\boldsymbol{s}^{(k)}}\right)^{\mathrm{T}} \boldsymbol{P}_{k-1}^{-1} \left(\boldsymbol{I} - \frac{\boldsymbol{y}^{(k)}\boldsymbol{s}^{(k)\mathrm{T}}}{\boldsymbol{y}^{(k)\mathrm{T}}\boldsymbol{s}^{(k)}}\right) + \frac{\boldsymbol{s}^{(k)}\boldsymbol{s}^{(k)\mathrm{T}}}{\boldsymbol{y}^{(k)\mathrm{T}}\boldsymbol{s}^{(k)}}. \tag{6.106}$$

Les préconditionneurs à mémoire limitée s'inspirent de cette formule pour construire $\boldsymbol{P}_k^{-1}$ à partir d'un préconditionneur de premier ordre $\boldsymbol{P}_0^{-1}$ et de k paires $(\boldsymbol{s}^{(k)}, \boldsymbol{y}^{(k)})$.

Supposons que les paires $(\boldsymbol{s}^{(k)}, \boldsymbol{y}^{(k)})$ vérifient l'équation de la sécante (6.97) et sont $\boldsymbol{A}$-conjuguées, c'est-à-dire que

$$\boldsymbol{s}^{(k)\mathrm{T}}\boldsymbol{y}^{(k)} = \boldsymbol{s}^{(k)\mathrm{T}}\boldsymbol{A}\boldsymbol{s}^{(k)} = \delta_{ij}, \tag{6.107}$$

la notation δ_{ij} faisant référence au symbole de Kronecker. Alors, une récurrence immédiate permet d'établir que

$$\begin{aligned} \boldsymbol{P}_k^{-1} &= \left(\boldsymbol{I} - \sum_{i=1}^{k} \frac{\boldsymbol{s}^{(i)}\boldsymbol{s}^{(i)\mathrm{T}}}{\boldsymbol{s}^{(i)\mathrm{T}}\boldsymbol{A}\boldsymbol{s}^{(i)}}\boldsymbol{A}\right)^{\mathrm{T}} \boldsymbol{P}_0^{-1} \left(\boldsymbol{I} - \sum_{i=1}^{k} \boldsymbol{A}\frac{\boldsymbol{s}^{(i)}\boldsymbol{s}^{(i)\mathrm{T}}}{\boldsymbol{s}^{(i)\mathrm{T}}\boldsymbol{A}\boldsymbol{s}^{(i)}}\right) \\ &+ \sum_{i=1}^{k} \frac{\boldsymbol{s}^{(i)}\boldsymbol{s}^{(i)\mathrm{T}}}{\boldsymbol{s}^{(i)\mathrm{T}}\boldsymbol{A}\boldsymbol{s}^{(i)}}. \end{aligned} \tag{6.108}$$

En notation matricielle, la formule (6.108) se réécrit

$$\begin{aligned} \boldsymbol{P}_k^{-1} &= [\boldsymbol{I} - \boldsymbol{S}(\boldsymbol{S}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{S})^{-1}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{A}]^{\mathrm{T}}\boldsymbol{P}_0^{-1}[\boldsymbol{I} - \boldsymbol{A}\boldsymbol{S}(\boldsymbol{S}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{S})^{-1}\boldsymbol{S}^{\mathrm{T}}] \\ &+ \boldsymbol{S}(\boldsymbol{S}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{S})^{-1}\boldsymbol{S}^{\mathrm{T}}, \end{aligned} \tag{6.109}$$

où $\boldsymbol{S}$ est la matrice dont les colonnes sont les $(\boldsymbol{s}^{(i)})_{i\in[\![1,k]\!]}$. Cette matrice $\boldsymbol{P}_k^{-1}$ est nommée préconditionneur à mémoire limitée (ou LMP pour *Limited Memory Preconditioner* en anglais). L'approche fut mise au point par Morales and Nocedal [2000] avant d'être reprise sous son nom moderne, par exemple par Gratton et al. [2011]. Remarque : le caractère diagonal de $\boldsymbol{S}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{S}$ résulte de la relation de conjugaison (6.107). Son inversion est donc triviale.

La méthode de Ritz-LMP consiste à définir la matrice $\boldsymbol{S}$ à partir des vecteurs de Ritz de $\boldsymbol{A}$, puis à injecter cette définition dans la formule (6.109). On obtient ainsi un préconditionneur du deuxième ordre symétrique défini positif, qu'on peut utiliser pour accélérer la convergence d'une méthode itérative comme le gradient conjugué ou les itérations de Chebyshev. En particulier, le recours à un préconditionneur à mémoire limitée équivaut à la déflation sous certaines conditions [Nabben and Vuik, 2006].

Les performances du LMP sont données dans le tableau 6.6.

6.4.3 Sensibilité au second membre

Précédemment, on a montré l'intérêt du préconditionnement du deuxième ordre pour accélérer la convergence de la méthode du gradient conjugué. Toutefois, les paires de Ritz utilisées pour définir ce précondionneur du deuxième ordre sont issues d'un premier système linéaire $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Le tableau 6.7 montre que cette information ne peut pas nécessairement être réutilisée pour résoudre un système linéaire $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}'$ dont seul le second membre diffère.

Voilà qui réduit l'utilitée de la déflation et du LMP. Toutefois, il n'est pas exclu que des optimisations des méthodes ou des préconditionneurs viennent

Méthode	Préconditionneur de premier ordre	Nombre de paires de Ritz	Itérations
Gradient Conjugué préconditionné	$\boldsymbol{C}_c^{nr}$	0	38
Gradient Conjugué avec LMP	$\boldsymbol{C}_c^{nr}$	30	17

FIGURE 6.6 – Performance du LMP pour l'inversion de l'opérateur de corrélation augmenté. Norme du résidu à convergence : inférieure à $\epsilon = 10^{-6}$. On remarque que dans nos expériences, l'utilisation du LMP ne permet pas d'améliorer le nombre d'itérations par rapport au simple préconditionnement du premier ordre.

changer ces résultats dans le futur. Notamment, il est envisageable d'étudier la sensibilité au second membre en définissant $\boldsymbol{b}' = \boldsymbol{b} + \delta\boldsymbol{b}$, où $\|\delta\boldsymbol{b}\|_2 \ll \|\boldsymbol{b}\|_2$.

Equation	Méthode	Type de préconditionneur	Itérations
$\boldsymbol{Ax} = \boldsymbol{b}$ ou $\boldsymbol{Ax} = \boldsymbol{b}'$	Gradient Conjugué	Premier ordre	38
$\boldsymbol{Ax} = \boldsymbol{b}$	Gradient Conjugué	Déflation	6
$\boldsymbol{Ax} = \boldsymbol{b}'$	Gradient Conjugué	Déflation	28
$\boldsymbol{Ax} = \boldsymbol{b}$	Gradient Conjugué	LMP	17
$\boldsymbol{Ax} = \boldsymbol{b}'$	Gradient Conjugué	LMP	30

FIGURE 6.7 – Performance du préconditionnement du deuxième ordre quand on change le second membre de l'équation.

6.5 Troncature des itérations

On achève ce chapitre en donnant quelques remarques sur l'arrêt des algorithmes itératifs avant convergence totale (*i.e.* résidu égal au zéro de la machine).

D'un point de vue algébrique, un algorithme qui résout le système linéaire $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ est une fonction linéaire du second membre $\boldsymbol{b}$, qui renvoie l'image $\boldsymbol{x}$ par l'application $\boldsymbol{A}^{-1}$. Cette propriété découle de la linéarité de la mutliplication matricielle par $\boldsymbol{A}^{-1}$, qui s'écrit

$$(\boldsymbol{x}_1 = \boldsymbol{A}^{-1}\boldsymbol{b}_1) \wedge (\boldsymbol{x}_2 = \boldsymbol{A}^{-1}\boldsymbol{b}_2) \tag{6.110}$$

$$\Rightarrow \quad (\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2) = \boldsymbol{A}^{-1}(\alpha_1\boldsymbol{b}_1 + \alpha_2\boldsymbol{b}_2), \tag{6.111}$$

cela pour tout couple de scalaires $(\alpha_1, \alpha_2) \in \mathbb{R}^2$.

Dans le cas des algorithmes itératifs, on peut se demander si la troncature précoce des itérations vient contrarier, ou non, cette linéarité. Considérons l'algorithme du gradient conjugué et les itérations de Chebyshev, qui sont deux méhodes itératives qu'on souhaite comparer.

Indirectement, la méthode du gradient conjugué fait intervenir le second membre $\boldsymbol{b}$ pour calculer les coefficients α_j et β_j à chaque itération. La relation entre les directions de descente successives $\boldsymbol{v}^{(j)}$ et $\boldsymbol{v}^{(j+1)}$ n'est donc pas linéaire. Une analyse rapide montre que, par conséquent, l'itéré $\boldsymbol{x}^{(k)}$ ne dépend lui-même pas linéairement de $\boldsymbol{b}$, à moins qu'il soit exactement égal à la solution exacte $\boldsymbol{x}$. Pour cette raison, on qualifie l'algorithme du gradient conjugué de méthode « non linéaire ».

Les itérations de Chebyshev, en revanche, dépendent linéairement du second membre $\boldsymbol{b}$. On peut donc parler de méthode « linéaire » et étudier les propriétés associées.

On peut montrer que cet algorithme est symétrique à condition que la matrice $\boldsymbol{A}$ le soit, puisqu'il correspond à la multiplication par un polynôme en $\boldsymbol{A}$. C'est une propriété très désirable en assimilation de données, où l'on cherche à définir des opérateurs symétriques (les matrices de covariance et leurs inverses), tout en cherchant à limiter le nombre d'itérations (afin de réduire les coûts d'application des opérateurs). Ceci reste vrai lorsqu'on utilise un préconditionneur. On peut ainsi utiliser un préconditionneur du deuxième ordre pour préconditionner les itérations de Chebyshev. La plupart du temps, on ne profite pas de l'absence de produit scalaires, puisque l'expression préconditionneur en contient, mais on profite toujours de la linéarité lors de la troncature des itérations.

Enfin, on peut écrire le code adjoint des itérations de Chebyshev. Il est utile lorsqu'on cherche à inverser une matrice $\boldsymbol{A}$ donc on dispose d'une factorisation de Cholesky.

6.6 L'essentiel du chapitre

Le raffinement de maillage permet de modéliser les opérateurs de corrélation en éléments finis tout en minimisant les sources d'erreurs numériques, comme la dépendance au maillage et la mauvaise répartition des degrés de liberté. Toutefois, l'opérateur de corrélation ainsi augmenté n'admet plus d'inverse trivial. Or, l'accès à l'inverse de l'opérateur de corrélation d'erreurs d'observation est une condition nécessaire dans les formulation populaires de l'assimilation de données.

Dans ce chapitre, on a étudié le réalisme du raffinement de maillage en quantifiant la facilité d'accès à l'inverse par une gamme d'algorithmes itératifs. Après une présentation brèves de ces algorithmes, les expériences ont montré que l'opérateur de corrélation augmenté pouvait être inversé en un petit nombre d'itérations, à condition de disposer d'un bon préconditionneur de premier ordre. Au final, il ressort que l'opérateur de corrélation non augmenté possède les propriétés spectral suffisantes pour être un bon préconditionneur du premier ordre.

Ce qu'il faut retenir :

- L'algorithme du gradient conjugué préconditionné par l'opérateur de corrélation sans raffinement est le plus performant pour inverser l'opérateur de corrélation augmenté.
- Cette performance peut s'améliorer en exploitant la déflation, stratégie de préconditionnement du deuxième ordre.
- La déflation est construite à partir des paires de Ritz d'un résolution antécédente.
- Cependant, cette information du deuxième ordre ne permet pas d'accélérer la résolution du même système linéaire lorsqu'on en change le second membre.
- Losqu'on tronque les itérations, les itérations de Chebyshev forment une alternative viable à la méthode du gradient conjugué.

Chapitre 7

Raffinement et stratégies alternatives

La définition de l'opérateur de corrélation augmenté s'appuie sur la résolution de l'équation de diffusion sur un maillage fin et le recours à des opérateurs de transfert depuis et vers le maillage des observations. Cette approche a l'avantage de rendre la modélisation des corrélations précise et indépendante de la distribution spatiale des observations. Toutefois, l'inversion de l'opérateur de corrélation augmenté requiert l'emploi de méthodes itératives, soumises à la disponibilité d'un préconditionneur. Par contraste avec l'approche initiale, qui consiste simplement à résoudre l'équation de diffusion discrétisée en éléments finis sur le maillage des observations, le raffinement de maillage apporte donc un certain nombre de contraintes.

Néanmoins, d'autres stratégies de modélisation permettent d'exploiter efficacement le formalisme associé au raffinement de maillage. Dans la section 7.1, on discute de la possibilité de « scinder » l'opérateur de corrélation en deux facteurs indépendamment inversibles. L'idée est de résoudre deux systèmes linéaires au lieu d'un seul, qui sont chacun mieux conditionnés que le système initial. Dans la section 7.2, on montre que le raffinement de maillage permet de redéfinir les matrices de masse et de raideur. En résulte une modélisation de l'opérateur de corrélation relativement facile à inverser, dont les performances indiquent un compromis entre l'usage du raffinement de maillage et l'approche sans raffinement. Enfin, l'étude se termine sur la section 7.3, qui envisage la sommation de plusieurs opérateurs de corrélation. Cette stratégie est motivée par les résultats de la sous-section 4.2.3.

Les résultats de ce chapitre sont majoritairement nouveaux. En outre,

ils ouvrent la porte à de nombreux développements théoriques et pratiques dont l'étude approfondie s'inscrit dans les perspectives de ce travail de thèse. Néanmoins, les expériences proposées permettent déjà de constater l'utilité de certaines approches, tout en indiquant les aspects qui devront être examinés en priorité.

7.1 Scission des opérateurs

Dans l'équation (5.27), l'opérateur de corrélation augmenté est construit « d'un bloc ». En effet, les opérateurs de transfert interviennent avant et après la résolution de l'équation de diffusion sur le maillage fin. Ici, on décide de faire appel aux opérateurs de transfert une fois supplémentaire quand on atteint la moitié des itérations de l'opérateur de diffusion. C'est possible si le paramètre de régularité m est pair, ce qu'on suppose dans la suite. On obtient donc une factorisation de l'opérateur de corrélation en matrices carrées, ce qui possède des avantages pour l'inversion.

7.1.1 Opérateur scindé

Le diagramme de dualité est l'outil qui permet de formaliser le lien entre un opérateur de corrélation et les applications qui le composent. Cela tient au fait que l'action d'un opérateur discret s'interprète comme un chemin dans le diagramme, qui suit les arêtes et le sens des flèches du graphe. Il est donc naturel de revisiter ce diagramme, en cherchant d'autres chemins, qui correspondent à des modélisations alternatives de l'opérateur de corrélation augmenté.

Considérons le chemin représenté sur le diagramme de la figure (7.1). Cette fois-ci, au lieu de résoudre les m itérations de la diffusion au cours d'un même cycle, on décide de revenir de $\mathbb{R}_q$ vers l'espace des observations $\mathbb{R}^p$ au bout de $m/2$ itérations, de telle sorte que l'opérateur de corrélation $\boldsymbol{C}$ s'écrive

$$\boldsymbol{C} = (\boldsymbol{F}^\star \boldsymbol{D}_f^{1/2} \boldsymbol{F})^2 \boldsymbol{M}_c^{-1}, \tag{7.1}$$

où $\boldsymbol{D}_f^{1/2}$ désigne l'opérateur de diffusion agissant sur $m/2$ pas de temps

$$\boldsymbol{D}_f^{1/2} = [(\boldsymbol{M}_f + \boldsymbol{K}_f)^{-1} \boldsymbol{M}_f]^{m/2}. \tag{7.2}$$

Conformément aux résultats de la section 5.3, on choisit de définir $\boldsymbol{F}^\star$ par

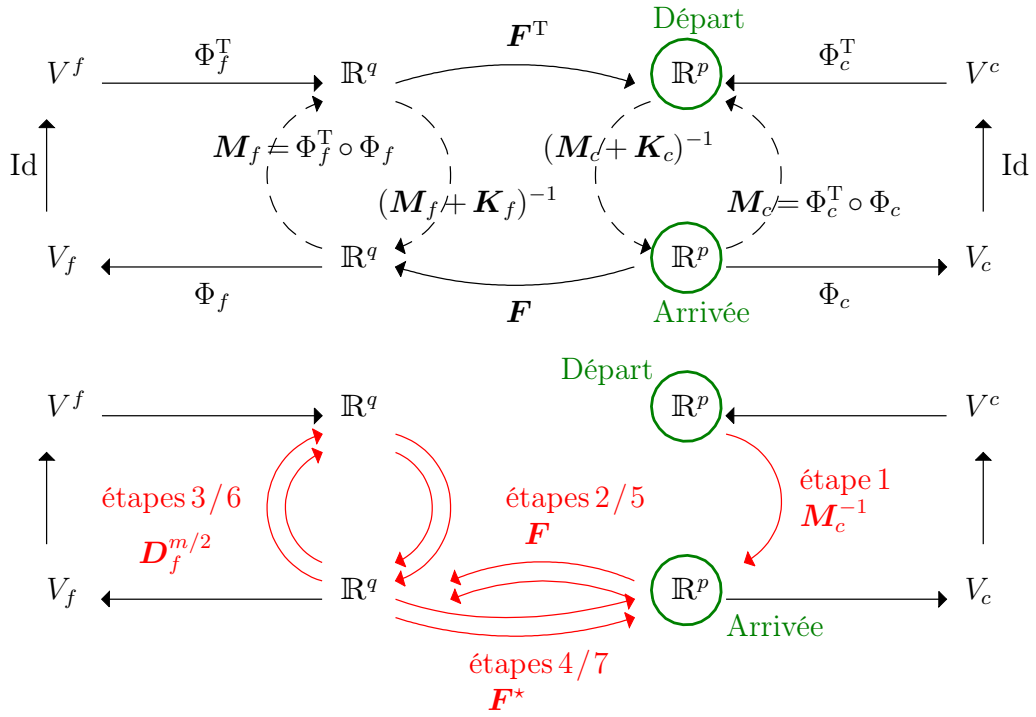


Figure 7.1 – Diagrammes de dualité. Le premier diagramme représente les applications, leurs espaces de départ et d'arrivée. Le deuxième donne le sens de lecture (en rouge) auquel correspond l'opérateur de corrélation à deux niveaux (7.1). Son application se compose de 7 étapes, les étapes 2/3/4 étant identiques aux étapes 5/6/7.

injection. En remplaçant $\boldsymbol{F}$ par son expression (5.23), on obtient que

$$\boldsymbol{C} = [\boldsymbol{F}^\star \boldsymbol{D}_f^{1/2} \boldsymbol{M}_f^{-1} (\boldsymbol{F}^\star)^{\mathrm{T}}] \boldsymbol{M}_c [\boldsymbol{F}^\star \boldsymbol{D}_f^{1/2} \boldsymbol{M}_f^{-1} (\boldsymbol{F}^\star)^{\mathrm{T}}]. \tag{7.3}$$

La figure 7.2 permettent de constater les erreurs d'amplitude associées à la formulation (7.3).

On constate que les erreurs d'amplitude sont localement très élevées, jusqu'à atteindre 50% sur des zones entières. Ce phénomène est dû à la présence du facteur $\boldsymbol{F}\boldsymbol{F}^\star$ au « milieu » de l'équation (7.3). Ce facteur correspond à un projecteur de taille $q \times q$ et de rang $p \ll q$ qui vient dégrader la qualité de l'approximation. Malheureusement, le rang de $\boldsymbol{F}\boldsymbol{F}^\star$ est indépendant du niveau de raffinement, empêchant toute convergence lorsque la taille des éléments diminue.

Toutefois, ce résultat très négatif est à relativiser en fonction de la valeur

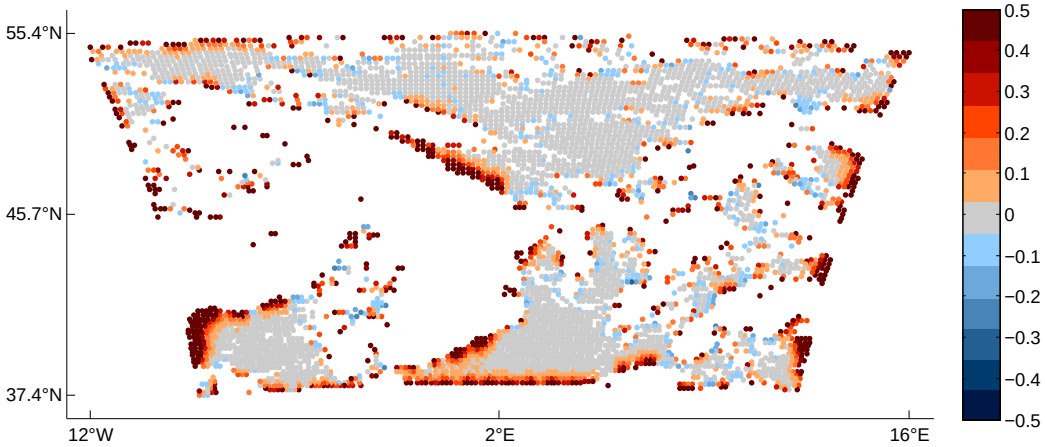


FIGURE 7.2 – Carte des erreurs d'amplitude associées à l'opérateur de corrélation scindé. L'erreur est bien plus importante que sans raffinement.

du paramètre m. En effet, si l'on résout la diffusion sur m pas de temps, et que m est suffisamment grand (typiquement supérieur à 8 dans nos expériences), l'approximation (7.3) donne des résultats excellents, quasiment aussi bons que le schéma de raffinement non scindé. Même si ces valeurs ne correspondent pas à notre cas d'étude, cette remarque est susceptible d'être utile dans d'autres applications.

7.1.2 Inversion et troncature

L'avantage de la formulation (7.3) est que $\boldsymbol{C}$ admet maintenant une factorisation sous la forme

$$\boldsymbol{C} = \boldsymbol{U}\boldsymbol{M}_c\boldsymbol{U}, \tag{7.4}$$

où $\boldsymbol{U}$ est la matrice carrée, symétrique définie positive d'expression

$$\boldsymbol{U} = \boldsymbol{F}^\star \boldsymbol{D}_f^{1/2} \boldsymbol{M}_f^{-1} (\boldsymbol{F}^\star)^{\mathrm{T}}. \tag{7.5}$$

Cela signifie que $\boldsymbol{C}$ peut être inversée facteur par facteur. On a donc

$$\boldsymbol{C}^{-1} = \boldsymbol{U}^{-1}\boldsymbol{M}_c^{-1}\boldsymbol{U}^{-1}. \tag{7.6}$$

Numériquement, le conditionnement de $\boldsymbol{U}$ est *a priori* plus petit que celui de $\boldsymbol{C}$, ce qui facilite l'inversion grâce à une méthode itérative. De façon similaire au chapitre 6, le système linéaire $\boldsymbol{U}\boldsymbol{x} = \boldsymbol{b}$ peut être préconditionné par $(\boldsymbol{C}_c^{nr})^{1/2}$, l'opérateur de diffusion construit sans raffinement sur $m/2$ pas de temps :

$$(\boldsymbol{C}_c^{nr})^{1/2} = [(\boldsymbol{M}_c + \boldsymbol{K}_c)^{-1}\boldsymbol{M}_c]^{m/2}\boldsymbol{M}_c^{-1}. \tag{7.7}$$

Si on utilise une méthode linéaire, comme les itérations de Chebyshev, on peut faire appel à son adjoint pour calculer $\boldsymbol{U}^{-\mathrm{T}}$. Dans ce cas, on force la symétrie de $\boldsymbol{C}^{-1}$ en posant $\boldsymbol{C}^{-1} = \boldsymbol{U}^{-1}\boldsymbol{M}_c^{-1}\boldsymbol{U}^{-\mathrm{T}}$.

Cette méthode a donc l'avantage de faciliter l'inversion de $\boldsymbol{C}$ en découpant le problème en plusieurs sous-problèmes mieux conditionnés.

7.2 Redéfinition des matrices de masse et de raideur

Précédemment, la recherche d'un espace adapté à la résolution en éléments finis de l'équation de diffusion a conduit à l'introduction d'un maillage secondaire, plus fin et contenant moins d'éléments malformés que le maillage des observations. Les matrices d'éléments finis sont construites et l'équation de diffusion est résolue sur le maillage fin, puis les résultats sont ramenés dans l'espace des observations à l'aide d'opérateurs de transfert.

Néanmoins, l'utilisation d'un maillage secondaire peut être utile à bien d'autres égards. Dans cette section, on étudie la possibilité de définir la matrice de masse $\boldsymbol{M}_c$ et la matrice de raideur $\boldsymbol{K}_c$ à partir du maillage fin, tout en résolvant l'équation de diffusion uniquement dans l'espace des observations. Cette approche est suggérée par plusieurs facteurs. Si la construction susmentionnée de $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$ découle naturellement d'une redéfinition des opérateurs de transferts, elle peut également s'interpréter comme une prolongation de l'interpolation de la masse ou une lecture exotique du diagramme de dualité.

On montre que l'opérateur de corrélation ainsi construit est facile à inverser, et offre un compromis des approches sans et avec raffinement proposées précédemment.

7.2.1 Critère d'interpolation

L'opérateur de transfert $\boldsymbol{F}$, lorsqu'il correspond à une interpolation linéaire, vérifie toujours $\boldsymbol{F}^\star\boldsymbol{F} = \boldsymbol{I}_c$ (sous-section 5.2.1). La seule condition est que l'espace grossier soit inclus dans l'espace fin, c'est-à-dire $V_c \subset V_f$. Dans ce cas, la matrice de masse $\boldsymbol{M}_c$ est reliée à $\boldsymbol{M}_f$ par la relation

$$\boldsymbol{M}_c = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{M}_f\boldsymbol{F}. \tag{7.8}$$

Peut-on obtenir une relation similaire à (7.8), qui relie les matrices de raideur $\boldsymbol{K}_c$ et $\boldsymbol{K}_f$?

La réponse est positive. Pour le voir, reprenons les notations de la sous-section 5.2.1 et considérons la fonctionnelle faisant intervenir la norme H^1 (et non la norme L^2 !) :

$$\mathcal{J}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \|\sum_{i=1}^{p}\alpha^i\varphi_i - \sum_{j=1}^{q}\beta^j\psi_j\|^2_{H^1}. \tag{7.9}$$

Matriciellement, cette fonctionnelle s'écrit

$$\mathcal{J}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \boldsymbol{\alpha}^{\mathrm{T}}(\boldsymbol{M}_c+\boldsymbol{K}_c)\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^{\mathrm{T}}\dot{\boldsymbol{P}}_c^f\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{M}_f+\boldsymbol{K}_f)\boldsymbol{\beta}, \tag{7.10}$$

où $\dot{\boldsymbol{P}}_c^f$ est la matrice de taille $p\times q$ de terme général

$$\begin{aligned}(\dot{\boldsymbol{P}}_c^f)_{\mathrm{ij}} &= \langle\varphi_i,\psi_j\rangle_{H^1}\\ &= \int_{\mathcal{D}}\varphi_i(\boldsymbol{z}')\psi_j(\boldsymbol{z}')\mathrm{d}\boldsymbol{z}' + \int_{\mathcal{D}}\nabla\varphi_i(\boldsymbol{z}')\cdot\nabla\psi_j(\boldsymbol{z}')\mathrm{d}\boldsymbol{z}'.\end{aligned} \tag{7.11}$$

La minimisation de $\mathcal{J}(\boldsymbol{\alpha},\boldsymbol{\beta})$ selon $\boldsymbol{\alpha}$ ou $\boldsymbol{\beta}$ permet de définir les opérateurs de transfert

$$\boldsymbol{F} = (\boldsymbol{M}_f+\boldsymbol{K}_f)^{-1}(\dot{\boldsymbol{P}}_c^f)^{\mathrm{T}} \quad\text{et}\quad \boldsymbol{F}^\star = (\boldsymbol{M}_c+\boldsymbol{K}_c)^{-1}\dot{\boldsymbol{P}}_c^f. \tag{7.12}$$

Encore une fois, on peut vérifier que la condition $V_c\subset V_f$ entraîne $\boldsymbol{F}^\star\boldsymbol{F}=\boldsymbol{I}_c$, puis

$$(\boldsymbol{M}_c+\boldsymbol{K}_c) = \boldsymbol{F}^{\mathrm{T}}(\boldsymbol{M}_f+\boldsymbol{K}_f)\boldsymbol{F}. \tag{7.13}$$

Dans (7.8) comme dans (7.13), $\boldsymbol{F}$ correspond à une interpolation linéaire. En effet, supposons que la base des $(\varphi_i)_{i\in[\![1,p]\!]}$ soit incluse dans la base des $(\psi_j)_{j\in[\![1,q]\!]}$. Alors pour minimiser (7.9), il suffit de choisir $\beta_j=\alpha_i$ lorsque $\psi_j=\varphi_i$ et $\beta_j=0$ sinon. Remarque : lorsque les espaces sont inclus, mais pas les bases, remplacer $\boldsymbol{F}$ par une interpolation linéaire commet une légère approximation.

Stratégie : modélisation de l'opérateur de corrélation

Supposons maintenant qu'il n'existe pas de relation de hiérarchie entre les maillage grossier et fin. On choisit d'imposer les relations (7.8) et (7.13), si bien qu'on ne construit plus les matrices $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$ à partir du maillage grossier. Conséquence immédiate : il n'est même plus nécessaire de construire le maillage grossier. Ce dernier est défini implicitement à partir du maillage fin, de la méthode d'interpolation qui définit $\boldsymbol{F}$ et des relations (7.8) et (7.13).

Ensuite, on modélise l'opérateur de corrélation comme dans le cas sans raffinement, en posant

$$\begin{aligned} \boldsymbol{C} &= [(\boldsymbol{M}_c + \boldsymbol{K}_c)^{-1}\boldsymbol{M}_c]^m \boldsymbol{M}_c^{-1} \\ &= [(\boldsymbol{F}^{\mathrm{T}}(\boldsymbol{M}_f + \boldsymbol{K}_f)\boldsymbol{F})^{-1}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{M}_f\boldsymbol{F})]^m (\boldsymbol{F}^{\mathrm{T}}\boldsymbol{M}_f\boldsymbol{F})^{-1}. \end{aligned} \tag{7.14}$$

On obtient ainsi un opérateur de corrélation relativement facile à inverser (les seuls inversions ont lieu dans l'espace des observations, qui profite néanmoins de l'existence d'un maillage fin. La formule de l'inverse est donnée par

$$\begin{aligned} \boldsymbol{C}^{-1} &= \boldsymbol{M}_c[\boldsymbol{M}_c^{-1}(\boldsymbol{M}_c + \boldsymbol{K}_c)]^m \\ &= (\boldsymbol{F}^{\mathrm{T}}\boldsymbol{M}_f\boldsymbol{F})[(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{M}_f\boldsymbol{F})^{-1}(\boldsymbol{F}^{\mathrm{T}}(\boldsymbol{M}_f + \boldsymbol{K}_f)\boldsymbol{F})]^m. \end{aligned} \tag{7.15}$$

Cette nouvelle démarche peut se deviner à partir du diagramme de dualité de la figure 7.3. En effet, le chemin bleu représente l'application $\boldsymbol{M}_c$ (resp. $\boldsymbol{M}_c + \boldsymbol{K}_c$), qui permet de « remonter » de l'espace primal vers l'espace dual. Le chemin rouge est un autre chemin qui partage le même espace de départ et le même espace d'arrivée que le chemin bleu, et qui correspond à la composition $\boldsymbol{F}^{\mathrm{T}}\boldsymbol{M}_f\boldsymbol{F}$ (resp. $\boldsymbol{F}^{\mathrm{T}}(\boldsymbol{M}_f + \boldsymbol{K}_f)\boldsymbol{F}$). Imposer les relations (7.8) et (7.13) revient à substituer systématiquement le chemin bleu par le chemin rouge.

7.2.2 Performances

On valide l'expression (7.14) en étudiant l'erreur par rapport à la solution analytique. On constate sur les figures 7.4 et 7.5 que les erreurs sont inférieures à celle du schéma sans raffinement (environ 20% d'erreurs en moins). La qualité de l'approximation est donc meilleure. Néanmoins, les erreurs restent visibles et on n'atteint pas la précision d'un opérateur totalement raffiné comme proposé dans le chapitre 5.

Ce résultat indique que le raffinement des matrices de masse et de raideur est une piste à suivre dans le futur. Cependant, la convergence d'un tel schéma (lorsque le raffinement s'intensifie) est encore mal comprise. Sa compréhension nécessite donc des développements théoriques plus poussés.

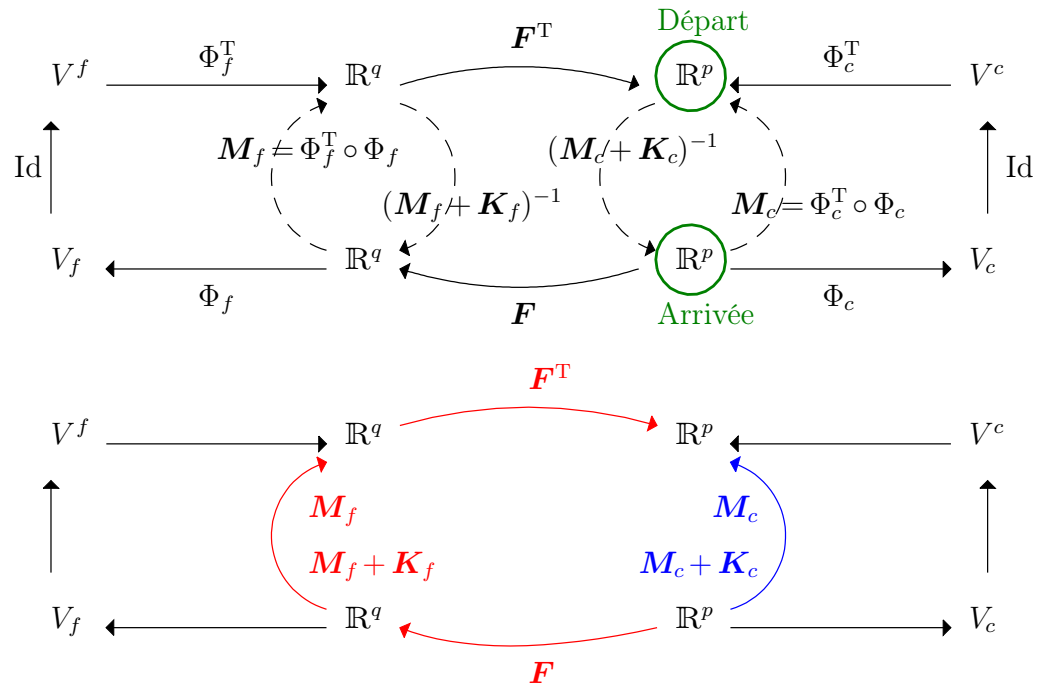


Figure 7.3 – Diagrammes de dualité. Le premier diagramme représente les applications, leurs espaces de départ et d'arrivée. Le deuxième représente deux chemins dans le diagramme. Habituellement, $\boldsymbol{M}_c$ et $\boldsymbol{M}_c + \boldsymbol{K}_c$ sont définis selon le chemin bleu. Ici, on se propose de les définir selon le chemin rouge, qui fait appel aux opérateurs du maillage fin.

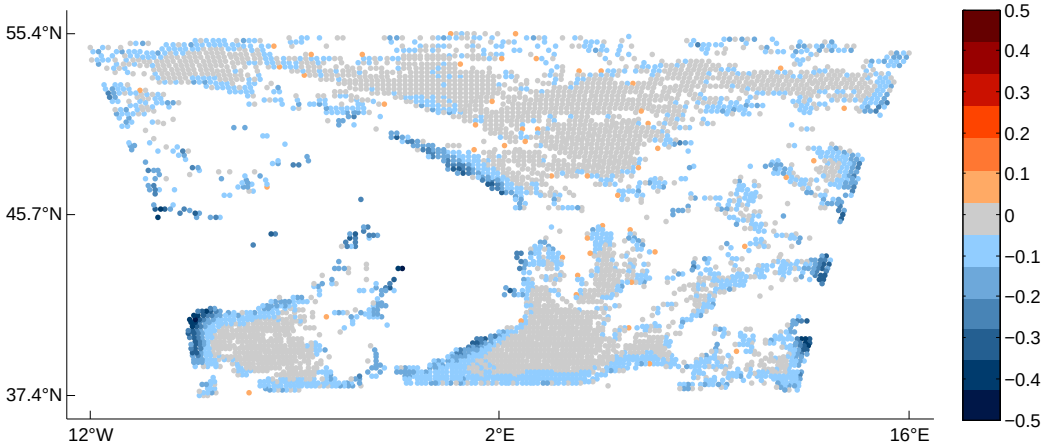


Figure 7.4 – Carte des erreurs d'amplitude associées à la l'opérateur de corrélation (7.14).

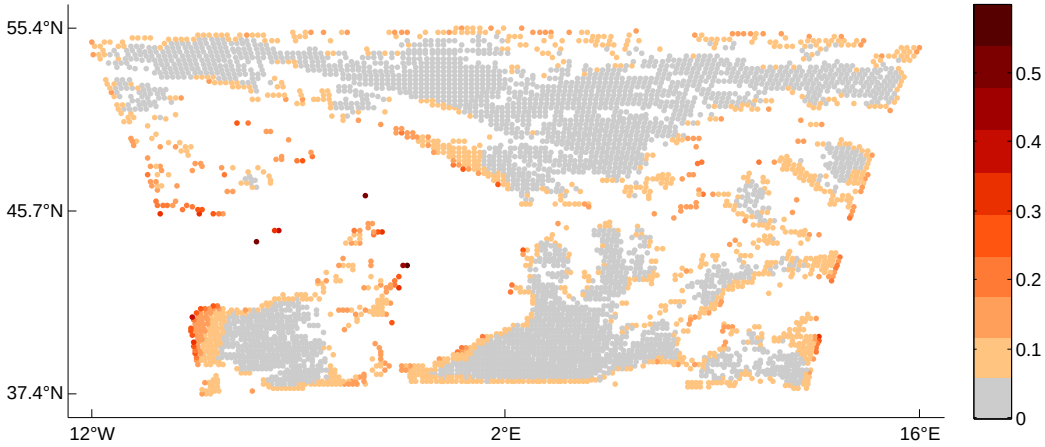


FIGURE 7.5 – Carte des erreurs de forme associées à la l'opérateur de corrélation (7.14).

7.2.3 Déraffinement et perspectives

L'interpolation des matrices de masse et de raideur à partir d'un maillage auxiliaire offre de nombreuses possibilités d'applications et de développements. Cette approche, au même titre que l'interpolation de la masse, rejoint la classe des méthodes qui consistent à optimiser le contenu des matrices $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$ pour l'application considérée. Il ne s'agit plus de définir les matrices d'éléments finis et de les utiliser « au mieux ». Il s'agit, au contraire, de déterminer la modélisation qui soit la plus robuste et la plus flexible de ces matrices, pour en faire une utilisation simple.

Suivant cette logique, on est tenté, au lieu d'utiliser le raffinement de maillage, d'introduire un maillage auxiliaire plus petit, obtenu par déraffinement de maillage. Les calculs y sont logiquement plus faciles à mener. Toutefois, d'autres aspects entrent en considération, notamment la défficience de rang en cas d'inversion. Dans cette sous-section, on discute qualitativement des stratégies envisageables qui définissent $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$ à partir d'un maillage auxiliaire, et on donne quelques pistes de recherche à suivre.

Supposons qu'on dispose d'un maillage auxiliaire, éventuellement égal au maillage des observations. Soient $\boldsymbol{M}_\pi$ et $\boldsymbol{K}_\pi$ les matrices de masse et de raideur discrétisées sur ce maillage. On propose de définir la matrice de raideur comme

$$\boldsymbol{K}_c = \boldsymbol{F}^{\mathrm{T}} \boldsymbol{K}_\pi \boldsymbol{F} \tag{7.16}$$

et la matrice de masse comme

$$\boldsymbol{M}_c = \alpha\tilde{\boldsymbol{M}}_c + (1-\alpha)\boldsymbol{F}^{\mathrm{T}}\boldsymbol{M}_\pi\boldsymbol{F}, \tag{7.17}$$

où le terme $\alpha\tilde{\boldsymbol{M}}_c$ est un terme de régularisation permettant d'assurer que $\boldsymbol{M}_c$ est inversible ($\alpha \in [0,1]$). Notons q la dimension du maillage auxiliaire. On distingue trois cas d'étude selon la valeur de q.

Cas n°1 : $q = p$

Supposons que le maillage auxiliaire soit exactement le maillage des observations. Lorsque $\alpha = 0$, on peut montrer que $\boldsymbol{K}_\pi = \boldsymbol{K}_c$ et que $\boldsymbol{M}_\pi = \boldsymbol{M}_c$. On se retrouve alors dans le cadre des éléments finis $\mathbb{P}_1$ standards. A l'inverse, si $\alpha = 1$, alors (7.17) se réécrit $\boldsymbol{M}_c = \tilde{\boldsymbol{M}}_c$. Le choix $\tilde{\boldsymbol{M}}_c = \boldsymbol{M}_c^{\#}$ conduit dans ce cas à la condensation de masse décrite dans le chapitre 3.

Lorsque $\alpha \in]0,1[$, alors l'équation (7.17) s'interprète comme un compromis entre l'utilisation de la masse condensée et l'utilisation de la masse consistante.

Cas n°2 : $q > p$

Dans le cas où $\boldsymbol{M}_\pi = \boldsymbol{M}_f$ et $\boldsymbol{K}_\pi = \boldsymbol{K}_f$ sont obtenus à partir d'un maillage fin, alors $\boldsymbol{F}^{\mathrm{T}}\boldsymbol{M}_\pi\boldsymbol{F}$ est inversible et on peut considérer $\alpha = 0$. On retrouve alors le cadre de la sous-section 7.2.1.

Cas n°3 : $q < p$

L'intérêt des définitions (7.17) et (7.16) est révélé lorsque le maillage auxiliaire résulte du déraffinement du maillage des observations. Dans ce cas, $\boldsymbol{F}^{\mathrm{T}}\boldsymbol{M}_\pi\boldsymbol{F}$ n'est pas inversible et il est nécessaire d'imposer $\alpha \neq 0$ afin que $\boldsymbol{M}_c$ et $\boldsymbol{M}_c + \boldsymbol{K}_c$ soient inversibles. Deux options sont envisageables dans ce contexte : $\tilde{\boldsymbol{M}}_c = \boldsymbol{I}$ ou $\tilde{\boldsymbol{M}}_c = \boldsymbol{M}_c^{\#}$.

La validation de nouveau schéma fait partie des études à mener dans le futur. En effet, il offre la possibilité d'exploiter un maillage déraffiné de bonne qualité pour définir les matrices $\boldsymbol{M}_c$ et $\boldsymbol{K}_c$, tout en proposant une formule explicite pour l'inverse de $\boldsymbol{C}$, puisque $\boldsymbol{C}$ et $\boldsymbol{C}^{-1}$ restent définis comme dans le chapitre 2.

7.3 Somme d'opérateurs

Les diagnostics de la sous-section 4.2.3 ont montré que pour décrire au mieux les corrélations réelles contenues dans les données de SEVIRI, il faut écrire l'opérateur de corrélation comme une somme d'opérateurs. Dans cette section, on donne quelques pistes d'inversion de ces opérateurs, faisant notamment usage de la formule de Sherman-Morrison-Woodbury. On précise que cet axe de recherche ne fait pas l'objet d'un développement poussé dans ce manuscrit. Il fait partie des travaux d'investigation à envisager dans le futur.

7.3.1 Inversion sur le maillage initial

Soit l'opérateur de corrélation

$$\boldsymbol{C}_r = \boldsymbol{C}_1 + \boldsymbol{C}_2, \tag{7.18}$$

où $\boldsymbol{C}_1$ et $\boldsymbol{C}_2$ désignent deux opérateurs de corrélations symétriques définis positifs. L'objectif est de calculer l'inverse de $\boldsymbol{C}_r$. Rappel : le nombre de colonnes de $\boldsymbol{C}_r$ est égal au nombre d'observations dans le jeu de données.

Les expériences de la sous-section 4.2.3 montrent que plusieurs couples de matrices $\boldsymbol{C}_1$ et $\boldsymbol{C}_2$ sont de bons candidats pour décrire les corrélations contenues dans les données de SEVIRI. Notamment, la figure 4.7 suggère de poser

$$\begin{aligned} \boldsymbol{C}_1 &= \alpha \boldsymbol{I} \\ \boldsymbol{C}_2 &= (1-\alpha)(\boldsymbol{M}_c + \boldsymbol{K}_c)^{-1} \boldsymbol{M}_c (\boldsymbol{M}_c + \boldsymbol{K}_c)^{-1}, \end{aligned} \tag{7.19}$$

alors que la figure 4.8 suggère le choix

$$\begin{aligned} \boldsymbol{C}_1 &= \alpha (\boldsymbol{M}_c + \boldsymbol{K}_c)^{-1} \boldsymbol{M}_c (\boldsymbol{M}_c + \boldsymbol{K}_c)^{-1} \\ \boldsymbol{C}_2 &= (1-\alpha)(\boldsymbol{M}_c + \boldsymbol{K}'_c)^{-1} \boldsymbol{M}_c (\boldsymbol{M}_c + \boldsymbol{K}'_c)^{-1}, \end{aligned} \tag{7.20}$$

où $\boldsymbol{K}_c$ et $\boldsymbol{K}'_c$ sont deux matrices de raideurs contenant des longueurs de portée différentes.

Comme toujours, il n'est pas envisageable de calculer explicitement les coefficients de $\boldsymbol{C}_r$ lorsque la taille du jeu de données devient important. Néanmoins, il est possible d'avoir une factorisation de Cholesky de $\boldsymbol{M}_c$ et $(\boldsymbol{M}_c + \boldsymbol{K}_c)$, de telle sorte que $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, $\boldsymbol{C}_1^{-1}$ et $\boldsymbol{C}_2^{-1}$ soient aisément calculables. Cependant, même sous ces conditions, il n'existe pas de formule simple pour

calculer l'inverse de la somme (7.18). L'inversion de $\boldsymbol{C}_r$ recquiert donc un effort calculatoire.

Quitte à fournir un effort pour inverser $\boldsymbol{C}_1+\boldsymbol{C}_2$, autant utiliser un raffinement de maillage. En effet, la modélisation de $\boldsymbol{C}_r$ comme une somme d'opérateurs s'inscrit dans une démarche de recherche de précision, en regard des expériences de la sous-section 4.2.3. Il est donc inconvenant d'introduire des erreurs numériques dues à discrétisation de la diffusion sur des maillages non structurés. L'inversion de $\boldsymbol{C}_r$ rejoint alors le cadre d'utilisation des méthodes itératives décrites dans les sections 6.1 à 6.4. La factorisation

$$\boldsymbol{F}^\star\boldsymbol{D}_{f,1}\boldsymbol{F}\boldsymbol{M}_c^{-1}+\boldsymbol{F}^\star\boldsymbol{D}_{f,2}\boldsymbol{F}\boldsymbol{M}_c^{-1}=\boldsymbol{F}^\star(\boldsymbol{D}_{f,1}+\boldsymbol{D}_{f,2})\boldsymbol{F}\boldsymbol{M}_c^{-1}. \tag{7.21}$$

offre une piste pour définir une préconditionneur de premier ordre. Ce travail fait partie des perspectives d'étude.

D'autre part, l'inverse de $\boldsymbol{C}_r$ peut être calculée par la formule de Sherman-Morrison-Woodbury. Sous réserve que $\boldsymbol{C}_1^{-1}+\boldsymbol{C}_2^{-1}$ est lui-même inversible et après simplification, on trouve que

$$\boldsymbol{C}_r^{-1}=\boldsymbol{C}_2^{-1}(\boldsymbol{C}_1^{-1}+\boldsymbol{C}_2^{-1})^{-1}\boldsymbol{C}_1^{-1}. \tag{7.22}$$

Ainsi, pour inverser (7.18), deux stratégies sont possibles. La première consiste à résoudre le système linéaire associé à $\boldsymbol{C}_1+\boldsymbol{C}_2$. La deuxième stratégie consiste à exploiter l'expression (7.22), de laquelle naît le système linéaire alternatif associé à $\boldsymbol{C}_1^{-1}+\boldsymbol{C}_2^{-1}$.

En pratique, le choix $\boldsymbol{C}_1=\boldsymbol{I}$ mène à l'expression

$$\boldsymbol{C}_r^{-1}=\boldsymbol{C}_2^{-1}(\boldsymbol{I}+\boldsymbol{C}_2^{-1})^{-1}. \tag{7.23}$$

L'expérience montre que la méthode du gradient conjugué appliquée au système linéaire

$$(\boldsymbol{I}+\boldsymbol{C}_2)\boldsymbol{x}=\boldsymbol{b} \tag{7.24}$$

converge en trois fois moins d'itérations ($\backsim$ 50 vs $\backsim$ 175 pour SEVIRI) que la même méthode appliquée au système linéaire (de même taille)

$$(\boldsymbol{I}^{-1}+\boldsymbol{C}_2^{-1})\boldsymbol{x}=\boldsymbol{b}. \tag{7.25}$$

A moins de trouver une méthode de préconditionnement efficace, on écarte donc la possibilité d'inverser $(\boldsymbol{I}+\boldsymbol{C}_2)$ par la formule (7.22).

7.3.2 Inversion en présence de raffinement

On repart de l'équation (7.18) en supposant cette fois que

$$\begin{aligned} \boldsymbol{C}_1 &= \alpha \boldsymbol{I} \\ \boldsymbol{C}_2 &= (1-\alpha)\boldsymbol{F}^\star \boldsymbol{D}_f \boldsymbol{F} \boldsymbol{M}_c^{-1}. \end{aligned} \tag{7.26}$$

La formule de Sherman-Morrison-Woodbury s'écrit

$$(\boldsymbol{A}+\boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{C}^{-1}+\boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}. \tag{7.27}$$

En posant

$$\begin{aligned} \boldsymbol{A} &= \alpha \boldsymbol{I} \\ \boldsymbol{U} &= \boldsymbol{F}^\star \\ \boldsymbol{V} &= \boldsymbol{M}_f \boldsymbol{F} \boldsymbol{M}_c^{-1} \\ \boldsymbol{C} &= (1-\alpha)\boldsymbol{C}_f, \end{aligned} \tag{7.28}$$

on trouve que

$$(\alpha \boldsymbol{I} + (1-\alpha)\boldsymbol{C}_2)^{-1} \tag{7.29}$$

$$= (\alpha \boldsymbol{I} + (1-\alpha)\boldsymbol{F}^\star \boldsymbol{C}_f \boldsymbol{M}_f \boldsymbol{F} \boldsymbol{M}_c^{-1})^{-1}$$

$$= \tfrac{1}{\alpha}\boldsymbol{I} - \tfrac{1}{\alpha^2}\boldsymbol{F}^\star \left(\tfrac{1}{1-\alpha}\boldsymbol{C}_f^{-1} + \tfrac{1}{\alpha}\boldsymbol{M}_f \boldsymbol{F} \boldsymbol{M}_c^{-1} \boldsymbol{F}^\star\right)^{-1} \boldsymbol{M}_f \boldsymbol{F} \boldsymbol{M}_c^{-1}. \tag{7.30}$$

On remarque que si $\boldsymbol{F}^\star$ définit un transfert par injection (sous-section 5.2.2), l'expression se simplifie en utilisant l'égalité (5.23) :

$$(\alpha \boldsymbol{I} + (1-\alpha)\boldsymbol{C}_2)^{-1} = \frac{1}{\alpha}\boldsymbol{I} - \frac{1}{\alpha^2}\boldsymbol{F}^\star \left(\frac{1}{1-\alpha}\boldsymbol{C}_f^{-1} + \frac{1}{\alpha}\boldsymbol{J}\right)^{-1} \boldsymbol{M}_f \boldsymbol{F} \boldsymbol{M}_c^{-1}, \tag{7.31}$$

où

$$\boldsymbol{J} = \begin{array}{|c|c|} \hline \boldsymbol{I}_c & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{0} \\ \hline \end{array}. \tag{7.32}$$

Ainsi, la formule de Sherman-Morrison-Woodbury permet d'inverser $\boldsymbol{C}_r$ en cherchant la solution du système linéaire

$$\left(\frac{1}{1-\alpha}\boldsymbol{C}_f^{-1} + \frac{1}{\alpha}\boldsymbol{J}\right)\boldsymbol{x} = \boldsymbol{b}. \tag{7.33}$$

Toutefois, la taille de ce système est égale au nombre de noeuds du maillage fin. Considérant le fait que le système est également difficile à résoudre (même principe que dans la sous-section 7.3.1), on rejette cette méthode d'inversion.

7.3.3 Inversion en présence de déraffinement

Précédemment, on a exploité la formule de Sherman-Morrison-Woodbury pour inverser $\boldsymbol{C}_r = \boldsymbol{C}_1 + \boldsymbol{C}_2$. En résultait un second système linéaire plus compliqué à résoudre, en raison de sa taille ou de sa complexité. Toutefois, on peut affaiblir les hypothèses sur $\boldsymbol{C}_2$ pour faire apparaître un système linéaire de taille plus petite, et donc plus facile à résoudre.

Supposons qu'au lieu de modéliser $\boldsymbol{C}_2$ par raffinement de maillage, on cherche au contraire à recourrir au déraffinement (sous-section 5.1.2). Alors on peut montrer que $\boldsymbol{C}_2$ s'écrit

$$\boldsymbol{C}_2 = (1-\alpha)\boldsymbol{F}^\star \boldsymbol{D}_s \boldsymbol{F} \boldsymbol{M}_c^{-1}, \tag{7.34}$$

où $\boldsymbol{F}$ désigne cette fois l'opérateur de transfert du maillage des observations vers le maillage surgrossier. On note respectivement $\boldsymbol{M}_s$, $\boldsymbol{D}_s$ et $\boldsymbol{C}_s$ la matrice de masse, l'opérateur de diffusion et l'opérateur de corrélation définis sur le maillage surgrossier. Alors $\boldsymbol{C}_s$ et $\boldsymbol{C}_r$ sont inversibles, même si $\boldsymbol{C}_2$ ne l'est pas forcément.

En posant

$$\begin{aligned} \boldsymbol{A} &= \alpha \boldsymbol{I} \\ \boldsymbol{U} &= \boldsymbol{F}^\star \\ \boldsymbol{V} &= \boldsymbol{M}_s \boldsymbol{F} \boldsymbol{M}_c^{-1} \\ \boldsymbol{C} &= (1-\alpha)\boldsymbol{C}_s, \end{aligned} \tag{7.35}$$

la formule de Sherman-Morrison-Woodbury permet d'écrire :

$$(\alpha \boldsymbol{I} + (1-\alpha)\boldsymbol{C}_2)^{-1} \tag{7.36}$$

$$= (\alpha \boldsymbol{I} + (1-\alpha)\boldsymbol{F}^\star \boldsymbol{C}_s \boldsymbol{M}_s \boldsymbol{F} \boldsymbol{M}_c^{-1})^{-1}$$

$$= \tfrac{1}{\alpha}\boldsymbol{I} - \tfrac{1}{\alpha^2}\boldsymbol{F}^\star \left(\tfrac{1}{1-\alpha}\boldsymbol{C}_s^{-1} + \tfrac{1}{\alpha}\boldsymbol{M}_s \boldsymbol{F} \boldsymbol{M}_c^{-1} \boldsymbol{F}^\star \right)^{-1} \boldsymbol{M}_s \boldsymbol{F} \boldsymbol{M}_c^{-1}. \tag{7.37}$$

En définissant $\boldsymbol{F}^\star$ comme une interpolation linéaire du maillage surgrossier vers le maillage des observations, on arrive à établir que

$$\frac{1}{1-\alpha}\boldsymbol{C}_s^{-1} + \frac{1}{\alpha}\boldsymbol{M}_s \boldsymbol{F} \boldsymbol{M}_c^{-1} \boldsymbol{F}^\star = \frac{1}{1-\alpha}\boldsymbol{C}_s^{-1} + \frac{1}{\alpha}\boldsymbol{I}. \tag{7.38}$$

On se retrouve donc à résoudre le système linéaire

$$\left(\frac{1}{1-\alpha}\boldsymbol{C}_s^{-1} + \frac{1}{\alpha}\boldsymbol{I} \right) \boldsymbol{x} = \boldsymbol{b}, \tag{7.39}$$

qui est cette fois de taille plus petite que le système linéaire initial associé à $\boldsymbol{C}_r$. Si cette taille est suffisamment faible, il n'est pas exclu de calculer explicitement les coefficients de la matrice pour l'inverser directement.

Qualitativement, puisque $\boldsymbol{F}^\star$ est une interpolation linéaire, les fonctions de corrélation contenues dans $\boldsymbol{C}_2$ s'obtiennent par interpolation linéaire des fonctions de corrélation contenues dans $\boldsymbol{C}_s$. Ces fonctions sont représentées sur la figure 7.6.

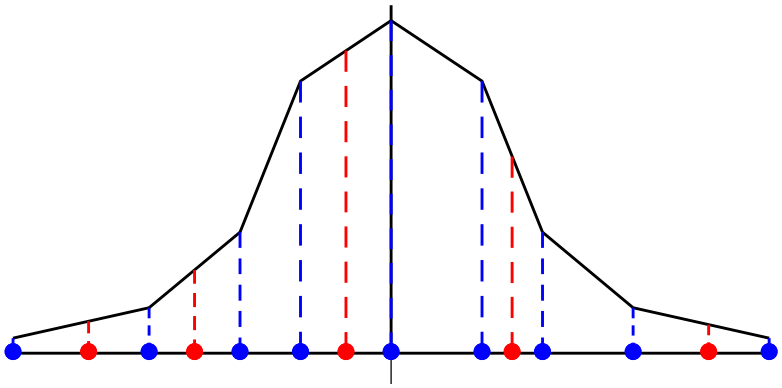


FIGURE 7.6 – Interpolation linéaire d'une fonction de corrélation en dimension 1. Tous les points sont des observations. Seuls les point bleus appartiennent au maillage surgrossier. Les valeurs correspondant aux points rouges sont calculées par interpolation linéaire.

Enfin, on termine par une analyse spectrale qualitative de la régularisation.

Posons $\boldsymbol{C}_2 = (1-\alpha)\boldsymbol{C}$. Soit $(\lambda, \boldsymbol{v})$ une valeur propre de $\boldsymbol{C}$ et son vecteur propre associé. On a

$$\boldsymbol{C}_r\boldsymbol{v} = (\alpha\boldsymbol{I} + (1-\alpha)\boldsymbol{C})\boldsymbol{v} = \alpha\boldsymbol{v} + (1-\alpha)\lambda\boldsymbol{v} = (\alpha + \lambda - \alpha\lambda)\boldsymbol{v}. \qquad (7.40)$$

Ainsi, $(\alpha+\lambda-\alpha\lambda)$ est la valeur propre de $\boldsymbol{C}_r$ associée à $\boldsymbol{v}$. Les valeurs propres de $\boldsymbol{C}_r$ sont donc supérieures ou égales à celle de $\boldsymbol{C}$. La figure 7.7 représente les spectres respectifs de $\boldsymbol{C}_r$ et de $\boldsymbol{C}$. On y visualise aussi la comparaison entre les spectres de $\boldsymbol{C}_r^{-1}$ et de $\boldsymbol{C}^{-1}$. On voit qu'en ajoutant le terme de

régularisation, on excite beaucoup moins les hautes fréquences qu'avec la simple diffusion.

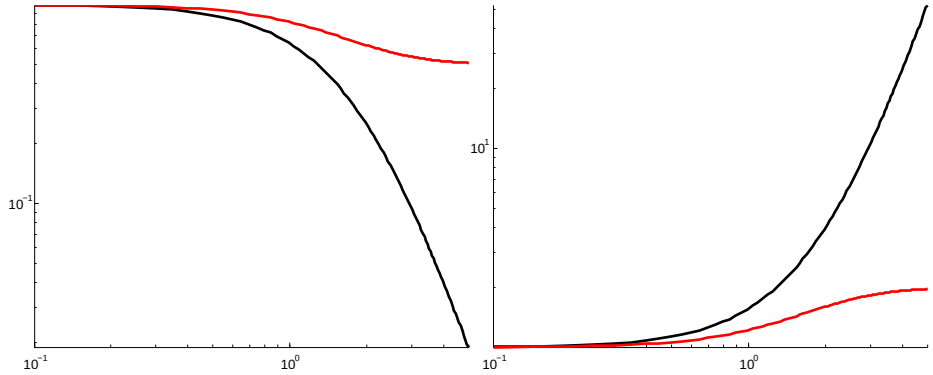


Figure 7.7 – Gauche : spectre de Matérn ($m = 2$, $l = 0.5$) en noir et spectre régularisé en rouge. Droite : spectres des inverses correspondant à la figure de gauche. La valeur de α est estimée dans la sous-section 4.2.3.

7.4 L'essentiel du chapitre

Grâce au diagramme de dualité, nous avons construit plusieurs modèles de $\boldsymbol{R}$ alternatifs. Le premier consiste à scinder l'opérateur de corrélation raffiné en deux opérateurs inversibles, construits à partir de l'équation de diffusion agissant sur $m/2$ pas de temps. Nous avons montré que, bien qu'attrayante au premier abord, cette formulation n'était pas satisfaisante en pratique, en raison des erreurs d'amplitude significatives, qui ne justifient pas son coût en calcul élevé. Ensuite, nous avons montré qu'en interpolant les matrices de masse et de raideur, on parvenait à améliorer le modèle de corrélation initial (*i.e.* sans raffinement), tout en conservant sa facilité d'inversibilité. Nous pensons que ce nouveau modèle possède de grands atouts qui justifieront une étude approfondie dans le futur. Enfin, nous avons discuté de la possibilité de modéliser la matrice $\boldsymbol{R}$ comme une somme faisant intervenir un terme de régularisation. Plusieurs formules d'inversion ont été proposées. Le couplage de cette approche avec le déraffinement de maillage semble être prometteur, à condition d'exhiber un critère objectif pour construire le maillage surgrossier.

Bilan et perspectives

Ce dernier chapitre vient clore la présentation du manuscrit. Il est composé de deux sections, dont les contenus respectifs se réfèrent d'une part au bilan de la thèse, et d'autre part aux études envisageables dans le futur.

Dans un premier temps, nous donnons une vue d'ensemble des travaux effectués sur la modélisation des corrélations d'erreurs d'observation. Ce bilan comporte deux parties. Tout d'abord, nous montrons comment les résultats théoriques et pratiques permettent de répondre aux objectifs fixés en début de thèse. Puis, un résumé de la méthodologie employée met en évidence les apports de la thèse qui viennent compléter ces objectifs initiaux.

Dans un second temps, un commentaire de la thèse vient mettre en perspective les différents résultats présentés dans le manuscrit vis-à-vis de l'état actuel des travaux en assimilation de données. Nous commençons par discuter des futurs développements à envisager en fonctions des préocupations scientifiques et opérationnelles. Enfin, nous évaluons de l'enjeu posé par la prise en compte des corrélations d'erreurs d'observation dans un contexte opérationnel.

8.1 Bilan de l'étude

Commençons par dresser un bilan des réponses scientifiques apportées aux différents défis identifiés en amont et au cours de la thèse. Cette section contient une revue des problématiques abordées au cours de l'étude et des méthodes construites pour y répondre.

8.1.1 Réponses aux objectifs principaux

Compte-tenu de la rareté des travaux concernant la modélisation (et même l'estimation) des corrélations d'erreurs d'observation, la première mission était de « dépoussiérer » le sujet et d'évaluer son intérêt pour la communauté météorologique et océanograhique. Les travaux de Michel [2018] et Waller et al. [2016a] ont permis de cibler une catégorie de données dont les erreurs sont spatialement corrélées : les observations des satellites géostationnaires, et notamment les données du sondeur infrarouge Seviri, disponibles grâce à Météo-France. En outre, ces observations sont assimilées quotidiennement dans les modèles de prévision opérationnels Arome et Arpege.

Partant de ces données et du constat de leur structure géographique, les objectifs de la thèse peuvent alors se formuler ainsi : proposer un modèle d'opérateur de corrélation $\boldsymbol{R}$ tel que l'inverse $\boldsymbol{R}^{-1}$ soit d'un coût numérique raisonnable, en mettant l'accent sur la robustesse de la modélisation vis-à-vis de la distribution spatiale des observations et la facilité de sa paramétrisation.

La stratégie visant à répondre à ces objectifs s'inspire de la littérature portant sur la modélisation des corrélations d'erreurs d'ébauche, en particulier Mirouze and Weaver [2010]. La question de l'accès à l'inverse fait de l'équation de diffusion (implicite) un candidat de choix pour modéliser la matrice $\boldsymbol{R}^{-1}$ sous la forme d'un opérateur différentiel (chapitre 1). L'avantage de cette formulation est double puisqu'il permet de s'orienter vers des schémas de discrétisation adaptés au maillages non structurés, comme ceux décrits par les observations satellitaires. Parmis ces options, la méthode des éléments finis s'impose comme une des approches les plus robustes pour notre application (chapitre 2).

En suivant cette stratégie, nous parvenons à modéliser des opérateurs $\boldsymbol{R}$ et $\boldsymbol{R}^{-1}$:

- **Adaptatifs**. En effet, la combinaison de la méthode des éléments finis avec des procédés de raffinement de maillage (partie III) permet de prendre en compte la distribution spatiale des observations de manière robuste.
- **Flexibles**. Le modèle de corrélation proposé dépend d'un petit nombre de paramètres : la variance, la longueur de portée et le paramètre de régularité. Il s'adapte également au cas où l'on dispose d'un tenseur de corrélation hétérogène dans l'espace, cela sans effort théorique supplémentaire.
- **Performants**. L'utilisation d'équations aux dérivées partielles est tout à fait approprié au traitement de systèmes de grande taille, qui apparaissent lorsque le nombre d'observations est très grand.

La réalisation des expériences à partir de données opérationnelles (Seviri) a permis de valider les opérateurs dans un contexte réaliste. Tout comme dans Michel [2018], nous avons eu recours à des images de résolution supérieure à l'opérationnel, exploitant une densité moyenne d'une observation tous les 12km (au lieu d'une tous les 70km, comme dans Arome), anticipant ainsi la réduction du *thinning* suite à la prise en compte des corrélations. Pour ce faire, les paramètres du modèle de corrélation sont estimés directement

à partir des données de l'imageur. D'autre part, ce contexte réaliste a permis d'estimer le coût de l'inversion de $\boldsymbol{R}$ lorsqu'on utilise des techniques de raffinement de maillage, démontrant incidemment leur faisabilité et la performance des préconditionneurs de second ordre.

8.1.2 Contributions notables

Les contributions de cette thèse à la modélisation des opérateurs de corrélation sont multiples. Un des développements essentiels est la mise au point d'un cadre théorique rassemblant le point de vue des distributions, qui est utile à l'étude des opérateurs de corrélation, et le cadre hilbertien, qui sert à définir la méthode des éléments finis.

Ce cadre théorique est utile à bien des fins, parmis lesquelles :

- L'établissement d'un lien clair entre les opérateurs de corrélation continus et discrets au travers l'introduction de familles de fonctions de pondération. La définition des espaces et des métriques rencontrées en assimilation de données est directement intégrée à la notion de discrétisation.
- La mise en place d'un outil théorique puissant, le diagramme de dualité, auquel on se réfère tout au long du manuscrit pour sa clarté et ses qualités de synthèse. L'application d'un opérateur de corrélation s'interprète comme un chemin dans le diagramme, qui comprend un nombre de boucles égal au nombre de pas de la diffusion.
- Le développement de toute une classe de méthodes basées sur l'équation de diffusion implicite, répondant aux objectifs posés en début de thèse. Selon la robustesse et la précision recherchées, on redéfinit au choix les matrices d'éléments finis ou bien l'opérateur de diffusion à partir d'un maillage auxiliaire. Ce maillage auxiliaire s'obtient en général par raffinement ou déraffinement du maillage des observations.
- La mise en relation des méthodes de diffusion avec d'autres méthodes, comme celle de Brankart et al. [2009]. L'interprétation de la matrice de raideur comme un produit de gradient donne un nouveau sens à la représentation et l'assimilation des gradients sur des maillages non structurés.

Le modèle de corrélation n'étant rien sans qu'on sache exprimer ses paramètres, une section du chapitre 4 est également dédiée à l'estimation des paramètres de corrélation à l'aide de la méthode de Desroziers et al. [2005].

Une analyse des différents choix de paramétrisation permet de faire ressortir plusieurs échelles de corrélation, dont on discute de la pertinence. Les expériences sont réalisées en considérant une échelle de corrélation intermédiaire, qui correspond à la valeur diagnostiquée dans Waller et al. [2016a].

8.2 Perspectives de recherche

Nous donnons maintenant des pistes d'étude pour poursuivre ce travail de thèse. Après avoir mentionné les axes de développement qui semblent les plus pertinents, nous évaluons rapidement les prochains défis qu'il faudra relever pour une description complète des corrélations spatiales d'erreurs d'observation.

8.2.1 Axes de développements privilégiés

Le travail présenté dans ce manuscrit ne décrit pas une approche de la modélisation des opérateurs de corrélation, mais une classe de méthodes, qui comportent chacune des avantages et des inconvénients, selon l'utilisation qu'on lui réserve. C'est pourquoi nous indiquons quelle méthode développer en priorité en fonction des contraintes pratiques.

Coût numérique :

L'avantage majeur de l'équation de diffusion implicite est la possibilité de modéliser directement $\boldsymbol{R}^{-1}$ à partir des matrices d'éléments finis $\boldsymbol{M}^{-1}$ et $\boldsymbol{M} + \boldsymbol{K}$. Dans cette approche, il suffit de savoir inverser $\boldsymbol{M}$ pour appliquer $\boldsymbol{R}^{-1}$, ce qui est facile puique $\boldsymbol{M}$ est diagonale par bande. De surcroît, le recours à la condensation de masse permet d'effacer cette dernière difficulté, au prix d'une légère perte de précision.

Pour aller plus loin, il convient donc de garder cette propriété d'« inversion facile » en recherchant les matrices $\boldsymbol{M}$ et $\boldsymbol{K}$ qui possèdent les meilleures propriétés pour la modélisation des corrélations. Autrement dit, il ne s'agit pas d'utiliser les matrices d'éléments finis au mieux, mais de faire en sorte que ces matrices n'induisent pas d'erreurs numériques significatives lors de leur utilisation. Ce paradigme fait l'objet de la section 8.2.

Dans la section 8.2, les matrices $\boldsymbol{M}$ et $\boldsymbol{K}$ sont construites à partir d'un maillage auxiliaire et d'opérateurs d'interpolation. Pour la suite, nous conseillons

d'étudier la sensibilité de cette modélisation au maillage auxiliaire et de considérer d'autres opérateurs d'interpolation.

Robustesse :

Supposons maintenant qu'on ne traite que des données extrêmement non structurées et qu'on recherche un modèle de $\boldsymbol{R}$ qui dépende le moins possible de la distribution spatiale des observations. Dans ce cas, le recours au raffinement de maillage, comme décrit dans le chapitre 6, est la solution la plus robuste.

Le modèle de $\boldsymbol{R}$ « augmenté », ainsi nommé car l'espace des observations est complété par l'ajout de points artificiels, se montre aussi précis que l'exige l'application. Toutefois, ce procédé de raffinement de maillage complique l'inversion. Les résultats du chapitre 7 indiquent que cette complication peut être partiellement contournée par l'introduction d'un préconditionneur du deuxième ordre.

La recherche de préconditionneurs est un sujet vaste dont le chapitre 7 ne fait qu'effleurer la surface. L'exploitation de modèles de $\boldsymbol{R}$ simplifiés pour préconditionner le système linéaire associé au $\boldsymbol{R}$ augmenté est néanmoins une piste prometteuse.

Précision :

La précision, telle qu'on la définit, est la capacité de reproduire le modèle de Matérn par le modèle de diffusion. Dans cette optique, le recours au raffinement de maillage est toujours de mise. Toutefois, il convient d'étudier l'effet de la troncature des itérations des algorithmes itératifs sur la résolution de l'équation de diffusion, vis-à-vis d'une éventuelle perte de précision subséquente.

En effet, la résolution d'un système linéaire de grande taille jusqu'à convergence peut s'avérer coûteux en pratique. La troncature des itérations autour d'une précision arbitraire est donc un sujet d'étude pertinent. Dans ce contexte, on conseille d'étudier les propriétés des algorithmes linéaires comme les itérations de Chebyshev, couplés à des préconditionneurs du deuxième ordre.

Réalisme :

Les points précédents ne font pas état de la réalité statistique des erreurs d'observation. Les diagnotics du chapitre 5 montrent que le modèle de Matérn n'est pas seul suffisant pour parfaitement décrire les erreurs estimées de l'imageur SEVIRI.

Pour décrire ces erreurs, plusieurs pistes sont proposées, allant de l'ajout d'un terme de régularisation à la sommation de plusieurs modèles de corrélation (chapitre 8). Ces pistes décrivent des nouvelles classes de modèles dont les complexités respectives ne sont pas encore bien comprises. A l'étude de la pertinence de ces modèles s'ajoute l'étude de leurs performances et de la modélisation de l'inverse.

Modification des prétraitements :

Enfin, nous avons mentionné la possibilité de modifier l'étape de *thinning* pour faciliter l'étape de maillage en améliorant la distribution spatiale des observations. Ce thème n'a pas été analysé en détail au cours de la thèse. Le *thinning* s'apparentant au procédé de déraffinement de maillage, nous avons présenté quelques méthodes jugées pertinentes dans la section 6.1.

Sujet très prometteur pour le futur de l'assimilation de données, la modification du *thinning* nécessite la définition d'un critère de sélection qui apparaît non trivial, puisqu'il dépend de la stratégie de déraffinement choisie et du nombre données qu'on désire conserver.

8.2.2 Enjeu de l'insertion dans l'opérationnel

Dans un futur relativement proche, nous envisageons de valider le modèle de $\boldsymbol{R}$ en éléments finis à partir d'un modèle de météorologie opérationnel, comme ARPEGE ou AROME, ou encore un modèle d'océanographie comme NEMO. Le but est de quantifier l'impact d'un « $\boldsymbol{R}$ corrélé » sur l'assimilation de données, en termes de performances et d'amélioration des résultats.

Pour commencer, nous nous en tiendrons au modèle le plus simple, en éléments finis sans maillage auxiliaire, tel qu'il est présenté dans les parties I et II du manuscrit. Si les premiers résultats s'avèrent concluants, alors d'autres formulations pourront être considérées.

Les outils requis pour le développement d'un module de corrélations d'erreurs d'observation sont standards : le maillage des observations est construit par l'algorithme de Delaunay et la matrice de masse est inversée par un solveur direct. On suggère donc d'utiliser CGAL ou GMSH pour le premier (même si un grand nombre d'algorithmes sont déjà présents dans BOOST) et une librairie comme MUMPS pour le deuxième. Ce module pourra être développé en C++ ou en FORTRAN, de telle sorte à être interfacé avec OOPS.

L'étape suivante est l'adaption du modèle en éléments finis aux cas des corrélations en trois dimensions (horizontales + verticales / intercanaux), ou en quatre dimensions si l'on prend en compte les corrélation temporelles. Néanmoins, la tâche n'est pas sans peine puisque la structure des observations change d'un niveau vertical à l'autre, en raison des nombreux prétraitements des données. Il est donc (*a priori*, hypothèse à confirmer) exclus de recourir à une approche « 2D+1D » qui couple un modèle de corrélations horizontal avec une modèle de corrélations vertical unidimensionnel. Fort heureusement, la méthode des éléments finis se transpose très facilement à la dimension trois, et la théorie associée reste inchangée.

Les méthodes développées au cours de la thèse ne s'appliquent pas eclusivement aux données de l'imageur SEVIRI. Dans le futur, nous espèrons appliquer ces méthodes à d'autres types de données contenant de fortes corrélations spatiales, comme les données des altimètres pour l'océanographie, des satellites défilants, des radars, ou encore des lidars. Afin de prendre en compte tous types de géométries, on s'intéressera à la résolution de l'équation de diffusion sur des variétés, qui sont des objets naturels quand on considère la courbure de la Terre ou encore les trajectoires des satellites défilants.

Pour conclure, donnons un commentaire sur les techniques mises en oeuvre au cours de la thèse. Notre contribution à l'assimilation de données n'est pas le développement d'une méthode isolée, mais plutôt le croisement d'une variété de méthodes issues de domaines généralement séparés. A l'avenir, l'assimilation couplée et la part grandissante du paradigme stochastique (entre autres !) feront naître de nouveaux défis dans l'assimilation de données. La complexification évidente des systèmes de prévision numérique du temps devra donc s'accompagner d'une réflexion profonde sur l'interdépendance entre les modèles et les données, tout en s'affranchissant des barrières traditionnelles séparant le travail des scientifiques.

Bibliographie

G. Allaire. *Analyse numérique et optimisation : une introduction à la modélisation mathématique et à la simulation numérique.* Mathématiques appliquées. Editions de l'Ecole polytechnique, 2005.

D. M. A. Aminou, H. J. Luhmann, C. Hanson, P. Pili, B. Jacquet, S. Bianchi, P. Coste, F. Pasternak, and F. Faure. Meteosat second generation - a comparison of on-ground and on-flight imaging and radiometric performances of SEVIRI on MSG-1, 2003.

I. Babuska. Error-bounds for finite element method. *Numerische Mathematik*, 16 :322–333, 1971.

R. E. Bank, T. F. Dupont, and H. Yserentant. The hierarchical basis multigrid method. *Numerische Mathematik*, 52 :427–458, 1988.

R. N. Bannister. A review of forecast error covariance statistics in atmospheric variational data assimilation. I : Characteristics and measurements of forecast error covariances. *Quaterly Journal of the Royal Meteorological Society*, 134 :1951–1970, 2008a.

R. N. Bannister. A review of forecast error covariance statistics in atmospheric variational data assimilation. II : Modelling the forecast error covariance statistics. *Quaterly Journal of the Royal Meteorological Society*, 134 :1971–1996, 2008b.

R. Barrett. Templates for the solution of linear systems : building blocks for iterative methods. In *Software, environments, tools*, 1994.

L. Beirao Da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L. D. Marini, and A. Russo. Basic principles of virtual element methods. *Mathematical Models and Methods in Applied Sciences*, 23 :199–214, 2013.

L. Beirao da Veiga, F. Brezzi, L. D. Marini, and A. Russo. The hitchhiker's guide to the virtual element method. *Mathematical Models and Methods in Applied Sciences*, 24 :1541–1573, 2014.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15 :1373–1396, 2003.

M. Belkin, J. Sun, and Y. Wang. Discrete laplace operator on meshed surfaces. In *Proceedings of the Twenty-fourth Annual Symposium on Computational Geometry*, pages 278–287. ACM, 2008.

M. Belo Pereira and L. Berre. The use of an ensemble approach to study the background error covariances in a global NWP model. *Monthly Weather Review*, 134 :2466–2489, 2006.

A. F. Bennett. Inverse modelling of the ocean and atmosphere. *Cambridge University Press*, 2002.

M. Benzi. Preconditioning techniques for large linear systems : A survey. *Journal of Computational Physics*, 182 :418–477, 2002.

M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14 :1–137, 2005.

H. Berger and M. Forsythe. Satellite wind superobbing. Technical report, Forecasting Research Technical Report 451, Met Office, Exeter UK, 2004.

K. H. Bergman and W. D. Bonner. Analysis Error as a Function of Observation Density for Satellite Temperature Soundings with Spatially Correlated Errors. *Monthly Weather Review*, 104 :1308–1316, 1976.

D. Bolin and F. Lindgren. A comparison between Markov approximations and other methods for large spatial data sets. *Comp. Statist. Data Anal.*, 61 :7–32, 2013.

N. Bormann and P. Bauer. Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I : Methods and application to ATOVS data. *Quaterly Journal of the Royal Meteorological Society*, 136 :1036–1050, 2010.

N. Bormann, A. Collard, and P. Bauer. Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. II : Application to AIRS and IASI data. *Quaterly Journal of the Royal Meteorological Society*, 136 :1051–1063, 2010.

N. Bourbaki, H. Eggleston, and S. Madan. *Elements of Mathematics*. Actualités scientifiques et industrielles. Springer-Verlag, Berlin, DE, 1987.

F. Bouttier and P. Courtier. Data assimilation concepts and methods. *ECMWF Meteorological Training Course Lecture Series*, 59, 1999.

J. M. Brankart, C. Ubelmann, C. E. Testut, E. Cosme, P. Brasseur, and J. Verron. Efficient parameterization of the observation error covariance matrix for square root or Ensemble Kalman Filters : Application to ocean altimetry. *Monthly Weather Review*, 137 :1908–1927, 2009.

S. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Texts in Applied Mathematics. Springer, New York, NY, 2013.

H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer New York, New York, USA, 2010.

T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler. A computational framework for infinite-dimensional Bayesian inverse problems. Part I : The linearized case, with application to global seismic inversion. *SIAM Journal on Scientific Computing*, 35 :A2494–A2523, 2013.

G. Burgers, P. J. Van Leeuwen, and G. Evensen. Analysis scheme in the ensemble kalman filter. *Monthly Weather Review*, 126 :1719–1724, 1998.

C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. Zang. *Spectral Methods in Fluid Dynamics*. Springer, New York, NY, 1987.

V. Chabot, M. Nodet, N. Papadakis, and A. Vidard. Accounting for observation errors in image data assimilation. *Tellus A : Dynamic Meteorology and Oceanography*, (1) :23629, 2015.

T. F. Chan, E. Chow, Y. Saad, and M. C. Yeung. Preserving symmetry in preconditioned krylov subspace methods. *SIAM Journal on Scientific Computing*, 20 :568–581, 1996.

H. Chi, L. Beirao da Veiga, and G. Paulino. Some basic formulations of the virtual element method (vem) for finite deformations. In *Computer Methods In Applied Mechanics And Engineering*, volume 318, pages 148–192, 2017.

M. Christon and D. Roach. The numerical performance of wavelets for pdes : the multi-scale finite element. *Computational Mechanics*, 25 :230–244, 2000.

P. Ciarlet. *The Finite Element Method for Elliptic Problems*. Classics in Applied Mathematics. SIAM, Philadelphia, PA, 2002.

S. Cohn, A. Da Silva, J. Guo, M. Sienkiewicz, and D. Lamich. Assessing the effects of data selection with the dao physical-space statistical analysis system. *Monthly Weather Review*, 126 :2913–2926, 1998.

P. Courtier. Dual formulation of four-dimensional variational assimilation. *Quaterly Journal of the Royal Meteorological Society*, 123 :2449–2461, 1997.

P. Courtier, J.-N. Thépaut, and A. Hollingsworth. A strategy for operational implementation of 4D-Var, using an incremental approach. *Quaterly Journal of the Royal Meteorological Society*, 120 :1367–1387, 1994.

P. Courtier, E. Andersson, W. Heckley, J. Pailleux, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier, and M. Fisher. The ecmwf implementation of three dimensional variational assimilation (3d-var). i : Formulation. *Quaterly Journal of the Royal Meteorological Society*, 124 :1783–1807, 1998.

M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary stokes equations i. *Revue francaise d'automatique, informatique, recherche opÃĩrationnelle. Mathématique, tome 7.*, 3 :33–75, 1973.

J. A. Cummings. Operational multivariate ocean data assimilation. *Quaterly Journal of the Royal Meteorological Society*, 131 :3583–3604, 2005.

A. Da Silva, J. Pfaendtner, J. Guo, M. Sienkiewicz, and S. Cohn. Assessing the effects of data selection with the dao physical-space statistical analysis system. In *Proceedings of the 2nd WMO Symposium on assimilation of observations in meteorology and oceanography*, pages 273–278, 1995.

R. Daley. *Atmospheric Data Analysis*. Cambridge Atmospheric and Space Sciences Series. Cambridge University Press, Cambridge, UK, 1991.

R. Daley and E. Barker. Navdas : Formulation and diagnostics. *Monthly Weather Review*, 120 :869–883, 2001.

M. L. Dando, A. J. Thorpe, and J. R. Eyre. The optimal density of atmospheric sounder observations in the Met Office NWP system. *Quaterly Journal of the Royal Meteorological Society*, 133 :1933–1943, 2007.

M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars. *Computational Geometry : Algorithms and Applications*. Springer, 2008.

A. Deckmyn and L. Berre. A wavelet approach to representing background error covariances in a limited-area model. *Monthly Weather Review*, 133 : 1279âĂŞ–1294, 2005.

D. P. Dee, S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kallberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart. The ERA-Interim reanalysis : configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137 :553–597, 2011.

J. Derber and F. Bouttier. A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus*, 51a : 195âĂŞ–221, 1999.

M. Desbrun, M. Meyer, P. Schröder, and A. H. Barr. Implicit fairing of irregular meshes using diffusion and curvature flow. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 317–324. ACM Press/Addison-Wesley Publishing Co., 1999.

G. Desroziers, G. Hello, and J.-N. Thépaut. A 4d-var re-analysis of fastex. *Quaterly Journal of the Royal Meteorological Society*, 129 :1301–1315, 2003.

G. Desroziers, L. Berre, B. Chapnik, and P. Poli. Diagnosis of observation, background and analysis-error statistics in observation space. *Quaterly Journal of the Royal Meteorological Society*, 131 :3385–3396, 2005.

Q. Du, V. Faber, and M. Gunzburger. Centroidal voronoi tessellations : Applications and algorithms. *SIAM Review*, 41 :637–676, 1999.

H. Edelsbrunner, T. Tan, and R. Waupotitsch. An $o(n^2 \log n)$ time algorithm for the minmax angle triangulation. *SIAM Journal on Scientific and Statistical Computing*, 13 :994–1008, 1992.

G. Egbert, A. Bennett, and F. M. Topex/Poseidon tides estimated using a global inverse model. *Journal on Geophysical Research*, 99 :24821–24852, 1994.

A. El Akkraoui and P. Gauthier. Convergence properties of the primal and dual forms of variational data assimilation. *Quaterly Journal of the Royal Meteorological Society*, 136 :107–115, 2010.

A. El Akkraoui, P. Gauthier, S. Pellerin, and S. Buis. Intercomparison of the primal and dual formulations of variational data assimilation. *Quaterly Journal of the Royal Meteorological Society*, 134 :1015–1025, 2008.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Mathematics and Its Applications. Springer Netherlands, 2000.

A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Series*. Springer, New York, NY, 2010.

L. C. Evans. *Partial Differential Equations*. Graduate studies in mathematics. American Mathematical Society, USA, 2010.

B. Fischer. *Polynomial based iteration methods for symmetric linear systems*. Wiley-Teubner series, advances in numerical mathematics. Wiley, 1996.

M. Fisher and S. Gürol. Parallelization in the time dimension of four-dimensional variational data assimilation. *Quaterly Journal of the Royal Meteorological Society*, 143 :1136–1147, 2017.

G. Gaspari and S. E. Cohn. Construction of correlation functions in two and three dimensions. *Quaterly Journal of the Royal Meteorological Society*, 125 :723–757, 1999.

E. Gaul, M. Gutknecht, O. Liesen, and R. Nabben. Deflated and augmented krylov subspace methods : Basic facts and a breakdown-free deflated minres. *SIAM Journal of Matrix Analysis and Applications*, pages 495âĂŞ–518, 2018.

P. Gauthier, M. Tanguay, S. Laroche, S. Pellerin, and J. Morneau. Extension of 3DVAR to 4DVAR : Implementation of 4DVAR at the meteorological service of canada. *Monthly Weather Review*, 135 :2339, 2007.

I. M. Gelfand and N. I. Vilenkine. *Generalized Functions : Applications of harmonic analysis*. Applications of Harmonic Analysis. Academic Press, New York, USA, 1964.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 3rd edition, 1996.

S. Gratton and J. Tshimanga. An observation-space formulation of variational assimilation using a Restricted Preconditioned Conjugate Gradient algorithm. *Quaterly Journal of the Royal Meteorological Society*, 135 : 1573–1585, 2009.

S. Gratton, A. Lawless, and N. Nichols. Approximate gauss-newton methods for nonlinear least-squares problems. *SIAM Journal on Optimization*, 18 : 106âĂŞ–132, 2007.

S. Gratton, A. Sartenaer, and J. Tshimanga. On a class of limited memory preconditioners for large scale linear systems with multiple right-hand sides. *SIAM Journal on Optimization*, 21 :912–935, 2011.

S. Gratton, S. GÃijrol, E. Simon, and P. L. Toint. On the use of the saddle formulation in weakly-constrained 4d-var data assimilation. *arXiv preprint arXiv :1709.06383*, 2017.

G. Green. *An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism.* (Göteborg : Wezäta-Melins 1958). author, 1828.

M. H. Gutknecht and S. Röllin. The chebyshev iteration revisited. *Parallel Comput.*, 28 :263–283, 2002.

P. Guttorp and T. Gneiting. Studies in the history of probability and statistics XLIX : On the Matérn correlation family. *Biometrika*, 93 :989–995, 2006.

W. Hackbusch. *Multi-Grid Methods and Applications*, volume 4. Springer Series in Computational Mathematics, 1985.

T. Hamill, S. Mullen, C. Snyder, Z. Toth, and D. Baumhefner. Ensemble forecasting in the short to medium range : Report from a workshop. *Bulletin of the American Meteorological Society*, 81 :2653âĂŞ–2664, 2000.

M. Hazewinkel. *Encyclopaedia of Mathematics : Supplement.* Encyclopaedia of Mathematics. Springer Netherlands, 2012.

M. Hein, J.-Y. Audibert, and U. von Luxburg. From graphs to manifolds - weak and strong pointwise consistency of graph laplacians. In *COLT*, 2005.

E. Hewitt and K. Stromberg. *Real and Abstract Analysis : A Modern Treatment of the Theory of Functions of a Real Variable.* Graduate Texts in Mathematics. Springer New York, 1975.

N. J. Higham. Computing the nearest correlation matrix - a problem from finance. *IMA journal of Numerical Analysis*, 22 :329–343, 2002.

H. Hoppe. Progressive meshes. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 99–108, 1996.

P. L. Houtekamer and H. L. Mitchell. Data assimilation using an ensemble kalman filter technique. *Monthly Weather Review*, 126 :796–811, 1998.

H. Järvinen, E. Andersson, and F. Bouttier. Variational assimilation of time sequences of surface observations with serially correlated errors. *Tellus A*, 51 :469–488, 1999.

G. H. Jones. *The Theory of Generalised Functions*. Cambridge University Press, Cambridge, UK, 3nd edition, 1982.

E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 2003.

P. Krysl, A. Trivedi, and B. Zhu. Object-oriented hierarchical mesh refinement with charms. *International Journal for Numerical Methods in Engineering*, 60 :1401–1424, 2004.

A. S. Lawless, S. Gratton, and N. K. Nichols. Approximate iterative methods for variational data assimilation. *International Journal for Numerical Methods in Fluids*, 47 :1129–1135, 2005.

P. D. Lax and A. N. Milgram. Parabolic equations. *Contributions to the theory of partial differential equations*, 33 :167–190, 1954.

F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields : the stochastic partial differential equation approach. *J. Roy. Stat. Soc. : Series B Stat. Method.*, 73 : 423–498, 2011.

Z.-Q. Liu and F. Rabier. The interaction between model resolution, observation resolution and observation density in data assimilation : A one-dimensional study. *Quaterly Journal of the Royal Meteorological Society*, 128 :1367–1386, 2002.

A. Lorenc. Analysis methods for numerical weather prediction. *Quaterly Journal of the Royal Meteorological Society*, 112 :1177–1194, 1986.

A. Lorenc. Four-dimensional variational data assimilation. In E. Blayo, M. Boquet, E. Cosme, and L. F. Cugliandolo, editors, *Advanced Data Assimilation for Geosciences*, pages 31–73. Oxford University Press, Oxford, UK, 2015.

M. Luby. A simple parallel algorithm for the maximal independent set problem. *SIAM*, 15 :1036–1055, 1986.

F. Magoules and F. X. Roux. *Calcul scientifique parallèle : Cours, exemples avec OpenMP et MPI, exercices corrigés.* Dunod, Paris, 2013.

S. Malardel. *Fondamentaux de météorologie : à l'école du temps*. Cépaduès, Toulouse, FR, 2005.

B. Ménétrier, T. Montmerle, Y. Michel, and L. Berre. Linear filtering of sample covariances for ensemble-based data assimilation. Part I : Optimality criteria and application to variance filtering and covariance localization. *Monthly Weather Review*, 143 :1622–1643, 2015a.

B. Ménétrier, T. Montmerle, Y. Michel, and L. Berre. Linear filtering of sample covariances for ensemble-based data assimilation. Part II : Application to a convective-scale NWP model. *Monthly Weather Review*, 143 : 1644–1664, 2015b.

Y. Michel. Estimating deformations of random processes for correlation modelling : methodology and the one-dimensional case. *Quaterly Journal of the Royal Meteorological Society*, 139 :771–783, 2013.

Y. Michel. Revisiting fisher's approach to the handling of horizontal spatial correlations of the observation errors in a variational framework. *Quaterly Journal of the Royal Meteorological Society*, 2018. (In press).

I. Mirouze and A. T. Weaver. Representation of correlation functions in variational assimilation using an implicit diffusion operator. *Quaterly Journal of the Royal Meteorological Society*, 136 :1421–1443, 2010.

D. Mitrovic and D. Zubrinic. *Fundamentals of Applied Functional Analysis*. Monographs and Surveys in Pure and Applied Mathematics. Taylor & Francis, 1997.

J. L. Morales and J. Nocedal. Automatic preconditioning by limited memory quasi-newton updating, 2000.

R. Nabben and C. Vuik. A comparison of deflation and the balancing preconditioner. *SIAM Journal on Scientific Computing*, 27 :1742–1759, 2006.

L. Nazareth. A relationship between the bfgs and conjugate gradient algorithms and its implications for new algorithms. *SIAM Journal on Numerical Analysis*, pages 794–800, 1979.

O. Nikodym. Sur une généralisation des intégrales de m. j. radon. *Fundamenta Mathematicae*, 15 :131–179, 1930.

J. Nocedal and J. S. Wright. *Numerical Optimization*. Springer, 2006.

C. J. Paciorek and M. J. Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17 :483–âĂŞ506, 2006.

O. Pannekoucke, L. Berre, and G. Desroziers. Filtering properties of wavelets for local background-error correlations. *Quaterly Journal of the Royal Meteorological Society*, 133 :363–379, 2007.

D. Parrish and J. Derber. The national meteorological center's spectral statistical interpolation analysis system. *Monthly Weather Review*, 120 :1747–1763, 1992.

P. Persson and G. Strang. A simple mesh generator in matlab. *SIAM Review*, 46 :329–345, 2004.

U. Pinkall and K. Polthier. Computing discrete minimal surfaces and their conjugates. *Experimental Mathematics*, 2 :15–36, 1993.

A. Quarteroni, R. Sacco, and F. Saleri. *Méthodes Numériques : Algorithmes, analyse et applications*. Springer Milan, 2008.

F. Rabier. Importance of data : A meteorological perspective. In E. P. Chassignet and J. Verron, editors, *Ocean Weather Forecasting : An Integrated View of Oceanography*, pages 343–360. Springer, Dordrecht, NL, 2006.

F. Rabier, H. Järvinen, E. Klinker, J.-F. Mahfouf, and A. Simmons. The ecmwf operational implementation of four-dimensional variational assimilation. i : Experimental results with simplified physics. *Quaterly Journal of the Royal Meteorological Society*, 126, 2000.

F. Rawlins, S. P. Ballard, K. J. Bovis, A. M. Clayton, D. Li, G. W. Inverarity, A. C. Lorenc, and T. J. Payne. The met office global four-dimensional variational data assimilation scheme. *Quaterly Journal of the Royal Meteorological Society*, 133 :347–362, 2007.

L. Raynaud, L. B., and G. Desroziers. Objective filtering of ensemble-based background-error variances. *Quaterly Journal of the Royal Meteorological Society*, 135 :1177–1199, 2009.

M. Reuter, S. Biasotti, D. Giorgi, G. Patanè, and M. Spagnuolo. Discrete laplace-beltrami operators for shape analysis and segmentation. *Computers and Graphics*, 33 :381–390, 2009.

G. A. Ruggiero, E. Cosme, J.-M. Brankart, and J. Le Sommer. An efficient way to account for observation error correlations in the assimilation of date from the future swot high-resolution altimeter mission. *Journal Atmospheric Oceanographic Technology*, 33 :2755–2768, 2016.

Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA, 2nd edition, 2003.

Y. Saad and M. H. Schultz. GMRES : A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7 :856–869, 1986.

Y. Saad, M. Yeung, J. Erhel, and F. Guyomarc'h. A deflated version of the conjugate gradient algorithm. *SIAM Journal on Scientific Computing*, 21 : 1909–1926, 1999.

L. Schwartz and G. Melese. *Theory of "distributions" and Some of Its Applications (Fourier Transformation)*. Johns Hopkins University, Virginia, 1951.

J. R. Shewchuk. What is a good linear finite element ? - interpolation, conditioning, anisotropy, and quality measures. Technical report, In Proc. of the 11th International Meshing Roundtable, 2002.

R. Sibson. A brief description of natural neighbor interpolation. *John Wiley and Sons*, pages 21–36, 1981.

D. Simpson, F. Lindgren, and H. Rue. In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics*, 23 :65–74, 2012.

I. Souopgui, H. E. Ngodock, A. Vidard, and F.-X. Le Dimet. Incremental projection approach of regularization for inverse problems. *Applied Mathematics & Optimization*, (2) :303–324, 2016.

M. L. Stein. *Interpolation of Spatial Data. Some Theory for Kriging*. Springer, New York, NY, 1999.

L. M. Stewart, S. L. Dance, N. K. Nichols, J. R. Eyre, and J. Cameron. Estimating interchannel observation-error correlations for IASI radiance

data in the Met Office system. *Quaterly Journal of the Royal Meteorological Society*, 140 :1236–1244, 2014.

R. S. Strichartz. *A Guide to Distribution Theory and Fourier Transforms*. Studies in advanced mathematics. World Scientific, USA, 2003.

O. J. Sutton. The virtual element method in 50 lines of matlab. *Numer. Algorithms*, 75 :1141–1159, 2017.

A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, Philadelphia, PA, 2005.

U. Trottenberg, C. W. Oosterlee, A. Schuller, and A. Brandt. *Multigrid*. Elsevier Science, 2001.

G. Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les parallélloèdres primitifs. *Journal für die reine und angewandte Mathematik*, 134 :198–287, 1908.

J. A. Waller, S. Ballard, S. L. Dance, G. Kelly, N. K. Nichols, and D. Simonin. Diagnosing horizontal and inter-channel observation error correlations for SEVIRI observations using observation-minus-background and observation-minus-analysis statistics. *Remote Sens.*, 8 :581, 2016a.

J. A. Waller, S. L. Dance, and N. K. Nichols. Theoretical insight into diagnosing observation error correlations using observation-minus-background and observation-minus-analysis residuals. *Quaterly Journal of the Royal Meteorological Society*, 142 :418–431, 2016b.

J. A. Waller, D. Simonin, S. L. Dance, N. K. Nichols, and S. Ballard. Diagnosing observation error correlations for Doppler radar radial winds in the Met Office UKV model using observation-minus-background and observation-minus-analysis statistics. *Monthly Weather Review*, 144 :3533–3551, 2016c.

A. J. Wathen. Preconditioning. *Acta Numerica*, 24 :329–376, 2015.

A. T. Weaver and P. Courtier. Correlation modelling on the sphere using a generalized diffusion equation. *Quaterly Journal of the Royal Meteorological Society*, 127 :1815–1846, 2001.

A. T. Weaver and I. Mirouze. On the diffusion equation and its application to isotropic and anisotropic correlation modelling in variational assimilation. *Quaterly Journal of the Royal Meteorological Society*, 139 :242–260, 2013.

P. P. Weston, W. Bell, and J. R. Eyre. Accounting for correlated error in the assimilation of high-resolution sounder data. *Quaterly Journal of the Royal Meteorological Society*, 140 :2420–2429, 2014.

J. S. Whitaker and T. M. Hamill. Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130 :1913–1924, 2002.

P. Whittle. Stochastic processes in several dimensions. *Bull. Inst. Internat. Statist.*, 40 :974–994, 1963.

J. Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20A :32–52, 1928.

X. Zhao, R. Conley, N. Ray, V. S. Mahadevan, and X. Jiao. Conformal and non-conformal adaptive mesh refinement with hierarchical array-based half-facet data structures. *Procedia Engineering*, 124 :304–316, 2015.

O. C. Zienkiewicz, R. L. Taylor, and J. Z. Zhu. *The Finite Element Method : Its Basis and Fundamentals*. Elsevier Science, 2005.

Annexe A

Méthode des volumes finis

Les équations aux dérivées partielles contenues dans ce manucrit sont exclusivement discrétisées par la méthode des éléments finis. Pourtant, ce choix n'est pas obligatoire. Dans cette première annexe, on discute de la méthode des volumes finis et des aspects qui nous ont fait écarter cette méthode pour notre application.

Considérons l'équation de diffusion sur un pas de temps, de condition initiale f et d'inconnue u :

$$u - \nabla \cdot \boldsymbol{\kappa}(\boldsymbol{z}) \nabla u = f. \tag{A.1}$$

Pour alléger la présentation, on suppose que ces fonctions sont suffisamment régulières pour que leurs dérivées sont dans $L^2(\mathcal{D})$. La formulation faible de l'équation (A.1) s'écrit

$$\langle u - \nabla \cdot \boldsymbol{\kappa}(\boldsymbol{z}) \nabla u, \varphi \rangle_{L^2} = \langle f, \varphi \rangle_{L^2}, \tag{A.2}$$

où φ désigne une fonction de pondération. Pour rappel, on souhaite que l'équation (A.2) soit vérifiée pour le plus grand nombre de fonctions de pondérations φ possible. L'application de la formule de Green et l'utilisation des conditions de Neumann au bords du domaine de définition $\mathcal{D}$ permettent d'écrire

$$\langle u, \varphi \rangle_{L^2} + \langle \boldsymbol{\kappa}(\boldsymbol{z}) \nabla u, \nabla \varphi \rangle_{L^2} = \langle f, \varphi \rangle_{L^2}. \tag{A.3}$$

La différence entre les éléments finis et les volumes finis tient au choix de ces fonctions de pondération. Supposons qu'on résolve l'équation (A.1) sur la triangulation $\mathcal{T}$. En volumes finis, on impose que

$$\varphi \in V_c = \mathrm{Vect}\{\chi_k, k \in [\![1, \mathrm{card}(\mathcal{T})]\!]\}, \tag{A.4}$$

où χ_k est la fonction constante par morceaux nulle en dehors de l'élément τ_k. Ainsi, le nombre de degrés de liberté de la méthode des volumes finis correspond au nombre d'éléments dans $\mathcal{T}$, et non au nombre de noeuds. Pour que le nombre de degrés de liberté soit égal au nombre d'observations, il faudrait donc construire $\mathcal{T}$ de telle sorte qu'à chaque observation soit associé un élément τ_k.

En pratique, cette condition est extrêmement difficile à satisfaire. En effet, il n'existe pas à notre connaissance, de méthode permettant de construire une triangulation à partir d'un nuage de points, qui comporte autant d'éléments triangulaires que de points dans le nuage. L'approche classique consisterait à construire la triangulation de Delaunay déduite des positions des observations, et de considérer la méthode des volumes finis sur le maillage dual (voir section 4.3). Toutefois, cette approche implique de calculer des volumes de polygones généraux (les cellules de Voronoï). De plus, à l'ordre le plus bas, la méthode des volumes finis sur le maillage dual est équivalente à la méthode des éléments finis sur le maillage primal.

Les multiples raisons énoncées précédemment expliquent pourquoi l'étude s'est naturellement orienté vers la méthode des éléments finis, plutôt que vers la méthode des volumes finis.

Annexe B

Méthode des éléments virtuels

La méthode des éléments virtuels est une alternative récente spécialement développée pour s'adapter à tous types de maillages [Chi et al., 2017]. Très attrayante au premier abord, ses résultats numériques ne sont toutefois pas suffisamment convaincants pour qu'on l'utilise à la place des éléments finis. En outre, la littérature concernant les éléments virtuels étant récente, nous avons peu de recul quant à ses véritables avantages théoriques. Néanmoins, on détaille succintement cette méthode en réponse à d'éventuelles questions, et afin d'évaluer son potentiel pour notre application.

Repartons du problème modèle

$$u - \nabla \cdot \boldsymbol{\kappa}(\boldsymbol{z}) \nabla u = f. \tag{B.1}$$

où u et f sont des fonctions suffisamment régulières, telles que leurs dérivées sont dans $L^2(\mathcal{D})$. La forme variationnelle de (B.1) s'écrit

$$a(u, \varphi) + b(u, \varphi) = f(\varphi),$$

où

$$\begin{aligned} a(u, \varphi) &= \langle u, \varphi \rangle_{L^2} \\ b(u, \varphi) &= \langle \boldsymbol{\kappa}(\boldsymbol{z}) \nabla u, \nabla \varphi \rangle_{L^2} \\ f(\varphi) &= \langle f, \varphi \rangle_{L^2}. \end{aligned} \tag{B.2}$$

La méthode des éléments virtuels peut être vue comme une généralisation de la méthode des éléments finis, dans laquelle la fonction-test φ appartient à un espace d'approximation qui inclus à la fois des fonctions polynomiales et des fonctions harmoniques définies implicitement.

Soit $\mathcal{T}$ une partition polygonale de $\mathcal{D}$. L'espace d'approximation global V_c est défini comme l'ensemble des fonctions de $H_0^1(\mathcal{D})$ dont la restriction à l'élément $\tau \in \mathcal{T}$ est dans le petit espace V_τ [Sutton, 2017, Beirao Da Veiga et al., 2013] :

$$V_c = \underset{\tau\in\mathcal{T}}{\oplus} V_\tau = \{\varphi \in H_0^1(\mathcal{D}), \varphi_{|\tau} \in V_\tau, \forall \tau \in \mathcal{T}\}. \tag{B.3}$$

On peut ainsi exprimer a, b et f à partir des contributions locales a_τ, b_τ et f_τ :

$$\begin{aligned} a(u,\varphi) &= \sum_{\tau\in\mathcal{T}} a_\tau(u_{|\tau}, \varphi_{|\tau}) \\ b(u,\varphi) &= \sum_{\tau\in\mathcal{T}} b_\tau(u_{|\tau}, \varphi_{|\tau}) \\ f(\varphi) &= \sum_{\tau\in\mathcal{T}} f_\tau(\varphi_{|\tau}) \end{aligned} \tag{B.4}$$

L'espace d'approximation local V_τ est quant à lui défini par

$$V_\tau = \{\phi \in H^1(\mathcal{D}), \Delta\phi = 0, \phi_{|\partial\tau} \in \mathcal{C}^0(\partial\tau) \text{ et } \phi_{|e} \in \mathcal{P}_1(e), \forall e \in \partial\tau\}. \tag{B.5}$$

Il contient les fonctions polynomiales sur chaque arête e de la frontière $\partial\tau$, qui vérifient $\Delta\phi = 0$ dans l'intérieur de τ. On peut vérifier que

$$\mathcal{P}_1(\tau) \subset V_\tau \tag{B.6}$$

et que $\dim(V_\tau) = n_e$, où n_e est le nombre de sommets de τ.

Lorsque les éléments de $\mathcal{T}$ sont des polygones généraux, expliciter une base de V_τ et *a posteriori* de V_c peut se révéler difficile, voire impossible. Partant de (B.6), la méthode des éléments virtuels consiste à introduire la projection

$$\Pi_\tau : V_\tau \to \mathcal{P}_1(\tau) \subset V_\tau \tag{B.7}$$

qui vérifie pour tout $\phi \in V_\tau$:

$$\int_{\partial\tau} (\Pi_\tau\phi)_{|\partial\tau} \mathrm{d}\sigma = \int_{\partial\tau} \phi_{|\partial\tau} \mathrm{d}\sigma \tag{B.8}$$

et pour tout $\phi \in V_\tau$ et $\eta \in \mathcal{P}_1(\tau)$:

$$a_\tau(\Pi_\tau\phi, \eta) + b_\tau(\Pi_\tau\phi, \eta) = a_\tau(\phi, \eta) + b_\tau(\phi, \eta). \tag{B.9}$$

D'un certaine façon, la projection Π_τ est « $(a_\tau + b_\tau)$-orthogonale », ce qui signifie que, pour tout couple de fonctions $(\phi_i, \phi_j) \in V_\tau^2$, la décomposition

$$\phi = \Pi_\tau \phi + (I - \Pi_\tau)\phi \tag{B.10}$$

entraîne

$$\begin{aligned} a_\tau(\phi_i, \phi_j) + b_\tau(\phi_i, \phi_j) &= a_\tau(\Pi_\tau\phi_i, \Pi_\tau\phi_j) \\ &+ b_\tau(\Pi_\tau\phi_i, \Pi_\tau\phi_j) \\ &+ a_\tau((I - \Pi_\tau)\phi_i, (I - \Pi_\tau)\phi_j) \\ &+ b_\tau((I - \Pi_\tau)\phi_i, (I - \Pi_\tau)\phi_j) \end{aligned} \tag{B.11}$$

les termes croisés étant nuls en vertu de (B.9). On peut donc traiter chaque terme séparément. Les deux premiers termes peuvent être traités de manière standard, puisque $\Pi_\tau\phi_i$ et $\Pi_\tau\phi_j$ sont dans $\mathcal{P}_1(\tau)$. Pour cela, on introduit la base de $\mathcal{P}_1(\tau)$ composée des monômes

$$\begin{aligned} m_1(\boldsymbol{z}) &= 1 \\ m_2(\boldsymbol{z}) &= \frac{x - \bar{x}}{|\tau|} \\ m_3(\boldsymbol{z}) &= \frac{y - \bar{y}}{|\tau|}, \end{aligned} \tag{B.12}$$

où $\boldsymbol{z} = (x, y) \in \tau$, $\bar{\boldsymbol{z}} = (\bar{x}, \bar{y})$ est le barycentre de τ et $|\tau|$ le diamètre de τ. Le calcul de $a_\tau(\Pi_\tau\phi_i, \Pi_\tau\phi_j)$ et $b_\tau(\Pi_\tau\phi_i, \Pi_\tau\phi_j)$ en découle facilement. Les deux autres termes, en revanche, sont plus difficiles à calculer puisque $(I - \Pi_\tau)\phi_i$ et $(I - \Pi_\tau)\phi_j$ sont dans $V_\tau \backslash \mathcal{P}_1(\tau)$, dont on ne connaît pas explicitement les fonctions. Pour contourner la difficulté, on fait les approximations

$$a_\tau((I - \Pi_\tau)\phi_i, (I - \Pi_\tau)\phi_j) \simeq |\tau| \sum_{\boldsymbol{z} \in \mathcal{S}} (I - \Pi_\tau)\phi_i(\boldsymbol{z})(I - \Pi_\tau)\phi_j(\boldsymbol{z}) \tag{B.13}$$

$$b_\tau((I - \Pi_\tau)\phi_i, (I - \Pi_\tau)\phi_j) \simeq \sum_{\boldsymbol{z} \in \mathcal{S}} (I - \Pi_\tau)\phi_i(\boldsymbol{z})(I - \Pi_\tau)\phi_j(\boldsymbol{z}). \tag{B.14}$$

Cette somme est calculable puisqu'elle ne fait intervenir que les valeurs aux sommets de l'élément [Beirao da Veiga et al., 2014].

Les figures B.1 et B.2 montrent respectivement l'erreur d'amplitude et l'erreur de forme associées à la méthode des éléments virtuel sur le cas test utilisé dans le manuscrit. On constate que ces erreurs sont du même ordre de grandeur, mais supérieures, aux erreurs numériques de la méthode des éléments finis.

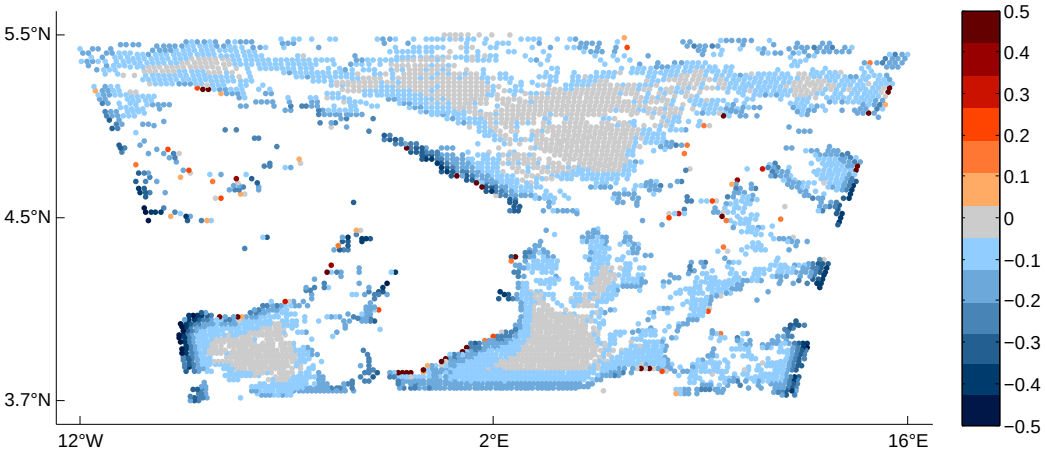


Figure B.1 – Carte des erreurs d'amplitude des éléments virtuels.

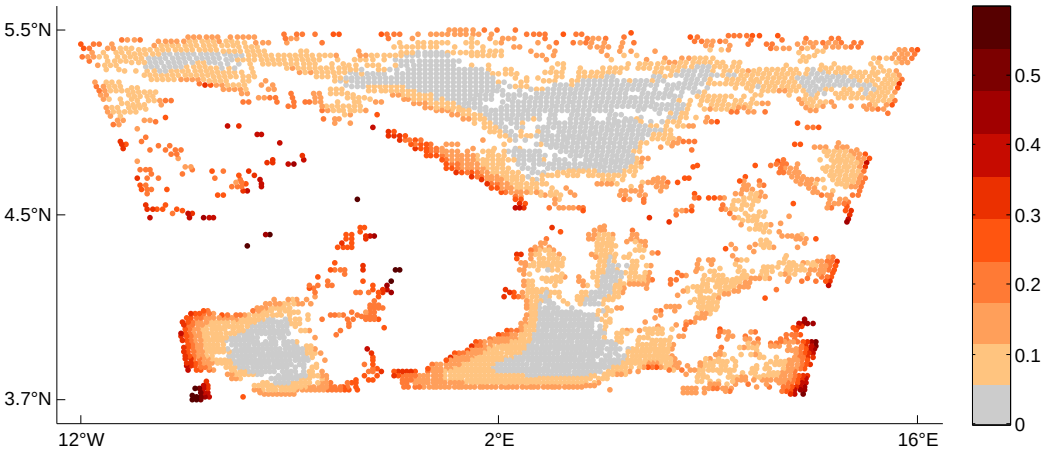


Figure B.2 – Carte des erreurs de forme des éléments virtuels.

Annexe C

Diffusion sur des graphes généraux

Les méthodes des éléments finis, des volumes finis, des différences finies et des éléments virtuels s'appuient comme bien d'autres sur la notion de maillage et d'éléments pour se définir. Implicitement, la structure géométrique des données est décrite par un graphe sous-jacent, qui décrit la connectivité et la dépendance entre les données. Etant donnée la difficulté de construire des maillages de qualité à partir de certaines observations, nous avons envisagé de résoudre l'équation de la diffusion directement sur des graphes généraux, qui relient les données par ordre de proximité. Dans cette annexe, on redéfinit les opérateurs différentiels sur ces graphes et on discute de la possibilité de les exploiter pour notre application.

Soient $V = (\boldsymbol{v}_i)_{i\in[\![1,p]\!]}$ un ensemble de points du plan, qu'on appelle « sommets », et $E \subset V^2$ un ensemble non vide contenant des couples d'éléments de V, appelés « arêtes ». Le graphe $\mathcal{G}$ est défini comme le couple (V, E).

On souhaite équiper $\mathcal{G}$ d'une structure permettant de définir l'opérateur laplacien, puis l'équation de diffusion. Pour ce faire, on munit l'ensemble des fonctions de V à valeurs dans $\mathbb{R}$, noté $\mathbb{R}^V$, du produit scalaire

$$\langle f, g\rangle_V = \sum_{i=1}^{p} f(\boldsymbol{v}_i)g(\boldsymbol{v}_i)\chi_i, \tag{C.1}$$

où les $(\chi_i)_{i\in[\![1,p]\!]}$ sont des coefficients strictement positifs. On munit ensuite l'ensemble des fonctions de E à valeurs dans $\mathbb{R}$, noté $\mathbb{R}^E$, du produit scalaire

$$\langle F, G\rangle_E = \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} F(\boldsymbol{v}_i, \boldsymbol{v}_j)G(\boldsymbol{v}_i, \boldsymbol{v}_j)\theta_{ij}, \tag{C.2}$$

où les $(\theta_{ij})_{(i,j)\in[\![1,p]\!]^2}$ sont également des coefficients strictement positifs, vérifiant pour tout $i \in [\![1,p]\!]$ et pour tout $j \in [\![1,p]\!]$:

$$\theta_{ij} = \theta_{ji}. \tag{C.3}$$

On peut montrer facilement que $\mathcal{H}_V = (V, \langle\cdot,\cdot\rangle_V)$ et $\mathcal{H}_E = (E, \langle\cdot,\cdot\rangle_E)$ sont des espaces de Hilbert [Hein et al., 2005].

Ce cadre hilbertien permet de redéfinir les méthodes de Galerkin sur des graphes quelconques. Au préalable, on doit méanmoins introduire l'opérateur de différentiation

$$\mathrm{d} : \mathcal{H}_V \to \mathcal{H}_E \tag{C.4}$$

défini pour toute fonction $f \in \mathbb{R}^V$ et tout $(\boldsymbol{v}_i, \boldsymbol{v}_j) \in E$ par

$$(\mathrm{d}f)(\boldsymbol{v}_i, \boldsymbol{v}_j) = \gamma_{ij}(f(\boldsymbol{v}_j) - f(\boldsymbol{v}_i)), \tag{C.5}$$

où les $(\gamma_{ij})_{(i,j)\in[\![1,p]\!]^2}$ sont strictement spositifs et vérifient $\gamma_{ij} = \gamma_{ji}$. Son adjoint est l'opérateur

$$\mathrm{d}^\star : \mathcal{H}_E \to \mathcal{H}_V \tag{C.6}$$

défini pour toute fonction $F \in \mathbb{R}^E$ et tout $\boldsymbol{v}_i \in V$ par

$$(\mathrm{d}^\star F)(\boldsymbol{v}_i) = \frac{1}{2\chi_i}\sum_{j=1}^{p}\gamma_{ij}\theta_{ij}(F(\boldsymbol{v}_j, \boldsymbol{v}_i) - F(\boldsymbol{v}_i, \boldsymbol{v}_j)). \tag{C.7}$$

On dispose alors de tous les outils permettant de définir une méthode des éléments finis sur graphes. Il suffit de considérer une base $(\varphi_k)_{k\in[\![1,p]\!]}$ de $\mathcal{H}_V$ et de calculer les matrices de masse et de raideur de coefficients :

$$\begin{aligned} \boldsymbol{M}_{kl} &= \langle\varphi_k, \varphi_l\rangle_V \\ &= \sum_{i=1}^{p}\varphi_k(\boldsymbol{v}_i)\varphi_l(\boldsymbol{v}_i)\chi_i \end{aligned} \tag{C.8}$$

et

$$\begin{aligned} \boldsymbol{K}_{kl} &= \langle\mathrm{d}\varphi_k, \mathrm{d}\varphi_l\rangle_E \\ &= \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\gamma_{ij}^2\theta_{ij}(\varphi_k(\boldsymbol{v}_j) - \varphi_k(\boldsymbol{v}_i))(\varphi_l(\boldsymbol{v}_j) - \varphi_l(\boldsymbol{v}_i)). \end{aligned} \tag{C.9}$$

On définit respectivement

$$\begin{aligned} \omega_{ij} &= \gamma_{ij}^2\theta_{ij} \\ \boldsymbol{V}_{kk} &= \sum_{j=1}^{p}\omega_{kj} \\ \boldsymbol{W}_{kl} &= \omega_{kl}, \end{aligned} \tag{C.10}$$

où $\boldsymbol{V}$ est une matrice diagonale. En supposant que pour tout $i \in [\![1, p]\!]$ et tout $j \in [\![1, p]\!]$, les fonctions $(\varphi_k)_{k\in[\![1,p]\!]}$ vérifient la propriété de Dirac

$$\varphi_i(\boldsymbol{v}_j) = \delta_{ij}, \tag{C.11}$$

on peut montrer que $\boldsymbol{M}$ est la matrice diagonale de coefficients

$$\boldsymbol{M}_{kk} = \chi_k. \tag{C.12}$$

et que

$$\boldsymbol{K} = \boldsymbol{V} - \boldsymbol{W}. \tag{C.13}$$

Pour définir une méthode des éléments finis sur desgraphes généraux, il suffit donc de spécifier les coefficients $(\chi_i)_{i\in[\![1,p]\!]}$ et $(\omega_{ij})_{(i,j)\in[\![1,p]\!]^2}$. Toutefois, de nombreux choix de paramètres n'ont pas de sens physique, et ne permettent pas de représenter des corrélations géométriquement pertinentes. On donne ici un récapitulatif des approches qui définissent un opérateur Laplacien « géométrique » et de leur applicabilité à la modélisation des corrélations spatiales d'erreurs d'observation.

Lorsque le graphe $\mathcal{G}$ décrit une triangulation, Pinkall and Polthier [1993] propose de choisir

$$\omega_{ij} = \frac{\operatorname{cotan}(\alpha_{ij}) + \operatorname{cotan}(\beta_{ij})}{2}, \tag{C.14}$$

où α_{ij} et β_{ij} sont les mesures des angles opposés à l'arrête (i, j) dans chacun des triangles contenant cette arête. Associée au choix

$$\chi_i = |\mathcal{V}(\boldsymbol{v}_i)| \tag{C.15}$$

(l'aire de la cellule de Voronoï entourant le noeud $\boldsymbol{v}_i$) suggéré par Desbrun et al. [1999], cette approximation donne un schéma équivalent aux éléments finis $\mathbb{P}_1$ avec condensation de masse. L'assemblage de la méthode sur graphe étant plus compliqué et la littérature moins abondante, on lui préfère donc les éléments finis standards.

Une alternative consiste à définir les coefficients ω_{ij} comme valeurs d'une fonction gaussienne, comme présentée dans Belkin and Niyogi [2003] et Belkin et al. [2008]. L'avantage est que cette méthode ne contraint pas la structure du maillage, en particulier sa connectivité, et les matrices obtenues sont toujours creuses. Cependant, la prise en compte des conditions aux limites n'est pas évidente et Reuter et al. [2009] laisse penser que l'approche n'est pas adaptée aux graphes fortement non structurés. Voilà pourquoi on ne s'est pas dirigé vers la diffusion sur graphe au premier abord.

Néanmoins, de nombreuses études commencent à s'intéresser à la convergence de ce types de schémas numériques, comme Hein et al. [2005]. On mentionne en particulier la théorie spectrale des graphes, qui semble être le cadre le mieux adapté à l'étude des propriétés mathématiques des laplaciens discrets.

Annexe D

Autres jeux de données

Dans cette annexe, on reproduit quelques expériences-clefs à partir de situations météorologiques variées. On étudie trois situations différentes. Pour chacune, sont représentés dans cet ordre :

- Le maillage des observations ;
- Les erreurs d'amplitude en éléments finis (sans condensation de masse) ;
- Le maillage raffiné (h- raffinement non hiérarchique) ;
- Les erreurs d'amplitude de l'opérateur raffiné (transfert par injection).

Situation n°1 :

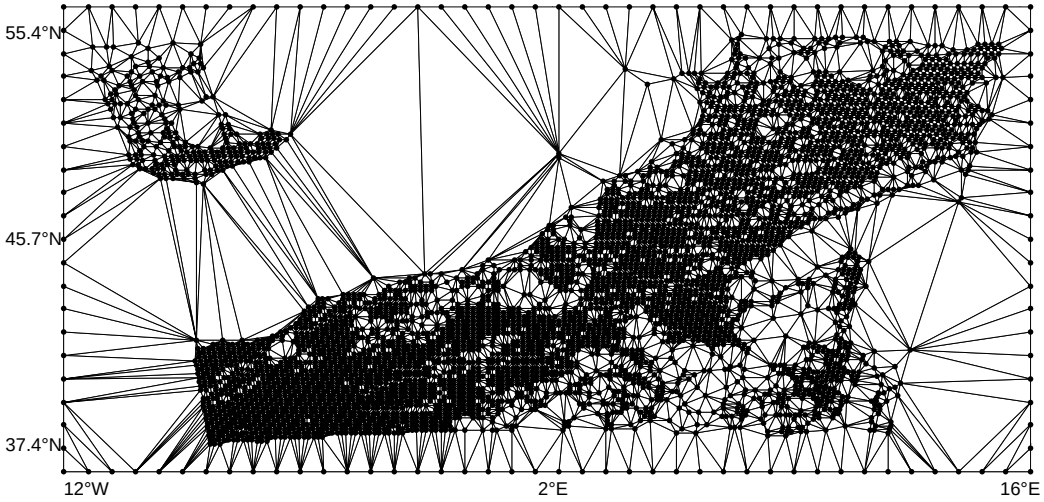


Figure D.1 – Maillage des observations.

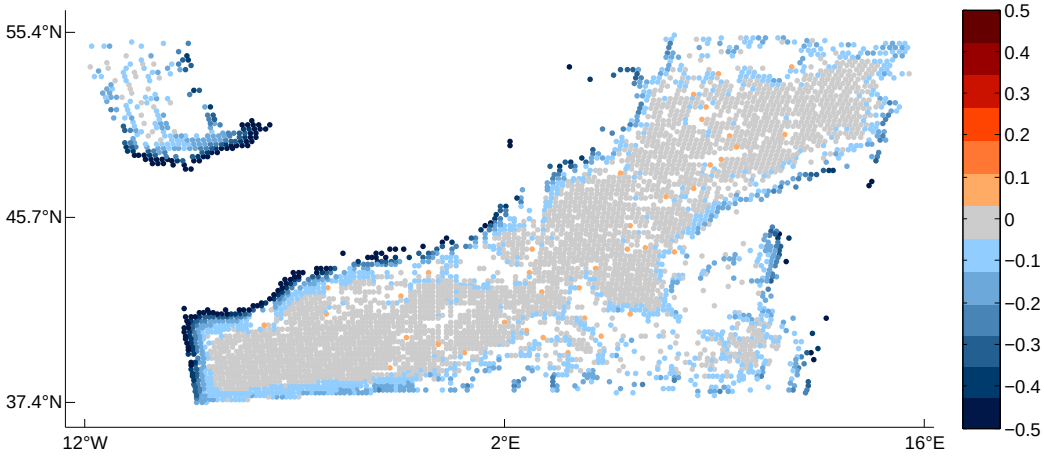


Figure D.2 – Carte d'erreurs d'amplitude en éléments finis (sans condensation de masse).

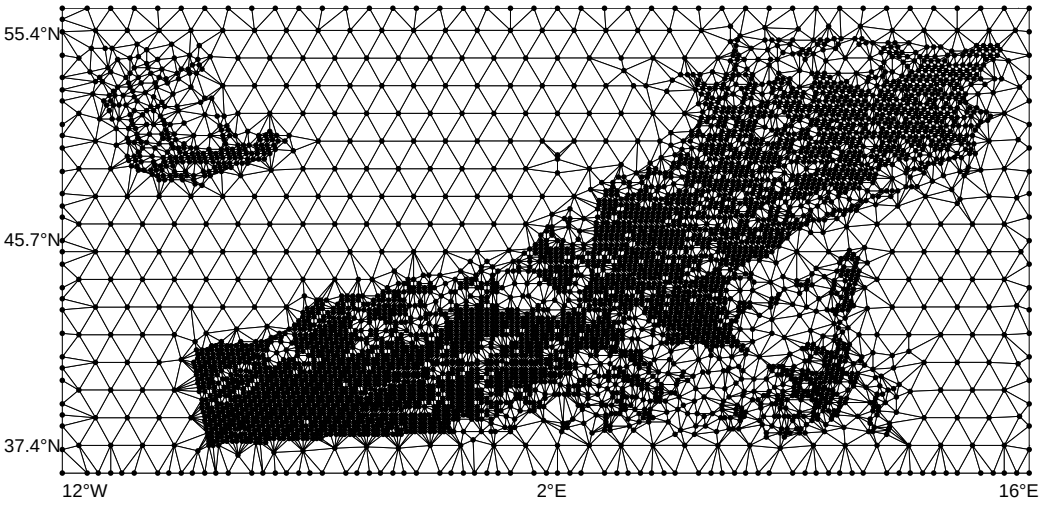


Figure D.3 – Maillage raffiné (h- raffinement non hiérarchique).

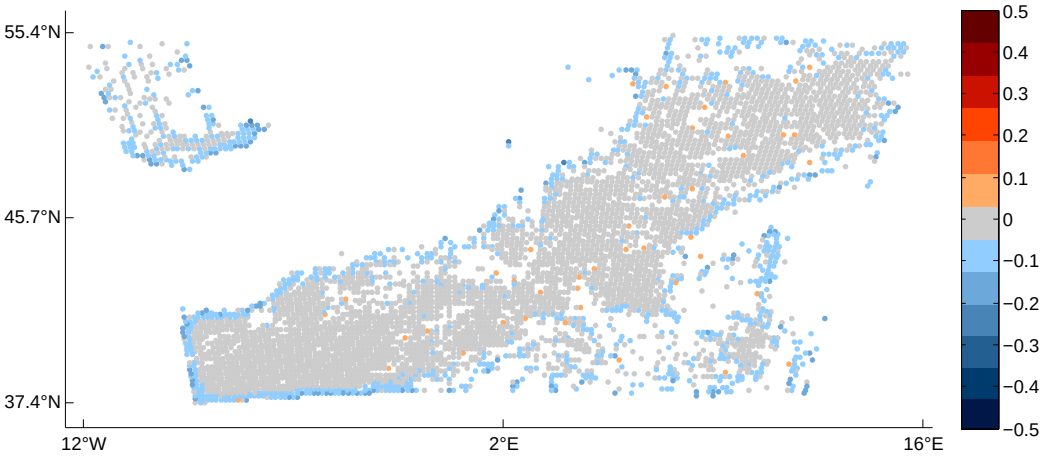


Figure D.4 – Carte d'erreurs d'amplitude de l'opérateur raffiné (transfert par injection).

Situation n°2 :

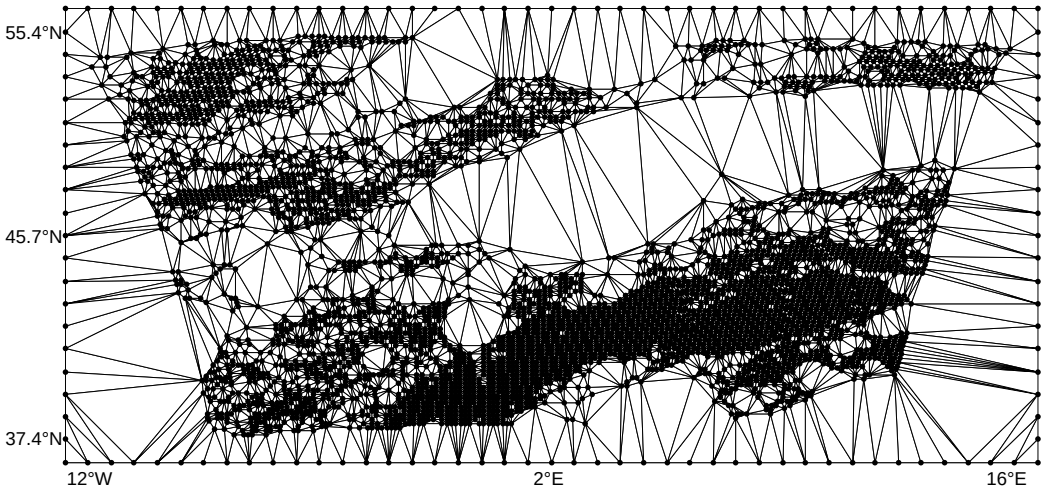


Figure D.5 – Maillage des observations.

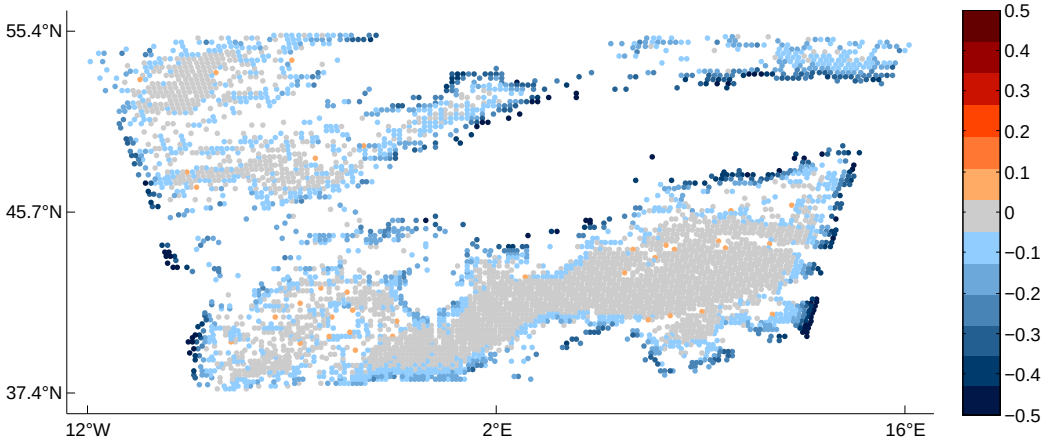


Figure D.6 – Carte d'erreurs d'amplitude en éléments finis (sans condensation de masse).

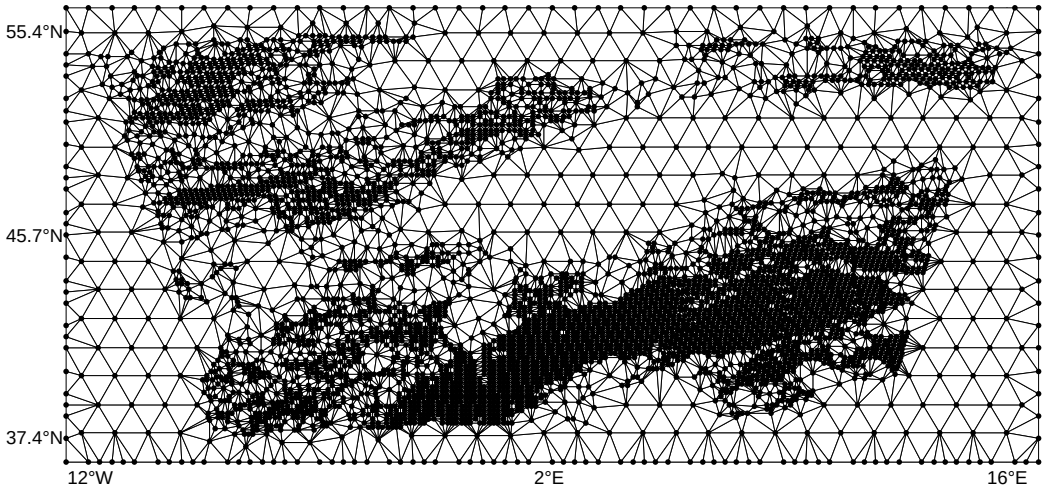


Figure D.7 – Maillage raffiné (h- raffinement non hiérarchique).

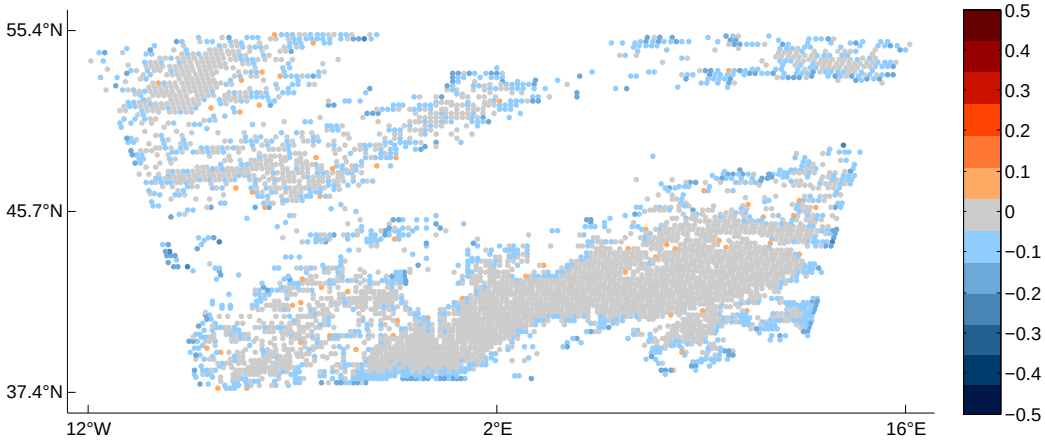


Figure D.8 – Carte d'erreurs d'amplitude de l'opérateur raffiné (transfert par injection).

Situation n°3 :

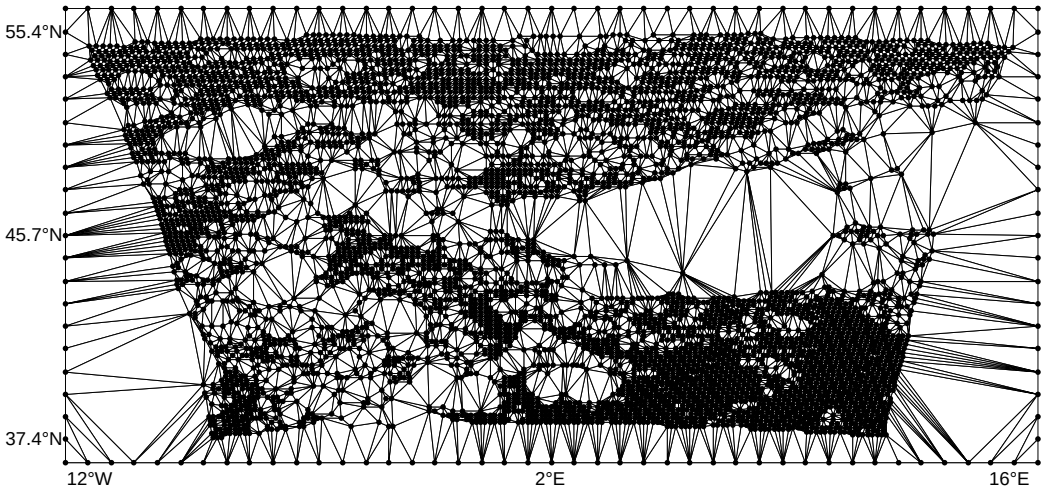


Figure D.9 – Maillage des observations.

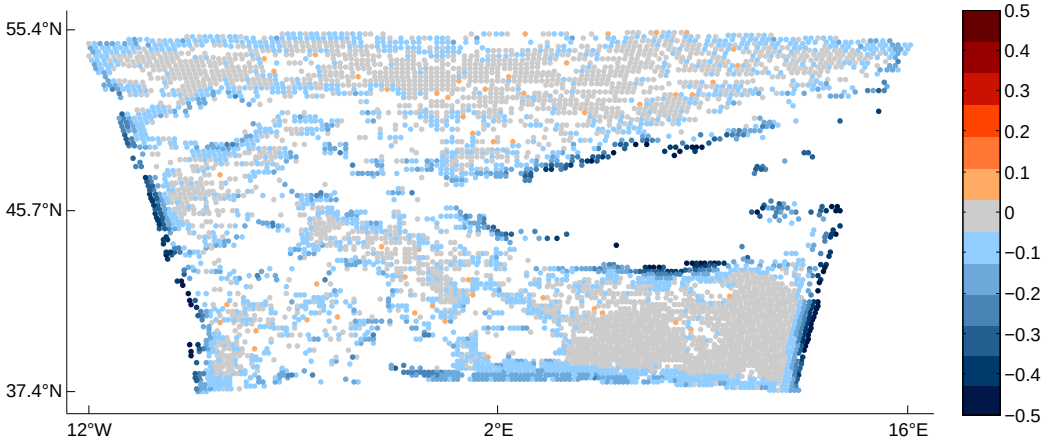


Figure D.10 – Carte d'erreurs d'amplitude en éléments finis (sans condensation de masse).

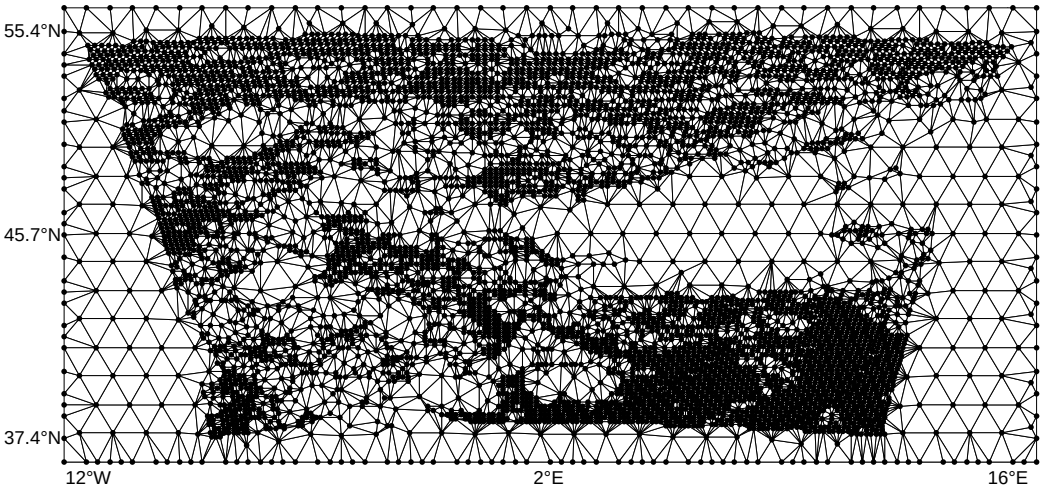


Figure D.11 – Maillage raffiné (h- raffinement non hiérarchique).

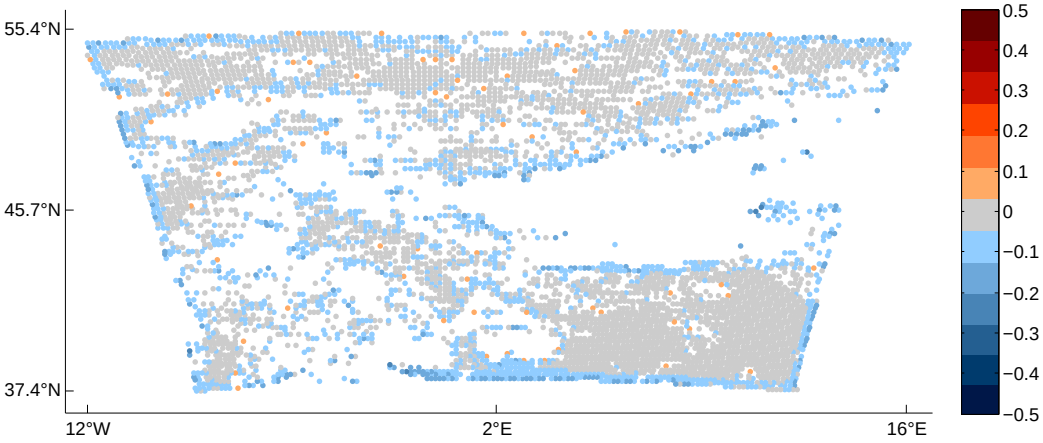


Figure D.12 – Carte d'erreurs d'amplitude de l'opérateur raffiné (transfert par injection).

Annexe E

Structure du code

Une grande partie du temps de travail de la thèse fut dédiée à la mise en place des expériences décrites dans le manuscrit. L'environnement d'expérimentation se devait d'être flexible, facilement maintenable et suffisamment complet pour incorporer des éléments d'assimilation de données, de génération de maillage, d'algèbre linéaire, et toute une panoplie d'outils annexes. Dans cette annexe, on décrit brièvement la structure de code retenue pour les expériences.

Le choix fut fait d'adopter le logiciel MATLAB comme outil de programmation pour deux raisons. Premièrement, ce logiciel offre de nombreuses fonctionnalités de génération de maillage et d'algèbre linéaire. Deuxièmement, il permet de développer le code par une approche orientée objet, offrant ainsi une bonne capacité de maintenance et un environnement flexible. On souligne que l'adoption de l'approche orientée objet était également motivée par la perspective d'implémenter un jour notre modèle de corrélation dans le système OOPS.

Le diagramme de classe du projet est représenté sur la figure E.1. Les emplacements des données sont rassemblés dans la classe `Locations`. Ces emplacements servent à définir deux types d'objets distincts : les `Data_Set` et les `Mesh_Structure`. La classe `Data_Set` contient les données et les tout ce qui les concerne (valeurs, manipulation, visualisation), tandis que `Mesh_Structure` regroupe tout ce qui concerne le maillage (structure, génération, outils de raffinement, visualisation).

La classe abstraite `Space`, qui hérite à des deux classes `Data_Set` et `Mesh_Structure`, sert à construire tous les vecteurs « informatifs » de l'assimilation de données. Ainsi, l'état vrai, l'ébauche, l'analyse, les observations,

les innovations... partagent tous la structure « de base » contenue dans `Space`. Cette classe contient notamment tous les tests (adjoint, validation des opérateurs de corrélation, systèmes linéaires, cartes d'erreurs) dont l'implémentation ne dépend pas mathématiquement du type de donnée. Pour définir ces tests, il est nécessaire que `Space` hérite simultanément de `Data_Set` et de `Mesh_Structure`, ce qui donne lieu à l'architecture en diamant représentée sur la figure E.1 [1].

Enfin, les méthodes virtuelles de la classe `Space` sont implémentées dans les classes filles `Obs_Space` et `State_Space`, qui décrivent respectivement l'espace des observations et l'espace du modèle. En particulier, l'appel de l'opérateur d'observation sur un objet de type `Obs_Space` ou `State_Space` se fait par appel générique de la méthode `apply_correlation_operator` dans `Space`. L'utilisation du polymorphisme (ou de l'équivalent en Maltab) permet alors de distinguer entre les classes filles. Un autre exemple est celui de la génération de maillage, qui n'est pas le même selon qu'on traite des observations (maillage non structuré) ou de l'ébauche (maillage structuré).

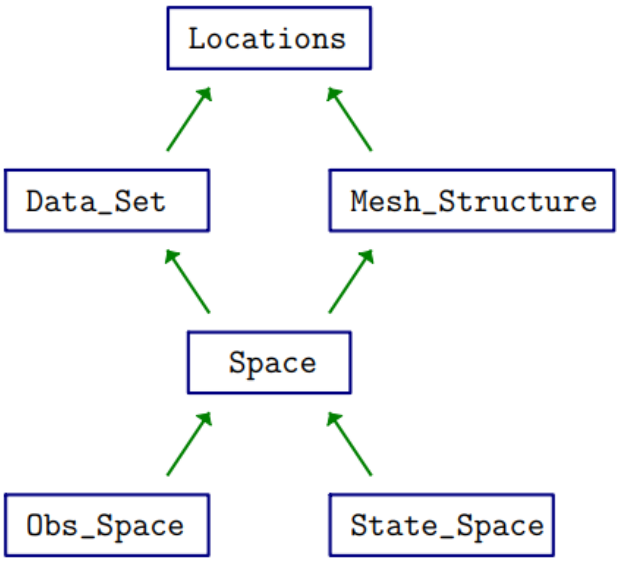


Figure E.1 – Diagramme de classe. Les cadres représentent les classes et les flèches les relations d'héritage.

1. Dans ce cas, l'héritage multiple est pratique et n'apparaît pas dangereux. On tient toutefois à signaler qu'il n'est pas indispensable et qu'il est tout à fait possible d'exploiter une autre structure, par exemple en attribuant à chaque objet de type `Space` un attribut de type `Data_Set` et un autre de type `Mesh_Structure`.

Le code est lancé à partir d'un fichier de configuration qui fixe toutes les variables d'entrée du système. Ces variables sont stockées par groupes dans des classes de type `Config_X`, ou le « `X` » permet de distinguer entre les données, le domaine de résolution de l'équation de diffusion, les maillages, les solveurs, les modèles de corrélation et l'affichage (figure E.2). Ces derniers sont représentés par des objets qu'on nomme `param_X`, placés en attribut de `Config_X`.

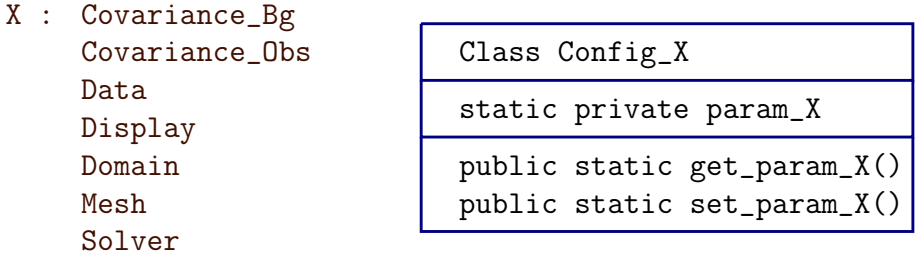


Figure E.2 – Modèle de classe de configuration.

Les différents algorithmes de résolution de sytèmes linéaires (`solver_PCG`, `solver_PGMRES`, `solver_deflated_PCG`) sont placés dans des fonctions indépendantes, pour assurer que les implémentation ne dépende de la définition d'aucune classe. Les opérations courantes (addition, soustraction, multiplication, division, produit externe, norme, comparaison...) sont surchargées dans `Space` pour que l'appel aux différents algorithmes soit transparent.

Enfin, les outils de maillage sont rassemblés dans une librairie, qu'on peut utiliser en dehors de l'assimilation de données.

Annexe F

Publication

Modelling spatially correlated observation errors in variational data assimilation using a diffusion operator on an unstructured mesh

O. Guillet[1,2*], A. T. Weaver[1], X. Vasseur[4], Y. Michel[2], S. Gratton[3] and S. Gürol[1]
[1] *CERFACS / CECI CNRS UMR 5318, 42 avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France*
[2] *CNRM UMR 3589, Météo-France and CNRS, 42 avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France*
[3] *INPT-IRIT, University of Toulouse and ENSEEIHT, 2 rue Camichel, BP 7122, 31071 Toulouse Cedex 7, France*
[4] *ISAE-SUPAERO, University of Toulouse, 10 avenue Edouard Belin, BP 54032, 31055 Toulouse Cedex 4, France*

*Correspondence to: Oliver Guillet, Météo-France, 42 avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France.
E-mail: oliver.guillet@meteo.fr

We propose a method for representing spatially correlated observation errors in variational data assimilation. The method is based on the numerical solution of a diffusion equation, a technique commonly used for representing spatially correlated background errors. The discretization of the pseudo-time derivative of the diffusion equation is done implicitly using a backward Euler scheme. The solution of the resulting elliptic equation can be interpreted as a correlation operator whose kernel is a correlation function from the Matérn family.

In order to account for the possibly heterogeneous distribution of observations, a spatial discretization technique based on the finite element method (FEM) is chosen where the observation locations are used to define the nodes of an *unstructured* mesh on which the diffusion equation is solved. By construction, the method leads to a convenient operator for the *inverse* of the observation error correlation matrix, which is an important requirement when applying it with standard minimization algorithms in variational data assimilation. Previous studies have shown that spatially correlated observation errors can also be accounted for by assimilating the observations together with their directional derivatives up to arbitrary order. In the continuous framework, we show that the two approaches are formally equivalent for certain parameter specifications. The FEM provides an appropriate framework for evaluating the derivatives numerically, especially when the observations are heterogeneously distributed.

Numerical experiments are performed using a realistic data distribution from the Spinning Enhanced Visible and InfraRed Imager (SEVIRI). Correlations obtained with the FEM-discretized diffusion operator are compared with those obtained using the analytical Matérn correlation model. The method is shown to produce an accurate representation of the target Matérn function in regions where the data are densely distributed. The presence of large gaps in the data distribution degrades the quality of the mesh and leads to numerical errors in the representation of the Matérn function. Strategies to improve the accuracy of the method in the presence of such gaps are discussed.

Keywords : observation errors; correlation functions; diffusion operator; variational assimilation; unstructured mesh; finite element method

Received . . .

© 2017 Royal Meteorological Society

1. Introduction

Specifying background and observation error covariance matrices ($\boldsymbol{B}$ and $\boldsymbol{R}$, respectively) that are accurate approximations of the true error covariance matrices is a challenging problem in operational data assimilation for the atmosphere and ocean. Over the past two decades, there has been considerable research devoted to the estimation of background error covariances, notably through the use of ensemble methods, and to the development of covariance models for representing them efficiently in $\boldsymbol{B}$ (*e.g.*, see the review articles by Bannister (2008a,b, 2017)). Comparatively fewer studies have addressed the estimation and modelling of observation error covariances, especially correlations. One key aspect of the problem for variational data assimilation is that standard minimization algorithms require an operator for the precision matrix $\boldsymbol{R}^{-1}$, either for the computation of the gradient of the cost function or for preconditioning (Michel 2018). Thus, even if we have an accurate $\boldsymbol{R}$ operator at our disposal, we still need to specify an efficient $\boldsymbol{R}^{-1}$ operator for computational purposes. Designing such an operator for large problems can be difficult.

In practice, certain assumptions are invoked that greatly simplify the structure of $\boldsymbol{R}$ that is specified in operational data assimilation systems. In particular, observation errors from one observing system are assumed to be uncorrelated with those from another observing system. In a multi-instrument observing system, this assumption is usually extended to the individual instruments themselves. As a result, $\boldsymbol{R}$ is defined as a block-diagonal matrix where the specification of the observation error covariances associated with each block can be treated independently for each observing system or instrument. Despite this simplification, each block typically corresponds to a large number of observations (*e.g.,* several millions for certain satellite observations).

Satellite radiance observations are well known to have correlated errors. For example, significant horizontal error correlations in radiances have been diagnosed by Bormann *et al.* (2010) for the Infrared Atmospheric Sounding Interferometer (IASI) and by Waller *et al.* (2016a) for the Spinning Enhanced Visible and InfraRed Imager (SEVIRI). Furthermore, highly correlated observation errors are expected from future satellite instruments, such as infrared sounders of the Meteosat Third Generation, which will provide high-resolution information about water vapour and temperature structures of the atmosphere (Stuhlmann *et al.* 2005).

For radiances, it is customary to separate the vertical (or inter-channel) correlations from the horizontal spatial correlations. In recent years, substantial progress has been made in representing inter-channel error correlations (Bormann *et al.* 2010; Bormann and Bauer 2010; Stewart *et al.* 2014; Weston *et al.* 2014; Waller *et al.* 2016a; Campbell *et al.* 2017). The size of the matrices required to represent these correlations is rather small (less than 10^3 rows or columns), which makes them straightforward to handle computationally using, for example, Cholesky decomposition. A similar technique has been used by Järvinen *et al.* (1999) to model temporal correlations in surface pressure observations.

Matrices associated with horizontal correlations are much larger than those associated with inter-channel or temporal correlations. Furthermore, due to the irregular nature of the spatial distribution of observations, they tend to have more complicated structure than those associated with correlated background error. These two features make horizontally correlated observation error difficult to handle computationally. For this reason, horizontal correlations are often neglected altogether, although this has to be done with caution, especially when considering high-density observations.

Rather than explicitly accounting for horizontal correlations in $\boldsymbol{R}$, mitigating strategies such as variance inflation, thinning and superobbing are typically employed (Rabier 2006). Inflating the observation error variances has the effect of downweighting the influence of the observations, as is the case when correlations are explicitly accounted for. Thinning is used to reduce the spatial and spectral resolution of the observations (and hence their error correlations) by selecting a reduced set of locations and channels. Superobbing combines locations or channels at different positions and can help reduce the observation error variances as well as their correlations. However, these procedures are ultimately suboptimal as they involve discarding potentially valuable observational information (Liu and Rabier 2002; Dando *et al.* 2007; Stewart *et al.* 2008).

Brankart *et al.* (2009) proposed a method to account for spatially correlated observation errors by focusing on modelling $\boldsymbol{R}^{-1}$ rather than $\boldsymbol{R}$. In particular, assuming that $\boldsymbol{R}$ is constructed from an exponential function, then $\boldsymbol{R}^{-1}$ is very sparse and can be accounted for indirectly by assimilating the observations together with their spatial derivatives, where the weights given to the spatial derivatives are related to the length-scale of the (exponential) correlation function. Chabot *et al.* (2015) discuss a related technique to account for spatially correlated errors in image observations. The method is appealing especially when the observations are sufficiently structured to simplify the computation of the spatial derivatives. For example, Ruggiero *et al.* (2016) used the Brankart *et al.* (2009) method to account for spatially correlated observation errors in simulated altimeter products from the future Surface Water and Ocean Topography (SWOT) satellite mission.

Following earlier work by M. Fisher at the European Centre for Medium-Range Weather Forecasts, Michel (2018) has shown that it is possible to carry out the main correlation operator computations on an auxiliary grid with simplified structure. The correlation operator in the space of the (possibly unstructured) observations is then obtained using an interpolation operator and its adjoint. While the method provides an efficient model for $\boldsymbol{R}$, it does not lead to a convenient and inexpensive expression for $\boldsymbol{R}^{-1}$, as required for variational data assimilation. Michel (2018) used a sequential Lanczos algorithm to build a low-rank approximation of $\boldsymbol{R}^{-1}$ in terms of its dominant eigenpairs. However, the method can be costly, as many eigenpairs may be required to obtain an adequate approximation of $\boldsymbol{R}^{-1}$.

In this article, we present an alternative method for modelling spatially correlated observation errors in variational data assimilation. Our starting point for modelling correlations in $\boldsymbol{R}$ is the framework for modelling correlations in $\boldsymbol{B}$ for which an extensive body of research exists. However, many of the standard methods used for modelling background error correlations, such as those based on spectral or (first-generation) wavelet transforms, require structured grids and thus are not appropriate for modelling $\boldsymbol{R}$. Diffusion operators can be used to model a class of correlation functions from the Matérn family (Guttorp and Gneiting 2006) and are popular for modelling $\boldsymbol{B}$ in ocean applications of variational data assimilation (Weaver and Courtier 2001; Carrier and Ngodock 2010). For numerical applications, the diffusion method provides useful flexibility regarding the choice of spatial and temporal discretization schemes. In particular, spatial discretization schemes based on the Finite Element Method (FEM) or Finite Volume Method (FVM) can be used to adapt the diffusion operator to an unstructured mesh, as desired for modelling $\boldsymbol{R}$. Furthermore, temporal discretization schemes based on backward Euler implicit methods provide immediate access to an inverse correlation operator, which greatly simplifies the specification of $\boldsymbol{R}^{-1}$.

A similar method for modelling spatial correlations on an unstructured mesh was developed by Lindgren *et al.* (2011) for

© 2017 Royal Meteorological Society

spatial interpolation (kriging) applications in geostatistics and by Bui-Thanh *et al.* (2013) for modelling prior (background) error covariances in a seismic inverse problem. Lindgren *et al.* (2011) (see also Simpson *et al.* (2012) and Bolin and Lindgren (2013)) use the fact that Gaussian fields with a specific covariance function are solutions to a linear stochastic partial differential equation (SPDE). Solving the SPDE is a convenient way of imposing this specific covariance structure on a random field. In fact, the SPDE can be interpreted as a stochastic diffusion equation and is related to the "square-root" of a diffusion-based covariance operator. In Lindgren *et al.* (2011), the SPDE is discretized on a triangular two-dimensional (2D) mesh, where the nodes of the mesh include the observation locations as well as other locations where interpolated values are desired. In our approach, the diffusion equation is also discretized on a triangular 2D mesh, which is built exclusively from the observation locations (*i.e.*, there are no additional nodes as in Lindgren *et al.* (2011)). This approach allows spatial correlations to be modelled directly between observation locations, as required for $\boldsymbol{R}$.

The structure of the article is as follows. Section 2 introduces the theoretical framework for correlation modelling with the diffusion equation. In particular, this section discusses the relationship between the diffusion equation and correlation functions from the Matérn family. Generalizations of the method are then introduced and discussed within the context of modelling $\boldsymbol{R}$ for certain observation types. Section 3 addresses the issue of discretizing the diffusion equation on unstructured grids using the FEM. Here, we derive explicit expressions for the matrix operators $\boldsymbol{R}$ and $\boldsymbol{R}^{-1}$ in terms of the components of the FEM-discretized diffusion operator. In Section 4, we establish a formal link between the diffusion-based approach and the method of Brankart *et al.* (2009) that involves assimilating successive derivatives of the observed field, up to a certain order. In Section 5, the diffusion method is applied to a realistic distribution of observations from SEVIRI and the accuracy of the method is assessed by comparing results with the analytical Matérn correlation model. Section 6 provides a summary and discusses future research directions for improving the accuracy of the method.

2. Correlation modelling with a diffusion operator

In this section, we introduce key aspects of the theory of diffusion-based correlation operators, as required for our study. We focus on correlation operators defined on the Euclidean space $\mathbb{R}^2$ and subdomains of $\mathbb{R}^2$ since the application considered in Section 5 concerns the modelling of 2D spatial observation error correlations on a plane. The reader can find a more general presentation in Weaver and Mirouze (2013), and references therein, where diffusion-based correlation operators are formulated on Euclidean spaces other than $\mathbb{R}^2$ and on the sphere $\mathbb{S}^2$.

In what follows, we will adopt the notation where continuous functions and operators are in italics, while vectors and matrices are in boldface.

2.1. *Correlation and covariance operators*

We consider correlation operators on the spatial domain Ω contained in $\mathbb{R}^2$. Let $f : \boldsymbol{z} \mapsto f(\boldsymbol{z})$ be a square-integrable function ($f \in L^2(\Omega)$) of the spatial coordinates $\boldsymbol{z} = (z_1, z_2)^{\mathrm{T}} \in \Omega$. A *correlation operator* $\mathcal{C} : f \mapsto \mathcal{C}[f]$ is an integral operator of the form

$$\mathcal{C}[f](\boldsymbol{z}) = \int_\Omega c(\boldsymbol{z}, \boldsymbol{z}') f(\boldsymbol{z}') \mathrm{d}\boldsymbol{z}', \tag{1}$$

where $\mathrm{d}\boldsymbol{z} = \mathrm{d}z_1 \mathrm{d}z_2$ is the Lebesgue measure on $\mathbb{R}^2$, and $c : (\boldsymbol{z}, \boldsymbol{z}') \mapsto c(\boldsymbol{z}, \boldsymbol{z}')$ is a correlation function where $(\boldsymbol{z}, \boldsymbol{z}') \in \Omega \times \Omega$. The correlation operator is symmetric and positive definite in the sense of the $L^2(\Omega)$-inner product:

$$\begin{aligned} \int_\Omega \mathcal{C}[f_1](\boldsymbol{z})\, f_2(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} &= \int_\Omega f_1(\boldsymbol{z})\, \mathcal{C}[f_2](\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}, \\ &\quad \forall f_1, f_2 \in L^2(\Omega), \\ \int_\Omega \mathcal{C}[f_1](\boldsymbol{z})\, f_1(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} &> 0, \quad \forall f_1 \neq 0 \in L^2(\Omega). \end{aligned} \tag{2}$$

Notably, the first of these properties implies that the correlation function is symmetric: $c(\boldsymbol{z}, \boldsymbol{z}') = c(\boldsymbol{z}', \boldsymbol{z})$ for any pair $(\boldsymbol{z}, \boldsymbol{z}') \in \Omega \times \Omega$. A correlation function also has unit amplitude ($c(\boldsymbol{z}, \boldsymbol{z}) = 1$).

For data assimilation, we need to define *covariance operators*. In particular, if $\bar{f} \in L^2(\Omega)$ then $\mathcal{R} : \bar{f} \mapsto \mathcal{R}[\bar{f}]$ is the observation error covariance operator defined as

$$\mathcal{R}[\bar{f}](\boldsymbol{z}) = \int_\Omega \bar{c}(\boldsymbol{z}, \boldsymbol{z}') \bar{f}(\boldsymbol{z}') \mathrm{d}\boldsymbol{z}', \tag{3}$$

where $\bar{c} : \boldsymbol{z}, \boldsymbol{z}' \mapsto \bar{c}(\boldsymbol{z}, \boldsymbol{z}') = \sigma(\boldsymbol{z})\sigma(\boldsymbol{z}')c(\boldsymbol{z}, \boldsymbol{z}')$ is the covariance function, and $\sigma(\boldsymbol{z}) = \sqrt{\bar{c}(\boldsymbol{z}, \boldsymbol{z})}$ is the standard deviation at the location $\boldsymbol{z}$, which we assume is non-zero so that $\mathcal{R}$ is strictly positive definite. Combining Eqs (1) and (3) yields the standard relationship

$$\left.\begin{aligned} f(\boldsymbol{z}') &= \sigma(\boldsymbol{z}')\bar{f}(\boldsymbol{z}') \\ \mathcal{R}[\bar{f}](\boldsymbol{z}) &= \sigma(\boldsymbol{z})\, \mathcal{C}[f](\boldsymbol{z}) \end{aligned}\right\}, \tag{4}$$

which allows us to separate the specification of $\sigma(\boldsymbol{z})$ and $\mathcal{C}$. In this study, we focus on computational aspects of specifying $\mathcal{C}$.

2.2. *Inverse correlation and inverse covariance operators*

The inverse of the correlation operator $\mathcal{C}$ is defined as the operator $\mathcal{C}^{-1} : g \mapsto \mathcal{C}^{-1}[g] = f$ where $g = \mathcal{C}[f]$. The inverse correlation operator is also symmetric and positive definite in the sense of the $L^2(\Omega)$-inner product:

$$\begin{aligned} \int_\Omega \mathcal{C}^{-1}[g_1](\boldsymbol{z})\, g_2(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} &= \int_\Omega g_1(\boldsymbol{z})\, \mathcal{C}^{-1}[g_2](\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}, \\ &\quad \forall g_1, g_2 \in L^2(\Omega), \\ \int_\Omega \mathcal{C}^{-1}[g_1](\boldsymbol{z})\, g_1(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} &> 0, \quad \forall g_1 \neq 0 \in L^2(\Omega). \end{aligned}$$

In general, $\mathcal{C}^{-1}$ is a differential operator, which cannot be expressed as an integral operator with an ordinary function as its kernel (as in Eq. (1)). However, it is possible to express $\mathcal{C}^{-1}$ as an integral operator if the kernel is considered to be a sum of generalized functions (Jones 1982).

If $\bar{g} = \mathcal{R}[\bar{f}]$ then the inverse of the covariance operator $\mathcal{R}$ is the operator $\mathcal{R}^{-1} : \bar{g} \mapsto \mathcal{R}^{-1}[\bar{g}]$ where

$$\left.\begin{aligned} g(\boldsymbol{z}') &= \frac{1}{\sigma(\boldsymbol{z}')}\, \bar{g}(\boldsymbol{z}') \\ \mathcal{R}^{-1}[\bar{g}](\boldsymbol{z}) &= \frac{1}{\sigma(\boldsymbol{z})}\, \mathcal{C}^{-1}[g](\boldsymbol{z}) \end{aligned}\right\}. \tag{5}$$

2.3. *Matérn correlation functions*

A well-known class of isotropic and homogeneous correlation functions is the Matérn class (Stein 1999; Guttorp and Gneiting 2006). Here, we are interested in a subclass of Matérn functions that have the specific form

$$c_{m,\ell}(r) = \frac{2^{2-m}}{(m-2)!} \left(\frac{r}{\ell}\right)^{m-1} K_{m-1}\left(\frac{r}{\ell}\right), \tag{6}$$

© 2017 Royal Meteorological Society

where $m > 1$ is an integer, $K_m(\cdot)$ is the modified Bessel function of the second kind of order m, $r = \|\boldsymbol{z} - \boldsymbol{z}'\|_2$ is the Euclidean distance between $\boldsymbol{z}$ and $\boldsymbol{z}'$, and ℓ is a scale parameter. The parameter m controls the scale-dependent smoothness properties of $c_{m,\ell}$, with larger values of m providing more selective damping at small scales. The parameter ℓ controls the spatial extent of the smoothing.

Matérn functions with $m > 2$ are differentiable at the origin ($r = 0$). For these functions, it is customary to define the length-scale D of $c_{m,\ell}$ in terms of the local curvature of the correlation function near the origin (the Daley length-scale). It is a quantity of practical interest since it can be estimated locally from derivatives of an ensemble of simulated errors (Belo Pereira and Berre 2006). In terms of ℓ and m, the Daley length-scale of Eq. (6) is given by (Weaver and Mirouze 2013)

$$D = \sqrt{2m - 4}\,\ell. \tag{7}$$

The Daley length-scale is the natural scale parameter of the Gaussian function defined by $c_g(r) = \exp(-r^2/2D^2)$. The Gaussian function can be derived as a limiting case of Eq. (6) as $m \to +\infty$ with ℓ simultaneously decreased to keep D fixed (Weaver and Mirouze 2013). The correlation functions with small values of m have fatter tails than those with larger values of m (for the same value of D), as illustrated in Figure 1.

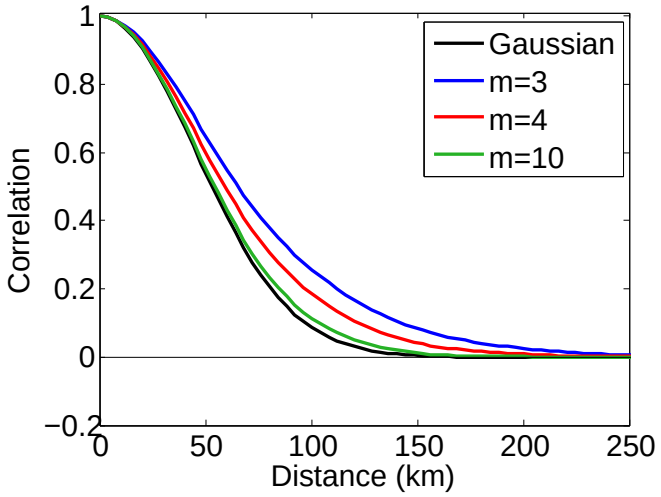


Figure 1. Cross section of a two-dimensional correlation function generated with Eq. (6) for different values of the parameter m and a fixed value of $D = 45$ km.

When $m = 2$, properties (2) still hold, but the correlation functions are no longer differentiable at the origin. For these functions, we can define the correlation length-scale as the scale parameter ℓ itself or some other characteristic measure. Correlation functions with $m = 2$ are displayed in Figure 2 for different values of ℓ. These functions have fat tails, and sharper correlations near the origin than those of the differentiable Matérn functions. We will come back to this point in Section 5 when considering the application to SEVIRI observations.

2.4. *The inverse correlation operator*

Let $\mathcal{C} : f \mapsto \mathcal{C}[f]$ be the correlation operator that has $c(\boldsymbol{z}, \boldsymbol{z}') = c_{m,\ell}(r)$ given by Eq. (6) as its kernel. Furthermore, suppose that Ω extends to include the whole of $\mathbb{R}^2$ and that $g = \mathcal{C}[f]$ and its derivatives vanish as $r \to \infty$. Since $c_{m,\ell}$ is homogeneous, $\mathcal{C}$ is a convolution operator,

$$g(\boldsymbol{z}) = \int_{\mathbb{R}^2} c_{m,\ell}(\boldsymbol{z} - \boldsymbol{z}') f(\boldsymbol{z}') \mathrm{d}\boldsymbol{z}', \tag{8}$$

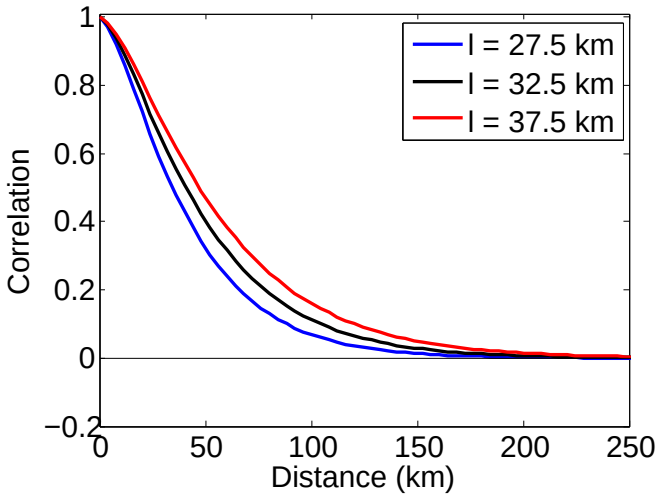


Figure 2. Cross section of a two-dimensional correlation function generated with Eq. (6) for different values of the parameter ℓ (see legend) and a fixed value of $m = 2$.

and we can use the Fourier Transform (FT) to derive an expression for $\mathcal{C}^{-1}$ (*e.g.*, see Jones (1982)).

Let $\hat{f} : \hat{\boldsymbol{z}} \mapsto \hat{f}(\hat{\boldsymbol{z}})$, $\hat{g} : \hat{\boldsymbol{z}} \mapsto \hat{g}(\hat{\boldsymbol{z}})$ and $\hat{c}_{m,\ell} : \hat{\boldsymbol{z}} \mapsto \hat{c}_{m,\ell}(\hat{\boldsymbol{z}})$ denote the FT of f, g and $c_{m,\ell}$, respectively, where $\hat{\boldsymbol{z}}$ is the vector of spectral wave-numbers. Taking the FT of Eq. (8) yields

$$\hat{g}(\hat{\boldsymbol{z}}) = \hat{c}_{m,\ell}(\hat{\boldsymbol{z}})\, \hat{f}(\hat{\boldsymbol{z}}) \tag{9}$$

where (Whittle 1963)

$$\hat{c}_{m,\ell}(\hat{\boldsymbol{z}}) = \frac{\gamma^2}{\left(1 + \ell^2\, \hat{\boldsymbol{z}}^2\right)^m} \tag{10}$$

and

$$\gamma^2 = 4\pi(m-1)\,\ell^2. \tag{11}$$

A necessary and sufficient condition for a homogeneous and isotropic function to yield a positive definite operator, in the sense of the second property in (2), is that its FT is non-negative (see Theorem 2.10 in Gaspari and Cohn (1999)). This is clearly satisfied by Eq. (10).

Let $\mathcal{C}^{-1} : g \mapsto \mathcal{C}^{-1}[g]$ be the inverse correlation operator and let $f = \mathcal{C}^{-1}[g]$. From Eq. (9), we have

$$\hat{f}(\hat{\boldsymbol{z}}) = \frac{1}{\hat{c}_{m,\ell}(\hat{\boldsymbol{z}})}\, \hat{g}(\hat{\boldsymbol{z}}). \tag{12}$$

Taking the inverse FT of Eq. (12) leads to the elliptic equation (Whittle 1963)

$$\frac{1}{\gamma^2}(I - \ell^2 \nabla^2)^m g(\boldsymbol{z}) = f(\boldsymbol{z}), \tag{13}$$

where I is the identity operator and $\nabla^2 \equiv \partial^2/\partial z_1^2 + \partial^2/\partial z_2^2$ is the 2D Laplacian operator. We can then identify $\mathcal{C}^{-1}$ as the elliptic operator in Eq. (13):

$$\mathcal{C}^{-1} = \frac{1}{\gamma^2}\left(I - \ell^2 \nabla^2\right)^m. \tag{14}$$

The constant γ^2 ensures that the correlation functions are properly normalized to have unit amplitude. Notice that γ^2 has physical units of length squared and can be interpreted as the natural "variance" of the Matérn covariance function associated with the (unnormalized) elliptic equation.

On a finite domain Ω, Eq. (13) must be supplied with appropriate boundary conditions. In this study, we use Neumann

© 2017 Royal Meteorological Society

conditions on the boundaries of Ω. Because of the boundary conditions, the correlation functions near the boundaries are not of precise Matérn form (Mirouze and Weaver 2010). This has implications on the normalization factor, which is no longer adequately described by a constant (Eq. (11)) near the boundaries. This point will be discussed further in Section 5.

2.5. *Computational aspects*

Consider a triangular mesh represented by a set of nodes $(\boldsymbol{z}_i)_{i\in[1,p]}$. In this study, $(\boldsymbol{z}_i)_{i\in[1,p]}$ are taken to be the observation locations at which spatial correlations need to be defined. Applying the inverse correlation operator $\mathcal{C}^{-1}$ numerically on the mesh requires m successive applications of a discretized representation of the operator $I-\ell^2\nabla^2$ (Eq. (13)). With appropriately chosen basis functions, this is a straightforward and computationally inexpensive operation since it involves multiplication by sparse matrices after discretization. In contrast, to apply the correlation operator $\mathcal{C}$ in the integral form (1) using the expression (6), one has to compute $c_{m,\ell}(\|\boldsymbol{z}_i-\boldsymbol{z}_j\|_2)$ for every pair $(\boldsymbol{z}_i,\boldsymbol{z}_j)$, which becomes unaffordable when the number of nodes (p) is large. For this reason, it is preferable to apply $\mathcal{C}$ by seeking a numerical solution of the elliptic equation (13) rather than attempting to integrate Eq. (1) numerically. Solving the elliptic equation requires solving m symmetric positive definite (SPD) linear systems in sequence, for which efficient methods are available (*e.g.*, see Saad (2003) for a general review). The numerical aspects of the solution algorithm will be discussed in Section 3.

2.6. *Interpretation as an implicit diffusion operator*

Equation (13) can be interpreted as a semi-discretized representation of a standard diffusion equation using a backward Euler temporal scheme. In particular, consider the 2D diffusion equation

$$\frac{\partial g}{\partial s} - \kappa\nabla^2 g = 0, \tag{15}$$

subject to the initial condition $g(\boldsymbol{z})\big|_{s=0}=\gamma^2 f(\boldsymbol{z})$, and to homogeneous Neumann boundary conditions $\nabla g\big|_{\partial\Omega}\cdot\hat{\boldsymbol{n}}=0$, where $\hat{\boldsymbol{n}}$ is the unit outward normal vector to the boundary $\partial\Omega$, ∇ is the 2D gradient operator, and $\cdot$ denotes the dot product. Here, s represents a non-dimensional pseudo-time coordinate, and κ is a constant pseudo-diffusion coefficient. Discretizing Eq. (15) using a backward Euler scheme with a pseudo-time step of unit size ($\Delta s=1$) leads to the semi-discrete elliptic equation

$$\left(I-\kappa\nabla^2\right)g_{n+1}(\boldsymbol{z}) = g_n(\boldsymbol{z}) \tag{16}$$

where n is the pseudo-time discretization index, and $g_0(\boldsymbol{z})=\gamma^2 f(\boldsymbol{z})$ is the "initial" condition. Considering the diffusion problem on the pseudo-time interval $n=[0,m-1]$ allows us to write Eq. (16) in the form of Eq. (13) with $\kappa=\ell^2$ and $g_m(\boldsymbol{z})=g(\boldsymbol{z})$. We can thus interpret the self-adjoint operator

$$\mathcal{L}^{-1} = \left(I-\kappa\nabla^2\right)^m \tag{17}$$

as an inverse diffusion operator acting backwards in pseudo-time over m-steps.

2.7. *More general functional shapes*

Numerically "time"-stepping an implicitly formulated diffusion equation is an efficient way to apply a correlation operator with Matérn kernel of the specific form (6). More general correlation functions than those from the Matérn family can be modelled by constructing $\mathcal{C}^{-1}$ as a linear combination of powers of the Laplacian operator (Weaver and Courtier 2001; Yaremchuk and Smith 2011; Weaver and Mirouze 2013):

$$\mathcal{C}^{-1} = \frac{1}{\widetilde{\gamma}^2}\left(I - a_1\ell^2\nabla^2 + a_2\ell^4\nabla^4 + \ldots + (-1)^p a_q\,\ell^{2q}\nabla^{2q}\right) \tag{18}$$

where p is a positive integer, $(a_k)_{k\in[1,q]}$ are constant coefficients, and $\widetilde{\gamma}$ is a normalization constant. The operator (14) is a special case of Eq. (18) with $\widetilde{\gamma}=\gamma$, $q=m$, and $(a_k)_{k\in[1,m]}$ defined by the binomial coefficient:

$$a_k \equiv b_k = \frac{m!}{k!(m-k)!}. \tag{19}$$

Following the procedure outlined in Section 2.4, we can easily derive, from Eq. (18), the FT of the kernel associated with $\mathcal{C}$:

$$\hat{c}(\hat{\boldsymbol{z}}) = \frac{\widetilde{\gamma}^2}{1+a_1\ell^2\,\hat{\boldsymbol{z}}^2+a_2\ell^4\,\hat{\boldsymbol{z}}^4+\ldots+a_q\ell^{2q}\,\hat{\boldsymbol{z}}^{2q}}. \tag{20}$$

Positiveness of $\hat{c}(\hat{\boldsymbol{z}})$ ensures that the kernel (the inverse FT of $\hat{c}$) is a valid correlation function. This is clearly guaranteed when the coefficients $(a_k)_{k\in[1,q]}$ are all positive. Negative coefficients can be used to generate functions with damped oscillatory behaviour about the zero-correlation axis, but special care is required to ensure that the resulting formulation leads to positive $\hat{c}(\hat{\boldsymbol{z}})$ (Weaver and Mirouze 2013; Barth *et al.* 2014). Negative correlations have been observed, for example, in the simulation of roll errors with wide-swath satellite altimeter measurements (Ruggiero *et al.* 2016).

2.8. *Anisotropy and inhomogeneity*

A further generalization is to replace $\ell^{2k}\nabla^{2k}$ in Eq. (18) with $\left(\nabla\cdot\boldsymbol{\kappa}\nabla\right)^k$, where $\nabla\cdot$ is the 2D divergence operator, and $\boldsymbol{\kappa}$ is a constant (anisotropic) diffusion tensor; *i.e.*, an SPD 2×2 matrix that allows the principal axes of the correlation functions to be stretched and rotated relative to the axes of the computational coordinates. This flexibility is desirable for representing spatial observation error correlations from polar orbiting satellites, whose principal axes may be preferentially aligned with the along- and across-track directions of the satellite path (Ruggiero *et al.* 2016).

For the Matérn family, the correlation functions associated with a constant diffusion tensor are still given by Eq. (6), but with the normalized distance measure r/ℓ replaced with $\sqrt{(\boldsymbol{z}-\boldsymbol{z}')^{\mathrm{T}}\boldsymbol{\kappa}^{-1}(\boldsymbol{z}-\boldsymbol{z}')}$. Furthermore, the parameter ℓ^2 in the normalization constant (11) for $\mathcal{C}^{-1}$ must be replaced with $\sqrt{\det(\boldsymbol{\kappa})}$ where det is the determinant.

Spatially constant correlation functions can be overly restrictive. This is particularly true when representing correlations of background error, which generally exhibit significant spatial variations due to the heterogeneous nature of atmospheric/ocean dynamics and of the observational network. Spatial variations can also be present in observation error correlations. For example, Waller *et al.* (2016c) showed that Doppler radar radial winds have error correlations that depend on both the height of the observation and on the distance of the observation away from the radar.

It is straightforward within the diffusion framework to account for inhomogeneous error correlations by making the diffusion tensor $\boldsymbol{\kappa}(\boldsymbol{z})$ spatially dependent. With this extension, the exact analytical form of the underlying correlation function is generally not known. However, when the spatial variation of $\boldsymbol{\kappa}(\boldsymbol{z})$ is sufficiently slow, the kernel of the integral solution of Eq. (13), with $\nabla\cdot\boldsymbol{\kappa}(\boldsymbol{z})\nabla$ used instead of $\ell^2\nabla^2$, can be expected to be approximately given by Eq. (6) in the vicinity of a given point (Mirouze and Weaver 2010; Weaver and Mirouze 2013; Yaremchuk and Nechaev 2013).

© 2017 Royal Meteorological Society

The exact normalization factors are no longer constant when the diffusion tensor is spatially dependent. For slowly varying $\boldsymbol{\kappa}(\boldsymbol{z})$, they can be approximated by (cf. Eq. (11))

$$\gamma(\boldsymbol{z}) \approx \sqrt{4\pi(m-1)\sqrt{\det(\boldsymbol{\kappa}(\boldsymbol{z}))}} \tag{21}$$

or a suitably smoothed version of Eq. (21) (Purser *et al.* 2003; Yaremchuk and Carrier 2012). Furthermore, to maintain symmetry of the correlation functions, they must be introduced symmetrically in the elliptic equation:

$$\frac{1}{\gamma(\boldsymbol{z})}\left(I - \nabla\cdot\boldsymbol{\kappa}(\boldsymbol{z})\nabla\right)^m \frac{1}{\gamma(\boldsymbol{z})}\, g(\boldsymbol{z}) = f(\boldsymbol{z}). \tag{22}$$

2.9. *Estimating parameters of the correlation model*

The Matérn correlation model requires specifying the smoothness parameter m and scale parameter ℓ. In the generalized correlation models described in Sections 2.7 and 2.8, the parameters to set are the coefficients $(a_k)_{k\in[1,q]}$ of the Laplacian operators up to order q (instead of the single parameter m), and the spatially dependent diffusion tensors $\boldsymbol{\kappa}(\boldsymbol{z})$ (instead of the single parameter ℓ).

The correlation model parameters need to be estimated from knowledge of the actual observation error statistics. For this purpose, the *a posteriori* diagnostic from Desroziers *et al.* (2005) is frequently used. This diagnostic provides an estimate of the total observation error covariances (*i.e.*, the combined components of measurement and representativeness error) from the cross-covariances between the analysis and the background residuals in observation space. The effectiveness of the technique for estimating observation error correlations is discussed by Waller *et al.* (2016b). The statistics are usually averaged in space and in time in order to increase the sample size and thus improve the robustness of the estimated covariances (Bormann *et al.* 2010; Bormann and Bauer 2010; Waller *et al.* 2016a; Michel 2018). Together with the fact that the method itself is based on some questionable assumptions, this suggests that this diagnostic should be used to provide only *coarse* estimates of the covariances. In this respect, the basic two-parameter Matérn correlation function (6) may be adequate for representing the statistics.

Some observation types may come equipped with an instrument error simulator. In particular, this is the case for the SWOT altimeter mission (Ubelmann *et al.* 2016). Assuming that the sources of measurement error are accurately modelled by the simulator, it can be used to provide more detailed sample estimates of the measurement component of the observation error covariances. Ruggiero *et al.* (2016) used the SWOT simulator to estimate parameters of the Brankart *et al.* (2009) covariance model. Complementary techniques for estimating the representativeness component of the observation error covariances are discussed in the recent review article by Janjić *et al.* (2018).

When reliable, comprehensive estimates of the observation error covariances are available, the multi-parameter formulations of the diffusion-based correlation model are appropriate. The approach considered in Sections 3 and 5 will focus on the two-parameter model, but can be adapted to the more general cases if necessary.

3. Finite element discretization

3.1. *Motivation*

We now investigate strategies to discretize the diffusion equation (13) in space. Given a set of observations at locations $(\boldsymbol{z}_i)_{i\in[1,p]}$, we wish to compute the solution of the diffusion equation at these same locations. Hence our choice for discretizing Eq. (13) is to build a computational mesh with nodes at observation locations, so that the solution can be computed directly at the nodes of this mesh. The FEM is one popular class of discretization strategies well known for handling such unstructured data distributions, and is the focus of this article.

Efficient solution techniques for partial differential equations (PDEs) convert a continuous operator problem to a discrete problem by a suitable projection onto a finite-dimensional subspace. Let Eq. (13) be written as

$$\left.\begin{array}{rcl} g_0(\boldsymbol{z}) &=& \gamma(\boldsymbol{z})\, f(\boldsymbol{z}) \\ \left(I-\ell^2\nabla^2\right) g_{n+1}(\boldsymbol{z}) &=& g_n(\boldsymbol{z}) \quad ; \quad n=[0,m-1] \\ g(\boldsymbol{z}) &=& \gamma(\boldsymbol{z})\, g_m(\boldsymbol{z}) \end{array}\right\}, \tag{23}$$

where the normalization factors have been introduced symmetrically as in Eq. (22). This is necessary for numerical applications, even with constant ℓ, since the exact normalization factors depend on the local accuracy of the numerical solution (which depends on the local quality of the mesh) and the boundary conditions. Here, we use Neumann boundary conditions on the spatial domain of interest Ω.

The FEM is a standard technique for solving PDEs numerically (Ciarlet 2002; Brenner and Scott 2013). The basic procedure involves defining a variational formulation of the infinite-dimensional continuous problem. This variational formulation is then solved by approximating the solution in a carefully chosen finite-dimensional subspace. Applying this procedure to Eq. (23) leads to a matrix formulation of the diffusion equation in the space of the observations. We will outline the procedure below and show how the resulting expressions can be used in formulations of the observation error covariance matrix and its inverse.

3.2. *Galerkin approximation*

Let $(\varphi_j)_{j\in\mathcal{I}}$ be an independent set of test functions used to discretize the diffusion equation. The $(\varphi_j)_{j\in\mathcal{I}}$ are called "degrees of freedom" and $\mathcal{I}$ is a set of indices of finite cardinality. Multiplying both sides of the elliptic equation in Eq. (23) by φ_i and integrating over Ω leads to the weak formulation of the PDE:

$$\int_\Omega \left(I-\ell^2\nabla^2\right) g_{n+1}(\boldsymbol{z})\,\varphi_j(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \;=\; \int_\Omega g_n(\boldsymbol{z})\,\varphi_j(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \tag{24}$$

for $j\in\mathcal{I}$ and $n=[0,m-1]$.

Now we introduce the Galerkin approximation in which $g_n(\boldsymbol{z})$ and $g_{n+1}(\boldsymbol{z})$ are represented by finite expansions in terms of $(\varphi_i)_{i\in\mathcal{I}}$:

$$\left.\begin{array}{rcl} g_n(\boldsymbol{z}) &=& \displaystyle\sum_{i\in\mathcal{I}} \alpha_n^{(i)}\,\varphi_i(\boldsymbol{z}) \\ \text{and}\quad g_{n+1}(\boldsymbol{z}) &=& \displaystyle\sum_{i\in\mathcal{I}} \alpha_{n+1}^{(i)}\,\varphi_i(\boldsymbol{z}) \end{array}\right\}. \tag{25}$$

Substituting these expressions into Eq. (24) and using Green's first identity together with Neumann boundary conditions yields the matrix equation

$$(\boldsymbol{M}+\boldsymbol{K})\boldsymbol{\alpha}_{n+1} \;=\; \boldsymbol{M}\boldsymbol{\alpha}_n, \tag{26}$$

where $\boldsymbol{\alpha}_n$ is a vector containing the coordinates $(\alpha_n^{(i)})_{i\in\mathcal{I}}$, $\boldsymbol{M}$ is the Gram mass matrix, and $\boldsymbol{K}$ is the stiffness matrix, with elements given by

$$\boldsymbol{M}_{ij} \;=\; \int_\Omega \varphi_i(\boldsymbol{z})\varphi_j(\boldsymbol{z})\mathrm{d}\boldsymbol{z}, \tag{27}$$

$$\text{and}\quad \boldsymbol{K}_{ij} \;=\; \ell^2\int_\Omega \nabla\varphi_i(\boldsymbol{z})\cdot\nabla\varphi_j(\boldsymbol{z})\mathrm{d}\boldsymbol{z}. \tag{28}$$

The stiffness matrix $\boldsymbol{K}$ is symmetric and positive semi-definite. The Gram mass matrix $\boldsymbol{M}$ is symmetric and positive definite

© 2017 Royal Meteorological Society

since $(\varphi_i)_{i\in\mathcal{I}}$ form an independent set of functions. It defines the weighting matrix for the $L^2(\Omega)$-inner product measured with respect to vectors $\boldsymbol{\alpha}_k$ and $\boldsymbol{\alpha}_l$ of basis coefficients; *i.e.*,

$$\int_\Omega g_k(\boldsymbol{z})\, g_l(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \;=\; \boldsymbol{\alpha}_k^{\mathrm{T}} \boldsymbol{M} \boldsymbol{\alpha}_l,$$

which using standard inner-product notation reads

$$\langle\, g_k,\, g_l \,\rangle \;=\; \langle\, \boldsymbol{\alpha}_k, \boldsymbol{\alpha}_l \,\rangle_{\boldsymbol{M}}. \tag{29}$$

Applying Eq. (26) on $n = [0, m-1]$ leads to a sequence of linear systems

$$\left.\begin{array}{rcl} \big(\boldsymbol{M}+\boldsymbol{K}\big)\boldsymbol{\alpha}_1 &=& \boldsymbol{M}\,\boldsymbol{\alpha}_0 \\ \big(\boldsymbol{M}+\boldsymbol{K}\big)\boldsymbol{\alpha}_2 &=& \boldsymbol{M}\,\boldsymbol{\alpha}_1 \\ & \vdots & \\ \big(\boldsymbol{M}+\boldsymbol{K}\big)\boldsymbol{\alpha}_m &=& \boldsymbol{M}\,\boldsymbol{\alpha}_{m-1} \end{array}\right\}. \tag{30}$$

After multiplying both sides of the equations in (30) by $\boldsymbol{M}^{-1}$, we can combine the resulting equations into a single equation

$$\big[\boldsymbol{M}^{-1}\big(\boldsymbol{M}+\boldsymbol{K}\big)\big]^m \boldsymbol{\alpha}_m \;=\; \boldsymbol{\alpha}_0,$$

which can be identified as the discretized weak formulation of $\mathcal{L}^{-1}$ in Eq. (17), defined for the vector $\boldsymbol{\alpha}$ of basis coefficients. We denote this matrix operator by

$$\boldsymbol{L}_{\boldsymbol{M}}^{-1} \;=\; \big[\boldsymbol{M}^{-1}\big(\boldsymbol{M}+\boldsymbol{K}\big)\big]^m \tag{31}$$

where the notation $\boldsymbol{L}_{\boldsymbol{M}}^{-1}$ indicates that this matrix is self-adjoint with respect to the $\boldsymbol{M}$-inner product (Eq. (29)); *i.e.*,

$$\boldsymbol{L}_{\boldsymbol{M}}^{-1} \;=\; \boldsymbol{M}^{-1}\big(\boldsymbol{L}_{\boldsymbol{M}}^{-1}\big)^{\mathrm{T}}\boldsymbol{M}. \tag{32}$$

The self-adjointness of $\boldsymbol{L}_{\boldsymbol{M}}^{-1}$ with respect to the $\boldsymbol{M}$-inner product corresponds to the self-adjointness of $\mathcal{L}^{-1}$ with respect to the $L^2(\Omega)$-inner product. The matrix $\boldsymbol{M}\boldsymbol{L}_{\boldsymbol{M}}^{-1}$ is symmetric in the usual sense.

3.3. *Discrete diffusion operator*

In this section, we drop the pseudo-time index n for clarity of notation. Let $\boldsymbol{g}$ be a vector of dimension card$(\mathcal{I})$, which contains the values of $\boldsymbol{g}$ at observation locations $(\boldsymbol{z}_i)$, $i = [1, p]$. Equation (25) describes the relation between the values $\boldsymbol{g}(\boldsymbol{z}_i)$ at observation locations and the coordinates $(\alpha^{(i)})_{i\in\mathcal{I}}$ of the basis functions $(\varphi_i)_{i\in\mathcal{I}}$. It can be written in matrix form as

$$\boldsymbol{g} \;=\; \boldsymbol{G}\,\boldsymbol{\alpha} \tag{33}$$

where the elements of $\boldsymbol{G}$ are defined through the relation

$$\boldsymbol{G}_{ij} \;=\; \varphi_j(\boldsymbol{z}_i) \tag{34}$$

with $j \in \mathcal{I}$ and $i = [1, p]$. In the following, we will only consider the standard $\mathbb{P}_1$-FEM approximation for which $\boldsymbol{G}$ is the identity matrix (card$(\mathcal{I}) = p$). Therefore, we will later omit $\boldsymbol{G}$. Nevertheless, we note that other approximations lead to more complex expressions for $\boldsymbol{G}$ (*e.g.*, when the $(\varphi_i)_{i\in\mathcal{I}}$ are harmonic functions and $\boldsymbol{G}$ is the corresponding spectral transform).

From now on, let us assume that a triangular mesh supporting the (observation) nodes is available, and that each node i in this triangulation corresponds to the point $\boldsymbol{z}_i$. Here, we choose the basis functions $(\varphi_i)_{i\in\mathcal{I}}$ to be continuous and linear inside each triangle, with the property (Ern and Guermond 2010, Chapter 8)

$$\varphi_i(\boldsymbol{z}_j) = \delta_{ij}, \tag{35}$$

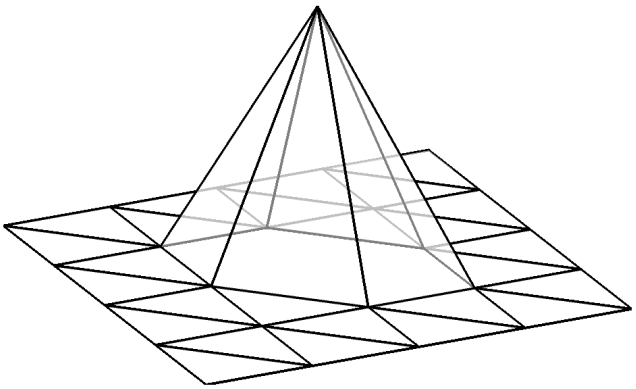


Figure 3. Representation of one $\mathbb{P}_1$-FEM basis function and its compact support. The function has a value equal to 1 at node $\boldsymbol{z}_i$ (the point (3,3) in the figure) and a value of 0 at other nodes.

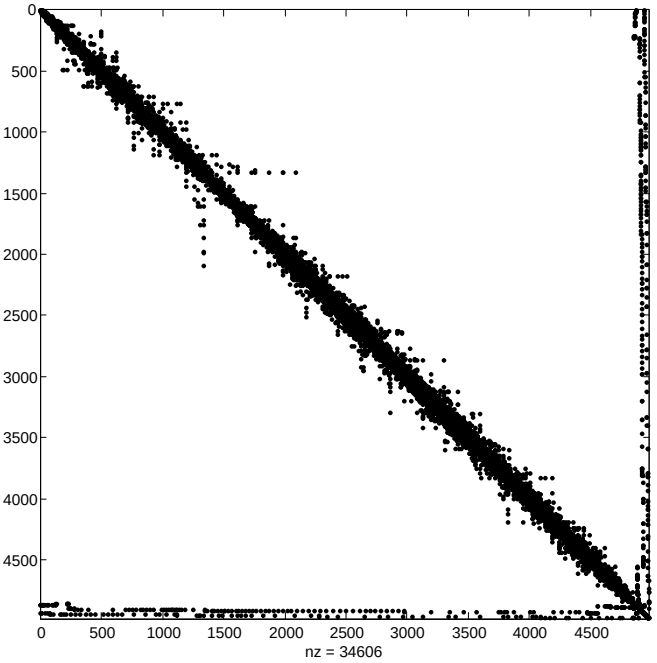


Figure 4. Profile of the mass matrix $\boldsymbol{M}$ for the unstructured satellite (SEVIRI) observations considered in Section 5. The profile of the stiffness matrix $\boldsymbol{K}$ is similar. The labelling on the horizontal and vertical axes corresponds to the column and row indices. The total number of non-zero ('nz') entries of the matrix is indicated below the horizontal axis. The non-zero entries that appear far from the diagonal correspond to boundary elements in the mesh (see section 3.6 for a discussion).

where δ_{ij} is the Kronecker delta. An immediate consequence of Eq. (35) is that $\boldsymbol{G}$ is the $p \times p$ identity matrix ($\boldsymbol{I}$). As already mentioned, this simple choice of basis functions corresponds to the $\mathbb{P}_1$-FEM and guarantees that every function has a local compact support (see Figure 3). In Eqs (27) and (28), the integrals can be computed exactly over the triangular nodes using standard integration techniques, since the integrands are polynomials of at most second order (Canuto *et al.* 1987). (Higher order finite elements may require using quadrature formula to evaluate the integrals.) Hence, the non-diagonal entries $\boldsymbol{M}_{ij}$ and $\boldsymbol{K}_{ij}$ equal zero as soon as $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ do not belong to the same triangle. Therefore, the choice of the $\mathbb{P}_1$ element is responsible for the sparsity of the matrices $\boldsymbol{M}$ and $\boldsymbol{K}$. The profile of $\boldsymbol{M}$ for the unstructured distribution of satellite observations considered in Section 5 is depicted in Figure 4.

© 2017 Royal Meteorological Society

We can write the discrete form of the $L^2(\Omega)$-inner product on the left-hand side of property (29) as

$$\langle \boldsymbol{g}, \boldsymbol{g} \rangle_{\boldsymbol{W}} \;=\; \boldsymbol{g}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{g} \tag{36}$$

where $\boldsymbol{W}$ is a symmetric and positive definite matrix of grid-dependent weighting factors. Since $\boldsymbol{G} = \boldsymbol{I}$ from our choice of basis functions, Eqs (29), (33) and (36) imply the simple relation

$$\boldsymbol{W} \;=\; \boldsymbol{M}. \tag{37}$$

3.4. *Formulation of $\boldsymbol{R}$ and $\boldsymbol{R}^{-1}$ in variational assimilation*

Consider the discrete form of the cost function in variational data assimilation, focusing on the observation term J_{o}. Let $\boldsymbol{d}$ be a p-dimensional vector where p is the number of assimilated observations. The ith component of $\boldsymbol{d}$ corresponds to the difference between the ith observation and its model counterpart. The observation term is formulated as the $\boldsymbol{R}^{-1}$-norm of the difference vector:

$$J_{\mathrm{o}} \;=\; \frac{1}{2} \boldsymbol{d}^{\mathrm{T}} \, \boldsymbol{R}^{-1} \, \boldsymbol{d}$$

where $\boldsymbol{R}$ is the (symmetric positive definite) observation error covariance matrix.

We can express $\boldsymbol{R}^{-1}$ in the standard factored form

$$\boldsymbol{R}^{-1} \;=\; \boldsymbol{\Sigma}^{-1} \, \boldsymbol{C}^{-1} \, \boldsymbol{\Sigma}^{-1}, \tag{38}$$

where $\boldsymbol{C}^{-1}$ is the inverse of the observation error correlation matrix, and $\boldsymbol{\Sigma}$ is a diagonal matrix containing the observation error standard deviations. Assuming that the errors are correlated and modelled by a discretized diffusion operator, then we can express $\boldsymbol{C}^{-1}$ as

$$\boldsymbol{C}^{-1} \;=\; \boldsymbol{\Gamma}^{-1} \boldsymbol{W} \, \boldsymbol{L}_{\boldsymbol{W}}^{-1} \, \boldsymbol{\Gamma}^{-1}, \tag{39}$$

where $\boldsymbol{\Gamma}^2$ is a diagonal matrix of normalization factors, and $\boldsymbol{L}_{\boldsymbol{W}}^{-1}$ is the inverse diffusion operator, which is self-adjoint with respect to the $\boldsymbol{W}$-inner product (Lorenc 1997; Weaver and Courtier 2001). The appearance of $\boldsymbol{W}$ in Eq. (39) comes from the discrete representation of spatial integration implied by the $L^2(\Omega)$-inner product. Here, we have implicitly assumed that $\boldsymbol{d}$ is the discrete representation of a square-integrable function.

When the diffusion operator is discretized using the FEM then, from Eq. (37), Eq. (39) becomes

$$\boldsymbol{C}^{-1} \;=\; \boldsymbol{\Gamma}^{-1} \boldsymbol{M} \, \boldsymbol{L}_{\boldsymbol{M}}^{-1} \, \boldsymbol{\Gamma}^{-1}. \tag{40}$$

Expressions for the covariance and correlation matrices follow directly from Eqs (31), (38) and (40):

$$\boldsymbol{R} \;=\; \boldsymbol{\Sigma} \, \boldsymbol{C} \, \boldsymbol{\Sigma} \tag{41}$$

where

$$\boldsymbol{C} \;=\; \boldsymbol{\Gamma} \, \boldsymbol{L}_{\boldsymbol{M}} \, \boldsymbol{M}^{-1} \, \boldsymbol{\Gamma} \tag{42}$$

and

$$\boldsymbol{L}_{\boldsymbol{M}} \;=\; \left[\left(\boldsymbol{M} + \boldsymbol{K} \right)^{-1} \boldsymbol{M} \right]^{m}. \tag{43}$$

3.4.1. *Square-root formulation*

It is convenient to construct $\boldsymbol{R}$ and $\boldsymbol{R}^{-1}$ as factored products

$$\left. \begin{aligned} \boldsymbol{R} \;&=\; \boldsymbol{V} \, \boldsymbol{V}^{\mathrm{T}} \\ \text{and} \quad \boldsymbol{R}^{-1} \;&=\; \left(\boldsymbol{V}^{-1} \right)^{\mathrm{T}} \boldsymbol{V}^{-1} \end{aligned} \right\}.$$

This factorization ensures that $\boldsymbol{R}$ and $\boldsymbol{R}^{-1}$ remain symmetric and positive definite in numerical applications. Furthermore, it gives us access to a "square-root" operator $\boldsymbol{V}$, which is a valuable tool for randomization applications; *i.e.*, for generating a spatially correlated random vector with covariance matrix equal to $\boldsymbol{R}$ given white noise as an input vector.

By restricting the number of diffusion steps m to be even, it follows from Eqs (41), (42) and (43), and Eqs (31), (38) and (39) that

$$\left. \begin{aligned} \boldsymbol{V} \;&=\; \boldsymbol{\Sigma} \, \boldsymbol{\Gamma} \, \boldsymbol{L}_{\boldsymbol{M}}^{1/2} \left(\boldsymbol{M}^{-1/2} \right)^{\mathrm{T}} \\ \text{and} \quad \boldsymbol{V}^{-1} \;&=\; \left(\boldsymbol{M}^{1/2} \right)^{\mathrm{T}} \boldsymbol{L}_{\boldsymbol{M}}^{-1/2} \, \boldsymbol{\Gamma}^{-1} \, \boldsymbol{\Sigma}^{-1} \end{aligned} \right\},$$

where

$$\left. \begin{aligned} \boldsymbol{L}_{\boldsymbol{M}}^{1/2} \;&=\; \left[\left(\boldsymbol{M} + \boldsymbol{K} \right)^{-1} \boldsymbol{M} \right]^{m/2} \\ \text{and} \quad \boldsymbol{L}_{\boldsymbol{M}}^{-1/2} \;&=\; \left[\boldsymbol{M}^{-1} \left(\boldsymbol{M} + \boldsymbol{K} \right) \right]^{m/2} \end{aligned} \right\}.$$

In deriving these expressions, we have used the relation

$$\boldsymbol{L}_{\boldsymbol{M}} \boldsymbol{M}^{-1} \;=\; \boldsymbol{L}_{\boldsymbol{M}}^{1/2} \left(\boldsymbol{M}^{-1/2} \right)^{\mathrm{T}} \boldsymbol{M}^{-1/2} \left(\boldsymbol{L}_{\boldsymbol{M}}^{1/2} \right)^{\mathrm{T}},$$

which follows from the identity (32) and the standard factorization $\boldsymbol{M} \;=\; \boldsymbol{M}^{1/2} \left(\boldsymbol{M}^{1/2} \right)^{\mathrm{T}}$ for an SPD matrix.

3.5. *Mass lumping*

The expressions for $\boldsymbol{C}$ (Eqs (42) and (31)) and $\boldsymbol{C}^{-1}$ (Eqs (40) and (43)) involve the inverse of the mass matrix, $\boldsymbol{M}^{-1}$. The choice of our basis functions renders $\boldsymbol{M}$ sparse and hence amenable to the use of a sparse direct SPD solver based on Cholesky decomposition. Nevertheless, it can be convenient to simplify computations further by using a technique known as *mass lumping*, which involves approximating the *consistent* mass matrix $\boldsymbol{M}$ in Eq. (31) with a diagonal matrix $\widetilde{\boldsymbol{M}}$ called the lumped mass matrix. For the case of the $\mathbb{P}_1$-FEM considered here, it is simply obtained by summing the coefficients of $\boldsymbol{M}$ on each row or column. This process is equivalent to computing the integral (27) using a low-order quadrature formula or replacing the basis functions $(\varphi_i)_{i \in \mathcal{I}}$ with piecewise constant functions $(\widetilde{\varphi}_i)_{i \in \mathcal{I}}$ on each element (Canuto *et al.* 1987). Hence the formula for the coefficients of $\widetilde{\boldsymbol{M}}$ becomes

$$\widetilde{\boldsymbol{M}}_{ii} \;=\; \sum_j \boldsymbol{M}_{ij} \;=\; \sum_j \int_{\Omega} \varphi_i(\boldsymbol{z}) \varphi_j(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}.$$

In Section 5, we evaluate this approximation in terms of its effect on the representation of Matérn correlations.

3.6. *Boundary nodes and the inverse correlation operator*

The nodes added at the artificial boundaries of the domain are required to make the elliptic problem well posed. However, they do not correspond to actual observation locations, so must be discarded once the correlation operator has been applied. This specific feature of the correlation operator has not been made explicit in the formulation of $\boldsymbol{R}$ presented in Section 3.4, but has implications for the specification of the inverse correlation operator as discussed in this section.

Let $\boldsymbol{C}_{\mathrm{b}}$ denote the correlation operator that includes the extra boundary nodes. It is formulated as an FEM-discretized diffusion operator according to Eq. (42). A correlation operator $\boldsymbol{C}$ associated with the actual observations can be obtained from $\boldsymbol{C}_{\mathrm{b}}$ using the formulation

$$\boldsymbol{C} \;=\; \boldsymbol{S} \, \boldsymbol{C}_{\mathrm{b}} \, \boldsymbol{S}^{\mathrm{T}} \tag{44}$$

where $\boldsymbol{S}$ is a selection matrix (a rectangular matrix of 0s and 1s) that, together with $\boldsymbol{S}^{\mathrm{T}}$, picks out the submatrix of $\boldsymbol{C}_{\mathrm{b}}$ whose

© 2017 Royal Meteorological Society

elements correspond to the correlations at the actual observation locations. If p_b denotes the number of boundary nodes, then $\boldsymbol{C}_\mathrm{b}$ is a full-rank $(p+p_\mathrm{b})\times(p+p_\mathrm{b})$ matrix, while $\boldsymbol{C}$ is a full-rank $p\times p$ matrix.

The inverse correlation matrix associated with Eq. (44) is

$$\boldsymbol{C}^{-1} \;=\; \big(\boldsymbol{S}\,\boldsymbol{C}_\mathrm{b}\,\boldsymbol{S}^\mathrm{T}\big)^{-1}, \tag{45}$$

which is not an explicit operator. To apply it requires solving a linear system. Rather than using Eq. (45), we can approximate the inverse as

$$\boldsymbol{C}^{-1} \;\approx\; \widetilde{\boldsymbol{C}}^{-1} \;=\; \boldsymbol{S}\,\boldsymbol{C}_\mathrm{b}^{-1}\,\boldsymbol{S}^\mathrm{T}, \tag{46}$$

which is straightforward to apply using the expressions presented in Section 3.4. However, the extent to which Eq. (46) is a good approximation of Eq. (45) is not obvious. As a trivial example, consider $\boldsymbol{C}_\mathrm{b}$ to be a 2×2 correlation matrix (*i.e.*, 2 nodes) with correlation coefficient (off-diagonal element) equal to ρ. If $\boldsymbol{S}$ and $\boldsymbol{S}^\mathrm{T}$ act to select one of the nodes then $\boldsymbol{C}$ and $\boldsymbol{C}^{-1}$ have a single element equal to one. For comparison, $\widetilde{\boldsymbol{C}}^{-1}$ has a single element equal to $1/(1-\rho)$, which shows that it is a good approximation of $\boldsymbol{C}^{-1}$ if the two points are weakly correlated ($\rho\approx 0$).

For the mesh used in our experiments, $\widetilde{\boldsymbol{C}}^{-1}$ and $\boldsymbol{C}^{-1}$ are practically equivalent: for a random vector $\boldsymbol{v}$, $\|\widetilde{\boldsymbol{C}}^{-1}\boldsymbol{C}\boldsymbol{v}-\boldsymbol{v}\|_\infty < 10^{-14}$. This is perhaps not surprising in view of the simple analysis above, since the distance between the artificial boundary nodes and the interior nodes is typically much larger than the correlation length-scale that is used in our experiments (see Section 5).

3.7. *Computational aspects*

The discrete inverse covariance matrix $\boldsymbol{R}^{-1}$ (taking into account the approximation (46)) is built from a combination of diagonal matrices $\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Gamma}^{-1}$ and $\widetilde{\boldsymbol{M}}$ (assuming a lumped mass matrix), and a product of m matrices involving the left-scaled, shifted stiffness matrix, $\widetilde{\boldsymbol{M}}^{-1}\big(\widetilde{\boldsymbol{M}}+\boldsymbol{K}\big)$. The resulting operator is well suited for a parallelization strategy based on domain decomposition in a distributed memory environment, where the observations are split between processors according to their spatial location, and Message Passing Interface communications are performed at the domain boundaries before each application of the stiffness matrix. As an operator, $\boldsymbol{R}^{-1}$ can be applied cheaply and is therefore ideal in variational data assimilation for minimization algorithms that require $\boldsymbol{R}^{-1}$, but not $\boldsymbol{R}$.

Applying $\boldsymbol{R}$ is computationally more demanding than applying $\boldsymbol{R}^{-1}$. The discrete covariance matrix $\boldsymbol{R}$ is built from a combination of diagonal matrices $\boldsymbol{\Sigma}$, $\boldsymbol{\Gamma}$ and $\widetilde{\boldsymbol{M}}^{-1}$, and a product of m matrices involving the right-scaled, *inverse* of the shifted stiffness matrix, $\big(\widetilde{\boldsymbol{M}}+\boldsymbol{K}\big)^{-1}\widetilde{\boldsymbol{M}}$. An efficient way to apply the latter is to solve, in sequence, each of the linear systems in (30) involving the sparse symmetric positive definite matrix $\widetilde{\boldsymbol{M}}+\boldsymbol{K}$.

The sparsity of $\widetilde{\boldsymbol{M}}+\boldsymbol{K}$ depends on the orthogonality of the basis functions $(\varphi_i)_{i\in\mathcal{I}}$ with respect to the $L^2(\Omega)$-inner product. We have chosen here to use compactly-supported piecewise polynomial functions, which results in a large number of zero entries in $\boldsymbol{K}$ (and $\boldsymbol{M}$). Choosing different types of functions would result in alternative covariance operators that would generally be more costly to apply.

For most applications related to 2D computational domains, the linear systems in (30) can be solved up to machine precision using a direct method based on Cholesky decomposition (Duff *et al.* 1989; Davis 2006). Iterative methods can be used to solve the linear system approximately when the size of the matrix is very large. Weaver *et al.* (2016) highlight the importance of using a *linear* iterative solver, together with the adjoint of the solver, in a square-root formulation of the correlation matrix in order to preserve numerical symmetry of the correlation matrix when using a modest convergence criterion. Linear iterative solvers based on multi-grid (Gratton *et al.* 2011) or the Chebyshev Iteration (Weaver *et al.* 2016, 2018) are particularly well suited for this problem.

For the experiments described in Section 5, a direct method has been used to solve the linear systems in (30) to an accuracy largely below the discretization error of the FEM.

4. Link between a diffusion-based covariance model and assimilating derivatives of observations

The method of Brankart *et al.* (2009) (referred to as the Brankart method hereafter) for accounting for correlated observation errors has gained popularity in recent years, particularly in oceanography for the assimilation of high-resolution altimeter data from SWOT (Ruggiero *et al.* 2016). The Brankart method involves assimilating the observations together with successive derivatives of the observations. This method can be viewed, under certain assumptions, as a diffusion-based approach for modelling correlated error. The purpose of this section is to establish a formal mathematical link between the two methods in order to help improve our understanding of the advantages and disadvantages of each method.

4.1. *Continuous formulation*

The approach presented in Brankart *et al.* (2009) involves linearly transforming the observations into an augmented set of observations. In this subsection, we consider the approach in a continuous framework before treating the discrete problem in the next subsection. Let the observations be denoted by a continuous function $y:\boldsymbol{z}\mapsto y(\boldsymbol{z})$ where $y\in L^2(\Omega)$. We introduce the linear transform operator $\mathcal{T}[y]$ such that the resulting function contains both y and successive derivatives of y. Brankart *et al.* (2009) focus mainly on assimilating the first-order derivatives of y, while Ruggiero *et al.* (2016) consider both first- and second-order derivatives of y. While it is possible to assimilate higher order derivatives, for reasons of clarity, we choose to consider only the first- and second-order derivative information, as in Ruggiero *et al.* (2016). This will be sufficient to illustrate the link with the diffusion approach. We adopt similar notation to that of Brankart *et al.* (2009) and, as in the previous section, we focus on the domain Ω contained in $\mathbb{R}^2$.

The Brankart method involves formulating the inverse observation error correlation operator as

$$\mathcal{R}_\mathrm{B}^{-1} = \mathcal{T}^\mathrm{T}\big(\mathcal{R}^+\big)^{-1}\mathcal{T} \tag{47}$$

where

$$\big(\mathcal{R}^+\big)^{-1} \;=\; \mathrm{diag}\big(\,a_0 I,\; a_1 I,\; a_2 I,\; a_3 I,\; a_4 I,\; a_5 I\,\big),$$

and the operator $\mathcal{T}$ and its transpose are defined as

$$\mathcal{T} \;=\; \begin{pmatrix} I \\ \partial/\partial z_1 \\ \partial/\partial z_2 \\ \partial^2/\partial z_1^2 \\ \partial^2/\partial z_2^2 \\ \partial^2/\partial z_1\partial z_2 \end{pmatrix} \tag{48}$$

and

$$\mathcal{T}^\mathrm{T} = \begin{pmatrix} I & -\dfrac{\partial}{\partial z_1} & -\dfrac{\partial}{\partial z_2} & \dfrac{\partial^2}{\partial z_1^2} & \dfrac{\partial^2}{\partial z_2^2} & \dfrac{\partial^2}{\partial z_1\partial z_2}, \end{pmatrix}.$$

© 2017 Royal Meteorological Society

The elements $(a_i)_{i=[0,5]}$ can be functions of $\boldsymbol{z}$ but here we consider them to be constant. The last component of $\mathcal{T}$ involving cross-derivatives is not considered by Brankart *et al.* (2009) or Ruggiero *et al.* (2016) but is required here to compare with the 2D diffusion-based formulation since the latter involves powers of the Laplacian operator in a general coordinate system $\boldsymbol{z} = (z_1, z_2)^{\mathrm{T}}$ where z_1 and z_2 are not necessarily aligned with the principal axes of the 2D correlation functions. The operator $\left(\mathcal{R}^{+}\right)^{-1}$ is to be interpreted as the inverse error covariance operator of the augmented set of observations y, $\partial y/\partial z_1$, $\partial y/\partial z_2$, $\partial^2 y/\partial z_1^2$, $\partial^2 y/\partial z_2^2$, $\partial^2 y/\partial z_1 \partial z_2 \in L^2(\Omega)$. The operator $\mathcal{R}_{\mathrm{B}}^{-1}$ is symmetric with respect to the $L^2(\Omega)$-inner product. Note that the components of the second derivatives are symmetric, while those of the first derivatives are anti-symmetric (Tarantola 2005, pp. 130-131).

Expanding Eq. (47) allows us to write

$$\mathcal{R}_{\mathrm{B}}^{-1} = a_0 - a_1 \frac{\partial^2}{\partial z_1^2} - a_2 \frac{\partial^2}{\partial z_2^2} + a_3 \frac{\partial^4}{\partial z_1^4} + a_4 \frac{\partial^4}{\partial z_2^4} + a_5 \frac{\partial^4}{\partial z_1^2 \partial z_2^2}. \tag{49}$$

Now consider the inverse of an implicit diffusion-based covariance operator assuming that the variance σ^2 is constant:

$$\mathcal{R}^{-1} = \frac{1}{\sigma^2 \gamma^2} \left(I - \ell^2 \nabla^2\right)^m \tag{50}$$

(see Eq. (14) for the corresponding inverse correlation operator). By comparing Eqs (49) and (50), it is easy to see that they are equivalent when $m = 2$ and when the elements of $\left(\mathcal{R}^{+}\right)^{-1}$ are chosen to be

$$\begin{aligned} a_0 &= \frac{1}{\sigma^2 \gamma^2}, \\ a_1 = a_2 &= \frac{2\,\ell^2}{\sigma^2 \gamma^2}, \\ a_3 = a_4 &= \frac{\ell^4}{\sigma^2 \gamma^2}, \\ \text{and} \quad a_5 &= \frac{2\,\ell^4}{\sigma^2 \gamma^2}. \end{aligned}$$

The equivalence of the two methods is easily generalized to account for an arbitrary value of m by augmenting Eq. (48) to include derivatives and cross-derivatives of y up to order m, and by extending $\left(\mathcal{R}^{+}\right)^{-1}$ to include additional coefficients defined appropriately in terms of the binomial coefficients (Eq. (19)).

Unlike the diffusion-based approach, the Brankart method does not distinguish the inverse of the correlation operator from the inverse of the covariance operator. The parameters σ^2, γ^2 and ℓ are defined jointly via the coefficients a_k. Procedures for estimating these coefficients as spatially-dependent quantities are described by Ruggiero *et al.* (2016) and Yaremchuk *et al.* (2018). In the diffusion-based approach, the parameters σ^2 and ℓ (or, in general, the diffusion tensor $\boldsymbol{\kappa}$), and spatially-dependent generalizations of these parameters, can be estimated separately based on knowledge of the underlying (Matérn) covariance function to which sample estimates of the covariances can be fit. The relationship between the two methods becomes more difficult to quantify as soon as the parameters are made spatially dependent.

4.2. *Discrete formulation*

Consider the expression for the inverse of the covariance matrix associated with an FEM diffusion-based formulation for $m = 2$. From Eqs. (31), (38) and (40)

$$\begin{aligned} \boldsymbol{R}^{-1} &= \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma}^{-1} \left(\boldsymbol{M} + \boldsymbol{K}\right) \boldsymbol{M}^{-1} \left(\boldsymbol{M} + \boldsymbol{K}\right) \boldsymbol{\Gamma}^{-1} \boldsymbol{\Sigma}^{-1} \\ &= \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma}^{-1} \left(\boldsymbol{M} + 2\boldsymbol{K} + \boldsymbol{K}\boldsymbol{M}^{-1}\boldsymbol{K}\right) \boldsymbol{\Gamma}^{-1} \boldsymbol{\Sigma}^{-1}, \end{aligned}$$

which can be written as

$$\boldsymbol{R}^{-1} = \boldsymbol{\Sigma}^{-1} \boldsymbol{\Gamma}^{-1} \widehat{\boldsymbol{T}}^{\mathrm{T}} \left(\widehat{\boldsymbol{R}}^{+}\right)^{-1} \widehat{\boldsymbol{T}} \boldsymbol{\Gamma}^{-1} \boldsymbol{\Sigma}^{-1} \tag{51}$$

where

$$\widehat{\boldsymbol{T}} = \begin{pmatrix} \left(\boldsymbol{M}^{1/2}\right)^{\mathrm{T}} \\ \left(\boldsymbol{K}^{1/2}\right)^{\mathrm{T}} \\ \boldsymbol{M}^{-1/2}\boldsymbol{K} \end{pmatrix}, \qquad \left(\widehat{\boldsymbol{R}}^{+}\right)^{-1} = \begin{pmatrix} \boldsymbol{I} & & \\ & 2\boldsymbol{I} & \\ & & \boldsymbol{I} \end{pmatrix},$$

$\boldsymbol{K} = \boldsymbol{K}^{1/2}\left(\boldsymbol{K}^{1/2}\right)^{\mathrm{T}}$ and $\boldsymbol{M} = \boldsymbol{M}^{1/2}\left(\boldsymbol{M}^{1/2}\right)^{\mathrm{T}}$.

The length-scales (diffusion tensor) are hidden in the definition of $\boldsymbol{K}$. If we assume a constant length-scale ℓ then we can make it explicit in the expressions above by writing $\boldsymbol{K} = \ell^2 \widehat{\boldsymbol{K}}$. If we assume further that $\Sigma = \sigma \boldsymbol{I}$ and $\Gamma = \gamma \boldsymbol{I}$ then Eq. (51) can be written in the Brankart form

$$\boldsymbol{R}^{-1} = \boldsymbol{T}^{\mathrm{T}} \left(\boldsymbol{R}^{+}\right)^{-1} \boldsymbol{T}$$

where

$$\boldsymbol{T} = \begin{pmatrix} \left(\boldsymbol{M}^{1/2}\right)^{\mathrm{T}} \\ \left(\widehat{\boldsymbol{K}}^{1/2}\right)^{\mathrm{T}} \\ \boldsymbol{M}^{-1/2}\widehat{\boldsymbol{K}} \end{pmatrix} \quad \text{and} \quad \left(\boldsymbol{R}^{+}\right)^{-1} = \frac{1}{\sigma^2\gamma^2} \begin{pmatrix} \boldsymbol{I} & & \\ & 2\ell^2\boldsymbol{I} & \\ & & \ell^4\boldsymbol{I} \end{pmatrix}.$$

Note that $\boldsymbol{M}$ does not contain any information about derivatives, while $\boldsymbol{K}$ contains products of gradients (see Eqs (27) and (28)). Therefore, multiplying by $\boldsymbol{M}^{1/2}$, $\left(\widehat{\boldsymbol{K}}^{1/2}\right)^{\mathrm{T}}$ and $\boldsymbol{M}^{-1/2}\widehat{\boldsymbol{K}}$ corresponds to differentiation to the zeroth, first and second order, respectively (*cf.* Eq. (48)).

The FEM diffusion-based approach has distinct advantages over the Brankart method for unstructured meshes resulting from sparse or heterogeneously-distributed observations. The discretization of operator $\mathcal{T}$ in Eq. (48) relies directly on the ability to estimate first- and second-order derivatives on the mesh supporting the observations. While this is straightforward when considering structured data on regular grids, it becomes difficult when gaps appear in the spatial distribution of the observations. The FEM discretization described in this study offers a natural framework for handling such difficulties. As the computations rely on the triangulation supporting the observations, the derivatives are estimated at each point using all the information in its neighbourhood. This approximation involves all neighbouring points, even those that are close but do not share exactly the same latitude or longitude.

5. Application to unstructured satellite observations

In this section, we consider a realistic distribution of satellite observations from SEVIRI to illustrate how the FEM-discretized diffusion operator can be used to represent spatially correlated errors. In doing so, we discuss the accuracy of the method by comparing results with those obtained using the theoretical reference (Matérn) correlation function that the diffusion model is intended to represent.

5.1. *SEVIRI observations*

SEVIRI is a radiometer on board the Meteosat Second Generation satellite, which measures radiances at the top of the atmosphere from 12 different spectral channels (Schmetz *et al.* 2002). SEVIRI radiances provide useful information about temperature and humidity in the troposphere and lower stratosphere. In global numerical weather prediction, SEVIRI radiances are usually assimilated through the clear-sky radiance product, which undergoes cloud-clearing as well as superobbing to 16 pixel

© 2017 Royal Meteorological Society

by 16 pixel squares (Szyndel *et al.* 2005). In the operational limited-area model AROME* at Météo-France, the raw SEVIRI radiances are assimilated as described by Montmerle *et al.* (2007) with some recent adjustments such as the use of a variational bias correction (Auligné *et al.* 2007). The infrared channels are assimilated in clear-sky conditions and above low clouds (Michel 2018)[Table 1]. In this study, we focus on radiances from Channel 5 (wavelength 6.2 μm), which provides information about humidity in the upper troposphere.

The SEVIRI measurements are known for having spatially correlated observation errors (Waller *et al.* 2016a; Michel 2018). Therefore, they are thinned at a spatial resolution of 70km before assimilation in AROME. This thinning, as well as the screening step to remove cloud-contaminated data, result in a large amount of observations that is discarded. It also causes gaps in the spatial distribution of the observations that depend on the meteorological situation. Those gaps can be responsible for the presence of ill-shaped triangular elements in the mesh supporting the observations.

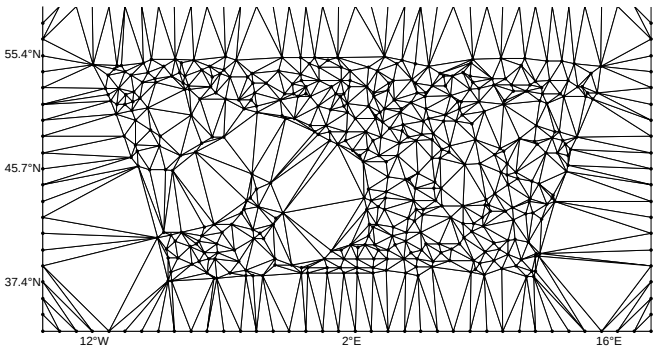
5.2. *Mesh generation*

The spatial domain is that of AROME. It covers France over an extended region between 12°W to 16°E and 37°N to 55°N. We define a rectangular domain containing the observations, with outer boundaries chosen far from the observation locations, relative to the correlation length-scale (see later), to minimize boundary effects on the solution of the diffusion equation in the interior of the domain. We impose Neumann boundary conditions as they have been naturally accounted for in Eq. (26) through the elimination of the boundary terms after integrating Eq. (24) by parts. The mesh is then built using a constrained Delaunay triangulation algorithm (Edelsbrunner *et al.* 1992), in such a way that the triangular nodes are (exactly) located at the observation locations. Figures 5(a) and 5(b) show examples of the mesh generated from SEVIRI observation locations resulting from two different levels of observation thinning. Figure 5(a) corresponds to the mesh resulting from the thinning algorithm used in the operational AROME model. Figure 5(b) corresponds to the mesh used for the experiments in this study. For this mesh, the total number of nodes is 4980, of which 124 are additional nodes at the artificial domain boundaries.

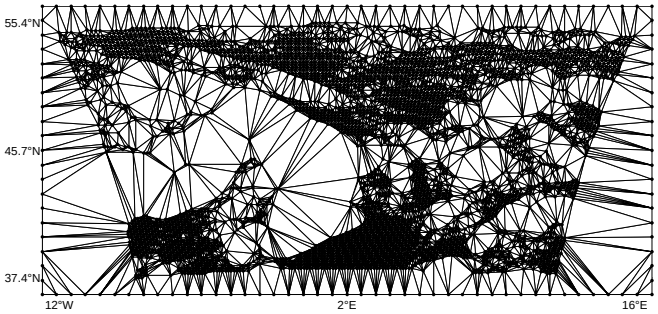
5.3. *Impulse response of the spatial correlation operator*

In this section, we evaluate the quality of the spatial correlations produced using the FEM-discretized diffusion operator by comparing them to the correlations produced using the analytical Matérn function. We choose constant values for m and ℓ to ensure consistency between the diffusion model and analytical Matérn model. With constant parameters, these models are expected, from theory, to give identical results. The procedure for estimating the values of the parameters m and ℓ is discussed below.

Actual estimates of observation error correlations for Channel 5 SEVIRI radiances have been computed by Waller *et al.* (2016a) and Michel (2018) using Desroziers diagnostics (Desroziers *et al.* 2005). A distinguishing feature of these estimates is the sharpness of the correlations near the origin and the rather slow decay of the correlations at large distances from the origin. As discussed in Section 2.3, this suggests that a Matérn function with a value of $m = 2$ is more appropriate than a Matérn function with a larger value of m. Therefore, we choose to use this value of m for the diffusion model. Furthermore, we use the Channel 5 correlation



(a) Mesh built from substantially thinned data



(b) Mesh built from mildly thinned data

Figure 5. Triangular mesh constructed from the locations of SEVIRI measurements that have undergone two different levels of thinning. The average distance between observations after thinning is approximately 70 km in (a) and 12 km in (b). The thinning used to generate the mesh in (a) is based on that used in the operational AROME configuration at Météo-France. The experiments in this article employ the mesh in (b). Thin or flat triangles are called "ill-shaped" because their presence is likely to induce numerical errors.

estimates from Figure 5a of Waller *et al.* (2016a) as a guideline for choosing a value of ℓ. In particular, we use their result that the distance at which the Channel 5 correlations drop to 0.2 is about 80 km. Correlations beyond 0.2 can be considered insignificant (Liu and Rabier 2003). Assuming that the correlation function is of Matérn type with $m = 2$, then we can invert Eq. (6) to determine ℓ such that $c_{m,\ell}(80\text{ km}) = 0.2$. This gives $\ell = 32.5$ km, which is the value of ℓ used in the following experiments.

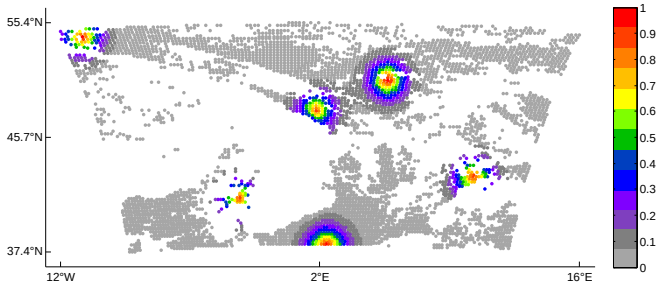
The spatial correlation function at a given point $\boldsymbol{z}_i$ corresponds to the ith column of the correlation matrix $\boldsymbol{C}$. It can be visualized by plotting the result of applying $\boldsymbol{C}$ to a vector that has a value of one at $\boldsymbol{z}_i$ and a value of zero at all other points. Figure 6(a) displays the result of applying the diffusion-based correlation operator (Eq. (42) *without* mass lumping) at six different points in the rectangular domain. The points have been selected to be sufficiently far apart so that the correlation functions do not intersect in any significant way. The points have also been chosen to sample different characteristics of the observation distribution. They include regions where the distribution is dense, sparse, near large gaps, and next to the artificial boundary nodes. A first, qualitative remark to make is that, for all points, the diffusion operator produces sensible, localized structures with spatial extent roughly consistent with the prescribed length-scale and with maximum amplitude close to one.

We can quantify the accuracy of the diffusion-modelled correlation functions by computing their difference with the corresponding Matérn function $c_{m,\ell}$. Denoting the difference field by $\varepsilon_i(\boldsymbol{z}_j)$, $j \in [1, p]$, then for each of the six points $\boldsymbol{z}_i$, $i = [1, 6]$, we have

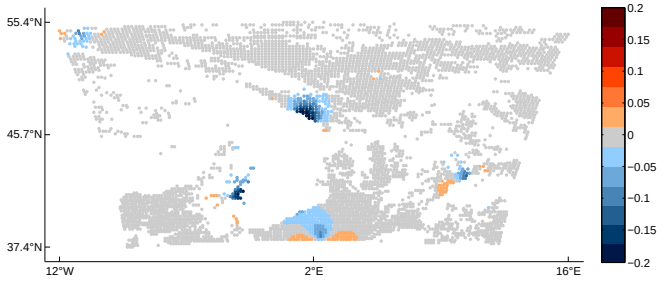
$$\varepsilon_i(\boldsymbol{z}_j) \;=\; C_{ij} - \left(C_{m,\ell}\right)_{ij}$$

*"Applications de la Recherche à l'Opérationnel à Méso-Echelle" (see Seity *et al.* (2011)).

© 2017 Royal Meteorological Society



(a) Impulse response



(b) Difference with respect to Matérn

Figure 6. (a) Impulse response of the FEM diffusion-based correlation operator at six points in the domain. (b) Difference between the numerical response in (a) and the values of the corresponding Matérn function centred on the same points.

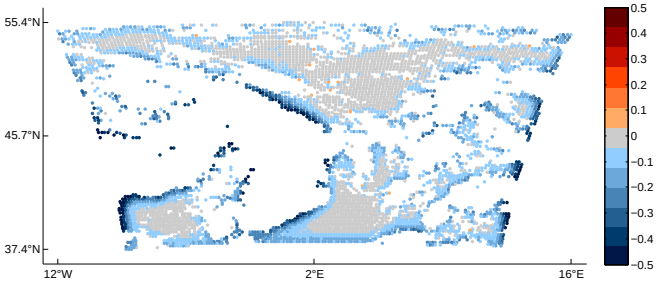


Figure 7. Amplitude error (Eq. (52)) at each node in the domain.

where $C_{ij} = c(\boldsymbol{z}_j, \boldsymbol{z}_i)$ and $\left(C_{m,\ell}\right)_{ij} = c_{m,\ell}(r_{ij})$, with $r_{ij} = \|\boldsymbol{z}_j - \boldsymbol{z}_i\|_2$, are the elements of the diffusion-modelled and Matérn correlation matrices, respectively. The difference field is displayed in Figure 6(b). The errors are small in magnitude (less than 5%) for the points in the densely observed regions, but are up to 20% for the points in sparsely observed regions and near large data gaps. The errors manifest themselves as inaccuracies in the diagonal and off-diagonal elements of the correlation matrix. They are mainly associated with ill-shaped elements in the mesh and the boundary conditions. In the following subsections, we present diagnostics to investigate these errors in more detail.

5.4. *Accuracy of the diagonal elements of* $\boldsymbol{C}$

The diagonal elements of the diffusion-based correlation matrix correspond to the amplitude (variance) of the correlation function at each node and should be equal to one. To quantify the amplitude errors of the actual estimates C_{ii}, we compute at each node i the difference

$$\varepsilon_i^{\text{amp}} = C_{ii} - 1. \tag{52}$$

The amplitude errors are shown in Figure 7. They appear to be minimal far from the boundaries and away from large data gaps. The errors associated with the latter are related to the quality of the mesh in these regions. This will be discussed further in Section 5.6.

The amplitude errors can be eliminated entirely by re-normalizing the diffusion operator at each point using the actual numerical values of the amplitude at each point. The square-root of the normalization factors are stored in the diagonal matrix $\boldsymbol{\Gamma}$ of Eq. (40). To diagnose the exact normalization factors requires as many applications of the square-root of the diffusion operator as the number of nodes on the mesh. These computations are expensive, so approximate methods are usually used instead (Weaver and Courtier 2001; Yaremchuk and Carrier 2012). Randomization is one such method, but requires a large number of random vectors to reduce the amplitude error to a satisfactory level (*e.g.*, 1000 vectors are required to reduce the errors to about 4%). Randomization is typically of interest for much larger problems than the one considered in this study.

Near the artificial boundary nodes, Neumann boundary conditions prevent flux exchanges across the boundary, causing the amplitude to double directly at points along straight boundaries (Mirouze and Weaver 2010, Appendix B) and to increase even more in the corners of the domain. The opposite occurs with Dirichlet boundary conditions (*i.e.*, the amplitude is diminished near the boundaries). Procedures to correct the amplitude near boundaries have been proposed in the literature. For example, Mirouze and Weaver (2010) show that the correct amplitude can be obtained by redefining the correlation operator as an average of two diffusion operators, one that employs Neumann boundary conditions and the other that employs Dirichlet boundary conditions. Their analysis was based on solutions of the continuous, one-dimensional (1D) diffusion equation in the presence of an isolated, straight boundary. The method has limitations, however, when applied in higher dimensions and in the presence of complex geometry. Furthermore, since the method involves a sum of diffusion operators, it results in a complicated expression for the inverse of the correlation operator. For this latter reason in particular, it is not considered appropriate for the problem at hand.

Building on the continuous, 1D theoretical analysis of Mirouze and Weaver (2010), Mirouze and Storto (2016) proposed a simple analytical correction to the normalization factor near the boundary as an alternative to the less practical "double-diffusion" approach of Mirouze and Weaver (2010). With Neumann boundary conditions, their analysis suggests that the normalization coefficient should be corrected by a factor $\xi = 1/(1 + c_{m,\ell}(r_{\text{b}}))$ where r_{b} is the Euclidean distance to the closest boundary point. For example, directly at the boundary, ξ equals $1/2$ to compensate for the doubling of the amplitude there with Neumann boundary conditions. The expression for the correction factor also suggests that nodes located at distances beyond the correlation length-scale (*i.e.*, such that $c_{m,\ell}(r_{\text{b}})$ is small) will be largely unaffected by the artificial boundaries. This point has been analyzed in mathematical detail in a recent article by Khristenko *et al.* (2018).

From Figure 7, it is interesting to notice that, apart from a few isolated points in the interior of the domain, the amplitude errors are negative; *i.e.*, the amplitude is mostly underestimated. This suggests that, for the points near the boundary nodes, the amplitude errors are dominated by the effects of large or ill-shaped triangular elements in the mesh, not the (Neumann) boundary conditions. Moving the boundary nodes closer to the interior nodes may reduce the mesh-related errors but at the expense of increasing the boundary condition-related errors. In such a case, corrections like those proposed by Mirouze and Storto (2016) would be needed.

5.5. *Accuracy of the re-normalized off-diagonal elements of* $\boldsymbol{C}$

The second kind of error concerns the overall shape of the correlation function, which is associated with the accuracy of

© 2017 Royal Meteorological Society

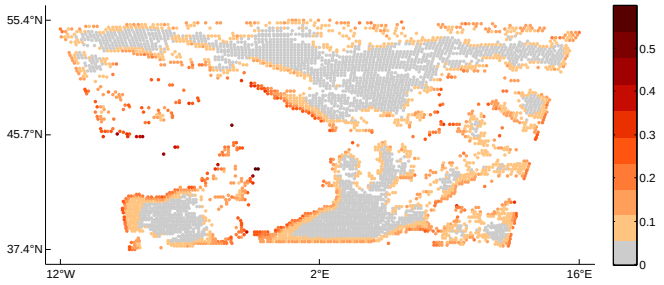


Figure 8. Normalized RMSE of the shape of the correlation function centered on each node in the domain (Eq. (53)).

the off-diagonal elements C_{ij} for $i \neq j$. Even if the amplitude at a particular node is correctly specified (*e.g.*, through re-normalization), the correlation of that node with other nodes might be underestimated or overestimated. We refer to these inaccuracies as shape errors to distinguish them from the amplitude errors discussed in the previous section.

To quantify the shape errors, we can compute the normalized root mean square error (RMSE) between the diffusion-modelled and analytical estimates of the off-diagonal elements:

$$\varepsilon_i^{\text{shape}} = \frac{\left(\sum_j \left|\widehat{C}_{ij} - \left(C_{m,l}\right)_{ij}\right|^2\right)^{1/2}}{\left(\sum_j \left|\left(C_{m,l}\right)_{ij}\right|^2\right)^{1/2}} \tag{53}$$

where

$$\widehat{C}_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}}\sqrt{C_{jj}}}$$

is the *exactly* normalized ijth element of matrix $\boldsymbol{C}$.

The shape errors are shown in Figure 8. They exhibit the same basic structure as the amplitude errors in Figure 7, with smallest errors at points where there is a high density of observations and largest errors at points near data gaps and the boundary regions. For the latter, the magnitude of the errors is generally between 10% and 30%, but reaches up to 50% at a few points. Errors within this range ($< 30\%$) can still provide a better approximation to $\boldsymbol{R}$ than assuming strictly uncorrelated errors (Stewart *et al.* 2013).

5.6. *Link between the accuracy of $\boldsymbol{C}$ and the quality of the mesh*

The accuracy of the diagonal and off-diagonal elements of the correlation matrix generated by the FEM-discretized diffusion operator is closely linked to the quality of the mesh. In this section, we provide an additional diagnostic to explore this link further.

The aspect ratio ($a(\tau)$) of a triangular element τ is used to compute error bounds in standard applications of the FEM (Ern and Guermond 2010, Section 1.5.1):

$$a(\tau) = \frac{h(\tau)}{\rho(\tau)} \tag{54}$$

where $h(\tau)$ is the size of the largest side of τ and $\rho(\tau)$ is the radius of its inscribed circle (see Figure 9). A large value of the aspect ratio indicates the presence of "flat" elements in the mesh (depicted schematically by the triangular element in the middle in Figure 9), which are typically responsible for causing interpolation errors. The aspect ratio has the property of being scale-invariant; *i.e.*, it only depends on the measure of the angles, but is not affected by the actual size of the edges. Therefore, in using the aspect ratio as a criterion for mesh quality, there is an implicit assumption that the mesh size is locally homogeneous; *i.e.*, any two elements found in the same region of the mesh are assumed to be approximately the same size, so that their "quality" only differs in their aspect ratio. This is generally ensured by mesh generators in standard applications of the FEM in numerical modelling.

In our application, however, the mesh is constrained by the observation locations, which can result in contiguous elements of very different size. As a consequence, the mesh generated from the observations does not satisfy the local homogeneous assumption required for the aspect ratio (54) to be a reliable indicator of mesh quality. Therefore, we seek an alternative indicator that detects the presence of overly-large elements as well as ill-shaped elements (depicted schematically by the triangular elements on the left and in the middle in Figure 9). Here, we propose the value of the circumradius $r(\tau)$ as one such indicator. It has the advantage of being high both when the triangles contain large angles (large aspect ratio) and when their size is significantly larger than others in the mesh.

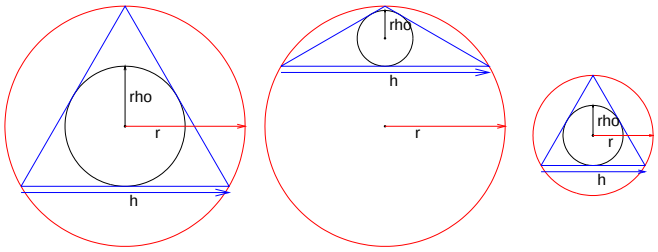


Figure 9. Inscribed circle (black) with radius "rho" (ρ in Eq. (54)) and circumcircle (red) with radius r for three different types of triangular elements whose largest side is denoted by h. The large and "flat" triangular elements correspond to elements of poor quality; they have a large circumradius. The small triangular element corresponds to an element of good quality; it has a small circumradius.

There is no guarantee that given a mesh constructed from an arbitrary distribution of observations, the FEM will lead to small errors in both shape and amplitude. On the contrary, heterogeneously distributed observation locations may cause the elements in the mesh to become ill-shaped, thus leading to increased errors in the FEM discretization. Figure 10 measures the mesh quality for the SEVIRI observation locations in terms of the circumradius of each triangle. Comparing this figure with the error maps (Figures 7 and 8) shows that the locations of the largest errors in both shape and amplitude are highly correlated with the presence of triangles with a large circumcircle radius.

One possible way to improve the quality of the mesh is to eliminate those observations that lead to ill-shaped elements in the FEM. Since the number of observations assimilated in current operational weather prediction systems is very small compared to the number of observations that is actually available (in some cases the number is smaller than 1% of the original

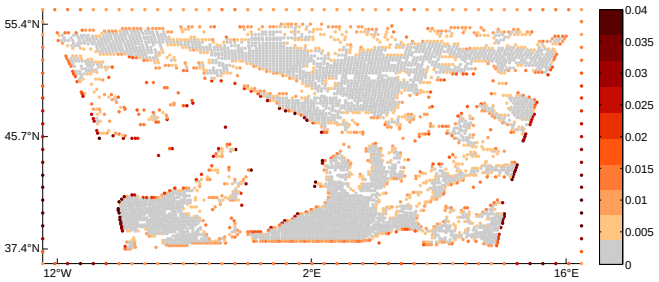


Figure 10. The value of the circumradius of each triangle at each node in the domain. For each node in the triangulation, the largest circumradius of the neighbouring triangles is taken. The map displays the values of these maximum circumradii at all nodes, including those on the artificial boundary.

© 2017 Royal Meteorological Society

set), performing an additional selection of observations based on a mesh-generation criterion is unlikely to deteriorate this ratio significantly. In such a case, removing a few observations would seem to be a reasonable compromise for building a suitable mesh. However, finding an objective criterion to do this correctly is non-trivial. Furthermore, in an operational environment, it would need to be automated and designed to reject as few observations as possible. While an interesting possibility, it is left as a future research direction to develop this idea further.

Another possibility to improve the accuracy of the method is to introduce artificial nodes in order to provide a mesh of better quality. This procedure could be automated using standard mesh-refinement techniques (Ern and Guermond 2010, Chapter 10), and thus seems particularly appealing. However, it leads to complicating issues similar to those encountered by Michel (2018) (and discussed in Section 3.6 for the particular case of the boundary nodes) concerning the representation of $\boldsymbol{C}^{-1}$.

Even in areas where the mesh-related errors are largest, the FEM-based diffusion operator produces a reasonable representation of the spatial correlations, which may be adequate for practical applications, especially in view of our typically inaccurate knowledge of the true observation error covariances. As pointed out by Stewart *et al.* (2008, 2013), it is generally better to have a slightly approximate model for the correlations in $\boldsymbol{R}$ than to neglect them altogether. Using simple analytical models, Fisher (2007) (see also Section 4.9 in Daley (1991)) examined the effects of mis-specifying *background* error covariance parameters on analysis error. His simple scalar example, which illustrates the effects of mis-specifying the background error variance, is equally applicable to the problem of mis-specifying the observation error variance. Specifically, his Figure 3 shows that the analysis error standard deviation is degraded by less than 5% when the background or observation error variance is mis-specified by a factor between 0.5 and 2. This is within the amplitude mis-specification bounds in our experiment, which are roughly between 0.5 and 1.2 when using an analytical estimate of the normalization factors (see Figure 7).

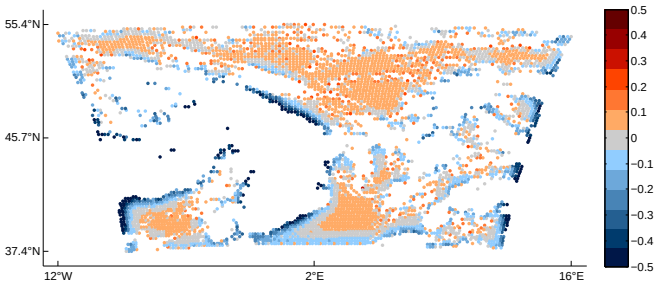
5.7. *Effect of mass lumping*

We recall from Section 3.5 that mass lumping results in a diagonal approximation of the mass matrix. In this section, we examine the effect of mass lumping on the representation of the Matérn correlation functions. The amplitude and shape errors that result from using a mass-lumped matrix are shown in Figures 11(a) and 11(b), respectively. These figures should be compared with Figures 7 and 8, which are the corresponding errors resulting from using the consistent (unapproximated) mass matrix.

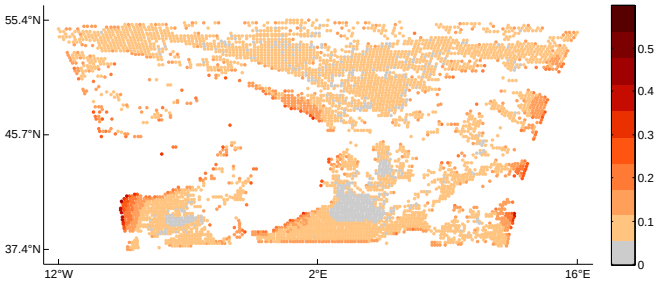
The amplitude and shape errors associated with the mass-lumped correlation matrix have similar structures to those associated with the consistent mass matrix. This implies that any mesh-dependent criterion used to predict the errors in the consistent mass formulation will also be relevant for the mass-lumped formulation. However, the magnitude of the errors is significantly larger at some points near data gaps and about 10% larger in areas where there is a high density of observations. Furthermore, mass lumping has a tendency to overestimate the amplitude, as evident by the large patches of positive error in Figure 11(a).

6. Summary and discussion

In this article, we addressed modelling and computational issues that arise when accounting for spatially correlated observation errors in variational data assimilation. Key requirements include the need to handle large covariance matrices, built from heterogeneously distributed observations, and the need to provide



(a) Amplitude error



(b) Shape error

Figure 11. (a) Amplitude error at each node in the domain (Eq. (52)), and (b) normalized RMSE of the shape of the correlation function centered on each node in the domain (Eq. (53)), when the mass lumping approximation is used (cf. Figures 7 and 8).

an efficient operator for the inverse of the observation error covariance matrix ($\boldsymbol{R}^{-1}$) as well as the observation error covariance matrix ($\boldsymbol{R}$) itself.

We showed how to construct a spatial correlation operator for observation error using a diffusion operator that is discretized with a finite element method (FEM) on a triangular 2D mesh whose nodes are defined by the observation locations. The basic technique has many similarities to the stochastic PDE approach developed by Lindgren *et al.* (2011) for spatial interpolation in geostatistical applications.

The theoretical basis of the diffusion operator approach to correlation modelling is well documented. Here, we considered a diffusion operator that results from integrating a diffusion equation over a finite number of steps with a backward Euler (implicit) scheme. The re-scaled solution of the resulting elliptic equation can be interpreted as a correlation operator whose kernel is a correlation function from the Matérn family. Crucially, the elliptic operator itself provides the corresponding inverse correlation operator, which can be used for defining $\boldsymbol{R}^{-1}$.

In the continuous framework of spatial correlation modelling, we established a formal link between the diffusion operator approach and the method of assimilating directional derivatives of the observations, up to arbitrary order, as proposed by Brankart *et al.* (2009). While the two methods are closely related, the diffusion framework offers better controllable flexibility and a clearer connection with theoretical correlation models. In the discrete framework, we showed how Brankart *et al.* (2009)'s method can be adapted to an unstructured mesh, by formulating the numerical representation of the derivative operators in terms of the mass and stiffness matrices of the FEM-discretized diffusion operator.

The correlation model based on a FEM-discretized diffusion operator was evaluated on an operational data-set from SEVIRI. To assess the accuracy of the method, results were compared to those produced using the analytical Matérn function, which should be identical for the constant correlation parameter settings considered. A qualitative assessment of the spatially correlated fields centred at various points in the domain showed that the

© 2017 Royal Meteorological Society

diffusion operator produced sensible structures, with amplitude close to one and spatial scale roughly consistent with the length-scale (square-root of the diffusion coefficient) that was specified. This was the case even in areas where the distribution of observations was extremely irregular.

To quantify the numerical errors, we evaluated separately the diagonal and off-diagonal elements of the diffusion-modelled correlation matrix. The diagonal elements are associated with the amplitude of the modelled correlation functions centered at each point. They were compared with their expected value of one. The errors were shown to be small ($< 5\%$) for points in densely observed regions, but reached 20% for points in sparsely observed regions and near large data gaps. In practice, amplitude errors can be corrected using a re-normalization procedure, although to do so accurately can be costly for very large data-sets.

The off-diagonal elements from the perfectly re-normalized diffusion-modelled matrix were compared with those from the correlation matrix built from the Matérn function. Discrepancies between the two are associated with inaccuracies in the overall shape of the modelled correlation function centred at each point. The shape errors were shown to have a similar spatial structure to the amplitude errors, with errors generally between 10% and 30%. Using the mass lumping approximation for the mass matrix led to a further increase in the magnitude of the errors of about 10%, but did not change the spatial structure of the errors.

Further analysis showed that the largest errors occurred predominantly in areas where the triangular elements in the mesh were ill-shaped; *i.e.*, either too "flat" or too "large". A diagnostic based on the radius of the circumcircle of an element was shown to be a reliable indicator of the quality of the mesh, with large (small) radii being well correlated with large (small) shape and amplitude errors. Although not explored in this study, one interesting possibility to improve the accuracy of the method is to reduce the number of ill-shaped elements in the mesh by using the circumcircle diagnostic in an objective data-thinning criterion prior to the data assimilation step.

Rather than thinning the data to improve the accuracy of the method, another possibility is to refine the mesh by adding extra nodes in areas where observations are missing, as in the Lindgren *et al.* (2011) approach. This procedure results in an auxiliary (higher resolution) mesh, different from the one supporting only the observations, on which the main computations are carried out. An interpolation operator and its adjoint are required to transfer fields between the original mesh and the auxiliary mesh, which results in an $\boldsymbol{R}$ operator with rectangular matrix components, as in the method described by Michel (2018). Therefore, although this approach provides a convenient and more accurate model for $\boldsymbol{R}$, it leads to difficulties in defining $\boldsymbol{R}^{-1}$, which can no longer be represented as an explicit operator.

The method described in this article is generic and could be adapted to other observation types, such as Doppler radar observations, and satellite-derived sea-surface temperature and altimeter observations, for which spatially correlated errors are known to be important.

Acknowledgement

The authors gratefully acknowledge the support of the French national programme LEFE/INSU. O. Guillet benefitted from an FCPLR (Formation Complémentaire Par la Recherche) grant from Météo-France. The authors thank two anonymous reviewers whose comments helped to improve an earlier version of the paper.

References

Auligné T, McNally AP, Dee DP. 2007. Adaptive bias correction for satellite data in a numerical weather prediction system. *Q. J. R. Meteorol. Soc.* **133**: 631–642, doi:10.1002/qj.56.

Bannister RN. 2008a. A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. *Q. J. R. Meteorol. Soc.* **134**: 1951–1970.

Bannister RN. 2008b. A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics. *Q. J. R. Meteorol. Soc.* **134**: 1971–1996.

Bannister RN. 2017. A review of operational methods of variational and ensemble-variational data assimilation. *Q. J. R. Meteorol. Soc.* **143**: 607–633.

Barth A, Beckers JM, Troupin C, Alvera-Azeárate A, Vandenbulcke L. 2014. divand-1.0: n-dimensional variational data analysis for ocean observations. *Geosci. Model Dev.* **7**: 225–241.

Belo Pereira M, Berre L. 2006. The use of an ensemble approach to study the background error covariances in a global NWP model. *Mon. Weather Rev.* **134**: 2466–2489.

Bolin D, Lindgren F. 2013. A comparison between Markov approximations and other methods for large spatial data sets. *Comp. Statist. Data Anal.* **61**: 7–32.

Bormann N, Bauer P. 2010. Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I: Methods and application to ATOVS data. *Q. J. R. Meteorol. Soc.* **136**: 1036–1050.

Bormann N, Collard A, Bauer P. 2010. Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. II: Application to AIRS and IASI data. *Q. J. R. Meteorol. Soc.* **136**: 1051–1063.

Brankart JM, Ubelmann C, Testut CE, Cosme E, Brasseur P, Verron J. 2009. Efficient parameterization of the observation error covariance matrix for square root or Ensemble Kalman Filters: Application to ocean altimetry. *Mon. Weather Rev.* **137**: 1908–1927.

Brenner S, Scott LR. 2013. *The Mathematical Theory of Finite Element Methods*. Texts in Applied Mathematics, Springer: New York, NY.

Bui-Thanh T, Ghattas O, Martin J, Stadler G. 2013. A computational framework for infinite-dimensional Bayesian inverse problems. Part I: The linearized case, with application to global seismic inversion. *SIAM J. Sci. Comput.* **35**: A2494–A2523.

Campbell WF, Satterfield E, Ruston B, Baker N. 2017. Accounting for correlated observation error in a dual formulation 4d-variational data assimilation system. *Mon. Weather Rev.* **145**: 1019–1032.

Canuto C, Hussaini MY, Quarteroni A, Zang T. 1987. *Spectral Methods in Fluid Dynamics*. Springer: New York, NY.

Carrier MJ, Ngodock H. 2010. Background-error correlation model based on the implicit solution of a diffusion equation. *Ocean Model.* **35**: 45–53.

Chabot V, Nodet M, Papadakis N, Vidard A. 2015. Accounting for observation errors in image data assimilation. *Tellus A* **67**: 23 629, doi:10.3402/tellusa.v67.23629.

Ciarlet P. 2002. *The Finite Element Method for Elliptic Problems*. Classics in Applied Mathematics, SIAM: Philadelphia, PA.

Daley R. 1991. *Atmospheric Data Analysis*. Cambridge Atmospheric and Space Sciences Series, Cambridge University Press: Cambridge, UK.

Dando ML, Thorpe AJ, Eyre JR. 2007. The optimal density of atmospheric sounder observations in the Met Office NWP system. *Q. J. R. Meteorol. Soc.* **133**: 1933–1943.

Davis TS. 2006. *Direct Methods for Sparse Linear Systems*. SIAM: Philadelphia, PA.

Desroziers G, Berre L, Chapnik B, Poli P. 2005. Diagnosis of observation, background and analysis-error statistics in observation space. *Q. J. R. Meteorol. Soc.* **131**: 3385–3396.

Duff IS, Erisman AM, Reid JK. 1989. *Direct Methods for Sparse Matrices*. Oxford University Press: Oxford, UK.

Edelsbrunner H, Tan T, Waupotitsch R. 1992. An $o(n^2 \log n)$ time algorithm for the minmax angle triangulation. *SIAM J. Sci. Comput.* **13**: 994–1008.

Ern A, Guermond JL. 2010. *Theory and Practice of Finite Elements*, *Applied Mathematical Series*, vol. 159. Springer: New York, NY.

Fisher M. 2007. The sensitivity of analysis errors to the specification of background error covariances. In: *Workshop on Flow-dependent Aspects of Data Assimilation*. ECMWF, Reading, UK, pp. 27–36.

Gaspari G, Cohn SE. 1999. Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.* **125**: 723–757.

Gratton S, Toint P, Tshimanga J. 2011. A comparison between conjugate gradients and multigrid solvers for covariance modelling in data

© 2017 Royal Meteorological Society

assimilation. *Q. J. R. Meteorol. Soc.* **139**: 1481–1487.

Guttorp P, Gneiting T. 2006. Studies in the history of probability and statistics XLIX: On the Matérn correlation family. *Biometrika* **93**: 989–995.

Janjić T, Bormann N, Bocquet M, Carton JA, Cohn SE, Dance SL, Losa SN, Nichols NK, Potthast R, Waller JA, Weston P. 2018. On the representation error in data assimilation. *Q. J. R. Meteorol. Soc.* **144**: 1257–1278, doi: 10.1002/qj.3130.

Järvinen H, Andersson E, Bouttier F. 1999. Variational assimilation of time sequences of surface observations with serially correlated errors. *Tellus A* **51**: 469–488.

Jones GH. 1982. *The Theory of Generalised Functions*. Cambridge University Press: Cambridge, UK, 3nd edn.

Khristenko U, Scarabosio L, Swierczynski P, Ullmann E, Wohlmuth B. 2018. Analysis of boundary effects on PDE-based sampling of Whittle-Matérn random fields. *ArXiv e-prints* 1809.07570.

Lindgren F, Rue H, Lindström J. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. Roy. Stat. Soc.: Series B Stat. Method.* **73**: 423–498.

Liu ZQ, Rabier F. 2002. The interaction between model resolution, observation resolution and observation density in data assimilation: A one-dimensional study. *Q. J. R. Meteorol. Soc.* **128**: 1367–1386.

Liu ZQ, Rabier F. 2003. The potential of high-density observations for numerical weather prediction: a study with simulated observations. *Q. J. R. Meteorol. Soc.* **129**: 3013–3035.

Lorenc AC. 1997. Development of an operational variational assimilation scheme. *J. Meteorol. Soc. Jpn.* **75**: 339–346.

Michel Y. 2018. Revisiting Fisher's approach to the handling of horizontal spatial correlations of the observation errors in a variational framework. *Q. J. R. Meteorol. Soc.* **144**: 2011–2025, doi:10.1002/qj.3249.

Mirouze I, Storto A. 2016. Handling boundaries with the one-dimensional first-order recursive filter. *Q. J. R. Meteorol. Soc.* **142**: 2478–2487.

Mirouze I, Weaver AT. 2010. Representation of correlation functions in variational assimilation using an implicit diffusion operator. *Q. J. R. Meteorol. Soc.* **136**: 1421–1443.

Montmerle T, Rabier F, Fischer C. 2007. Relative impact of polar-orbiting and geostationary satellite radiances in the ALADIN/France numerical weather prediction system. *Q. J. R. Meteorol. Soc.* **133**: 655–671, doi:10.1002/qj.34.

Purser RJ, Wu WS, Parrish DF, Roberts NM. 2003. Numerical aspects of the application of recursive filters to variational statistical analysis. Part II: spatially inhomogeneous and anisotropic general covariances. *Mon. Weather Rev.* **131**: 1536–1548.

Rabier F. 2006. Importance of data: A meteorological perspective. In: *Ocean Weather Forecasting: An Integrated View of Oceanography*, Chassignet EP, Verron J (eds), Springer: Dordrecht, NL, pp. 343–360.

Ruggiero GA, Cosme E, Brankart JM, Le Sommer J. 2016. An efficient way to account for observation error correlations in the assimilation of date from the future swot high-resolution altimeter mission. *J. Atmospheric Ocean. Technol.* **33**: 2755–2768.

Saad Y. 2003. *Iterative Methods for Sparse Linear Systems*. SIAM: Philadelphia, PA, 2nd edn.

Schmetz J, Pili P, Tjemkes S, Just D, Kerkmann J, Rota S, Ratier A. 2002. An introduction to Meteosat Second Generation (MSG). *Bull. Am. Meteorol. Soc.* **83**: 977–992.

Seity Y, Brousseau P, Malardel S, Hello G, Bénard P, Bouttier F, Lac C, Masson V. 2011. The AROME-France convective-scale operational model. *Mon. Weather Rev.* **139**: 976–991, doi:10.1175/2010MWR3425.1.

Simpson D, Lindgren F, Rue H. 2012. In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics* **23**: 65–74.

Stein ML. 1999. *Interpolation of Spatial Data. Some Theory for Kriging*. Springer: New York, NY.

Stewart LM, Dance SL, Nichols. 2013. Data assimilation with correlated observation errors: experiments with a 1-D shallow water model. *Tellus A.* **65**: 1, doi:10.3402/tellusa.v65i0.19546.

Stewart LM, Dance SL, Nichols NK. 2008. Correlated observation errors in data assimilation. *Int. J. Numer. Methods Fluids.* **56**: 1521–1527.

Stewart LM, Dance SL, Nichols NK, Eyre JR, Cameron J. 2014. Estimating interchannel observation-error correlations for IASI radiance data in the Met Office system. *Q. J. R. Meteorol. Soc.* **140**: 1236–1244.

Stuhlmann R, Rodriguez A, Tjemkes S, Grandell J, Arriaga A, Bézy JL, Aminou D, Bensi P. 2005. Plans for EUMETSAT's Third Generation Meteosat geostationary satellite programme. *Adv. Space Res.* **36**(5): 975–981, doi:https://doi.org/10.1016/j.asr.2005.03.091.

Szyndel M, Kelly G, Thépaut JJ. 2005. Evaluation of potential benefit of assimilation of SEVIRI water vapour radiance data from Meteosat-8 into global numerical weather prediction analyses. *Atmos. Sci. Letters* **6**: 105–111.

Tarantola A. 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM: Philadelphia, PA.

Ubelmann C, Gaultier L, Fu LL. 2016. SWOT simulator for ocean science. Online documentation. Available at https://github.com/SWOTsimulator/swotsimulator/commits/master/doc/source/science.rst.

Waller JA, Ballard S, Dance SL, Kelly G, Nichols NK, Simonin D. 2016a. Diagnosing horizontal and inter-channel observation error correlations for SEVIRI observations using observation-minus-background and observation-minus-analysis statistics. *Remote Sens.* **8**: 581, doi:10.3390/rs8070581.

Waller JA, Dance SL, Nichols NK. 2016b. Theoretical insight into diagnosing observation error correlations using observation-minus-background and observation-minus-analysis residuals. *Q. J. R. Meteorol. Soc.* **142**: 418–431.

Waller JA, Simonin D, Dance SL, Nichols NK, Ballard S. 2016c. Diagnosing observation error correlations for Doppler radar radial winds in the Met Office UKV model using observation-minus-background and observation-minus-analysis statistics. *Mon. Weather Rev.* **144**: 3533–3551.

Weaver AT, Courtier P. 2001. Correlation modelling on the sphere using a generalized diffusion equation. *Q. J. R. Meteorol. Soc.* **127**: 1815–1846.

Weaver AT, Gürol S, Tshimanga J, Chrust M, Piacentini A. 2018. "Time"-parallel diffusion-based correlation operators. *Q. J. R. Meteorol. Soc.* **144**: 2067–2088, doi:10.1002/qj.3302.

Weaver AT, Mirouze I. 2013. On the diffusion equation and its application to isotropic and anisotropic correlation modelling in variational assimilation. *Q. J. R. Meteorol. Soc.* **139**: 242–260.

Weaver AT, Tshimanga J, Piacentini A. 2016. Correlation operators based on an implicitly formulated diffusion equation solved with the Chebyshev iteration. *Q. J. R. Meteorol. Soc.* **142**: 455–471.

Weston PP, Bell W, Eyre JR. 2014. Accounting for correlated error in the assimilation of high-resolution sounder data. *Q. J. R. Meteorol. Soc.* **140**: 2420–2429.

Whittle P. 1963. Stochastic processes in several dimensions. *Bull. Inst. Internat. Statist.* **40**: 974–994.

Yaremchuk M, Carrier M. 2012. On the renormalization of the covariance operators. *Mon. Weather Rev.* **140**: 637–649.

Yaremchuk M, D'Addezio JM, Panteleev G, Jacobs G. 2018. On the approximation of the inverse error covariances of high resolution satellite altimetry data. *Q. J. R. Meteorol. Soc.* **144**: 1927–1932, doi:10.1002/qj.3336.

Yaremchuk M, Nechaev D. 2013. Covariance localization with the diffusion-based correlations models. *Mon. Weather Rev.* **141**: 848–860.

Yaremchuk M, Smith S. 2011. On the correlation functions associated with polynomials of the diffusion operator. *Q. J. R. Meteorol. Soc.* **137**: 1927–1932.

© 2017 Royal Meteorological Society