Modélisation des structures locales de covariance des erreurs de prévision à l'aide des ondelettes

Olivier Pannekoucke

UNIVERSITE TOULOUSE III - PAUL SABATIER

THESE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE DE TOULOUSE
délivré par l'Université Toulouse III - Paul Sabatier

Discipline : Physique de l'atmosphère

Présentée et soutenue par :

Olivier PANNEKOUCKE

Sujet de la thèse :

Modélisation des structures locales de covariance des erreurs de prévision à l'aide des ondelettes

Soutenue le 20 Mars 2008 devant le jury composé de :

M. Mohamed MASMOUDI	Président du jury
M. Andrew LORENC	Rapporteur
M. Pierre BRASSEUR	Rapporteur
M. Mike FISHER	Examinateur
M. Éric BRUN	Examinateur
M. Gérald DESROZIERS	Directeur de thèse
M. Loïk BERRE	Co-directeur de thèse

Météo-France CNRM-GAME/GMAP
42 Av. G. Coriolis - 31057 Toulouse Cedex 1

Table des matières

Remerciements

Tout d'abord, je tiens à remercier mes directeurs de thèse, Gérald Desroziers et Loïk Berre, pour leur soutien et la confiance qu'ils m'ont accordée. Ils ont toujours su être disponibles pour moi. J'ai été très heureux de travailler avec eux.

Je remercie Florence Rabier qui a bien voulu assurer la direction des premières années de cette thèse. Sa présence, son soutien et son enthousiasme par rapport à mon travail m'ont été d'une grande aide.

Je voudrais remercier chaleureusement Marie Farge, marraine de cette fcplr, qui m'a beaucoup apporté. Ses nombreuses remarques constructives ont grandement contribué à améliorer ce travail. Sa grande connaissance des ondelettes et l'utilisation qu'elle en fait en turbulence m'ont énormément influencé.

Merci au jury fcplr pour m'avoir accordé sa confiance et pour m'avoir permis d'entreprendre cette formation complémentaire par la recherche dans les meilleurs conditions qui soient.

Je remercie l'ensemble des membres du jury : Mohamed Masmoudi qui a bien voulu présider le jury, Andrew Lorenc et Pierre Brasseur qui ont accepté d'être rapporteurs, ainsi que Miker Fisher et Éric Brun.

J'ai été très heureux de passer ces trois années au sein du groupe GMAP du CNRM et plus particulièrement l'équipe PROC qui m'a accueilli.

Mon initiation à l'enseignement fut une activité passionnante et enrichissante. Ainsi je remercie chaleureusement Olivier Thual, Olivier Eiff, Alain Sevrain et Wladimir Bergez pour l'enseeiht ; François Lalaurette, Bernard Iché, Arnaud Méquigon et Isabelle Beau pour l'enm.

Je remercie chaleureusement Anthony Weaver, Serge Gratton, Sébastien Massart, Nicolas Daget, Sophie Ricci, Philippe Rogel, avec qui j'ai beaucoup appris lors de mon passage au CERFACS et depuis.

Je voudrais enfin remercier chaleureusement mes amis ainsi que ma famille pour leur soutien.

Résumé

En météorologie opérationnelle, il est nécessaire de connaître l'état de l'atmosphère à un instant donné pour en déduire ses états ultérieurs. Cette étape, appelée assimilation de données, est effectuée à l'aide d'observations. Elle est capitale, avec une influence directe sur la qualité des prévisions. Le processus d'assimilation aboutit à la création d'une représentation de l'atmosphère appelée analyse, qui doit être la plus proche possible de l'état réel de l'atmosphère. Le nombre d'observations disponibles à l'heure actuelle est de l'ordre de $\mathcal{O}(10^6)$, ce qui est insuffisant pour déterminer de manière univoque l'ensemble des degrés de liberté du modèle qui est de l'ordre de $\mathcal{O}(10^7)$.

Par conséquent, une prévision réalisée pour l'instant considéré est ajoutée et utilisée comme prédicteur. Le problème, ainsi fermé, est soluble de manière univoque (pour une méthode d'assimilation donnée).

La difficulté majeure vient alors de la nécessité de spécifier une matrice de covariance des erreurs de prévisions $\boldsymbol{B}$. Le rôle de cette matrice est de filtrer et de propager spatialement l'information observée. Cette matrice est trop grande pour être représentée directement ($\mathcal{O}(10^{14})$ éléments), et elle est donc plutôt modélisée sous la forme d'une cascade d'opérateurs relativement simples.

Dans la formulation actuelle, un modèle "simple" de cette matrice est implémenté, basé sur une formulation homogène. Ce modèle est imparfait car il ne représente pas les variations géographiques des fonctions de corrélation. En particulier, cela ne permet pas de rendre compte des variations de la longueur de portée (distance caractérisant le lien statistique entre deux points de l'atmosphère).

Pour représenter ces variations géographiques, une modélisation basée sur les ondelettes est utilisée au Centre Européen de Prévision Météorologique à Moyen Terme (CEPMMT). Dans cette configuration, les ondelettes sont utilisées pour représenter les variations "climatologiques" des statistiques d'erreur d'ébauche : les statistiques utilisées sont moyennées sur une période de l'ordre d'un mois.

Dans cette thèse, les ondelettes sont utilisées pour représenter non plus les variations géographiques moyennes, mais les variations à la fois spatiales et temporelles des corrélations. En outre, la calibration des statistiques est basée sur le recours à un ensemble d'assimilations perturbées, qui permet de simuler l'évolution des erreurs, mais dont la taille finie induit un bruit d'échantillonnage. Les propriétés des ondelettes pour filtrer ce bruit d'échantillonnage sont donc également examinées.

Les principaux résultats sont, d'une part, que les ondelettes sont capables de filtrer spatialement une grande partie de l'erreur d'échantillonnage. En effet, l'utilisation de l'hypothèse diagonale dans l'espace des ondelettes permet de réaliser une moyenne spatiale locale de la matrice estimée. En particulier, les portées modélisées ont des variations géographiques plus lisses : la cohérence spatiale est améliorée.

Ces résultats ont été confortés par l'étude des statistiques de l'échantillonnage de la portée, dans un cadre académique. En effet, il est apparu que le bruit d'échantillonnage affecte particulièrement les petites échelles des cartes de portée. Ainsi, un filtrage spatial local apparaît pertinent pour éliminer les erreurs de petite échelle et extraire les variations de grande échelle moins bruitées. Les ondelettes apparaissent alors comme un outil efficace à la fois pour la modélisation mais aussi pour le diagnostic de la portée. Ces résultats ont été établis en introduisant des expressions économiques d'estimation de la portée.

Ces propriétés ont été mises en oeuvre pour extraire les variations spatiales et temporelles des portées dans un modèle opérationnel de prévision numérique. Cela a permis de confirmer les différences de structure spatiale entre le signal recherché et le bruit d'échantillonnage, ainsi que la pertinence du moyennage spatial réalisé par les ondelettes. Les variations spatiales et journalières peuvent ainsi être estimées de façon relativement robuste. Cette partie de l'étude révèle aussi que la dynamique de la portée est complexe et structurée, et qu'elle est en partie liée à la situation météorologique.

Il est à noter enfin que des travaux complémentaires ont été menés au cours de cette thèse. Il s'agit du filtrage spatial des écarts types des erreurs de prévision, issus d'un ensemble d'assimilations. Une optimisation de ce filtrage a été proposée et testée, en se basant sur la théorie de l'estimation linéaire et sur l'estimation de ratios signal/bruit.

MOT CLÉS: Assimilation de données, ensemble d'analyses perturbées, estimation des covariances d'erreur d'ébauche, dépendance à l'écoulement, longueur de portée, ondelette sphérique, modélisation ondelette des covariances d'erreur d'ébauche.

Chapitre 1

Introduction

Pour décrire l'évolution de l'atmosphère sur une fenêtre temporelle donnée, il est nécessaire de déterminer son état initial. L'assimilation de données a pour objet d'estimer cet état initial, appelé *analyse*. Cette analyse est construite à partir des observations disponibles en les combinant avec une ébauche. Cette ébauche correspond à un état *a priori*, proche de l'état recherché et il s'agit généralement d'une prévision récente. Cette méthode de type prédiction/correction permet de filtrer une partie des erreurs d'observation. De plus, l'ébauche permet de fournir un information sur les régions et les petites structures qui échappent au réseau d'observation.

La construction de l'analyse nécessite la connaissance de la matrice de covariance d'erreur d'ébauche, notée $\boldsymbol{B}$. Cette matrice est fondamentale, en particulier parce qu'elle participe au filtrage des erreurs d'observation et à la propagation des corrections de l'ébauche apportées par les observations. Les fonctions de covariance d'erreur d'ébauche caractérisées par $\boldsymbol{B}$ présentent des variations spatiales et temporelles, qui dépendent notamment de la situation météorologique.

Cependant, la matrice $\boldsymbol{B}$ n'est pas connue et ne peut être qu'estimée. De plus, sa taille énorme (de l'ordre de $\mathcal{O}(10^{14})$) implique qu'elle ne peut pas être représentée directement en mémoire, même sur un supercalculateur : elle doit être approximée et modélisée.

La formulation actuelle de $\boldsymbol{B}$ ne permet pas de modéliser les variations géographiques des fonctions de corrélation : comme elle est basée sur l'hypothèse diagonale spectrale, seule la moyenne spatiale des fonctions de corrélation peut être représentée. Cette moyenne spatiale est généralement estimée à partir de statistiques moyennées sur une période longue (de l'ordre du mois), construites à l'aide d'un ensemble d'analyses perturbées.

Une nouvelle formulation de $\boldsymbol{B}$ basée sur l'hypothèse diagonale dans l'espace des ondelettes sphériques permet de représenter les variations géographiques des fonctions de corrélation. Il est alors possible de faire varier cette information locale au cours du temps pour obtenir une variation spatio-temporelle des corrélations. Pour déterminer cette dynamique, un ensemble d'analyses perturbées est utilisé. Cette méthode permet de construire les statistiques relatives à une journée particulière, mais la taille finie de l'échantillon induit une erreur d'estimation. Pour estimer correctement les statistiques relatives à une journée particulière, il faut donc être capable de filtrer ce bruit d'échantillonnage.

Sachant que les ondelettes contiennent à la fois une information de position et une information d'échelle, l'idée est d'étudier leur capacité à représenter les variations spatio-temporelles de $\boldsymbol{B}$ et à filtrer le bruit d'échantillonnage. Cette étude constitue le travail de thèse présenté ici.

Le manuscrit de thèse se compose de la manière suivante.

Les chapitres 2 et 3 sont introductifs. Le chapitre 2 présente l'assimilation et le rôle de la matrice $\boldsymbol{B}$. Le chapitre 3 décrit la manière dont $\boldsymbol{B}$ est estimée et représentée dans un schéma

d'analyse opérationnel. La méthode des ensembles d'assimilations perturbées est présentée en particulier, ainsi que les limitations de la formulation homogène actuelle.

Les trois chapitres suivants correspondent aux travaux menés pendant cette thèse. Ils correspondent aux traductions d'articles paru, soumis ou à soumettre prochainement. On peut résumer le cheminement entre ces trois papiers de la manière suivante, vis-à-vis de la problématique du filtrage spatial (en ondelette) du bruit d'échantillonnage.

Le chapitre 4 consiste d'une part en une mise en évidence formelle de la notion de moyenne spatiale locale associée à la formulation ondelette. Ce concept suggère d'autre part des propriétés de filtrage spatial du bruit d'échantillonnage (qui affecte les corrélations estimées via un ensemble), qui sont illustrées dans un cadre essentiellement académique.

Le chapitre 5 concerne les propriétés statistiques de l'estimation des portées de corrélation (à partir d'un ensemble). Il permet en particulier de mettre en évidence, dans un cadre académique, le fait que la structure spatiale du bruit d'échantillonnage est plutôt de petite échelle (et ce même si le signal étudié est de grande échelle). Ce résultat justifie la pertinence et l'efficacité de la moyenne spatiale réalisée par les ondelettes.

Le chapitre 6 porte sur la structure spatiale et la dynamique des portées, dans le cadre d'un ensemble réel d'assimilations (basées sur le système global Arpège). Les résultats dans ce cadre réaliste confortent ceux obtenus dans un cadre académique. Il apparaît d'une part que les petites échelles des cartes des portées brutes (non filtrées par les ondelettes) sont plus bruitées que les grandes échelles. D'autre part, la prédominance des structures de grande échelle dans le signal recherché est également mise en évidence. Ces deux caractéristiques expliquent la robustesse des estimations en ondelette, qui permet au final de mener une étude sur la dynamique spatio-temporelle des portées.

Le chapitre 7 apporte les conclusions de ce travail, ainsi que les perspectives qui s'en dégagent.

Les chapitres suivant correspondent aux annexes. Il s'agit d'une part des versions originales des articles paru ou soumis : l'article paru discutant les propriétés de filtrage des ondelettes, et l'article soumis décrivant les caractéristiques statistiques de l'estimation des portées. L'annexe C correspond à une contribution à un workshop du CEPMMT, qui porte essentiellement sur l'optimisation du filtrage spatial des écarts types issus d'un ensemble d'assimilations. Il est à noter en particulier que certaines des méthodes introduites dans l'annexe C sont appliquées à l'étude des portées dans le chapitre 6. L'annexe D permet d'approfondir les notions introduites dans le chapitre 4 : elle permet de décrire plus en détail les ondelettes et leur utilisation dans la modélisation des covariances d'erreur de prévision. Cette annexe est également suivie d'un approfondissement de la notion de "frame" (également introduite dans le chapitre 4).

Chapitre 2

Assimilation de données et rôle de B

RÉSUMÉ

Ce chapitre introduit la problématique de l'assimilation de données associée à la prévision opérationnelle. Le problème de la recherche d'un état initial est résolu par une méthode de prédiction/correction. Ainsi l'analyse objective est obtenue en corrigeant une ébauche à partir des observations. La correction est donnée par l'équation du BLUE. En particulier, cette équation fait intervenir la matrice de covariance d'erreur d'ébauche.

La matrice de covariance d'erreur d'ébauche joue un rôle fondamental dans les algorithmes actuels : elle participe au filtrage de l'erreur d'observation et à la propagation des corrections issues des observations. Cette matrice est caractérisée par ses écart types, ses corrélations et leurs longueurs de portée. L'évolution temporelle d'un ensemble d'états, dont la dispersion est caractérisée par la matrice de covariance d'erreur d'ébauche, permet d'introduire les équations du filtre de Kalman-Bucy.

Les algorithmes variationnels usuels pour résoudre l'équation du BLUE, et sa généralisation temporelle, sont également décrits.

2.1 L'assimilation de données

2.1.1 Problématique de l'assimilation de données

L'atmosphère est observée à l'aide de mesure. Les techniques de mesure les plus anciennes sont par exemple les mesures de la pression de surface, celles de la température ou encore celles du vent (direction et amplitude). Il y a aussi les techniques plus récentes telles que les mesures satellitaires de radiance, à partir desquelles il est possible de déduire le profil vertical de température, ou encore l'occultation GPS permettant de déduire le profil vertical d'humidité à partir de la réfraction dans l'atmosphère des signaux GPS.

Avec les notations usuelles, $\boldsymbol{y}^o$ désigne le vecteur constitué de toutes ces observations, et $\boldsymbol{x}^t$ désigne le vecteur d'état de l'atmosphère à un instant donné. L'indice t (pour *true*) indique ici qu'il s'agit de l'état exact, ou état vrai, pour l'instant considéré. Ce vecteur d'état désigne la représentation de l'atmosphère en machine. Un vecteur d'état $\boldsymbol{x}$ correspond donc à une représentation finie des champs sur la sphère et sur la verticale. Le passage d'un état du modèle à celui des observations est effectué à l'aide d'un opérateur d'observation $\mathcal{H}$. Cet opérateur est généralement non linéaire. Il permet de transformer les quantités décrites par le modèle numérique (tourbillon, divergence, température, *etc*) en quantités mesurées (radiance, vent, *etc*).

Le problème de la spécification de l'état de l'atmosphère est celui de la détermination de l'état $\boldsymbol{x}$ tel que, pour les observations de l'atmosphère $\boldsymbol{y}^o$ (que l'on peut supposer parfaites dans un premier temps), la relation suivante soit vérifiée :

$$\boldsymbol{y}^o = \mathcal{H}(\boldsymbol{x}). \tag{2.1}$$

Ce problème est également désigné sous le nom de *problème inverse*. En effet, l'état vrai $\boldsymbol{x}^t$ recherché est formellement donné par $\boldsymbol{x}^t = \mathcal{H}^{-1}(\boldsymbol{y}^o)$, en supposant que l'opérateur $\mathcal{H}$ est directement inversible, ce qui n'est souvent pas le cas. Plusieurs difficultés rendent effectivement cette inversion difficile ou peu précise.

Une première difficulté provient de la modélisation de l'atmosphère. L'atmosphère étant un milieu continu, sa représentation finie sous la forme d'un vecteur $\boldsymbol{x}$ induit des *erreurs de représentativité*, associées à la conversion (via l'opérateur $\mathcal{H}$) de l'état $\boldsymbol{x}$ en vecteur d'observation $\boldsymbol{y}$. Les erreurs de représentativité peuvent être associées à des problèmes d'interpolation, mais aussi à des défauts de modélisation (au niveau du transfert radiatif, par exemple).

Une deuxième difficulté vient du fait que les observations sont entachées d'erreurs de mesure, induisant une erreur sur la détermination de l'état vrai.

Un troisième problème est associé à l'hétérogénéité du réseau d'observation. En effet la densité spatio-temporelle des zones observées est inégale : il y a moins d'observations dans l'hémisphère Sud que dans l'hémisphère Nord, les zones océaniques sont très peu mesurées en surface (seulement quelques bateaux ou bouées). La conséquence est qu'il est plus difficile de déterminer l'état vrai dans l'hémisphère Sud ou sur les océans, qu'en Europe ou sur les États-Unis.

Par conséquent l'état vrai a peu de chances d'être effectivement obtenu, et le problème inverse se réduit à la détermination de l'état le plus proche de la réalité. Cet état est appelé *état analysé*, ou tout simplement *analyse* ; il est noté $\boldsymbol{x}^a$. L'erreur d'analyse $\boldsymbol{\varepsilon}^a$ est définie par $\boldsymbol{x}^a = \boldsymbol{x}^t + \boldsymbol{\varepsilon}^a$.

L'analyse $\boldsymbol{x}^a$ correspond donc à l'état le plus vraisemblable de l'atmosphère au vu des observations. Il n'y a pas si longtemps encore, cet état était construit à la main par les prévisionnistes (ou plutôt par les analystes). Ces professionnels ajustaient ainsi les isobares pour les faire correspondre au mieux avec les observations. Il s'agissait alors d'une *analyse subjective* de l'état de l'atmosphère.

L'assimilation de données est utilisée dans la prévision météorologique pour *déterminer de manière objective l'état analysé*. Pour un jeu d'observations donné, l'analyse est l'unique état vérifiant certaines contraintes. Elle est généralement obtenue comme la solution d'un problème d'optimisation d'une certaine fonctionnelle. Ainsi, l'état résultant du processus d'assimilation de données est naturellement appelé *analyse objective*, en opposition à l'analyse subjective. Dans la suite, le terme analyse correspond à l'analyse objective. Le paragraphe suivant s'intéresse à la manière de construire cet état.

2.1.2 Estimation objective : le BLUE

Une méthode de prédiction/correction

Il existe différentes techniques pour déterminer l'état analysé. Pour les problèmes opérationnels, rencontrés en météorologie ou en océanographie, l'analyse est obtenue par une méthode de prédiction/correction. Ainsi, une *ébauche* (correspondant généralement à la dernière prévision) est ajustée à partir des observations. La raison de l'utilisation d'une approche prédiction/correction est double.

D'une part, c'est une question de fermeture : le nombre d'observations disponibles est de l'ordre de $p = \mathcal{O}(10^5)$ pour un vecteur d'état dont la taille est de l'ordre de $n = \mathcal{O}(10^7)$. Il y a donc trop d'inconnues pour fixer l'état, *le problème n'est pas fermé*.

D'autre part, c'est une question de valeur ajoutée : il apparaît en pratique que l'ébauche apporte effectivement de l'information. Ceci est associé au fait que les prévisions sont en général de bonne qualité, et fournissent effectivement une information utile dans la recherche de l'analyse.

C'est au fond cette deuxième raison qui prévaut sur la première. En effet, si l'ébauche n'apportait pas d'information supplémentaire, alors elle ne serait d'aucune utilité. L'unique solution serait de se ramener à un problème fermé : par exemple, ajuster la résolution du modèle pour éviter l'indétermination.

L'introduction de nouvelle inconnues

L'utilisation de cette approche de prédiction/correction induit de nouvelles inconnues qui sont la matrice de covariance d'erreur d'observation et la matrice de covariance d'erreur d'ébauche.

L'erreur d'observation $\boldsymbol{\varepsilon}^o$ est définie comme l'écart entre la vérité vue dans l'espace des observations et les observations :

$$\boldsymbol{y}^o = \mathcal{H}(\boldsymbol{x}^t) + \boldsymbol{\varepsilon}^o. \tag{2.2}$$

En pratique, cette erreur est supposée nulle en moyenne, ce qui s'exprime formellement par $\mathbb{E}(\boldsymbol{\varepsilon}^o) = 0$ où $\mathbb{E}$ désigne l'espérance mathématique. L'origine de cette erreur provient de deux contributions.

La première contribution correspond à l'erreur de mesure : deux thermomètres, même identiques, ne donnent jamais la même température. En pratique les origines de l'incertitude peuvent être multiples. Cela peut venir par exemple de la présence d'eau au voisinage d'un capteur embarqué à bord d'un satellite.

L'autre contribution correspond à l'erreur de représentativité : la résolution du modèle est finie, et il n'est pas rare de réaliser une mesure en un point ne correspondant pas à un point de grille du modèle. Cette erreur est donc associée aux interpolations, et aux autres défauts de l'opérateur d'observation $\mathcal{H}$. Par exemple, un satellite observe la présence d'un nuage en un point intérieur à la maille du modèle, et pas de nuage ailleurs. Si le modèle est sec sur la maille, comme il ne "voit" pas ce qui est inférieur à sa maille, l'opérateur d'observation indiquera que l'intérieur de la maille est sèche, ce qui n'est pas le cas dans la réalité.

La matrice de covariance d'erreur d'observation

$$\boldsymbol{R} = \mathbb{E}\left(\boldsymbol{\varepsilon}^o \boldsymbol{\varepsilon}^{oT}\right) \tag{2.3}$$

caractérise l'amplitude et la structure spatiale de l'erreur d'observation. En pratique, une hypothèse simplificatrice est souvent utilisée : cette matrice est supposée diagonale. Cela signifie que les erreurs de deux observations distinctes ne sont pas corrélées. Si cette hypothèse semble réaliste pour deux thermomètres distants, elle n'est plus vraie dans le cas de mesures satellitaires réalisées avec le même instrument.

L'erreur d'ébauche $\boldsymbol{\varepsilon}^b$ est définie comme l'écart entre la vérité et l'ébauche :

$$\boldsymbol{x}^b = \boldsymbol{x}^t + \boldsymbol{\varepsilon}^b. \tag{2.4}$$

En pratique, cette erreur est supposée nulle en moyenne. Cela s'exprime formellement par $\mathbb{E}(\boldsymbol{\varepsilon}^b) = 0$. L'origine de l'erreur de prévision est double.

D'une part, cette erreur provient de la croissance, au cours de la prévision, de l'erreur $\boldsymbol{\varepsilon}^a$ de l'analyse dont est issue l'ébauche. D'autre part, une autre contribution correspond à l'erreur de modèle. Il s'agit des erreurs associées aux défauts de représentation des phénomènes

dynamiques et physiques (paramétrisations, constantes du modèle mal ajustées, et autres hypothèses physiques simplificatrices permettant de prévoir le temps dans un délai raisonnable). Ces notions trouveront leurs explications aux travers des équations du filtre de Kalman-Bucy (1960).

La matrice de covariance d'erreur d'ébauche

$$\boldsymbol{B} = \mathbb{E}\left(\boldsymbol{\varepsilon}^b \boldsymbol{\varepsilon}^{b^T}\right) \tag{2.5}$$

caractérise l'amplitude et la structure spatiale de l'erreur d'ébauche. La signification ensembliste de cette matrice sera étudiée ultérieurement dans ce chapitre.

Dans la suite de la présentation, il est supposé que les erreurs d'observations et d'ébauche ne sont pas corrélées. Cela se traduit par

$$\mathbb{E}\left(\boldsymbol{\varepsilon}^o \boldsymbol{\varepsilon}^{b^T}\right) = 0, \tag{2.6}$$

Dans le cas où une telle relation ne serait pas vérifiée, il est possible de s'y ramener, en effectuant un changement de base (Talagrand, 2002).

L'équation du BLUE

La différence entre les observations et l'ébauche vue dans l'espace des observations est appelé *l'innovation* (Daley, 1991). Elle est notée $\boldsymbol{d} = \boldsymbol{y}^o - \mathcal{H}(\boldsymbol{x}^b)$. Dans la suite du paragraphe, $\mathcal{H}$ est supposé linéaire, noté $\boldsymbol{H}$. La détermination de $\boldsymbol{x}^a$ est équivalente à la recherche d'un estimateur linéaire pour l'incrément d'analyse $\delta\boldsymbol{x}^a = \boldsymbol{x}^a - \boldsymbol{x}^b$, qui s'écrit ainsi $\delta\boldsymbol{x}^a = \tilde{\boldsymbol{K}}\boldsymbol{d}$. En notant $\boldsymbol{A}_{\tilde{\boldsymbol{K}}} = \mathbb{E}\left(\boldsymbol{\varepsilon}^a \boldsymbol{\varepsilon}^{a^T}\right)$, la variance totale de l'erreur d'analyse est $Trace(\boldsymbol{A}_{\tilde{\boldsymbol{K}}})$. En remarquant que $\delta\boldsymbol{x}^a = \boldsymbol{\varepsilon}^a - \boldsymbol{\varepsilon}^b$ et $\boldsymbol{d} = \boldsymbol{\varepsilon}^o - \boldsymbol{H}\boldsymbol{\varepsilon}^b$, on peut montrer que la matrice de covariance d'erreur d'analyse associée à la matrice $\tilde{\boldsymbol{K}}$ est alors

$$\boldsymbol{A}_{\tilde{\boldsymbol{K}}} = (\boldsymbol{I} - \tilde{\boldsymbol{K}}\boldsymbol{H})\boldsymbol{B}(\boldsymbol{I} - \tilde{\boldsymbol{K}}\boldsymbol{H})^T + \tilde{\boldsymbol{K}}\boldsymbol{R}\tilde{\boldsymbol{K}}^T. \tag{2.7}$$

À partir de la trace de cette expression, et après un peu de calcul matriciel, l'unique matrice $\boldsymbol{K}$, permettant de rendre minimale la variance totale de l'erreur d'analyse, s'écrit alors

$$\boldsymbol{K} = \left(\boldsymbol{B}^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T\boldsymbol{R}.$$

Il s'agit de l'équation du BLUE (Best Linear Unbiased Estimator). La matrice $\boldsymbol{K}$ est appelée *matrice de gain*. En utilisant la formule de Sherman-Morisson-Woodbury, la matrice de gain s'écrit également sous la forme

$$\boldsymbol{K} = \boldsymbol{B}\boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^T + \boldsymbol{R}\right)^{-1}.$$

Dans ce cas, la matrice de covariance d'erreur d'analyse $\boldsymbol{A}_{\tilde{\boldsymbol{K}}}$, que l'on note simplement $\boldsymbol{A}$, peut être simplifiée en

$$\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H})\boldsymbol{B},$$

traduisant une diminution de la variance d'erreur d'ébauche par l'introduction de l'information issue de l'innovation.

Les équations du BLUE sont donc

$$\begin{cases} \boldsymbol{x}^a = \boldsymbol{x}^b + \boldsymbol{K}(\boldsymbol{y}^o - \boldsymbol{H}\boldsymbol{x}^b), \\ \boldsymbol{K} = \boldsymbol{B}\boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^T + \boldsymbol{R}\right)^{-1}, \\ \boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H})\boldsymbol{B}. \end{cases} \tag{2.8}$$

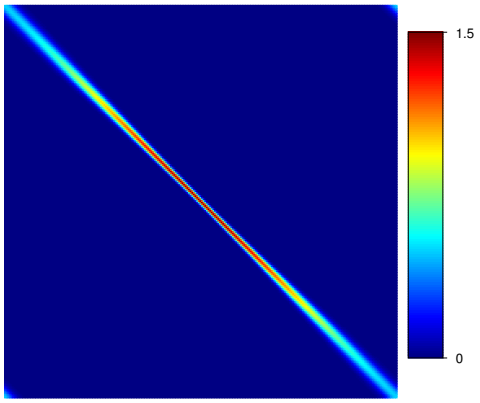


FIG. 2.1 – Représentation d'une matrice de covariance.

2.2 Information contenue dans les covariances d'erreur d'ébauche

La matrice de covariance $\boldsymbol{B}$ contient des informations complexes sur les liens statistiques existant pour l'erreur de prévision en chaque degré de liberté du modèle (par exemple la valeur d'un champ physique en un point de la grille du modèle est un degré de liberté du modèle). Ces liens statistiques sont décrits par les fonctions de covariance, également appelées fonctions de structure. Une fonction de structure caractérise ainsi le lien statistique existant entre un degré de liberté donné et tous les autres degrés de liberté du modèle. En particulier, cette fonction est tridimensionnelle spatialement.

Pour faciliter la compréhension des caractéristiques de $\boldsymbol{B}$, un champ unique sur le cercle est considéré. Sa matrice de covariance (définie ici de manière arbitraire) est représentée sur la figure 2.1. Les fonctions de covariance et d'autres composantes sont représentées sur la figure 2.2.

2.2.1 Fonctions de covariance

Une fonction de covariance $f_x(y)$ relative à un point x est définie comme étant la distribution des covariances $cov(x,y) = \mathbb{E}\left\{\varepsilon^b(x)\varepsilon^b(y)\right\}$ pour un point courant y sur le domaine. Dans le cas de l'exemple de matrice de covariance de la figure 2.1, les fonctions de covariance correspondent aux colonnes de la matrice.

Quelques fonctions de covariance sont représentées sur la figure 2.2-(a). Sur cette figure, chaque courbe colorée est une fonction de covariance. Le domaine circulaire étant périodique, une partie de la covariance associée à la position $0\,°$ (courbe colorée en bleu) se prolonge au voisinage de $360\,°$.

En comparant les fonctions de covariance entre elles, il apparaît qu'elles ont des maxima et des extensions spatiales différentes. Les maxima correspondent aux variances, tandis que l'extension spatiale est caractérisée par la longueur de portée. En effet, les fonctions de covariance au centre du domaine (vers $180\,°$) sont plus étroites et d'amplitude plus forte que les fonctions au début du domaine (vers $0\,°$). Décrivons les deux informations de variance et de portée.

2.2.2 Variance et écart type

La variance d'un paramètre, en un point de grille, quantifie la dispersion statistique de l'erreur et donc l'amplitude de l'incertitude (en valeur absolue). Pour un degré de liberté du modèle $\alpha_{k,i}$ (valeur d'un champ α_k du modèle en un point i de la grille), la variance de ce point est la

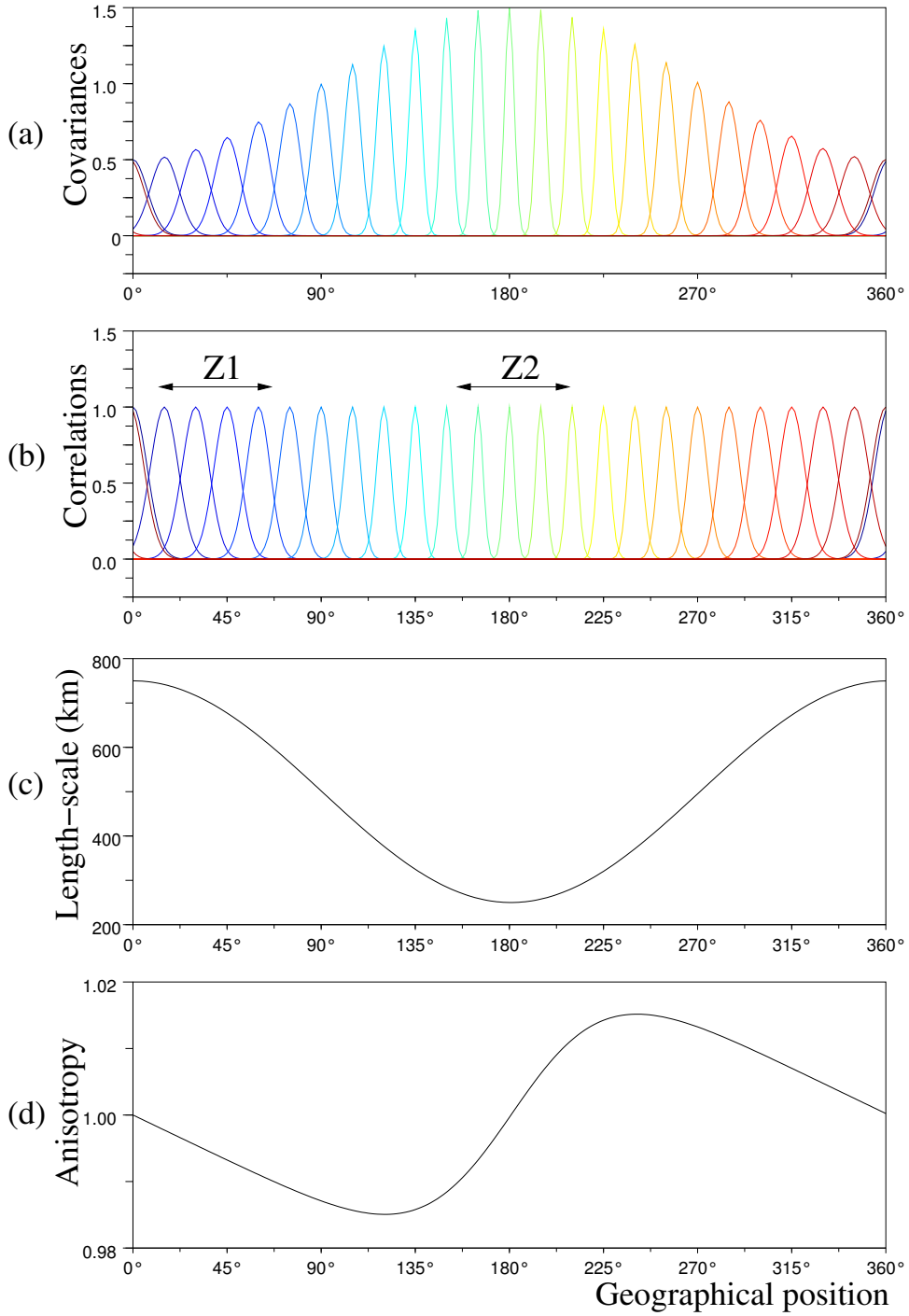


FIG. 2.2 – Représentation de quelques fonctions de covariance (a) associées à la matrice de la figure 2.1. Chaque courbe colorée correspond à une fonction de covariance relative au point à la verticale de son maximum. Les corrélations associées (b) sont caractérisées par des longueurs de portée (c) (cette longueur caractérise l'étendue spatiale de la fonction de corrélation). L'anisotropie (d) permet de diagnostiquer la dissymétrie des fonctions de corrélation : elle est supérieure à 1 pour une fonction de corrélation plus large à droite qu'à gauche.

quantité $\sigma^2_{k,i} = \mathbb{E}({\boldsymbol{\varepsilon}^b}^2_{k,i})$. Cette variance se situe sur la diagonale de la matrice de covariance $\boldsymbol{B}$. Ainsi, la carte des variances, *i.e.* le champ constitué de la variance des erreurs de prévision pour chaque champ modèle et pour chaque point de grille, est la diagonale de la matrice $\boldsymbol{B}$.

La carte de variance module l'amplitude des fonctions de covariance sur le domaine. Sur la figure 2.2-(a), les variances s'échelonnent de 0.5 au voisinage de $0\,^\circ$ à 1.5 au voisinage de $180\,^\circ$. Un autre paramètre classique est l'écart type $\sigma_{k,i}$, qui correspond à la racine carrée de la variance, et qui peut être vu comme une amplitude moyenne de l'erreur.

L'information d'amplitude est caractérisée par la variance (ou l'écart type), mais elle est indépendante de l'information d'extension spatiale des fonctions de covariance. Cette information est contenue dans les corrélations.

2.2.3 Fonctions et matrice de corrélation

Les fonctions de corrélation $f^c_x(y)$ sont déduites des fonctions de covariance d'après $f^c_x(y) = \frac{f_x(y)}{\sigma_x \sigma_y} = cor(x, y)$.

La matrice de corrélation $\boldsymbol{C}$ se déduit de la matrice de covariance $\boldsymbol{B}$ après normalisation de cette dernière à l'aide des écarts types. En notant $\boldsymbol{\Sigma}$ la matrice diagonale constituée des écarts types $(\sigma_{k,i})$, la matrice de corrélation est donnée par $\boldsymbol{C} = \boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{\Sigma}^{-T}$. Chaque colonne de cette matrice $\boldsymbol{C}$ est une fonction de corrélation associée à un degré de liberté.

Dans le cas où la fonction de corrélation est la même en tout point du domaine, alors la matrice de corrélation est dite *homogène*, sinon elle est dite *hétérogène*.

La figure 2.2-(b) représente quelques fonctions de corrélation associées à la matrice représentée sur la figure 2.1. Cette fois-ci, il n'y a plus de modulation d'amplitude : le maximum des fonctions de corrélation est 1. Les variations de l'extension spatiale sont alors plus visibles. Ainsi les fonctions au voisinage de $0\,^\circ$ sont plus larges que les fonctions étroites au voisinage de $180\,^\circ$.

Cette extension spatiale est caractérisée par une échelle de longueur : la longueur de portée.

2.2.4 Longueur de portée

La longueur de portée caractérise l'influence d'une erreur en un point sur les autres points. Si une erreur de prévision est commise en un point, la structure spatiale de l'écoulement implique une cohérence spatiale de cette erreur qui est plus ou moins locale, selon le paramètre et la situation météorologique en jeu. Le degré de localisation dépend du degré de liberté considéré : ainsi, les longueurs de portée sont plus grandes pour les champs de grande échelle (tels que la pression de surface) que pour les champs de petite échelle (tels que la température de surface).

Daley (1991) propose une définition de la longueur de portée similaire à la définition de la micro-échelle turbulente (micro-échelle de Taylor). Formellement, cette échelle s'écrit

$$L^2_D = -\frac{2\rho(0)}{\nabla^2\rho(0)}, \tag{2.9}$$

dans le cas 2D, avec ρ la fonction de corrélation. Dans le cas 1D, cette longueur est $L^2_D = -\frac{\rho(0)}{\nabla^2\rho(0)}$. Il est possible d'en donner une interprétation géométrique, à l'aide de la parabole osculatrice de la fonction de corrélation à l'origine. En effet, comme l'illustre la figure 2.3, cette longueur correspond à la distance pour laquelle cette parabole (courbe tiretée) vaut $1/2$. Ainsi définie, cette longueur de portée permet un diagnostic de la forme de la fonction de corrélation

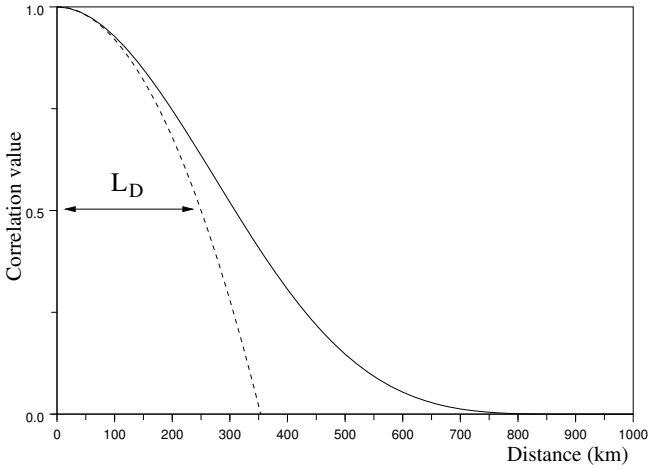


FIG. 2.3 – Représentation de la longueur de portée définie par Daley. La corrélation (courbe continue) est approximée à l'origine par sa parabole osculatrice (courbe tiretée). La portée correspondante est représentée par la flèche L_D.

en son origine. C'est un diagnostic important qui sera utilisé pour caractériser l'hétérogénéité spatiale des fonctions de corrélation.

En effet, l'hétérogénéité d'une matrice de corrélation est caractérisée par une variation géographique de la portée. Dans le cas homogène, la portée est constante (ou homogène) sur le domaine. La réciproque n'est pas vraie : ce n'est pas parce que la portée est homogène que la matrice de covariance l'est.

La figure 2.2-(c) représente le champ de longueur de portée pour la matrice $\boldsymbol{B}$ considérée. Ainsi, la variation de l'extension spatiale déjà mentionnée se retrouve bien sur ce diagnostic : la portée au voisinage de $0\,°$, de l'ordre de $750\,km$ est plus large que celle au voisinage de $180\,°$, de l'ordre de $250\,km$. Ces variations géographiques confirment que la matrice de covariance considérée est bien hétérogène.

Il est à noter que, en pratique, cette définition de la portée, donnée par l'équation (2.9), est d'usage limité. En effet, le calcul du laplacien nécessite, en théorie, la connaissance de l'ensemble de la fonction de corrélation. Il est donc nécessaire de connaître en principe la fonction de corrélation dans son intégralité.

Le chapitre 5 de ce manuscrit s'intéresse en particulier à la formulation d'approximations de cette longueur, pour la diagnostiquer dans un contexte opérationnel, et pour des géométries complexes (comme celles rencontrées en océanographie). En particulier, ces diagnostics permettent de définir une portée directionnelle à gauche Lp^- et à droite Lp^+, à partir desquelles une caractérisation de l'anisotropie sur le cercle est possible (par exemple le ratio $\frac{Lp^+}{Lp^-}$ représenté fig. 2.2-(d)). Il est à noter que le diagnostic de la portée peut servir à estimer le tenseur local de diffusion (Pannekoucke et Massart, 2008) dans la modélisation des covariances d'erreur d'ébauche basée sur un opérateur de diffusion (Weaver et Courtier, 2001).

2.3 Filtrage et propagation de l'innovation

2.3.1 Assimilation d'une seule observation

Pour illustrer les propriétés de l'analyse, un exemple classique est de considérer l'assimilation d'une seule observation.

En considérant une observation unique y^o, localisée en un point de la grille du modèle où l'ébauche vaut x^b, l'opérateur d'observation associé est une projection de l'espace du modèle sur le point de grille observé. Soit $\boldsymbol{\rho}$ la fonction de corrélation associée au point de grille, σ_b^2 la

variance de l'erreur de prévision en ce point et σ_o^2 la variance de l'erreur d'observation. Ainsi, l'incrément d'analyse $\delta \boldsymbol{x}^a = \boldsymbol{x}^a - \boldsymbol{x}^b$ s'écrit, d'après l'équation (2.8), $\delta \boldsymbol{x}^a = \boldsymbol{\rho}\,(y^o - x^b)/(1 + \frac{\sigma_o^2}{\sigma_b^2})$. Dans cette expression, l'innovation $y^o - x^b$ est filtrée par $\tau = 1/(1 + \frac{\sigma_o^2}{\sigma_b^2})$, avec le facteur τ variant de 0 à 1 comme coefficient de filtrage. Ce filtrage de l'innovation produit l'incrément de correction de l'ébauche au point d'observation. La corrélation ρ propage aux autres points du modèle l'innovation filtrée, donnant la correction finale. Dans le cas où $\tau = 1$, on fait plus confiance à l'observation qu'à l'ébauche : l'analyse vérifie au point d'observation $x^a = y^o$. A l'inverse, dans le cas où $\tau = 0$, on fait davantage confiance à l'ébauche qu'à l'observation et l'analyse vérifie au point d'observation $x^a = x^b$: l'innovation est entièrement filtrée. Dans le cas général, l'analyse vérifie au point d'observation $x^a = (1 - \tau)\, x^b + \tau y^o$ donnant une valeur intermédiaire entre l'ébauche et l'observation.

On perçoit ainsi l'enjeu important que revêt la modélisation de $\boldsymbol{B}$: si la modélisation est erronée, la qualité spatiale de la correction est dégradée.

2.3.2 Filtrage et propagation dans le formalisme du BLUE

De façon générale, la matrice $\boldsymbol{K}$ définie par l'équation (2.8) apparaît comme une matrice de filtrage et de propagation du vecteur d'innovation $\boldsymbol{d} = \boldsymbol{y}^o - \boldsymbol{H}\boldsymbol{x}^b$. Plus formellement, en introduisant la matrice $\boldsymbol{HBH}^T$ dans l'expression de $\boldsymbol{K}$ (Hollingsworth, 1987), il vient

$$\delta \boldsymbol{x}^a = \boldsymbol{B}\boldsymbol{H}^T(\boldsymbol{HBH}^T)^{-1}(\boldsymbol{HBH}^T)(\boldsymbol{HBH}^T + \boldsymbol{R})^{-1}\boldsymbol{d}. \tag{2.10}$$

Ainsi, dans un premier temps, l'innovation est filtrée par l'opérateur $(\boldsymbol{HBH}^T)(\boldsymbol{HBH}^T + \boldsymbol{R})^{-1}$ (correspondant au facteur τ mis en évidence dans l'exemple à une observation), puis propagée par l'opérateur $\boldsymbol{B}\boldsymbol{H}^T(\boldsymbol{HBH}^T)^{-1}$ (correspondant au rôle de la corrélation ρ dans le cas à une observation). La position finale de la matrice $\boldsymbol{B}$ dans l'expression indique que c'est elle qui propage l'information d'ajustement issue de l'innovation $\boldsymbol{d}$. En particulier, dans le cas où le vecteur d'état est représenté en points de grille, c'est $\boldsymbol{B}$ qui propage spatialement la correction.

2.3.3 Commentaires sur l'opérateur KH

$\boldsymbol{KH}$ est un autre opérateur, qui se retrouve dans l'équation de la matrice de covariance d'erreur d'analyse (2.8). Il s'interprète comme une matrice (en supposant l'opérateur linéaire) variant de 0 à $\boldsymbol{I}_n$ (Bouttier, 1994b).

Si $\boldsymbol{KH} = 0$, alors $\boldsymbol{x}^a = \boldsymbol{x}^b$: dans ce cas, aucune confiance aux observations n'est faite, ou bien elles ne sont pas assez nombreuses pour être prises en compte ; c'est ce qui arrive dans les zones pauvres en observations (océan Pacifique, Atlantique Sud, pôles).

Si $\boldsymbol{KH} = \boldsymbol{I}_n$, alors dans ces régions, $\boldsymbol{x}^a = \boldsymbol{K}\boldsymbol{y}^o$: la contribution des observations est plus grande que celle apportée par l'ébauche.

Dans la pratique, dans les zones denses en observations (Europe, Amérique du nord), $\boldsymbol{KH} = \mathbf{I}_n/2$: les observations et l'ébauche ont un poids semblable.

2.3.4 Illustration sur le cercle

Les points précédents s'illustrent simplement en considérant un exemple 1D, avec une matrice $\boldsymbol{B}$ analytique hétérogène en longueur de portée, et une variance homogène unité. La matrice $\boldsymbol{R}$ est définie par $\boldsymbol{R} = \sigma^{o2}\boldsymbol{I}_p$, avec $\sigma^o = 0.6$. Les observations sont supposées coïncider

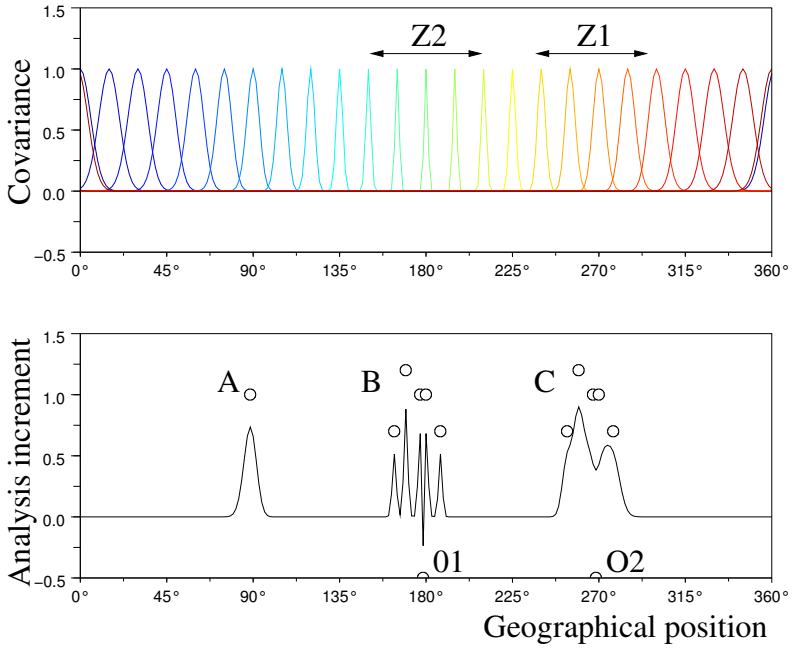


FIG. 2.4 – Expérience d'assimilation de données sur le cercle avec une matrice $\boldsymbol{B}$ hétérogène. En haut : représentation de quelques covariances de la matrice $\boldsymbol{B}$ utilisée. En bas : incrément d'analyse pour trois motifs d'observation : A désigne une observation unique, B et C correspondent à un même motif, mais positionné dans deux zones différentes $Z2$ (zone de portées courtes) et $Z1$ (zone de portées larges). Les ronds symbolisent la valeur de l'observation dont la position correspond à l'abscisse.

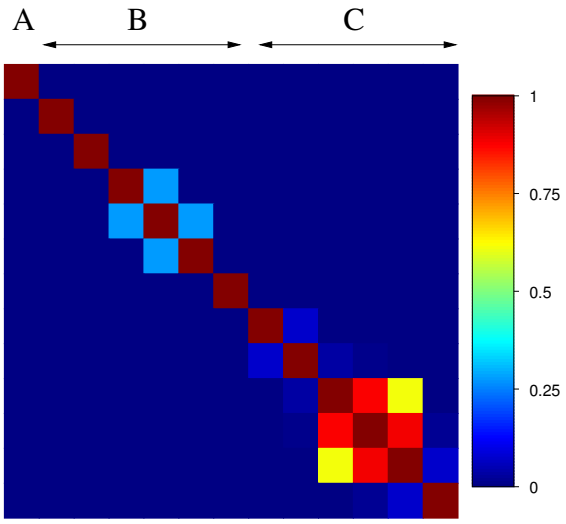


FIG. 2.5 – Représentation de la matrice $\boldsymbol{HBH}^T$ correspondant à la matrice de covariance d'erreur d'ébauche vue dans l'espace des observations. Ces 13 colonnes correspondent aux 13 positions d'observations représentées sur le graphique du bas de la figure 2.4. La première colonne correspond à l'observation isolée désigné par A, les six colonnes suivantes sont associées au motif B et les six dernières sont associées au motif C.

avec les composantes du vecteur d'état, de sorte que $\boldsymbol{H}$ correspond à un simple opérateur de projection. Les résultats sont reportés sur la figure 2.4.

Le graphique du haut sur la figure 2.4 représente les fonctions de corrélation en différents points répartis de manière uniforme sur le cercle (1 à 239 par pas de 20). La variation de longueur de portée est caractérisée par des fonctions de corrélation étroites dans la zone de courtes longueurs de portée (zone Z1) et par des fonctions plus larges dans la zone de larges portées (zone Z2).

Le graphique du bas sur la figure 2.4 représente le résultat d'une expérience d'assimilation d'observations. Si les groupes d'observations sont suffisamment éloignés, alors chaque groupe peut être considéré comme indépendant des autres dans l'assimilation. Dans cette expérience, on se donne un vecteur d'innovation $\boldsymbol{d}$ (les petits ronds de la figure b) considéré comme un vecteur d'observations si l'on se place dans le cas $\boldsymbol{x}^b = 0$, et on représente l'incrément d'analyse $\delta\boldsymbol{x}^a = \boldsymbol{K}\boldsymbol{d}$. Les lettres A, B et C représentent différentes zones de l'expérience. Du fait de leur éloignement, elles peuvent être considérées comme 3 expériences indépendantes (ceci est également justifié dans la suite avec l'étude approfondie de la matrice $\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^T$).

Le résultat de l'assimilation d'une seule observation est caractérisé par la lettre A. L'incrément d'analyse est alors proportionnel à la covariance au point d'observation (c'est ce qui a été montré au paragraphe 2.3.1). Les zones B et C sont le résultat de l'assimilation d'un même motif de répartition d'observations mais dans des zones de longueurs de portée différentes : $Z1$ pour B et $Z2$ pour C. La longueur de portée étant courte pour B, l'incrément d'analyse est très oscillant, comme si les observations étaient prises séparément. La correction apportée par l'incrément d'analyse est très locale et reste proportionnelle à la covariance en chaque point observé. Dans ce cas, la matrice $\boldsymbol{K}$ n'a pas beaucoup filtré l'innovation (la correction descend fortement vers l'observation $O1$ par exemple). Par contre, pour C, localisé dans une zone de portées larges, la correction est plus étendue et n'est plus localisée de façon prononcée en chaque observation. En particulier, les variations rapides sont lissées (la correction descend très faiblement vers l'observation $O2$ notamment), ce qui est dû au filtrage apporté par $\boldsymbol{K}$, et en particulier à la matrice $\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^T$.

L'analyse de ce filtrage peut donc être faite en examinant la matrice $\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^T$, représentée sur la figure 2.5.

Il apparaît que cette matrice est diagonale par bloc, avec un bloc associé à l'observation unique A, et aux deux motifs B et C. Le bloc associé au motif B, dans la zone Z2 où les portées sont courtes, est visiblement très proche d'une diagonale. Ainsi, chaque emplacement d'observation est vu comme étant décorrélé des autres pour l'erreur d'ébauche. L'analyse des observations prises dans leur ensemble revient (en première approximation) à l'analyse de chaque observation prise séparément.

Ce n'est plus le cas pour le bloc associé au motif C. En effet, ce motif, situé dans la zone Z1 de portées larges, présente de fortes corrélations, en particulier pour les localisations des trois observations rapprochées (sous-diagonale rouge et jaune). Ces trois localisations sont donc vues comme étant couplées les unes avec les autres, et l'analyse résultante des observations en ces points apparaît comme une moyenne pondérée par ces couplages, induisant le filtrage spatial. Il n'est plus possible de considérer l'analyse des trois localisations comme étant trois analyses séparées.

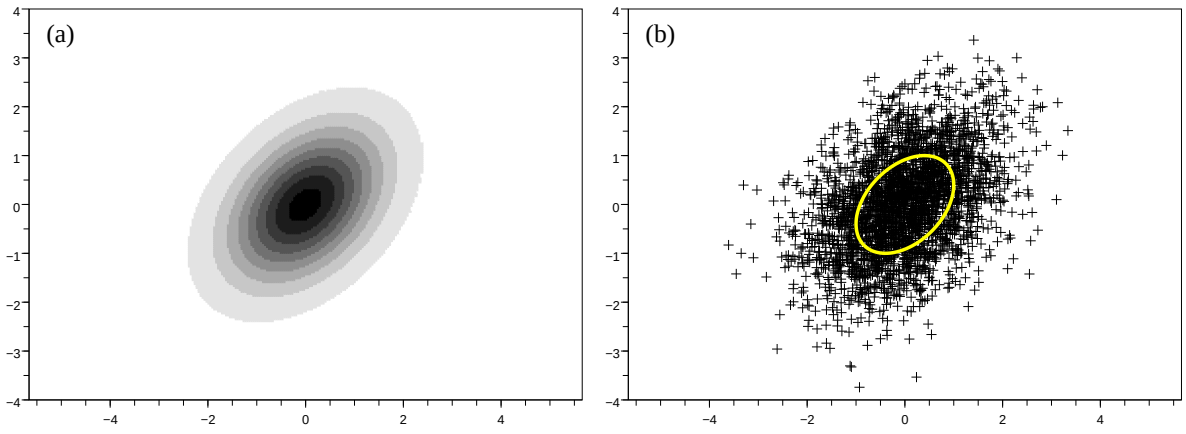


FIG. 2.6 – Représentation de l'allure d'une distribution gaussienne (a) et sa discrétisation à l'aide d'un échantillon de taille finie ($N_s = 3000$) (b). En (a), l'échelle de niveau de gris correspond à l'intensité de la probabilité : noir pour très probable, et blanc pour très peu probable. En (b) l'ellipse jaune correspond à l'ellipse caractéristique de la dispersion associée à $\boldsymbol{B}$ (analogue au cercle de rayon σ^b dans le cas d'une matrice de covariance $\boldsymbol{B} = \sigma^{b^2}\boldsymbol{I}_2$).

2.4 Équations du filtre de Kalman-Bucy

Une manière d'introduire l'évolution de l'erreur d'analyse (et de percevoir l'aspect ensembliste, qui sera évoqué ultérieurement) est de décrire l'échantillonnage associé à une distribution. L'ensemble ainsi constitué évolue par intégration du modèle, ce qui complète les équations du BLUE, pour conduire aux équations du filtre de Kalman-Bucy (Kalman, 1960). Ces étapes sont décrites maintenant.

2.4.1 Distribution de probabilité et échantillonnage

Dans le cas où la distribution de l'erreur d'ébauche est gaussienne, alors $\boldsymbol{B}$ caractérise entièrement cette distribution. La densité de probabilité est alors donnée par

$$p(\boldsymbol{\varepsilon}^b) = p(\boldsymbol{x}^b/\boldsymbol{x}^t) = \frac{1}{(2\pi\,|det\boldsymbol{B}|)^{n/2}} exp\left\{-\frac{1}{2}||\boldsymbol{\varepsilon}^b||^2_{\boldsymbol{B}^{-1}}\right\}. \tag{2.11}$$

Le graphique de gauche sur la figure 2.6-(a) représente l'allure en 2D de ce type de distribution. Une autre manière d'aborder une distribution de probabilité est de la considérer comme la limite infinie d'un échantillonnage. Par exemple, la figure 2.6-(b) représente un échantillonnage de la distribution gaussienne représentée sur la figure 2.6-(a), avec un échantillon comprenant $N_s = 3000$ points.

D'autre part, pour générer un ensemble d'états vérifiant une certaine distribution gaussienne caractérisée par une matrice $\boldsymbol{B}$ (Fisher et Courtier, 1995), il suffit de considérer une racine carré $\boldsymbol{B}^{1/2}$ de $\boldsymbol{B}$, définie telle que $\boldsymbol{B} = \boldsymbol{B}^{1/2}\boldsymbol{B}^{T/2}$. Une erreur d'ébauche $\boldsymbol{\varepsilon}^b$ compatible avec $\boldsymbol{B}$ est obtenue en générant un vecteur $\boldsymbol{\zeta}$ suivant une loi gaussienne de matrice de covariance $\boldsymbol{I}_n$ [1], puis en le transformant à l'aide de $\boldsymbol{B}^{1/2}$:

$$\boldsymbol{\varepsilon}^b = \boldsymbol{B}^{1/2}\boldsymbol{\zeta}. \tag{2.12}$$

[1] La génération d'un tel vecteur $\boldsymbol{\zeta}$ est aisée : si ce vecteur contient n composantes, alors il suffit de générer n réalisations indépendantes d'une variable normale. Le vecteur résultant suit une loi gaussienne de matrice de covariance $\boldsymbol{I}_n$.

C'est de cette manière que la figure 2.6-(b) a été obtenue.

Le volume d'information représenté par la dispersion de l'ensemble d'états peut être également associé à la notion d'entropie. Plus le volume est important, plus l'entropie est grande.

Dans le cas où la distribution n'est pas gaussienne, cette correspondance entre distribution analytique et échantillonnage est toujours valable, quand la taille de l'échantillon tend vers l'infini. Il est à noter, cependant, que générer une loi complexe dans des espaces de dimension supérieure à 4 ou 5 devient rapidement difficile (Tarantola, 2005). Dans ce cas, d'autres stratégies doivent être mises en oeuvre, *e.g.* l'utilisation d'une marche aléatoire orientée, telle que l'algorithme de Metropolis.

Dans la suite de ce chapitre, pour les applications réelles, les distributions considérées sont supposées gaussiennes.

2.4.2 Évolution temporelle de l'erreur d'analyse et l'erreur modèle

Évolution dans l'approche modèle parfait

Pour décrire l'évolution de l'erreur d'analyse, un point de départ est la correspondance entre une distribution de probabilité et son échantillonnage sous la forme d'un ensemble d'états. Ainsi, pour décrire l'évolution temporelle d'une distribution initiale, soumise à une dynamique, il suffit d'observer l'évolution temporelle de chaque élément de l'échantillon. La dynamique correspond ici à l'intégration par le modèle atmosphérique non linéaire, noté formellement sous la forme d'un opérateur $\mathcal{M}_{0\to T}$, faisant passer un état $\boldsymbol{x}(0)$ à l'instant $t = 0$ à un état $x(T) = \mathcal{M}_{0\to T}\{\boldsymbol{x}(0)\}$ à l'instant $t = T$, en résolvant une équation du type

$$\frac{d\boldsymbol{x}}{dt} = \mathcal{F}(\boldsymbol{x}, t). \tag{2.13}$$

Le schéma représenté sur la figure 2.7, illustre la dynamique de l'erreur d'analyse, supposée gaussienne à l'instant $t = 0$ et sphérique ; *i.e.* la matrice de covariance d'erreur d'analyse est de la forme $\boldsymbol{A} = \sigma^{a2}\boldsymbol{I}$. La dispersion de cette distribution est caractérisée par la matrice $\boldsymbol{A}$, et elle est représentée sur le graphique par le cercle en gras associé à la distribution (a).

La correspondance densité-échantillon est utilisée pour décrire la distribution à des instants ultérieurs. En examinant l'évolution temporelle, sur un temps court, de chaque état compatible avec la distribution initiale et correspondant à l'échantillonnage, il apparaît que la distribution sphéroïdale initiale est déformée en une distribution ellipsoïdale : il s'agit de l'ellipse en gras associée à la distribution (b). Cette déformation correspond au stade linéaire pour lequel la différence $\boldsymbol{x}^b(T) - \boldsymbol{x}^t(T) = \mathcal{M}_{0\to T}\left[\boldsymbol{x}^a(0)\right] - \mathcal{M}_{0\to T}\left[\boldsymbol{x}^t(0)\right]$ peut être approximée par

$$\mathcal{M}_{0\to T}\left[\boldsymbol{x}^a(0)\right] - \mathcal{M}_{0\to T}\left[\boldsymbol{x}^t(0)\right] \approx \boldsymbol{M}_{0\to T}\boldsymbol{\varepsilon}^a(0), \tag{2.14}$$

où $\boldsymbol{M}_{0\to T}$ désigne le modèle linéaire tangent à la trajectoire non linéaire $\{\mathcal{M}_{0\to t}\left[\boldsymbol{x}^a(0)\right]\}_{t\in[0,T]}$. Or $\boldsymbol{\varepsilon}^b(T) = \boldsymbol{x}^b(T) - \boldsymbol{x}^t(T)$. Ainsi, $\boldsymbol{\varepsilon}^b(T)$ est de distribution gaussienne, puisque résultant de la transformation linéaire d'un vecteur aléatoire gaussien, caractérisée par la matrice de covariance $\boldsymbol{B}(T) = \mathbb{E}\left[\boldsymbol{\varepsilon}^b(T)\boldsymbol{\varepsilon}^b(T)^T\right]$:

$$\boldsymbol{B}(T) = \boldsymbol{M}_{0\to T}\boldsymbol{A}{\boldsymbol{M}_{0\to T}}^T. \tag{2.15}$$

Pour un temps ultérieur plus long ($T \to T'$), l'approximation donnée par l'équation (2.14) n'est plus valable, et les non linéarités déforment alors l'ellipsoïde (b) en une forme délimitée par la courbe fermée en gras (c). La distribution de l'erreur n'est plus gaussienne. Il est tout de même possible de rechercher la distribution gaussienne approximant au mieux cette distribution

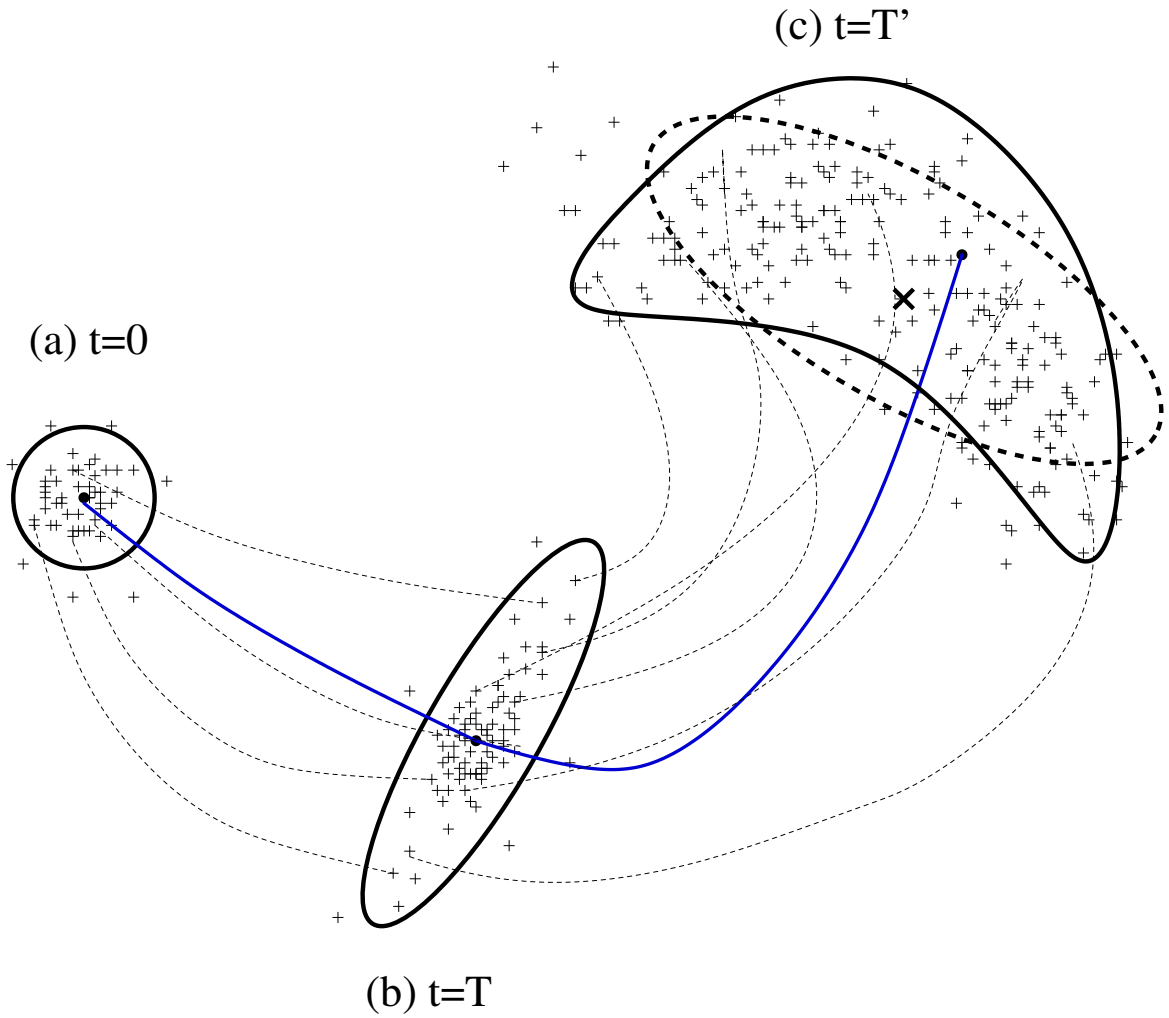


FIG. 2.7 – Schéma de l'évolution de l'erreur d'analyse au cours de l'intégration non linéaire $\mathcal{M}_{0\to t}$ du modèle, en observant l'évolution des nuages de points correspondant aux distributions, dans l'espace des phases (espace du modèle). Au temps $t = 0$ (a), la dispersion de l'erreur caractérisée par la matrice $\boldsymbol{A}$ est supposée sphérique. Au stade linéaire, cette dispersion est dilatée suivant des axes privilégiés donnant une dispersion elliptique caractérisée par $\boldsymbol{B}(T)$ au temps $t = T$ (b). Au stade non linéaire, pour le temps $t = T'$ (c), la dispersion est alors déformée, avec un contour en forme de croissant. La dispersion elliptique de la gaussienne équivalente est représentée en tireté gras, centrée sur la croix en gras.

complexe. Le résultat est représenté sur le schéma par l'ellipsoïde en tireté gras, dont le centre est noté par la croix en gras.

Pour les temps encore plus longs, dans le cas d'une dynamique chaotique, le nuage de points éclate en des structures complexes à dimensions fractales (Bergé *et al.* , 1997 ; Manneville, 2004).

Erreur modèle

Cette approche de l'évolution de l'erreur d'analyse est cependant incomplète. En effet, elle ne prend pas en compte les défauts du modèle qui entraînent des dérives par rapport à la réalité.

Or l'estimation de cette erreur modèle est difficile à obtenir, et la question de sa modélisation est encore une question ouverte. Certains préconisent de perturber la physique, d'autres imaginent une modélisation stochastique du type "marche au hasard".

Cette dernière consiste à ajouter à la prévision $\boldsymbol{x}^b(T)$ un terme de correction $\boldsymbol{\eta}(T)$, tel que ce vecteur suit une loi gaussienne de moyenne nulle et de matrice de covariance $\boldsymbol{Q}(T)$. Ainsi, l'équation de $\boldsymbol{B}(T)$ est alors complétée d'après

$$\boldsymbol{B}(T) = \boldsymbol{M}_{0\to T}\boldsymbol{A}{\boldsymbol{M}_{0\to T}}^T + \boldsymbol{Q}(T), \tag{2.16}$$

ce qui correspond à une augmentation de la dispersion caractérisant la croissance linéaire de l'erreur $\boldsymbol{M}_{0\to T}\boldsymbol{A}{\boldsymbol{M}_{0\to T}}^T$. Ainsi le volume caractérisant le nuage de points $\boldsymbol{M}_{0\to T}\boldsymbol{A}{\boldsymbol{M}_{0\to T}}^T$ augmente, ce qui correspond également à une augmentation de l'entropie. Cette équation d'évolution linéaire de $\boldsymbol{A}$ complète les équations du BLUE (2.8).

2.4.3 Filtre de Kalman-Bucy

On peut donc considérer le cycle suivant. Une ébauche est utilisée pour fournir une analyse suivant les équations du BLUE (2.8). Le volume de dispersion associé à $\boldsymbol{B}$ se réduit à un volume plus petit $\boldsymbol{A}$. Puis l'intégration par le modèle amplifie ces erreurs résiduelles d'après un schéma de la forme Eq. (2.16). La figure 2.8 représente schématiquement cette évolution de l'incertitude entre les phases d'analyse (contraction de la dispersion) et de prévision (dilatation de la dispersion).

Le filtre de Kalman-Bucy (Kalman, 1960) n'est autre que ce cycle, valable pour une évolution linéaire. Il s'écrit

$$\left\{\begin{array}{l} {\boldsymbol{x}^a}_k = {\boldsymbol{x}^b}_k + \boldsymbol{K}_k(\boldsymbol{y}^o{}_k - \boldsymbol{H}_k{\boldsymbol{x}^b}_k), \\ \boldsymbol{K}_k = \boldsymbol{B}_k{\boldsymbol{H}_k}^T\left(\boldsymbol{H}_k\boldsymbol{B}_k{\boldsymbol{H}_k}^T + \boldsymbol{R}_k\right)^{-1}, \\ \boldsymbol{A}_k = (\boldsymbol{I}_k - \boldsymbol{K}_k\boldsymbol{H}_k)\boldsymbol{B}_k, \\ {\boldsymbol{x}^b}_{k+1} = \boldsymbol{M}_{k\to k+1}\,{\boldsymbol{x}^a}_k, \\ \boldsymbol{B}_{k+1} = \boldsymbol{M}_{k\to k+1}\,\boldsymbol{A}_k\,{\boldsymbol{M}_{k\to k+1}}^T + \boldsymbol{Q}_{k+1}, \end{array}\right. \tag{2.17}$$

en considérant l'instant initial d'indice k, et l'instant ultérieur d'indice $k+1$, et la dynamique par le modèle linéaire $\boldsymbol{M}_{k\to k+1}$.

Le cas où le modèle d'évolution n'est plus linéaire correspond au filtre de Kalman étendu (Evensen, 1994), *i.e.* ${\boldsymbol{x}^b}_{k+1} = \mathcal{M}_{k\to k+1}({\boldsymbol{x}^a}_k)$.

2.4.4 Exemple de dynamique des covariances sur le cercle

Il est facile de construire un filtre de Kalman pour un modèle simple.

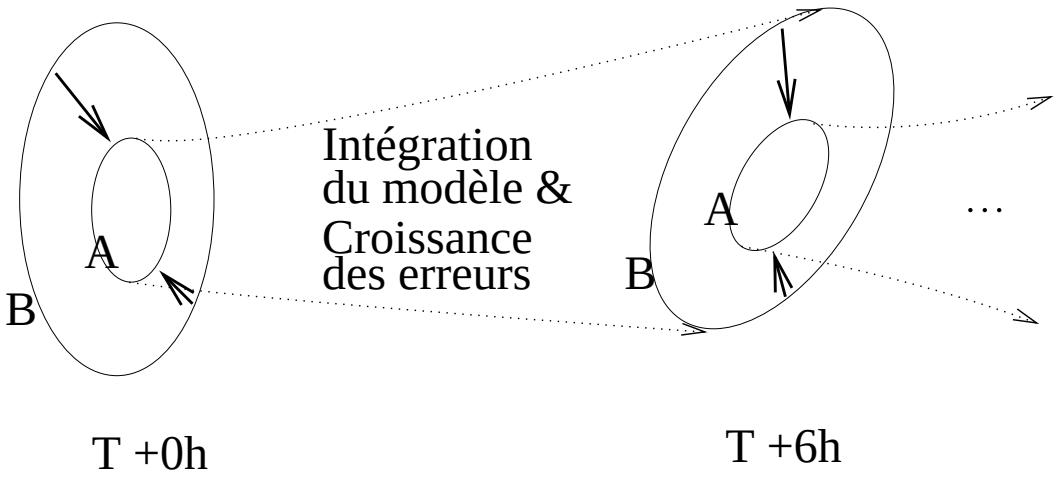


FIG. 2.8 – Représentation schématique de l'évolution de l'incertitude au cours des cycles d'analyse/prévision. L'analyse diminue l'incertitude, tandis que la prévision l'augmente.

Ainsi, sur le cercle discrétisé à l'aide de $N_g = 301$ points, une dynamique de transport par advection est un modèle simple mais déjà réaliste dans une certaine mesure. Pour fixer les idées, le cercle est supposé correspondre au cercle équatorial de la Terre. La vitesse d'advection est $U = 20\,m.s^{-1}$ et la durée de l'intégration est de $T = 6\,h$. Ainsi, la translation sur la période est de $\mathcal{L}_{U,T} = 430\,km$. Pour représenter une croissance de l'erreur, un coefficient d'inflation $\alpha = 1.02$ complète le modèle et permet d'amplifier les covariances d'un facteur α^2 toutes les six heures. Ainsi, le modèle appliqué à un état $f(x)$ s'écrit

$$\mathcal{M}_{t\to t+T}(f) = \alpha f(x - \mathcal{L}_{U,T}).$$

Ce modèle étant linéaire, les équations du filtre de Kalman peuvent être utilisées pour décrire la dynamique des covariances. Le modèle est supposé parfait : $\boldsymbol{Q} = 0$.

Le réseau d'observation considéré est hétérogène, avec une observation en chaque point de grille mais uniquement dans la zone de $90\,^\circ$ à $270\,^\circ$. La matrice de covariance d'erreur d'observation est $\boldsymbol{R} = {\sigma^o}^2 \boldsymbol{I}$, avec $\sigma^o = 0.95$. Comme condition initiale pour cette évolution, une matrice de covariance d'erreur d'ébauche $\boldsymbol{B}_{k=0}$ est crée. Cette matrice est construite avec une portée globalement homogène de valeur $L_H = 500\,km$, sauf au voisinage de $40\,^\circ$, où elle présente un raccourcissement des portées. Cette structure de courtes portées est notée L par la suite.

La figure 2.9 représente la dynamique des covariances d'erreur débauche. Cette dynamique est observée sur quelques fonctions de covariance (colonne de gauche), ainsi que sur quelques fonctions de corrélation (colonne de droite). Le diagnostic des portées, calculé à partir de la définition de Daley, permet de faire une synthèse de l'évolution.

Un premier point remarquable est la diminution des portées au dessus de la zone observée (colonne de droite). Après un cycle d'analyse/prévision, la portée des covariances d'erreur d'ébauche a diminué de 25% ($500\,km$ initialement pour atteindre $375\,km$ après le premier cycle $k = 1$). Cette décroissance se prolonge au cours des autres cycles avec un décrément du même ordre.

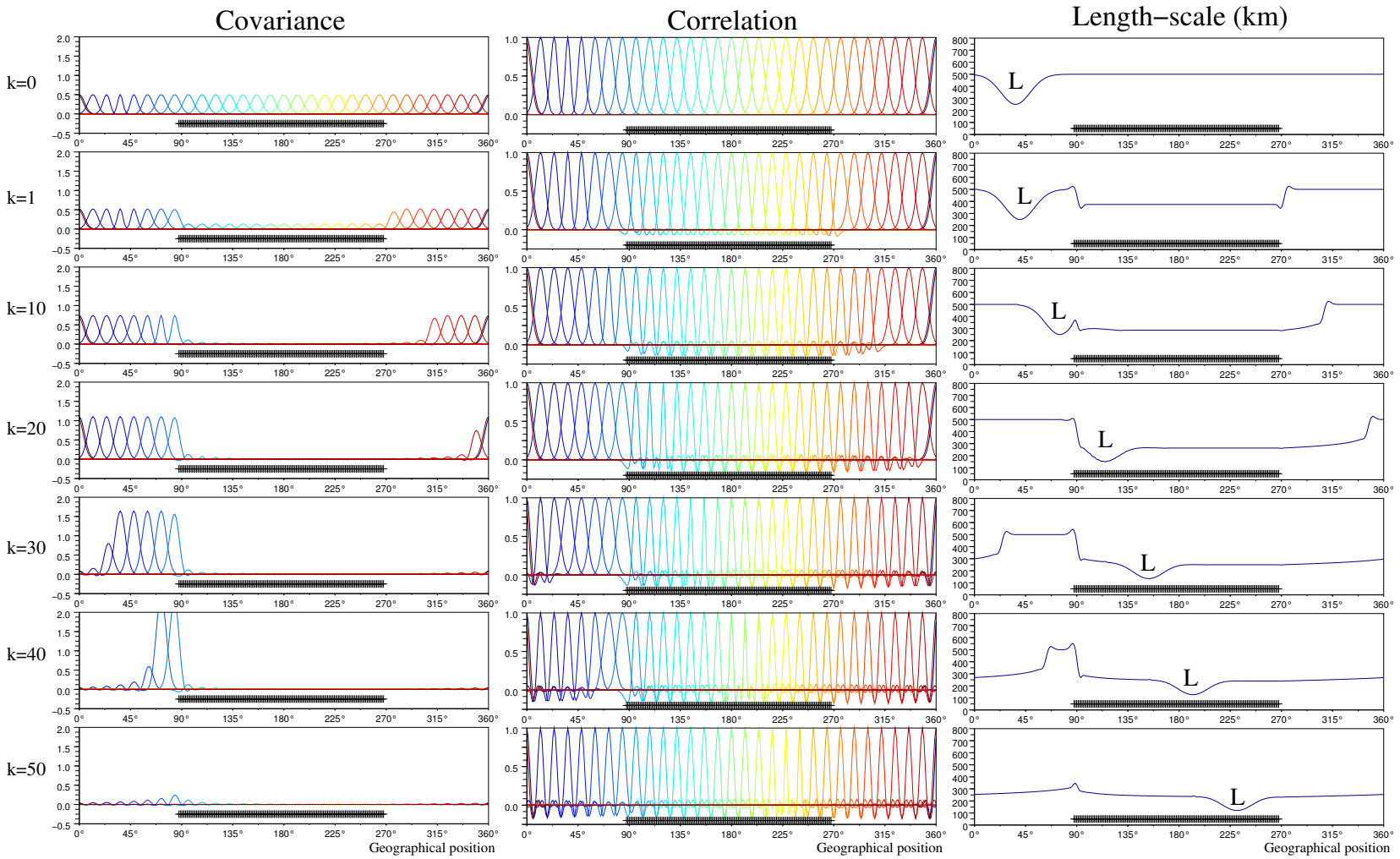


FIG. 2.9 – Exemple de dynamique des covariances au cours de cycles k pour un modèle parfait d'advection-amplification. Les covariances d'erreur d'ébauche initiale ($k = 0$) évoluent au cours des cycles de prévision ($k > 0$). Quelques fonctions de covariance sont représentées sur la colonne de gauche. Chaque courbe colorée est une fonction de covariance relative au point à la verticale du maximum. Au centre quelques fonctions de corrélation associés aux même emplacement que les fonctions de covariance de la colonne de gauche. À droite la portée diagnostiquée à l'aide de la formule de Daley Eq. (2.9) pour le cas 1D. L désigne une structure de portée plus courte. Il est possible de suivre cette structure au cours des cycles : il s'agit d'une structure cohérente de portée. Sur l'ensemble des graphiques, la localisation du réseau d'observation est désignée par des croix entre 90 ° et 270 °.

De plus, la structure de portée plus courte, repérée par L, peut être suivie au cours de l'ensemble des cycles. Elle évolue en se translatant sur le domaine et en subissant une atténuation (sans disparaître pour autant) dans la zone observée. Il est possible d'établir une analogie avec la notion de structure cohérente (McWilliams, 1984 ; Farge, 1992) telle que rencontrée en dynamique des tourbillons et de l'atmosphère. En effet, pouvant suivre cette structure au cours du temps, le terme de structure cohérente de portée peut être employé pour désigner ce type de structure évoluant dans l'écoulement. Des structures de ce type ont été observées en étudiant des ensembles de prévisions perturbées Arpège ; elles seront vues au chapitre 6.

Enfin, l'évolution des covariances (colonne de gauche) présente une croissance de la variance dans les zones non-observées et une diminution de la variance dans les zones observées. La croissance de la variance est de la forme α^{2k}.

Cet exemple académique simple permet de mettre en évidence cette dynamique de la structure spatiale des erreurs, bien diagnostiquée par la longueur de portée. Cependant, pour les modèles de prévision opérationnels, il n'est pas possible, en pratique, d'utiliser les équations du filtre de Kalman et d'obtenir l'évolution des matrices de covariance. Ainsi d'autres stratégies doivent être mises en oeuvre pour décrire cette évolution, comme le filtre de Kalman d'ensemble, ou encore des ensembles d'analyses perturbées. C'est en particulier cette dernière méthode qui sera utilisée ici pour l'estimation des statistiques d'erreur de prévision.

2.5 Des stratégies pour résoudre le BLUE

La matrice de gain $\boldsymbol{K}$ fait intervenir en pratique des matrices de très grande taille. En effet, pour un vecteur d'état $\boldsymbol{x}$ de l'ordre de $\mathcal{O}(10^6)$, la matrice $\boldsymbol{B}$ est de l'ordre de $\mathcal{O}(10^{12})$. L'inversion directe d'une telle matrice est impossible. C'est pourquoi, la formule du BLUE n'est pas utilisée telle quelle dans les applications. Il existe différentes stratégies pour approcher la solution sans inverser directement ces matrices.

Une première stratégie est de résoudre localement l'équation du BLUE, en regroupant des observations par paquets éloignés géographiquement. Ainsi, l'assimilation de ces paquets est alors considérée comme indépendante et les matrices à inverser sont de taille réduite. Cette méthode est décrite par Houtekamer et Mitchell (2001). Elle est naturellement adaptée à une résolution sur un ordinateur en parallèle. Par ailleurs elle est mise en oeuvre par le Service Météorologique du Canada, où un filtre de Kalman d'ensemble est utilisé (Houtekamer et Mitchell, 2005).

Une deuxième stratégie est d'exprimer le problème sous une forme variationnelle. Cela revient à minimiser une certaine fonction coût. Un cas particulier de fonction coût remarquable est le cas d'une fonction coût quadratique. Une telle fonction coût est de la forme $J(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{x}$, avec $\boldsymbol{\Gamma}$ une matrice symétrique définie positive. Ainsi, $\boldsymbol{\Gamma}$ peut être vue comme la matrice d'un produit scalaire. En notant $||\cdot||^2_{\boldsymbol{\Gamma}^{-1}}$ la norme associée, la fonction coût s'écrit $J(\boldsymbol{x}) = ||\boldsymbol{x}||^2_{\boldsymbol{\Gamma}^{-1}}$. Ce type de fonctionnelle correspond géométriquement à une surface en forme de cuvette, dont les courbes de niveaux sont de type ellipsoïdale.

La résolution analytique d'un tel problème est simple : le minimum est atteint au point $\boldsymbol{x}^*$ qui annule le gradient de J, $\boldsymbol{\nabla} J = 0$. La résolution numérique est effectuée de manière séquentielle en construisant un chemin allant vers le fond de la cuvette. Une méthode simple est celle de la plus grande pente. Cependant, la trajectoire est en zigzag et la convergence est peu rapide. Une méthode plus efficace est la méthode du gradient conjugué, qui revient à déformer

la cuvette de sorte que les courbes de niveau deviennent de type sphéroïdale. Ainsi, le chemin construit converge sans zigzag vers le minimum.

2.6 Schémas de résolution variationnelle

Dans cette section, les principaux algorithmes utilisés dans les schémas d'assimilation sont décrits. Il s'agit des algorithmes 3D-inc, 4D-inc et 3D-FGAT.

Il existe d'autre algorithmes, *e.g.* le 3D-PSAS, 4D-PSAS (Courtier, 1997). N'étant pas utilisés dans la suite, il ne sont pas décrits ici.

2.6.1 Schéma 3D-Var/3D-Inc

3D-Var

Le 3D-Var est une méthode pour obtenir l'analyse comme l'argument minimisant une fonction coût. Cette approche est équivalente à la détermination du maximum de vraisemblance pour des distributions gaussiennes de l'erreur (Lorenc, 1988). La fonction coût à minimiser est

$$2J(\boldsymbol{x}) = ||\boldsymbol{x} - \boldsymbol{x}^b||^2_{\boldsymbol{B}^{-1}} + ||\boldsymbol{y}^o - \mathcal{H}(\boldsymbol{x})||^2_{\boldsymbol{R}^{-1}}. \tag{2.18}$$

Dans cette expression, l'opérateur d'observation non linéaire implique que cette fonction n'est pas quadratique et peut donc posséder plusieurs minima. En tout cas, il n'existe pas d'expression analytique de la solution.

3D-Inc

Une version infinitésimale du 3D-Var peut quant à elle fournir une solution analytique au problème. Cette méthode est appelée *incrémentale*. En supposant que l'analyse est une correction proche de l'ébauche sous la forme $\boldsymbol{x}^a = \boldsymbol{x}^b + \delta\boldsymbol{x}^a$ avec $||\delta\boldsymbol{x}^a|| << 1$, il est possible de linéariser la fonction coût du 3D-Var (2.18) sous la forme

$$2J(\boldsymbol{x}) = ||\delta\boldsymbol{x}||^2_{\boldsymbol{B}^{-1}} + ||\boldsymbol{y}^o - \mathcal{H}(\boldsymbol{x}^b) - \boldsymbol{H}\delta\boldsymbol{x}||^2_{\boldsymbol{R}^{-1}}. \tag{2.19}$$

Cette fois, la fonctionnelle est quadratique. Étant de plus convexe, elle admet un unique minimum qui est obtenu au point où le gradient s'annule. Ici, le gradient est $\boldsymbol{\nabla} J(\delta\boldsymbol{x}) = \boldsymbol{B}^{-1}\delta\boldsymbol{x} - \boldsymbol{H}^T\boldsymbol{R}^{-1}(\boldsymbol{y}^o - \mathcal{H}(\boldsymbol{x}^b) - \boldsymbol{H}\delta\boldsymbol{x})$. Ainsi, la nullité du gradient en $\delta\boldsymbol{x}^a$ est équivalente à $(\boldsymbol{B}^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H})\delta\boldsymbol{x}^a = \boldsymbol{H}\boldsymbol{R}^{-1}\left(\boldsymbol{y}^o - \mathcal{H}(\boldsymbol{x}^b)\right)$ soit

$$\boldsymbol{x}^a = \boldsymbol{x}^b + (\boldsymbol{B}^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{R}^{-1}\left(\boldsymbol{y}^o - \mathcal{H}(\boldsymbol{x}^b)\right). \tag{2.20}$$

Ceci n'est autre que l'équation du BLUE. En effet, avec les notations introduites précédemment, $\boldsymbol{x}^a = \boldsymbol{x}^b + \boldsymbol{K}\left(\boldsymbol{y}^o - \mathcal{H}(\boldsymbol{x}^b)\right)$, avec $\boldsymbol{K} = \boldsymbol{B}\boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^T + \boldsymbol{R}\right)^{-1}$ la matrice de gain.

Limitations des méthodes 3D-Var/3D-Inc

Dans les applications opérationnelles, les observations ne sont pas toujours disponibles exactement au temps de l'ébauche. De plus, cette méthode ne prend pas en compte l'évolution temporelle de l'écoulement. Pour ce faire, la dimension temporelle est ajouté à l'algorithme 3D-Var/3D-Inc, ce qui correspond à la méthodes 4D-Var/4D-Inc.

2.6.2 Schéma 4D-Var/4D-Inc

4D-Var

Pour prendre en compte l'évolution temporelle de l'écoulement, l'opérateur d'observation de la méthode 3D-Var est modifié à l'aide de l'opérateur $\mathcal{M}_{t\to t+T}$:

$$2J(\boldsymbol{x}) = ||\boldsymbol{x} - \boldsymbol{x}^b||^2_{\boldsymbol{B}^{-1}} + \sum_k ||{\boldsymbol{y}^o}_k - \mathcal{H}_k \circ \mathcal{M}_{0\to k}(\boldsymbol{x})||^2_{\boldsymbol{R}_k^{-1}}. \tag{2.21}$$

Ainsi, la trajectoire d'intégration s'ajuste pour être compatible avec les observations et l'ébauche. Cet algorithme permet de faire évoluer de manière implicite les covariances d'erreur d'ébauche à la manière du filtre de Kalman (sous l'hypothèse que le modèle est parfait) (Thépaut *et al.* , 1995). L'erreur modèle peut être prise en compte avec une formulation à contrainte faible.

Cette fonctionnelle non quadratique n'admet pas de solution analytique. Les stratégies usuelles pour sa minimisation font intervenir l'utilisation du gradient (d'autre stratégies peuvent être utilisées, comme les méthodes métaheuristiques de recuit-simulé, *etc*mais elles ne sont pas bien adaptées aux problèmes de cette taille pour lesquels le coût de l'évaluation de la fonctionnelle est important). Le gradient de cette fonctionnelle est donné par

$$\boldsymbol{\nabla} J(\boldsymbol{x}) = \boldsymbol{B}^{-1}\boldsymbol{x} - \sum_k {\boldsymbol{M}_{0\to k}}^T {\boldsymbol{H}_k}^T \boldsymbol{R}_k^{-1} \left({\boldsymbol{y}^o}_k - \mathcal{H}_k \circ \mathcal{M}_{0\to k}(\boldsymbol{x})\right),$$

avec ${\boldsymbol{M}_{0\to k}}^T$ l'adjoint du modèle linéarisé le long de la trajectoire $\mathcal{M}_{0\to k}(\boldsymbol{x})$. L'utilisation de l'adjoint est coûteuse (l'intégration linéaire d'un terme quadratique demande le calcul de deux termes à la place d'un seul). De plus, il est nécessaire de stocker la trajectoire (souvent des stratégies associant stockage d'état intermédiaire/recalcul de trajectoire sont utilisés). Ainsi cet algorithme n'est pas utilisé en opérationnel, mais approximé par une suite de problèmes quadratiques de type 4D-Inc.

4D-Inc

Dans ce cas, et par analogie au 3D-Inc, la résolution du problème variationnel initial est recherchée en résolvant une succession de problèmes quadratiques locaux. A chaque étape, l'optimal obtenu est noté ${\boldsymbol{x}^a}_p$ où p est l'indice du numéro de l'itération. Cet état est utilisé pour générer une nouvelle trajectoire de référence, autour de laquelle le problème est linéarisé. Cette démarche n'assure pas de trouver le minimum absolu de la fonctionnelle 4D-Var, mais on espère ainsi en trouver une bonne approximation.

Ainsi les étapes se font de manière récursive selon

$$\begin{cases} 2J_{{\boldsymbol{x}^a}_p}(\delta\boldsymbol{x}) = ||\delta\boldsymbol{x}||^2_{\boldsymbol{B}^{-1}} + \sum_k ||\boldsymbol{d}_k^p - \boldsymbol{H}_k^p \boldsymbol{M}_{0\to k}^p \delta\boldsymbol{x}||^2_{\boldsymbol{R}_k^{-1}}, \\ {\delta\boldsymbol{x}^a}_p = ArgMin\left\{J_{{\boldsymbol{x}^a}_p}(\delta\boldsymbol{x}), \delta\boldsymbol{x} \in \mathbb{R}^n\right\}, \\ {\boldsymbol{x}^a}_{p+1} = {\boldsymbol{x}^a}_p + {\delta\boldsymbol{x}^a}_p, \\ {\boldsymbol{x}^a}_0 = \boldsymbol{x}^b, \end{cases} \tag{2.22}$$

avec $\boldsymbol{d}_k^p = {\boldsymbol{y}^o}_k - \mathcal{H}_k \circ \mathcal{M}_{0\to k}({\boldsymbol{x}^a}_p)$ l'innovation par rapport à la trajectoire issue de ${\boldsymbol{x}^a}_p$, $\boldsymbol{M}_{0\to k}^p$ le modèle linéarisé autour de la trajectoire d'état initial ${\boldsymbol{x}^a}_p$, et $\boldsymbol{H}_k^p$ l'opérateur d'observation linéarisé autour de $\mathcal{M}_{0\to k}({\boldsymbol{x}^a}_p)$. Le gradient de la fonctionnelle s'écrit à chaque étape

$$\boldsymbol{\nabla} J_{{\boldsymbol{x}^a}_p}(\delta\boldsymbol{x}) = \boldsymbol{B}^{-1}\delta\boldsymbol{x} - \sum_k \boldsymbol{H}_k^{pT} {\boldsymbol{M}_{0\to k}^p}^T \boldsymbol{R}_k^{-1} \left(\boldsymbol{d}_k^p - \boldsymbol{H}_k^p \boldsymbol{M}_{0\to k}^p \delta\boldsymbol{x}\right),$$

La fonctionnelle $J_{\boldsymbol{x}^a{}_p}$ étant quadratique , la solution du problème est atteinte quand le gradient s'annule, correspondant à un état unique. Cela s'exprime donc par

$$\delta \boldsymbol{x}^a{}_p = Arg \left\{ \boldsymbol{\nabla} J_{\boldsymbol{x}^a{}_p}(\delta \boldsymbol{x}) = 0, \delta \boldsymbol{x} \in \mathbb{R}^n \right\}.$$

Cette approche incrémentale nécessite la dérivation du modèle non linéaire à partir de son code informatique, puis l'adjointisation de ce code.

En pratique, l'algorithme 4D-Inc est utilisé avec un incrément à basse résolution pour diminuer le coût par rapport à un 4D-Var. Cependant, cet algorithme 4D-Inc reste relativement coûteux dans le cas où il est nécessaire de réaliser plusieurs assimilations différentes en peu de temps, avec des contraintes opérationnelles. Ce type d'utilisation intervient quand une approche ensembliste est mise en oeuvre (Belo Pereira and Berre, 2006). Ainsi, pour ce type de situation, il peut être avantageux d'utiliser un autre type d'algorithme : le 3D-FGAT.

2.6.3 Schéma 3D-FGAT

L'algorithme 3D-FGAT (First Guess at Appropriate Time), correspond à une simplification du 4D-Inc, pour lequel l'opérateur linéaire $\boldsymbol{M}^p_{0\to k}$ est remplacé par l'identité. Cela permet de réduire le coût numérique : il n'y a plus besoin de l'adjoint du modèle linéaire dans l'expression du gradient. De même que pour le 4D-Inc, il s'agit de minimiser de manière récursive un ensemble de fonctionnelles quadratiques. L'algorithme est de la forme

$$\left\{ \begin{array}{l} 2J_{\boldsymbol{x}^a{}_p}(\delta \boldsymbol{x}) = ||\delta \boldsymbol{x}||^2_{\boldsymbol{B}^{-1}} + \sum_k ||\boldsymbol{d}^p_k - \boldsymbol{H}^p_k \delta \boldsymbol{x}||^2_{\boldsymbol{R}^{-1}_k}, \\ \delta \boldsymbol{x}^a{}_p = ArgMin \left\{ J_{\boldsymbol{x}^a{}_p}(\delta \boldsymbol{x}), \delta \boldsymbol{x} \in \mathbb{R}^n \right\}, \\ \boldsymbol{x}^a{}_{p+1} = \boldsymbol{x}^a{}_p + \delta \boldsymbol{x}^a{}_p, \\ \boldsymbol{x}^a{}_0 = \boldsymbol{x}^b, \end{array} \right. \tag{2.23}$$

avec $\boldsymbol{d}^p_k = \boldsymbol{y}^o{}_k - \mathcal{H}_k \circ \mathcal{M}_{0\to k}(\boldsymbol{x}^a{}_p)$ l'innovation par rapport à la trajectoire issue du guess $\boldsymbol{x}^a{}_p$, et $\boldsymbol{H}^p_k$ l'opérateur d'observation linéarisé autour de $\mathcal{M}_{0\to k}(\boldsymbol{x}^a{}_p)$. Ainsi le gradient de chaque fonctionnelle quadratique intermédiaire est alors

$$\boldsymbol{\nabla} J_{\boldsymbol{x}^a{}_p}(\delta \boldsymbol{x}) = \boldsymbol{B}^{-1} \delta \boldsymbol{x} - \sum_k \boldsymbol{H}^{pT}_k \boldsymbol{R}^{-1}_k \left(\boldsymbol{d}^p_k - \boldsymbol{H}^p_k \delta \boldsymbol{x} \right),$$

où n'interviennent plus le modèle linéaire et son adjoint. En pratique, c'est l'ensemble de la trajectoire de l'ébauche qui est ajustée par le même incrément.

2.7 Description de la prévision à Météo-France

2.7.1 Modèle global Arpège

L'atmosphère est un fluide modélisé sous la forme d'un milieu continu. L'air est considéré comme étant un gaz parfait, mélange de deux gaz : l'air sec et la vapeur d'eau. Ce gaz est compressible et il est gouverné par les équations de Navier-Stokes. Dans un modèle de prévision global, l'atmosphère est décomposée en niveaux verticaux. Pour la plupart des modèles globaux, on résout les équations primitives, qui sont obtenues par simplification des équations de Navier-Stokes.

Un niveau donné de la grille du modèle est une sphère. La sphère est un domaine périodique pour lequel les fonctions de carré intégrable admettent deux représentations classiques équivalentes :

- la valeur de la fonction sur une grille de collocation,
- la représentation spectrale de la fonction dans la base des harmoniques sphériques.

Les champs sont alors représentés en harmoniques sphériques. Ce sont des fonctions

$$Y_n^m(\lambda, \phi) = P_n^m(\sin\phi)\, e^{im\lambda},$$

où λ est la longitude et ϕ la latitude ; les P_n^m désignent les polynômes de Legendre de première espèce. On appelle nombre d'onde total l'entier n. En notant N_T l'indice de troncature, un signal f est représenté par $f(\lambda, \phi) = \sum_{n=0}^{N_T} \sum_{m=-n}^{n} f_n^m\, Y_n^m(\lambda, \phi)$ où l'ensemble (f_n^m) correspond aux coefficients spectraux. Le passage d'une représentation à l'autre s'effectue par une transformation espace spectral - point de grille, semblable à la transformation de Fourier dans le cas du cercle.

L'avantage de l'espace spectral est que les harmoniques sphériques sont fonctions propres de l'opérateur laplacien. Ainsi, la résolution des équations primitives est plus simple et plus précise. Le modèle français Arpège [2] (Courtier *et al.* , 1991) est un modèle global étiré, pseudo-spectral, avec un schéma temporel semi-implicite semi-lagrangien, sur une grille linéaire.

Les champs dynamiques du modèle sont : le tourbillon relatif ζ, la divergence du vent η, la température T, la pression de surface P_S et l'humidité spécifique [3] q. L'état de l'atmosphère est ainsi modélisé par la valeur de ces cinq paramètres en chaque point de la grille de gauss et pour chaque niveau du modèle. L'atmosphère est donc représentée par un vecteur de la forme $\boldsymbol{x} = (\zeta, \eta, T, P_S, q)$. Chaque composante de ce vecteur est un degré de liberté dont les valeurs sont contraintes à respecter certains équilibres imposés par les équations diagnostiques (équation de continuité, relation hydrostatique,...) et dont l'évolution temporelle est donnée par les équations pronostiques (équation d'évolution du tourbillon, équation d'évolution de la divergence,...). Depuis l'utilisation de la microphysique de Lopez (2002), quatres nouvelles variables pronostiques supplémentaires ont été ajoutées : eau liquide, eau solide, neige et pluie.

2.7.2 Cycle journalier d'analyses et de prévisions

Le cycle opérationnel journalier, utilisé à Météo-France, comprend quatre cycles d'analyse et de prévision, avec un découpage sur des fenêtres de 6h. Le haut de la figure 2.10 illustre l'enchainement des étapes du cycle d'assimilation.

Il y a ainsi quatre analyses journalières : 00H UTC, 06H UTC, 12H UTC et 18H UTC. À chacune de ces analyses est associée une prévision à 6h pour fournir l'ébauche de l'analyse suivante. Par ailleurs, du fait de contraintes opérationnelles, un cycle de production avec des analyses à cut-off court est mis en oeuvre. Il est présenté en bas de la figure 2.10.

2.7.3 Réseau d'observations

Le réseau d'observations évolue au cours du temps et n'est pas le même suivant les cycles.

D'une part, le nombre d'observations et la qualité des observations ne sont pas constants. Par exemple, les radiosondages, observations de référence, ne sont pas disponibles à 06H et 18H, mais uniquement à 00H et 12H. D'autre part, les satellites défilants échantillonnent différentes régions du globe au cours du temps. De plus, un contrôle de qualité permet d'éliminer les observations considérées comme aberrantes car non compatibles avec l'ébauche.

[2] Action de Recherche Petite Echelle Grande Echelle

[3] $q = \dfrac{\text{masse de vapeur d'eau}}{\text{masse de vapeur d'eau + masse d'air sec}}$.

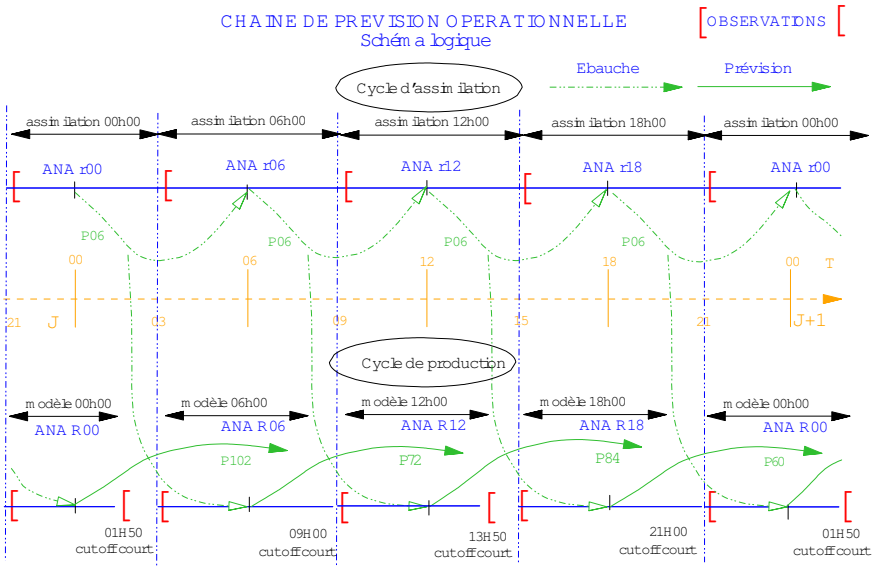


FIG. 2.10 – Représentation schématique des cycles d'analyse/prévision opérationnels en Avril 2007. (Source : Bulletin trimestriel de contrôle des modèles numériques, premier trimestre 2007, Météo-France/Dprévi/COMPAS)

Par ailleurs, il existe une grande hétérogénéité spatiale dans la couverture du réseau d'observations. Les océans sont moins observés que les continents, entraînant un contraste terre/mer. L'hémisphère Sud est moins observé que l'hémisphère Nord, ce qui est également à l'origine d'un contraste Nord/Sud.

Malgré cette hétérogénéité spatio-temporelle, le nombre d'observations disponibles est globalement du même ordre pour les quatre réseaux.

2.7.4 Schéma d'assimilation opérationnel

Le schéma d'assimilation Arpège est de type 4D-Var multi-incrémental (4D-Inc). Deux boucles externes, correspondant à deux minimisations autour de deux états successifs, sont utilisées avec des résolutions différentes : $T107$ pour la première et $T149$ pour la seconde. L'initialisation est assurée directement dans le schéma d'assimilation, par l'ajout d'un terme dans la fonction coût qui pénalise les ondes rapides. Ce schéma d'assimilation est opérationnel depuis juillet 2000.

2.7.5 Spécification des statistiques d'erreur

Les matrices de covariance d'erreur d'observation $\boldsymbol{R}_k$ pour les différents instants k sont supposées diagonales. Il existe des méthodes de détermination objective, pour ajuster au mieux ces variances (Desroziers et Ivanov, 2000 ; Chapnik *et al.* , 2006).

La matrice de covariance d'erreur d'ébauche est relativement constante temporellement. Il s'agit d'une formulation non-séparable (la propagation de la correction d'un niveau à l'autre dépend de l'échelle horizontale à corriger), et multivariée (la correction d'un paramètre physique induit celle des autres, en vertu d'équilibres représentés dans le modèle). Cette matrice est organisée sous la forme de produit d'opérateurs. Sa description est donnée dans le chapitre suivant.

2.8 Conclusions

Au long de ce chapitre, les techniques usuelles de l'assimilation de données ont été introduites. En particulier, l'équation du BLUE a été complétée par l'évolution temporelle de l'erreur d'analyse pour obtenir les équations du filtre de Kalman. Dans ces équations, l'importance de la matrice de covariance d'erreur d'ébauche $\boldsymbol{B}$ s'est avérée capitale, pour le filtrage et la propagation de la correction issue de l'innovation. Ces propriétés de filtrage et de propagation sont fortement influencées par l'étendue spatiale des fonctions de corrélation. Cette étendue est caractérisée par le diagnostic de la longueur de portée. De plus, la correspondance entre la distribution de probabilité, son échantillonnage et la dispersion de l'échantillon définis par $\boldsymbol{B}$ a permis une première incursion dans les approches ensemblistes.

Ces rôles et caractérisations ont été développés au travers d'exemples illustratifs. Ces exemples ont permis d'introduire la notion de structure cohérente de portée. Le diagnostic de la portée est ainsi un outil permettant de caractériser simplement une matrice de covariance d'erreur d'ébauche.

Il a été mentionné que la très grande taille de $\boldsymbol{B}$ ne permet une représentation et une manipulation directe en machine. Ceci rend nécessaire la modélisation de la matrice $\boldsymbol{B}$ (à laquelle cette thèse est consacrée). Il s'agit d'un problème difficile du fait de la taille de cette matrice, mais aussi parce que peu d'informations sont disponibles sur ces covariances.

D'autre part, les stratégies variationnelles permettant de résoudre l'équation du BLUE, avec leur extension temporelle 4D-Var, ont été décrites. Les algorithmes 4D-Inc et 3D-FGAT, utilisés dans cette thèse via des ensembles d'analyses perturbées, sont des algorithmes mis en place dans les grands centres de prévision numérique du temps (Météo-France, CEPMMT, *etc*).

La manière dont est représentée la matrice $\boldsymbol{B}$, ou encore la manière de l'estimer, n'ont pas encore été abordées. Naturellement ces deux problèmes sont des questions majeures pour l'assimilation de données. Dans le chapitre suivant, l'estimation de la matrice à partir d'un ensemble d'analyses perturbées est décrite, ainsi que quelques modélisations usuelles pour cette matrice, en particulier celle basée sur l'hypothèse diagonale spectrale. Cependant, cette modélisation souffre d'une limitation intrinsèque : elle s'avère incapable de représenter les variations géographiques des fonctions de corrélation. Une manière de parvenir à représenter de telles variations est de considérer une représentation ondelette à la place de la représentation spectrale.

Chapitre 3

Estimation et modélisation de la matrice B

RÉSUMÉ

Le chapitre précédent a permis d'introduire les notions de base pour l'assimilation de données. Formellement, la méthode est relativement simple. En pratique, tout est plus complexe : la taille des matrices, leur estimation, leur modélisation, les contraintes opérationnelles de production. Ce chapitre se consacre à l'estimation et à la modélisation de $\boldsymbol{B}$.

Dans un premier temps, la méthode basée sur un ensemble d'assimilations perturbées est décrite. Elle permet de simuler l'évolution spatio-temporelle des erreurs du système d'assimilation. Puis les caractéristiques de la matrice $\boldsymbol{B}$ sont présentées (hétérogénéité, non séparabilité horizontale-verticale, aspects multivariés, et dépendance à l'écoulement). La modélisation spectrale des auto-covariances et ses limitations sont par ailleurs exposées : la non représentation des variations géographiques des corrélations justifie le recours à une formulation ondelette.

3.1 Estimation de la matrice B

Il existe différentes techniques pour estimer la matrice $\boldsymbol{B}$. Les plus anciennes correspondent aux méthodes de Hollingsworth et Lönnberg (1986) ou encore à la méthode du NMC (Parrish et Derber, 1992 ; Bouttier, 1994b). Les plus récentes sont issues des méthodes basées sur un ensemble d'assimilations perturbées.

3.1.1 Principe de la méthode basée sur un ensemble d'assimilations perturbées

Le procédé d'un ensemble d'assimilations est illustré par la figure 3.1. À un instant donné, on dispose d'un ensemble d'analyses perturbées, dont la dispersion reflète l'erreur d'analyse. Chacune de ces analyses est intégrée sur 6h à l'aide du modèle non linéaire, pour produire un ensemble de prévisions perturbées, dont la dispersion reflète l'erreur d'ébauche. Lors de l'analyse suivante, pour chaque membre de l'ensemble, une nouvelle analyse est obtenue à partir de l'ébauche perturbée et des observations perturbées.

À un instant k et en considérant le membre l, les perturbations des observations dont il est question ici correspondent à l'ajout aux observations ${\boldsymbol{y}^o}_k$ de perturbations ${\boldsymbol{\varepsilon}^o}_{k,l} = \boldsymbol{R}_k^{1/2}\boldsymbol{\zeta}_{k,l}$ ($\boldsymbol{\zeta}_{k,l}$ est une réalisation d'un vecteur aléatoire de moyenne nulle et de matrice de covariance la matrice identitée) compatibles avec la matrice de covariance d'erreur d'observations $\boldsymbol{R}_k$ (voir la section 2.4.1).

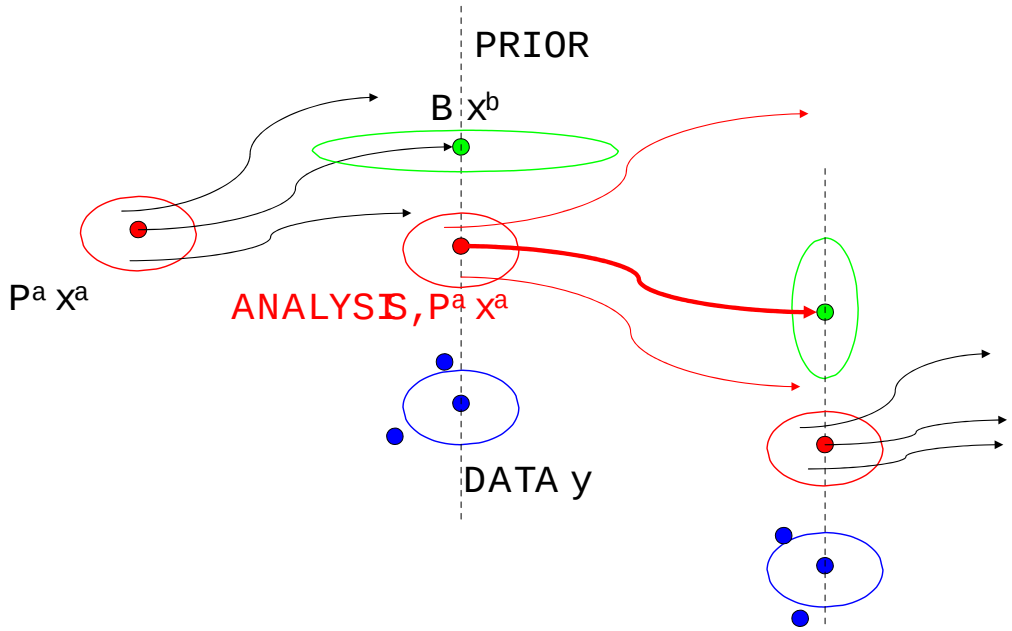


FIG. 3.1 – Schéma de l'évolution d'un ensemble d'assimilations perturbées (d'après Ehrendorfer, présentation au workshop de 2006 sur les méthodes adjointes).

3.1.2 Formalisme des ensembles d'assimilations perturbées

Formellement, pour un temps k et en notant l le numéro d'un membre de l'ensemble, l'estimation de la matrice $\boldsymbol{B}_k$ est donnée par $\overline{\boldsymbol{B}_k}$ telle que

$$\left\{ \begin{array}{l} \overline{\boldsymbol{x}^b{}_{k,l}} = \frac{1}{N_e} \sum_l \boldsymbol{x}^b{}_{k,l}, \\ \overline{\boldsymbol{B}_k} = \frac{1}{N_e - 1} \sum_l \left(\boldsymbol{x}^b{}_{k,l} - \overline{\boldsymbol{x}^b{}_{k,l}} \right) \left(\boldsymbol{x}^b{}_{k,l} - \overline{\boldsymbol{x}^b{}_{k,l}} \right)^T . \end{array} \right. \tag{3.1}$$

De même, l'estimation de la matrice $\boldsymbol{A}_k$ est donnée par $\overline{\boldsymbol{A}_k}$ telle que

$$\left\{ \begin{array}{l} \boldsymbol{y}^o{}_{k,l} = \boldsymbol{y}^o{}_k + \boldsymbol{R}_k^{1/2} \boldsymbol{\zeta}_{k,l}, \\ \boldsymbol{x}^a{}_{k,l} = \boldsymbol{x}^b{}_{k,l} + \tilde{\boldsymbol{K}}_k \left(\boldsymbol{y}^o{}_{k,l} - \mathcal{H}_k(\boldsymbol{x}^b{}_{k,l}) \right) , \\ \overline{\boldsymbol{x}^a{}_{k,l}} = \frac{1}{N_e} \sum_l \boldsymbol{x}^a{}_{k,l}, \\ \overline{\boldsymbol{A}_k} = \frac{1}{N_e - 1} \sum_l \left(\boldsymbol{x}^a{}_{k,l} - \overline{\boldsymbol{x}^a{}_{k,l}} \right) \left(\boldsymbol{x}^a{}_{k,l} - \overline{\boldsymbol{x}^a{}_{k,l}} \right)^T . \end{array} \right. \tag{3.2}$$

Le nuage analysé évolue avec $\boldsymbol{x}^b{}_{k+1,l} = \mathcal{M}_{k \to k+1}(\boldsymbol{x}^a{}_{k,l})$, où $\mathcal{M}_{k \to k+1}$ correspond à une intégration du modèle non linéaire sur $6h$. On peut montrer que les équations de l'analyse et du modèle sont utilisées pour faire évoluer les perturbations $\boldsymbol{\varepsilon}_{k,l} = \boldsymbol{x}_{k,l} - \overline{\boldsymbol{x}_{k,l}}$ du système :

$$\left\{ \begin{array}{l} \boldsymbol{\varepsilon}^a{}_{k,l} = \boldsymbol{\varepsilon}^b{}_{k,l} + \tilde{\boldsymbol{K}}_k (\boldsymbol{\varepsilon}^o{}_{k,l} - \boldsymbol{H} \boldsymbol{\varepsilon}^b{}_{k,l}) \\ \boldsymbol{\varepsilon}^b{}_{k+1,l} \approx \boldsymbol{M}_{k \to k+1} \boldsymbol{\varepsilon}^a{}_{k,l} \end{array} \right. , \tag{3.3}$$

où $\boldsymbol{M}_{k \to k+1}$ est le modèle linéaire tangent de $\mathcal{M}_{k \to k+1}$ (en réalité, c'est plutôt le modèle non linéaire qui est utilisé (à bon escient) pour faire évoluer les perturbations d'analyse ; cela permet de représenter des effets non linéaires, tels que la saturation de la croissance des erreurs dans les petites échelles).

Dans l'expression de l'équation d'analyse, $\tilde{\boldsymbol{K}}_k$ représente la matrice de gain du système déterministe opérationnel. On se propose d'expliciter son contenu et les deux raisons qui justifient le recours à cette matrice dans la section suivante.

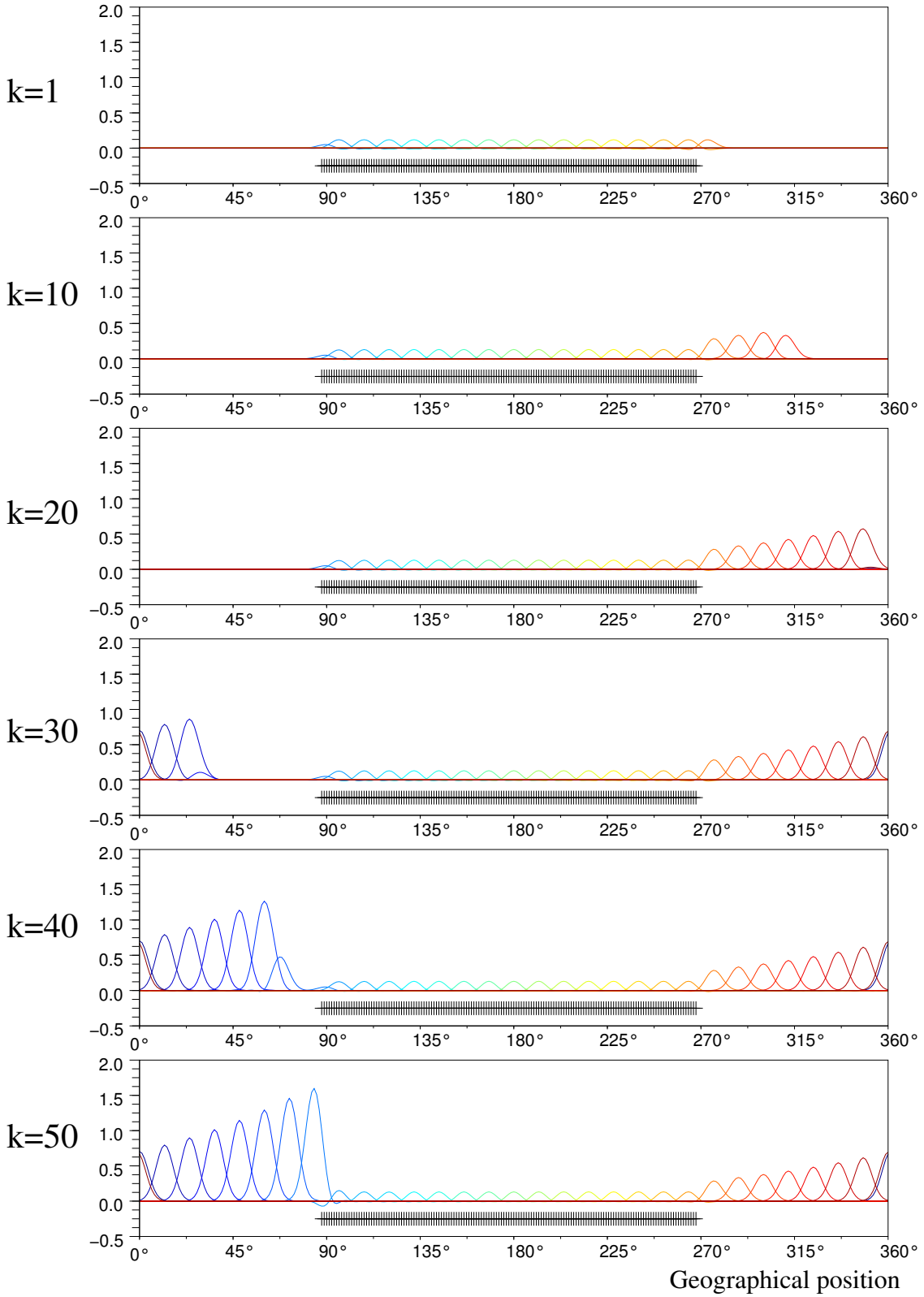


FIG. 3.2 – Exemple de l'évolution des covariances d'erreur de prévision estimées dans un ensemble d'analyses perturbées. Les itérations représentées correspondent à $k = 1$, $k = 10$, $k = 20$, $k = 30$, $k = 40$ et $k = 50$.

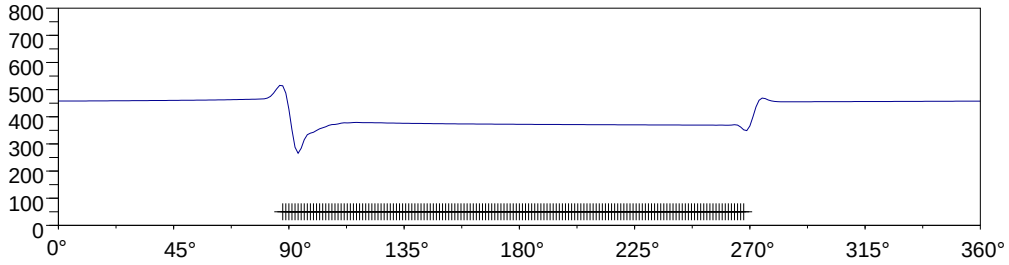


FIG. 3.3 – Carte de portée obtenue à l'itération $k = 50$, dans l'expérience décrite sur la figure 3.2

La figure 3.2 représente l'évolution des covariances ensemblistes formalisées dans les équations (3.1) et (3.2). Le modèle utilisé ici correspond au modèle d'advection/amplification présenté dans la section 2.4.4. Le schéma d'analyse est réalisé avec une matrice de corrélation $\boldsymbol{B}$ homogène de portée $L_h = 500\,km$. On considère le cas où seules les observations sont perturbées (*i.e.* pas l'ébauche) dans la toute première analyse (k=1), pour visualiser la façon dont les perturbations se propagent et s'amplifient dans l'espace et le temps. Le cas $k = 1$ correspond à la matrice $\boldsymbol{B}_e = \boldsymbol{M}\tilde{\boldsymbol{K}}\boldsymbol{R}\tilde{\boldsymbol{K}}^T\boldsymbol{M}^T$. Ainsi les covariances sont nulles partout sauf dans la région observée (et sa translatée). Puis les perturbations sont transportées le long du domaine : pour $k = 20$ les covariances au voisinage de 360 °commencent à être non nulles. À l'itération $k = 50$, les covariances transportées atteignent la zone observée à 90 °. La portée résultante dans ce cas est représentée sur la figure 3.7. Il apparaît que la portée est plus large dans la zone non observée que dans la zone observée : la portée est de l'ordre de $450\,km$ de 315 °à 65 °(en passant par 0 °) et de $400\,km$ de 90 °à 270 °, avec une oscillation de raccordement au bord de ce domaine.

Dans cet exemple, il apparaît que la matrice ensembliste ainsi construite est hétérogène. La carte de variance est d'amplitude plus faible dans la zone observée qu'ailleurs. De même, la portée est plus courte dans la zone observée qu'ailleurs.

3.1.3 Liens entre la matrice de gain utilisée et le filtrage spatio-temporel des covariances de l'ensemble

Comme explicité dans la section 3.1.2, l'ensemble d'assimilations perturbées inclut l'utilisation de la matrice de gain $\tilde{\boldsymbol{K}}_k$ du système déterministe opérationnel. Une première justification de l'utilisation de cette matrice est que cela permet en effet de simuler l'évolution des erreurs du système déterministe, et en particulier l'effet de sa matrice de gain (Berre et al 2006).

Par ailleurs, cette matrice de gain du système déterministe est en partie dépendante de l'écoulement, du fait par exemple de la non linéarité des opérateurs de couplage masse/vent (Fisher 2003) et des opérateurs d'observation. Certaines composantes de la matrice B du système déterministe sont cependant issues d'une moyenne spatiale globale et d'une moyenne temporelle longue (sur plusieurs semaines), appliquées aux covariances issues d'un ensemble d'assimilations perturbées (Fisher 2003, Belo Pereira et Berre 2006).

Cette moyenne spatiale et temporelle peut être vue comme un filtrage spatio-temporel des covariances issues de l'ensemble. Cela permet en effet de réduire les effets du bruit d'échantillonnage, inhérent à l'estimation des covariances à partir d'un ensemble de taille finie.

En résumé, une deuxième justification de l'utilisation de la matrice de gain du système déterministe est donc la possibilité d'appliquer un filtrage spatio-temporel, pour réduire le bruit

d'échantillonnage qui affecte les covariances issues de l'ensemble.

Un des buts des recherches en cours et de la présente thèse est d'optimiser ce filtrage spatio-temporel, en adaptant son degré de localisation spatiale et temporelle. Il s'agit en effet de trouver le meilleur compromis possible entre le souhait de filtrer le bruit d'échantillonnage, et le souhait de représenter des variations spatio-temporelles importantes et accessibles (en fonction de la taille de l'ensemble disponible).

3.1.4 Articulation entre l'optimisation du filtrage et l'utilisation des nouvelles covariances filtrées

L'objet de cette section est d'expliciter la stratégie d'articulation, dans cette thèse puis en opérationnel, entre l'optimisation du filtrage et l'utilisation effective des nouvelles covariances filtrées.

Au cours de cette thèse, une première étape consiste à diagnostiquer les variations spatio-temporelles des covariances d'erreur qui sont représentatives du système déterministe opérationnel.

Une deuxième étape consiste à mettre au point un filtrage spatio-temporel optimisé des corrélations à l'aide des ondelettes. On peut également mentionner que l'annexe C introduit une technique de filtrage spatial optimisé pour les écarts types.

Une troisième étape consiste à appliquer ce filtrage aux covariances de l'ensemble, et à diagnostiquer les caractéristiques des nouvelles covariances filtrées ainsi obtenues (par exemple en termes de variations spatio-temporelles, de spectres d'énergie, etc).

Une dernière étape consiste à utiliser ces nouvelles covariances filtrées, d'une part au sein de l'assimilation déterministe, et d'autre part au sein de l'ensemble d'assimilations perturbées.

Bien que cette dernière étape soit naturelle et essentielle pour l'opérationnel, la présente thèse s'est concentrée sur les trois premières étapes évoquées ci-dessus. Cela constitue en effet un pré-requis indispensable à la dernière étape, ainsi que la part de travail la plus conséquente en termes de recherche et de développement.

Cela n'est naturellement pas contradictoire avec une stratégie d'évolution continue et cohérente entre les covariances estimées et filtrées à partir de l'ensemble, et celles qui sont effectivement utilisées dans les systèmes déterministe et ensembliste.

En effet, à terme, la stratégie visée en opérationnel consiste bien à :

- produire un ensemble de prévisions perturbées valides pour une date t donnée, et issues d'un ensemble d'analyses valides à $t - 6h$;
- calculer les covariances d'ébauche associées, en leur appliquant un filtrage spatio-temporel adéquat (et optimisé en fonction de la taille de l'ensemble disponible) ;
- utiliser ces covariances dans l'analyse relative à la date t, à la fois au niveau du système déterministe et au sein de l'ensemble d'assimilations perturbées.

On peut mentionner que cette stratégie est en cours d'application pour les écarts types (Berre 2008, communication personnelle). Les premiers résultats indiquent que l'utilisation effective des écarts types du jour (avec un filtrage spatial optimisé), au sein même de l'ensemble, conduit à des résultats réalistes (par exemple, on ne constate pas du tout un effondrement de l'ensemble, et on obtient des structures spatiales relativement proches de ceux décrits en annexe C, tout en retrouvant de légères différences attendues en amplitude et en position).

Cette stratégie sera donc également appliquée aux corrélations, filtrées spatialement à l'aide des ondelettes.

3.1.5 Mise en oeuvre avec un ensemble 3D-FGAT

La construction et l'évolution d'un ensemble d'analyses perturbées suivent les cycles d'analyse/prévision décrits dans la section 2.7.2 et représentés sur la figure 2.10.

Pour chaque réseau, les observations disponibles sont perturbées et injectées dans le schéma d'analyse de chaque membre de l'ensemble. Cette opération est réalisée pour les quatre cycles journaliers. Différents schémas sont disponibles : le schéma 4D-Var multi-incrémental et le schéma 3D-FGAT. Ce dernier schéma est bien moins coûteux que le premier. Il permet par exemple de faire évoluer un ensemble d'une dizaine de membres en temps réel.

Ainsi, en opérationnel il parait possible d'estimer, en temps réel, la matrice $\boldsymbol{B}$ associée à un réseau, pour une date donnée.

3.2 Caractéristiques de la matrice B dans l'atmosphère

3.2.1 Expression formelle de la matrice B

La matrice $\boldsymbol{B}$ est une matrice pleine représentée par bloc suivant les paramètres physiques que l'on considère. En notant $\boldsymbol{B}[\alpha;\beta]$ la matrice de covariance entre les champs α et β, et $\boldsymbol{B}[\alpha]$ la matrice de covariance pour le seul champ α, il vient

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}[\zeta] & \boldsymbol{B}[\zeta;\eta] & \boldsymbol{B}\left[\zeta;(T,P_S)\right] & \boldsymbol{B}[\zeta;q] \\ \boldsymbol{B}[\zeta;\eta]^T & \boldsymbol{B}[\eta] & \boldsymbol{B}\left[\eta;(T,P_S)\right] & \boldsymbol{B}[\eta;q] \\ \boldsymbol{B}\left[\zeta;(T,P_S)\right]^T & \boldsymbol{B}\left[\eta;(T,P_S)\right]^T & \boldsymbol{B}\left[(T,P_S)\right] & \boldsymbol{B}\left[(T,P_S);q\right] \\ \boldsymbol{B}[\zeta;q]^T & \boldsymbol{B}[\eta;q]^T & \boldsymbol{B}\left[(T,P_S);q\right]^T & \boldsymbol{B}[q] \end{pmatrix}. \tag{3.4}$$

Chacun des blocs ci-dessus comprend à la fois les composantes horizontales et verticales des fonctions de structures. Les matrices $\boldsymbol{B}$ des différents grands centres opérationnels présentent un certain nombre de caractéristiques similaires. D'une part, les fonctions de structure sont *hétérogènes* pour un niveau donné. D'autre part, les corrections horizontales et verticales sont *non-séparables*. Enfin, les différents paramètres physiques sont couplés et donc corrélés : les fonctions de structures sont *multivariées*.

3.2.2 Diagnostic de l'hétérogénéité

Dans la réalité, les statistiques des erreurs d'ébauche sur la sphère ne sont pas homogènes. En effet, la dynamique a tendance à déformer les fonctions de structure (Bouttier, 1993 et 1994 ; Veersé et Thépaut, 1998). Différentes études le confirment, en particulier Ingleby (2001) pour le MetOffice, ou encore Belo Pereira et Berre (2006) qui, dans le cas d'Arpège, présentent un diagnostic de l'anisotropie des fonctions de structure "climatologiques".

Pour Arpège, cette hétérogénéité est illustrée par quelques résultats rapportés sur la figure 3.4. Sur cette figure, la carte de variance (a) et celle de longueur de portée (b) sont représentées pour le champ de pression de surface. Ces cartes correspondent à des caractéristiques moyennes issues des statistiques sur une période de plusieurs semaines. Ainsi, ces statistiques présentent des caractéristiques "climatologiques" et ne sont pas représentatives de la variabilité journalière.

Sur la carte de variance (a), trois bandes principales apparaîssent en fonction de la latitude (ainsi que des zone de faible variance sur les régions bien observées, comme l'Europe et les États-Unis). Cependant, la variance est relativement homogène en moyenne. Inversement, pour

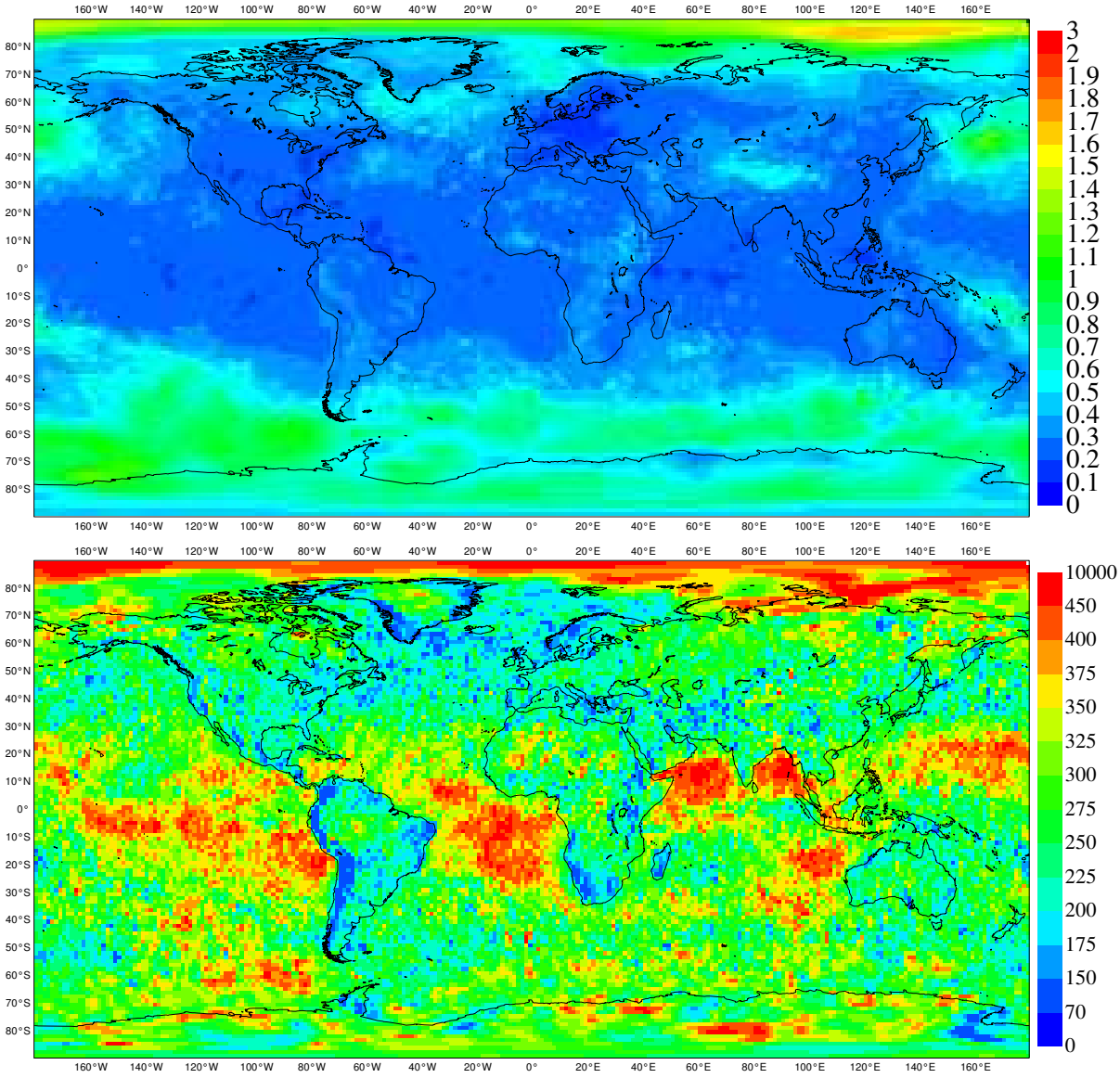


FIG. 3.4 – Représentation des cartes de variance (en hPa) (a) et de longueur de portée (en km) (b) pour la pression de surface.

une date particulière, nous verrons que les phénomènes météorologiques significatifs (tempêtes, cyclones,...) modifient de manière importante la répartition spatiale (dépendance à l'écoulement).

D'autre part, la carte de longueur de portée (b) présente globalement une certaine symétrie par rapport à l'équateur, avec de fortes longueurs de portée sur les pôles et à l'équateur. De vastes bandes de longueurs de portée moyennes couvrent les régions tropicales et tempérées. En regardant plus finement, il existe une disymétrie entre les deux hémisphères. Il apparaît que les longueurs de portée sont plus grandes dans l'hémisphère Sud que dans l'hémisphère Nord. En particulier, il y a une vaste zone de courtes longueurs de portée au niveau de l'Europe. Cette disparité s'explique par le fait que l'hémisphère Sud est plus pauvre en observations que l'hémisphère Nord. En particulier, la densité du réseau d'observation est grande sur l'Europe. Il est également à noter que, les longueurs de portée sont plus faibles sur terre que sur mer, et que le rail des dépressions contribue à diminuer la portée sur l'Atlantique Nord. Le relief joue aussi un rôle puisque se détachent distinctement la chaîne himalayenne et la cordillière des Andes.

3.2.3 Non-séparabilité verticale

Les statistiques des erreurs d'ébauche sont par ailleurs tridimensionnelles. Ainsi, un point important de ces statistiques est que la correction verticale apportée par une correction à un niveau donné dépend, entre autres, de l'échelle horizontale de la correction. C'est une propriété de *non-séparabilité* entre les directions horizontale et verticale.

L'étude des statistiques d'erreurs permet d'illustrer cette propriété. La figure 3.5 représente la corrélation moyenne entre l'erreur d'ébauche pour la température sur le niveau modèle 49 (proche de 850hPa) et les autres niveaux modèle, en fonction du nombre d'onde total, pour le modèle du CEPMMT (qui comporte 60 niveaux). Il a été vu, en section 2.3.1, que la correction apportée par l'innovation est propagée par les fonctions de covariance. Ainsi, ce graphique illustre comment la correction verticale varie en fonction de l'échelle horizontale observée. Il apparaît que pour les grandes échelles (petits nombres d'onde n), la correction est profonde : pour $n \leq 5$, la correction se propage entre les niveaux 45 et 60. A l'inverse, pour les petites échelles, la correction est peu profonde : pour $n = 100$, seuls les niveaux proches du niveau 49 sont corrigés. Le cas séparable correspondrait à une correction sur la verticale indépendante de la correction sur l'horizontale, et donc indépendante du nombre d'onde n.

Il est tentant de faire le lien entre la non-séparabilité et les modélisations limites classiques des écoulements fluides dans les conditions "eau peu profonde" (Shallow Water) et "eau profonde" (Deep Water).

Dans le premier cas, toute la colonne fluide est influencée. Cette approche correspond bien à l'approximation très grande échelle de l'atmosphère, où l'hypothèse de couche mince est bien vérifiée.

Dans le second cas, seule une fine épaisseur du fluide est influencée. L'épaisseur est associée à une échelle horizontale caractéristique, pilotant une décroissance exponentielle de l'influence d'une profondeur sur les autres.

Naturellement, la non-séparabilité ainsi présentée sous sa forme dépendante à l'échelle modale est trop simple. En effet, ce diagnostic est global et ne rend pas pleinement compte des structures locales de l'écoulement. Or un écoulement atmosphérique se caractérise par la présence de structures bien identifiées, telles que les dépressions, les fronts, *etc*. Ainsi, suivant la localisation géographique et la situation météorologique, la non-séparabilité peut se caractériser par des corrections obliques (Rabier *et al.* , 2000).

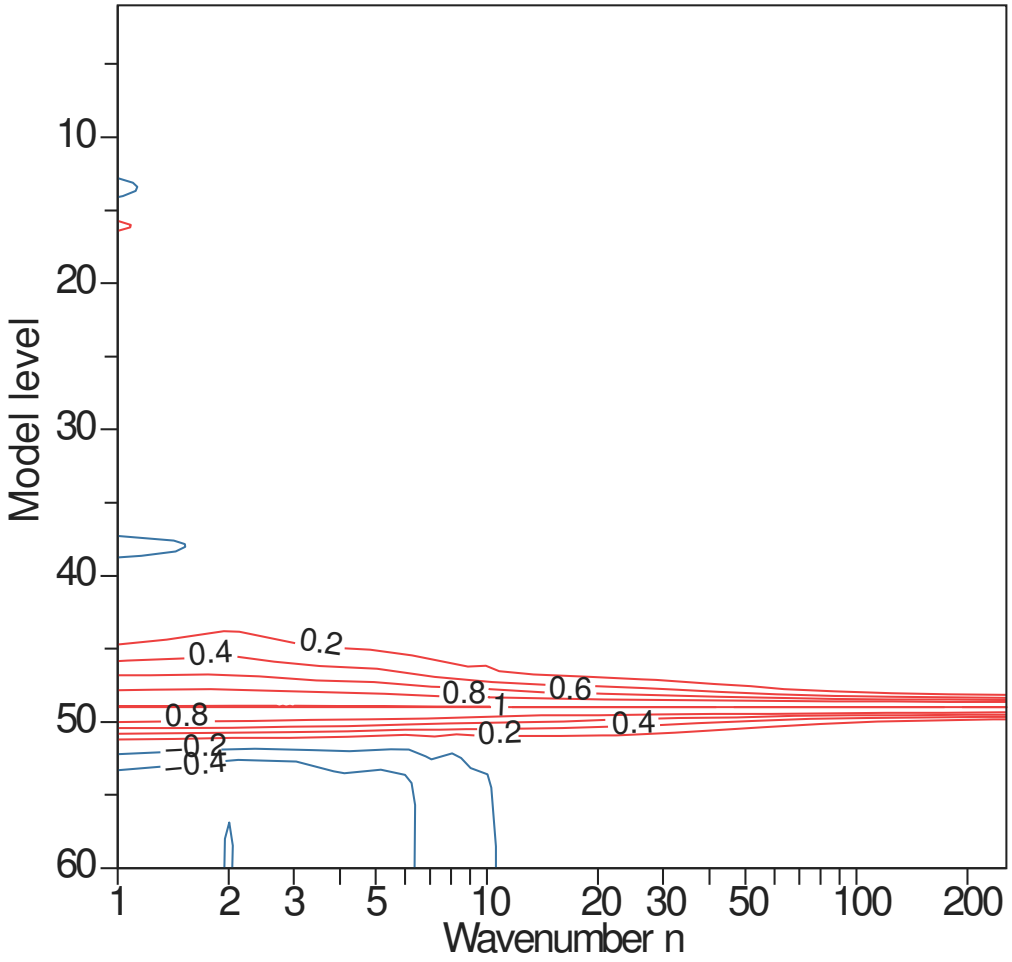


FIG. 3.5 – Corrélation moyenne entre l'erreur d'ébauche pour la température sur le niveau modèle 49 (proche de 850hPa) et les autres niveaux modèles, en fonction du nombre d'onde total en harmonique sphérique, pour le modèle du CEPMMT. (D'après Fisher, ECMWF Newsletter No. 106 - Winter 2005/06, p23-28)

3.2.4 Aspects multivariés : influence des relations de balance

Des interactions existent également entre les différents champs physiques. Cela explique le caractère *multivarié* de la matrice $\boldsymbol{B}$.

Dans l'atmosphère, des *équilibres* ou *balances* de grande échelle relient les champs entre eux. L'*équilibre géostrophique* ou *balance géostrophique* est l'équilibre qui aux moyennes latitudes relie le gradient du champ de pression avec le vent horizontal, appelé *vent géostrophique*, suivant

$$\mathbf{v}_g = \frac{1}{\rho f}\,\mathbf{k} \times \boldsymbol{\nabla} P_S, \tag{3.5}$$

avec $f = 2\Omega sin(\phi)$ le paramètre de Coriolis, ρ la masse volumique, Ω la vitesse angulaire de la terre, ϕ la latitude et $\mathbf{k}$ le vecteur normal à la surface du globe au point considéré.

De cette manière, le modèle prenant en compte ce type de dépendance, il vient qu'une erreur sur le champ de pression P_s se répercute sur le vent géostrophique, donc sur le tourbillon relatif ζ et sur la divergence η. Naturellement, le vent n'est pas exactement géostrophique, mais l'approximation du vent horizontal par le vent géostrophique reste à nos latitudes une bonne approximation [1]. Le vent réel $\mathbf{v}$ peut être considéré comme la superposition d'une *composante balancée* (le vent géostrophique $\mathbf{v}_g$), et d'une *composante non balancée* (le vent agéostrophique $\mathbf{v}_a$), tel que $\mathbf{v} = \mathbf{v}_g + \mathbf{v}_a$. De même, le champ de pression se décompose en une partie balancée $P_{S,b}$ et une partie non balancée $P_{S,ub}$. Par ailleurs, alors que les parties balancées du vent et de la pression sont reliées par l'équilibre géostrophique, les parties non balancées tendent à être décorrélées.

La formulation multivariée de Derber et Bouttier (1999) prend en compte ces équilibres (ainsi que les composantes balancées et non balancées associées), à l'aide de regressions linéaires spectrales entre les paramètres. Une fois que ces relations de balance sont prises en compte, il reste à modéliser les auto-covariances, c'est-à-dire les covariances propres à chaque paramètre. Dans la suite de ce manuscrit, on se concentrera sur la modélisation de ces auto-covariances.

3.2.5 Dépendance à l'écoulement

On sait par ailleurs que la matrice $\boldsymbol{B}$ devrait évoluer au cours du temps. Cette caractéristique a déjà été mise en évidence dans des contextes simples comme l'exemple de la dynamique sur le cercle de la section 2.4.4. Elle se retrouve dans des systèmes plus complexes tels que l'atmosphère.

Cette dépendance a en particulier été étudiée au cours de cette thèse par l'étude et le filtrage des variations journalières de l'écart type (annexe C) et de la longueur de portée (chapitre 6).

3.3 Modélisation de la matrice B (auto-covariances)

Dans cette section, la modélisation sur la sphère utilisée dans le modèle Arpège est présentée, ainsi que son extension à la formulation ondelette.

[1] Il est possible de montrer avec une analyse en ordre de grandeur que l'approximation est en $\mathcal{O}(Ro)$, avec Ro le nombre de Rossby. En considérant que la vitesse du vent est de l'ordre de $U = 10m.s^{-1}$, que les phénomènes météorologiques considérés sont d'échelle caractéristique $L = 1000\,km = 10^6 m$ et que la latitude considérée est d'environ $\phi = 45°$ donnant $f = 10^{-4}s^{-1}$, le nombre de Rossby est le quotient adimensionné défini par $Ro = \frac{\text{accélération inertielle}}{\text{accélération de Coriolis}} = \frac{U}{fL} = 0.1$. L'approximation du vent par le vent géostrophique est donc valable à 10% près.

3.3.1 Hypothèse diagonale spectrale

De manière générale, une matrice de covariance réelle est *symétrique positive*. Dans le cas où les composantes sont indépendantes, elle est de plus *définie* (*i.e.* ses valeurs propres sont strictement positives). En particulier, étant symétrique réelle, il existe une base orthonormale dans laquelle la représentation des covariances est donnée par une matrice diagonale. Il s'agit d'une *base de Karhunen-Loève* (Mallat, 2001).

Si la base de Karhunen-Loève est connue, alors la connaissance des covariances se réduit à celle des valeurs diagonales, correspondant au spectre du tenseur de covariance. Au lieu d'avoir à connaître n^2 composantes de la matrice, seulement n suffisent.

Il existe différentes stratégies permettant d'approximer la base de Karhunen-Loève, en déployant des dictionnaires de bases et en recherchant celle qui permet de se rapprocher le plus d'une diagonale. Ce type d'approche est basé *e.g.* sur l'utilisation des ondelettes orthogonales et en particulier leur filtre passe-bas/passe-haut associé, qui permet de construire de nouvelles bases orthogonales : les paquets d'ondelettes.

Dans le cas de la sphère, ces outils ne sont pas disponibles, mais cette notion d'approximation de la base de Karhunen-Loève est tout de même utilisée. Elle consiste alors à se donner une base et à supposer que la matrice de covariance $\boldsymbol{B}$ y est diagonale. Il s'agit alors de *l'hypothèse diagonale*. Courtier *et al.* (1998) ont fondé leur modèle pour $\boldsymbol{B}$ sur l'hypothèse que les corrélations sont diagonales dans la base des harmoniques sphériques, espace spectral associé à la sphère. Il s'agit alors de *l'hypothèse diagonale spectrale*.

3.3.2 Illustration de l'hypothèse diagonale sur le cercle

Un exemple 1D est utilisé pour visualiser l'impact de l'hypothèse diagonale en fonction du choix de la base. La figure 3.6 représente en (a1) les covariances d'une matrice $\boldsymbol{B}$ modélisant de manière schématique les variations physiques associées à un front météorologique positionné à la pseudo-latitude 180 °, caractérisé par une faible variance dans la zone de grande longueur de portée (zone Z1) et une variance forte dans la zone de petite longueur de portée (zone Z2). En (a2) l'incrément d'analyse est représenté pour deux observations localisées aux points A et B. Le point A est caractéristique de la zone Z1, tandis que B est caractéristique de Z2. Dans cette expérience, l'écart type de l'erreur d'observation est constant, pris égale à $\sigma_o = 0.5$. En A, dans la zone stable de l'écoulement, la variance de l'erreur d'ébauche étant faible par rapport à celle de l'erreur d'observation, l'incrément d'analyse a une amplitude plus faible. D'autre part, il est très étalé, propageant la correction à grande échelle. A l'inverse, en B, dans la zone instable de l'écoulement, la variance de l'erreur d'ébauche est grande par rapport à celle de l'observation : l'incrément d'analyse tend à déformer fortement l'ébauche pour coller davantage à l'observation. Par contre, la propagation de la correction reste très localisée au voisinage de l'observation. Ces résultats correspondent à la référence que l'on cherche à obtenir par l'application de l'hypothèse diagonale dans une base choisie.

Hypothèse diagonale en points de grille

En se plaçant dans l'espace point de grille, les vecteurs de la base constituée de diracs sont notés $\boldsymbol{\Delta}_i$ avec i l'indice de point de grille. L'hypothèse diagonale s'exprime

$$\boldsymbol{B} = \sum_i \sigma_i^2 \boldsymbol{\Delta}_i \boldsymbol{\Delta}_i^T = Diag(\sigma_i^2).$$

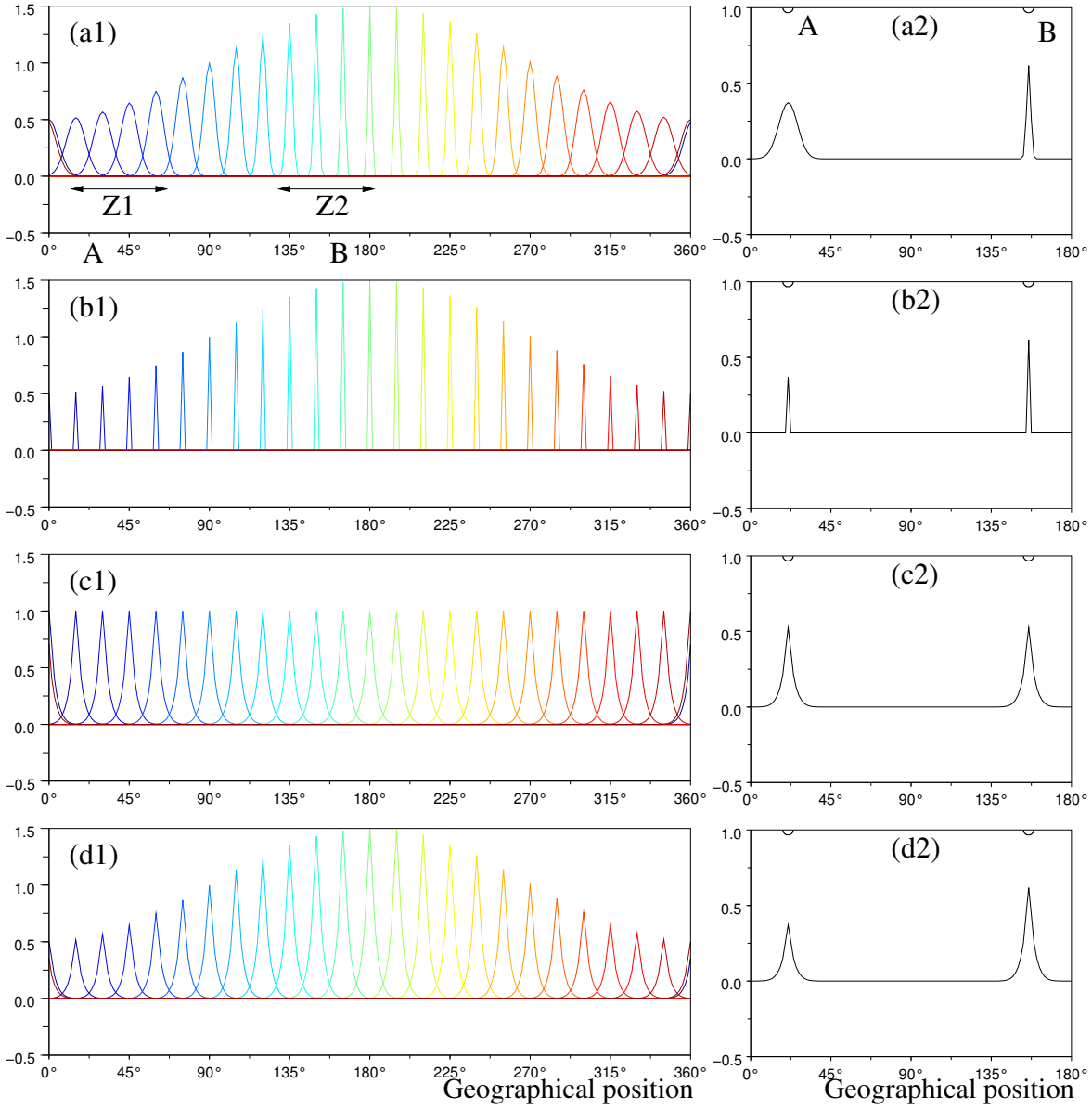


FIG. 3.6 – Expériences d'assimilation comparées à partir d'une matrice de référence exacte, hétérogène. Chaque figure représente une matrice de covariance associée dont on représente quelques fonctions de covariance en (a1,b1,c1,d1) et l'incrément d'analyse obtenu avec cette matrice en (a2,b2,c2,d2). La matrice de covariance de référence $\boldsymbol{B}$ est en (a1,a2), sa modélisation diagonale point de grille en (b1,b2), sa modélisation diagonale spectrale (homogène au sens des corrélations) en (c1,c2) et enfin sa modélisation diagonale hybride (points de grille / spectral) en (d1,d2).

Dans ce cas, σ_i^2 représente la variance de l'erreur de prévision associée au point de grille considéré. En notant $\mathbf{\Sigma} = Diag(\sigma_i)$, il vient $\boldsymbol{B} = \mathbf{\Sigma I \Sigma}^T$. La matrice de corrélation associée à $\boldsymbol{B}$ est dans ce cas l'identité. Ainsi, les fonctions de corrélation sont des diracs, ce qui signifie qu'une erreur en un point n'implique aucune contribution sur d'autres points. En particulier, les longueurs de portée sont nulles Ce choix de modélisation est représenté sur les panels (b1) et (b2) de la figure 3.6. Dans ce modèle, la carte de variance est correctement représentée, mais pas la longueur de portée ni *a fortiori* sa variation géographique. En particulier, d'après (b2), il n'y a pas de propagation de la correction, alors qu'elle existe en (a2). Par contre, la modulation de l'amplitude de la correction au point d'observation est exacte, égale à celle de (a2).

L'hypothèse diagonale dans l'espace point de grille permet donc une bonne représentation de la variance locale, mais elle n'est pas adaptée pour la propagation de la correction.

Hypothèse diagonale dans l'espace spectral

En se plaçant maintenant dans l'espace spectral sur la sphère, les vecteurs de la base spectrale sont notés $\mathbf{Y}_{n,m}$. Pour cet espace, l'hypothèse diagonale s'exprime

$$\boldsymbol{B} = \sum_{n,m} \lambda_n^2 \mathbf{Y}_{n,m} \mathbf{Y}_{n,m}^T.$$

Dans ce cas, il est possible de montrer que la covariance est homogène dans l'espace point de grille. Ce résultat se retrouve dans le cas du cercle, ce qui permet de l'illustrer simplement, en considérant l'expérience 1D de la figure 3.6-(c1). Dans ce modèle, la longueur de portée n'est pas nulle. Par contre elle ne varie pas. De plus, la carte de variance homogène ne représente pas les variations réelles, mais caractérise la variance moyenne. En particulier, d'après (b2) il y a propagation de la correction, mais cette propagation est trop courte en A et pas assez localisée en B. De plus, comme la variance est uniforme, l'amplitude de la correction n'est pas modulée comme en (a2) ou (b2).

L'hypothèse diagonale dans l'espace spectral permet donc de représenter une propagation moyenne de la correction, mais elle n'est pas adaptée pour la représentation de la variance locale. La propagation moyenne de l'information n'est par ailleurs pas satisfaisante localement.

Formulation hybride

Les deux modélisations précédentes permettent de modéliser certains aspects du signal réel. On envisage maintenant une formulation dans laquelle on tente de profiter au mieux de leurs points forts respectifs. Ainsi, une formulation hybride pour $\boldsymbol{B}$ est construite sous la forme $\boldsymbol{B} = \mathbf{\Sigma} \boldsymbol{C} \mathbf{\Sigma}^T$ où $\mathbf{\Sigma}$ est la carte exacte d'écart type (hypothèse diagonale point de grille) et $\boldsymbol{C}$ est la matrice de corrélation modélisée par l'hypothèse diagonale spectrale. Ainsi, $\boldsymbol{C} = \mathbf{S}^{-1}\mathbf{D}\mathbf{S}^{-T}$ où $\boldsymbol{D}$ est une matrice diagonale et $\mathbf{S}$ la matrice de passage de l'espace point de grille à l'espace spectral. Ce choix de modélisation est représenté sur la figure 3.6-(d1). Dans ce modèle, la carte de variance est la carte exacte. De plus, la longueur de portée n'est pas nulle, mais par contre elle reste homogène. D'après (d2), les déficiences de la formulation homogène relevées dans la partie précédente sont résolues pour la modulation de l'amplitude de la correction, mais pas pour la modulation de la propagation de la correction. Cette formulation $\boldsymbol{B} = \mathbf{\Sigma} \mathbf{S}^{-1} \mathbf{D} \mathbf{S}^{-T} \mathbf{\Sigma}^T$ sous la forme d'un produit d'opérateurs en cascade est classique. Comme on n'est pas capable de spécifier directement des structures complexes, on passe par des étapes intermédiaires plus simples aboutissant globalement à une structure complexe.

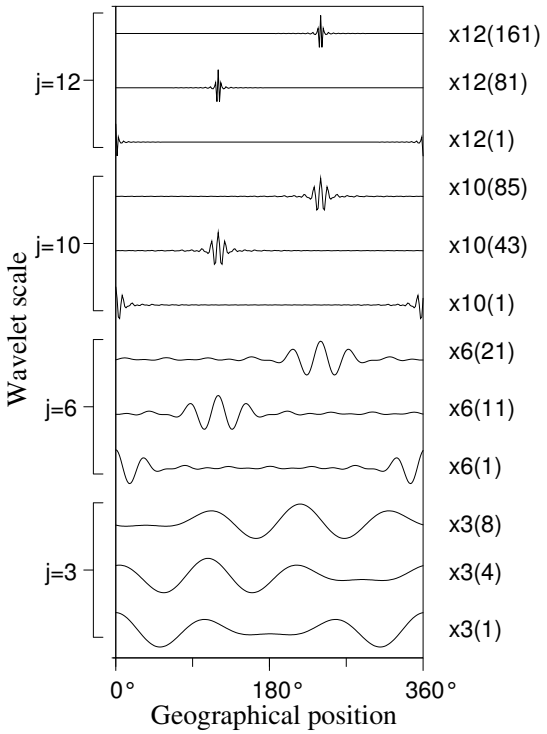


FIG. 3.7 – Quelques fonctions ondelettes $\psi_j(x - x_j(i))$ pour différentes échelles j et pour différentes positions $x_j(i)$.

La formulation hybride apparaît comme une alternative simple pour modéliser de manière précise la carte de variance, tout en ayant une longueur de portée non nulle. Cependant, on n'a pas résolu le problème de la sur/sous-estimation de la propagation. Une manière de répondre à cette question est d'envisager l'hypothèse diagonale dans une base plus favorable. Cela peut être par exemple une base ondelette qui allie à la fois une localisation spatiale et une localisation spectrale. Une telle formulation a été introduite par Fisher (2003). Elle est résumée dans la section suivante.

3.3.3 Formulation ondelette

Une fonction ondelette est une fonction oscillante qui est à la fois localisée en espace et en spectre. Une telle fonction est notée $\psi_j(x)$ où j correspond à l'indice d'échelle et x désigne la position géographique. Dans cette présentation la fonction ψ_j est d'autant plus localisée que l'indice j est grand. La fonction ondelette ψ_j localisée au point $x_j(i)$ est $\psi_j(x - x_j(i))$, elle est notée $\psi_{j,i}$.

La figure 3.7 représente de telles fonctions sur un cercle 1D. Il apparaît que pour un indice petit tel que $j = 3$ le support de la fonction ondelette correspond à l'ensemble du domaine. À ces échelles, les fonctions ondelettes sont alors proches des fonctions sinus/cosinus. À l'inverse, pour des indices grands tels que $j = 10$, le support de la fonction est plus restreint. Plus j devient grand et plus la fonction ondelette est localisée, se rapprochant ainsi d'un dirac.

Les spectres des fonctions ondelettes utilisées ici sont représentés sur la figure 3.8. Le spectre de la fonction $j = 10$ est localisé : il est nul partout sauf pour les nombres d'onde compris entre 30 et 63. Ainsi, la convolution avec une fonction ondelette correspond à un filtrage passe-bande. Des détails sur les ondelettes sont présentés en annexe D.

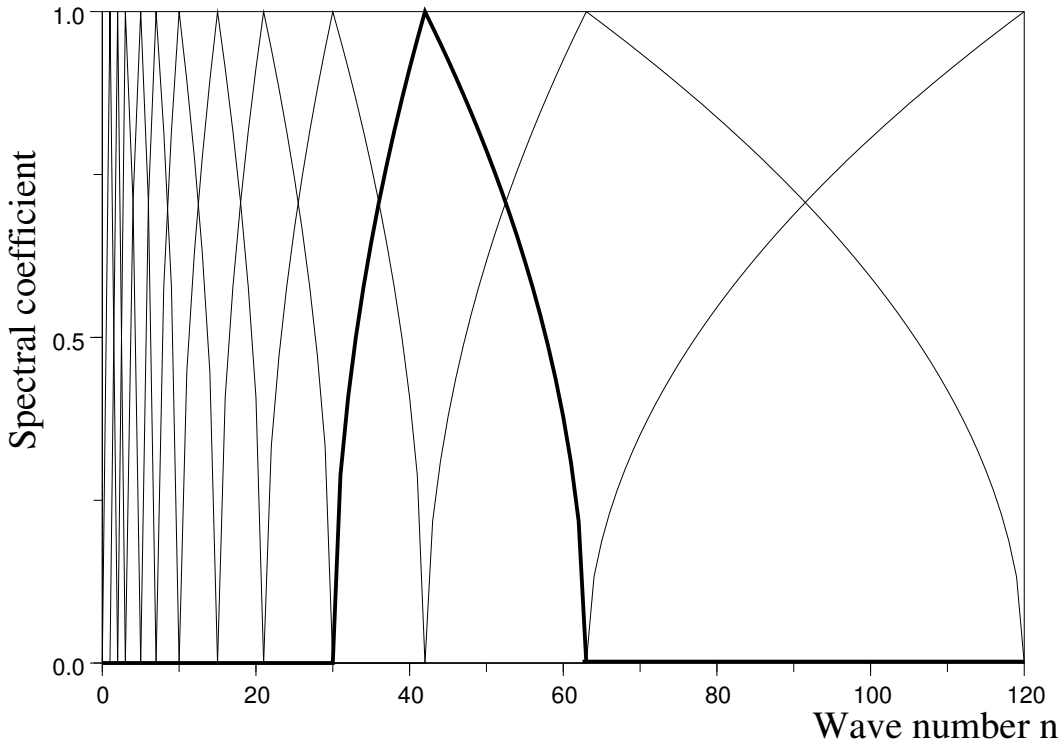


FIG. 3.8 – Coefficients spectraux des fonctions ondelettes $\psi_j(x)$ pour différentes échelles j (il y a une courbe par fonction) et pour la troncature spectrale $T = 120$. Le spectre associé à une fonction $\psi_j(x)$ particulière ($j = 10$) est représenté en gras.

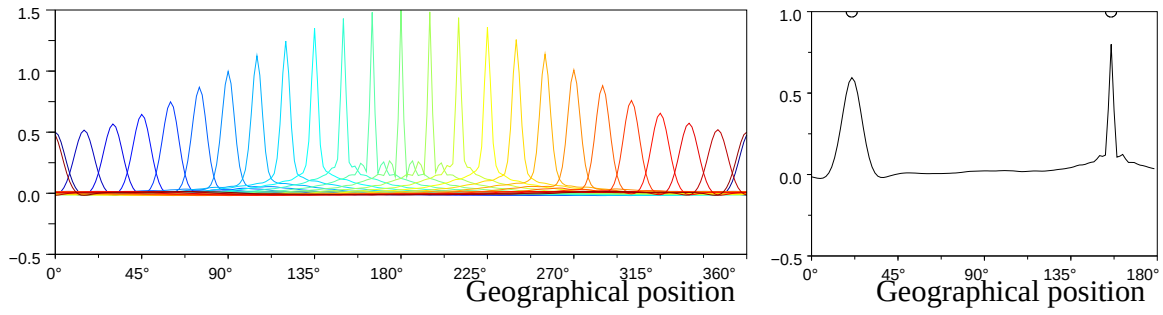


FIG. 3.9 – Similaire à la figure 3.6 pour la modélisation de la matrice de covariance à l'aide de l'hypothèse diagonale ondelette associée à une normalisation spectrale pour les corrélations, et avec une normalisation par l'écart-type de la matrice de référence.

La décomposition en ondelette correspond à la transformation d'un signal $\boldsymbol{\varepsilon}^b$, défini en points de grille, en un ensemble de coefficients ondelettes ${\boldsymbol{\varepsilon}^b}_{j,i}$. Ces coefficients correspondent aux produits scalaires entre $\boldsymbol{\varepsilon}^b$ et les fonctions $\psi_{j,i}$. Ainsi, un coefficient ondelette ${\boldsymbol{\varepsilon}^b}_{j,i}$ s'interprète comme la moyenne locale de $\boldsymbol{\varepsilon}^b$ avec la pondération $\psi_{j,i}$. Contrairement aux coefficients de Fourier résultant d'une moyenne globale du signal, la localisation spatiale des ondelettes permet d'identifier une information locale spécifique.

L'utilisation des ondelettes pour la modélisation des fonctions de corrélation à l'aide de l'hypothèse diagonale permet de représenter les variations géographiques des corrélations. En effet, on s'attend par exemple à ce que les variances des coefficients ondelettes de petite échelle soient plus grandes dans les régions où les portées de corrélation sont courtes.

La figure 3.9 représente quelques fonctions de covariance de la modélisation par l'hypothèse diagonale ondelette, associée à une normalisation spectrale de la corrélation de la matrice représentée sur la figure 3.6 (a1-a2). Cette approche ondelette hybride permet de représenter les variations des fonctions de corrélation, ce que ne parvenait pas à faire la formulation hybride de la figure 3.6 (d1-d2). En particulier, les fonctions modélisées à l'aide des ondelettes sont plus larges (resp. plus étroites) au voisinage de $0\,^\circ$ (resp. $180\,^\circ$) que les fonctions modélisées hybrides.

D'autre part, le fait que les ondelettes soient basées sur des moyennes spatiales locales suggère des propriétés de filtrage spatial du bruit d'échantillonnage associé à l'utilisation d'un ensemble de prévisions.

3.4 Conclusions

Dans ce chapitre, la modélisation opérationnelle de la matrice de covariance d'erreur de prévision a été exposée. En particulier, l'hypothèse diagonale spectrale a été présentée. Cette hypothèse constitue le coeur de la modélisation des corrélations.

L'estimation de la matrice de covariance d'erreur du modèle Arpège a été décrite. En particulier, la méthodologie des ensembles d'analyses perturbées permet de fournir un ensemble de prévisions perturbées, utilisées pour l'estimation de $\boldsymbol{B}$.

Les limites de l'hypothèse diagonale spectrale ont été illustrées dans un exemple analytique : cette hypothèse ne permet pas la représentation des variations spectrale géographiques des corrélations. L'hypothèse diagonale dans l'espace des ondelettes permet au contraire ces variations géographiques. C'est cette propriété remarquable qui sera en particulier exploitée dans cette thèse. Par ailleurs, l'utilisation d'un ensemble de taille finie induit un bruit d'échantillonnage. La possibilité de filtrer ce bruit avec les ondelettes constitue le deuxième axe majeur de cette thèse.

Chapitre 4

Propriétés de filtrage des ondelettes pour les corrélations locales d'erreur de prévision

Traduction d'un article publié : **Pannekoucke O., Berre L. and Desroziers G., 2007.** ***Filtering properties of wavelets for local background-error correlations*****, QJRMS, 133, 363–379.**

RÉSUMÉ

Les covariances d'erreur de prévision peuvent être estimées à partir d'un ensemble de différences de prévisions. La taille finie de l'ensemble utilisé induit un bruit d'échantillonnage qui affecte les estimations statistiques. Il est montré formellement que l'approche diagonale ondelette revient à effectuer une moyenne locale des corrélations. Sa capacité à filtrer le bruit d'échantillonnage est étudiée expérimentalement.

Ces propriétés sont tout d'abord étudiées dans un cadre analytique 1D simplifié. Ce contexte permet d'illustrer la capacité de l'approche diagonale ondelette à représenter des variations géographiques de portée. De plus, le bruit d'échantillonnage s'avère mieux filtré que si l'on n'utilise qu'un filtrage de Schur, en particulier pour les ensembles de petite taille.

Les propriétés de filtrage sont illustrées pour un ensemble de prévisions Arpège (le modèle de prévision global à Météo-France). Deux cas sont étudiés : la représentation des corrélations moyennées temporellement et le cas des corrélations relatives à une journée particulière. Les ondelettes s'avèrent capables d'extraire des variations géographiques de portée pertinentes, et reliées à la situation météorologique du jour considéré.

MOT CLÉS: Assimilation d'ensemble, covariance d'erreur d'ébauche, filtre de Kalman d'ensemble, bruit d'échantillonnage, ondelettes sphériques, ondelettes sur le cercle.

4.1 Introduction

La plupart des schémas d'assimilation cherchent à fournir une combinaison optimale des observations et d'une ébauche (donnée par une prévision à courte échéance). L'analyse optimale dérive de la théorie de l'estimation statistique. Dans cette théorie, les deux types ensembles d'information sont caractérisés par leur matrice de covariance correspondant à leurs erreurs respectives. Les matrices de covariance d'erreur déterminent le poids respectif de chaque source d'information dans l'analyse. Cependant, la spécification correcte de ces statistiques reste un défi essentiel des systèmes d'assimilation de données.

L'estimation des covariances d'erreur d'ébauche est un problème particulièrement difficile,

puisque dans les applications opérationnelles l'ébauche est un vecteur de dimension de l'ordre de 10^5 à 10^7. Dans ce cas, du fait de sa taille énorme, il est impossible de manipuler une telle matrice de covariance d'erreur. D'autre part, il est impossible de la spécifier de manière exacte, parce que l'information statistique disponible est insuffisante (Dee, 1995).

Pour pallier ces difficultés, un modèle statistique pour les covariances d'erreur d'ébauche doit être défini. Un tel modèle est souvent basé sur l'hypothèse que les corrélations d'erreur d'ébauche sont homogènes et isotropes (Gaspari et Cohn, 1999). Cette hypothèse revient à considérer que la matrice de corrélation d'erreur d'ébauche est diagonale dans l'espace spectral (Courtier *et al.* , 1998), ce qui facilite la représentation des ces statistiques.

Une technique pour spécifier la matrice de covariance d'erreur d'ébauche est d'utiliser un ensemble d'assimilations, obtenues en perturbant les observations et l'ébauche (Houtekamer *et al.* , 1996). Cette procédure a récemment été utilisée au CEPMMT et à Météo-France pour déterminer la composante moyenne des covariances d'erreur de prévision (Fisher, 2003 ; Belo Pereira et Berre, 2006). Dans ce cas, les covariances sont moyennées sur plusieurs semaines, avec en plus une hypothèse d'homogénéité. Une telle approche est en partie reliée à la méthode du filtre de Kalman d'ensemble (Evensen, 1994), pour lequel des covariances dépendantes de l'écoulement sont déduites directement d'un tel ensemble.

Cependant, les hypothèses d'homogénéité et d'isotropie sont trop simplificatrices pour représenter les "vraies" structures d'erreurs. En effet, Lönnberg (1988) a suggéré que les corrélations horizontales et verticales varient géographiquement : les échelles horizontales tendent à être plus larges aux tropiques que dans les moyennes latitudes, et ce du fait de la dynamique de l'atmosphère (Ingleby, 2001). Bouttier (1993, 1994) a également montré que les échelles de corrélation dépendent de la situation météorologique et de la densité du réseau d'observation.

En utilisant un ensemble d'assimilations, Belo Pereira et Berre (2006) ont montré de telles hétérogénéités et anisotropies dans les corrélations d'erreur d'ébauche. Pour cela, ils ont utilisé une nouvelle méthode peu coûteuse pour estimer les échelles de corrélation.

Le filtre de Kalman d'ensemble est une approche pour obtenir des corrélations d'erreur d'ébauche hétérogènes et dépendantes de la situation météorologique. Cependant, les ensembles utilisés pour l'estimation étant finis et de petite taille, les covariances sont alors bruitées et doivent être filtrées par un traitement supplémentaire. Le produit de Schur (Houtekamer et Mitchell, 2001) est une méthode de filtrage généralement utilisée pour filtrer les statistiques brutes. L'utilisation du produit de Schur a été discutée par Lorenc (2003).

En utilisant un autre point de vue, Fisher (2003) a récemment introduit l'idée d'utiliser des ondelettes sur la sphère pour améliorer la représentation des corrélations d'erreur d'ébauche. En particulier, une telle approche autorise la représentation d'hétérogénéités dans la description de ces statistiques. Une telle formulation est désormais opérationnelle au CEPMMT pour représenter des corrélations stationnaires (temporellement) mais hétérogènes, avec par exemple des corrélations plus larges dans les tropiques. Une approche similaire a été testée par Deckmyn et Berre (2005) dans le cas d'un modèle à aire limitée.

La représentation ondelette des covariances d'erreur de prévision implémentée au CEPMMT a été obtenue en utilisant un ensemble d'analyses sur plusieurs semaines. De plus, il est envisageable de combiner l'utilisation des ondelettes et d'ensembles pour obtenir des corrélations d'erreur d'ébauche hétérogènes et dépendantes de la situation.

L'objectif de ce papier est de montrer que les ondelettes fournissent un outil permettant de représenter des variations spatio-temporelles des corrélations, mais aussi de filtrer le bruit d'échantillonnage induit par l'utilisation d'ensembles de taille finie. Ces propriétés de filtrage sont étudiées en utilisant le diagnostic des longueurs de portée locales proposé par Belo Pereira et Berre (2006). D'autre part, on peut mentionner que des effets de filtrage analogues ont été

étudiés au Service Météorologique du Canada (Buehner et Charron, 2006), au travers d'un procédé de localisation spectrale et spatiale.

La structure de ce chapitre est la suivante. Dans la section 4.2, nous expliquons comment l'approche diagonale ondelette permet une moyenne spatiale locale des fonctions de covariances. Ceci permet une augmentation implicite de la taille de l'ensemble, tout en préservant la représentation des variations géographiques. La capacité des ondelettes à extraire l'information "utile" d'un ensemble est tout d'abord discutée dans la section 4.3 pour un problème académique d'assimilation sur le cercle. La section 4.4 montre l'application des mêmes ondelettes dans le cas de la sphère 2D, avec le système Arpège. Les résultats d'estimation des statistiques, respectivement sur une période longue et pour une date donnée, sont montrés. Les conclusions et perspectives sont présentées dans la section 4.5.

4.2 Moyennage spatial en ondelette des fonctions de covariance

4.2.1 Fonctions de covariance

Pour faciliter les explications, nous considérons dans cette section un domaine circulaire. Les développements sont également valides dans le cas d'un domaine bi-périodique en considérant les positions horizontales sous la forme de vecteurs à deux composantes (x et y) périodiques. Le formalisme peut aussi se généraliser au cas de la sphère ou dans un contexte tri-dimensionnel (les ajustements devant être faits en considérant la métrique particulière au domaine considéré, et intervenant dans les formules de calcul d'intégrale).

La valeur de séparation s entre deux points x et x" est définie comme la différence $s = x" - x$. À noter que s peut être positive ou négative : la valeur absolue $|s|$ est la distance de séparation, tandis que le signe de s correspond à l'orientation de la séparation[1].

La fonction locale de covariance d'erreur $f^x(s)$, relative à un point de référence x, est souvent calculée à partir d'un ensemble de N_e différences de prévision ε, d'après l'équation suivante (dans le cas horizontal) :

$$f^x(s) = \overline{\varepsilon(x)\varepsilon(x+s)} = \frac{1}{N_e}\sum_l \varepsilon(x,l)\varepsilon(x+s,l),$$

où $\varepsilon(x,l)$ est la réalisation d'une erreur de prévision à la position x, s désigne la séparation entre les deux points considérés, l est l'indice du numéro de membre de l'ensemble, et la barre correspond à la moyenne d'ensemble. Cette moyenne d'ensemble est ainsi calculée sur les N_e membres de l'ensemble.

La taille de l'échantillon est donc égale à N_e. Un premier problème est que la taille finie de l'ensemble induit un bruit d'échantillonnage, ce qui oblige à utiliser par exemple un filtre de Schur (Houtekamer et Mitchell, 2001). Un deuxième problème important est que la matrice de covariance ainsi estimée est de rang réduit (en pratique pour 100 membres le rang est 100 ; pour un vecteur d'état de dimension 10^6, il faudrait au moins 10^6 membres pour obtenir une matrice de plein rang). Ainsi, l'incrément d'analyse ne pourra scruter qu'un sous-espace vectoriel de petite dimension. Dans la suite de l'étude, on se concentre sur le premier problème (à savoir le bruit d'échantillonnage).

[1] Dans le cas 2D bi-périodique, la direction de séparation est donnée par l'argument de s, quand s est vue comme un nombre complexe : $s = s_x + is_y = |s|\,exp(i\,arg(s))$, avec s_x et s_y les composantes x et y de s.

4.2.2 Approche diagonale spectrale : une moyenne spatiale sur tout le domaine

Il est possible de représenter la matrice de covariance d'erreur d'ébauche $\boldsymbol{B}$ dans l'espace spectral (*e.g.* Courtier *et al.* , 1998). Sous l'hypothèse d'homogénéité horizontale, les fonctions de covariance vues dans l'espace spectral sont représentées par une matrice diagonale (ou diagonale par bloc dans le cas tri-dimensionnel, avec les covariances verticales dans les blocs diagonaux). La diagonale de $\boldsymbol{B}$ contient les variances des coefficients spectraux de l'erreur .

De manière équivalente, il est possible de montrer (voir annexe A de ce chapitre) que cette approche diagonale spectrale revient à calculer $\frac{1}{N_g}\sum_{x'} f^{x'}(s)$, qui correspond à la moyenne spatiale, effectuée sur l'ensemble du domaine, des fonctions de covariance[2] (N_g étant ici le nombre de points de grille du domaine). Les fonctions de covariance résultant de cette moyenne spatiale sont notées $f^x_{\boldsymbol{S}}(s)$, où l'indice $\boldsymbol{S}$ fait référence à l'approche diagonale spectrale. Elles sont toutes égales à cette moyenne globale

$$f^x_{\boldsymbol{S}}(s) = \frac{1}{N_g}\sum_{x'} f^{x'}(s),$$

ou de manière équivalente

$$f^x_{\boldsymbol{S}}(s) = \frac{1}{N_g\, N_e}\sum_{x'}\sum_{l} \varepsilon(x', l)\varepsilon(x' + s, l).$$

La fonction de covariance moyennée sur le domaine est ainsi estimée comme une moyenne sur un nombre de paires de réalisations d'erreurs, qui est égal à $\mathcal{N} = N_g N_e$ (au lieu de $\mathcal{N} = N_e$ dans le cas de l'estimation de la fonction de covariance locale).

Une telle augmentation de la taille de l'échantillon a une contre-partie évidente : la *moyenne spatiale sur l'ensemble du domaine* ne permet de représenter aucune variation géographique. En conséquence, on peut se demander s'il est possible de se restreindre à une *moyenne spatiale locale*, de sorte que la taille effective de l'échantillon augmente, tout en représentant des variations géographiques.

4.2.3 Approche diagonale ondelette : une série de moyennes spatiales locales

La transformation ondelette directe et inverse, telle que celle utilisée par Fisher (2003), est définie comme suit : $\hat{\boldsymbol{\varepsilon}}_j = \boldsymbol{\varepsilon} \otimes \boldsymbol{\psi}_j$ et $\boldsymbol{\varepsilon} = \sum_j \hat{\boldsymbol{\varepsilon}}_j \otimes \boldsymbol{\psi}_j$, où, dans les deux expressions, $\otimes$ désigne le produit de convolution dans l'espace physique, et $\boldsymbol{\psi}_j$ sont des fonctions passe-bandes définies pour plusieurs échelles différentes j (voir Courtier *et al.* , 1998 pour la caractérisation du produit de convolution par une fonction radiale sur la sphère, ainsi que l'annexe D sur les ondelettes). On peut également mentionner que les fonctions $\boldsymbol{\psi}_j$ ne sont pas orthogonales, contrairement à d'autres ondelettes utilisées en traitement du signal.

Les fonctions $\hat{\boldsymbol{\varepsilon}}_j$ et $\boldsymbol{\psi}_j$ peuvent être représentées en points de grille. Quelques exemples de fonctions ondelettes $\boldsymbol{\psi}_j$ sont représentées en 1D (Fig. 4.1) et peuvent être comparées avec les fonctions spectrales (sinus ou cosinus qui ne sont pas des fonctions localisées sur le domaine). Ainsi, il vient que chaque fonction ondelette a une position spécifique et une échelle spécifique.

[2] On reconnaît le théorème de Wiener-Khintchine utilisé dans le traitement de la turbulence et qui permet de faire le lien entre la fonction de corrélation et le spectre de puissance du signal.

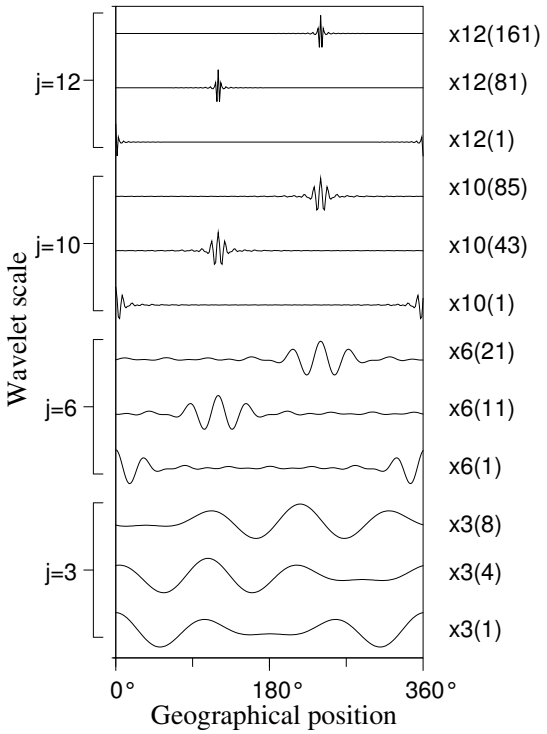


FIG. 4.1 – Quelques fonctions ondelettes $\psi_j(x - x_j(i))$ pour différentes échelles j et pour différentes positions $x_j(i)$.

La position caractérise la localisation spatiale tandis que l'échelle caractérise la fréquence locale. Les coefficients $\hat{\varepsilon}_{j,x_j(i)}$ associés au champ d'erreur transformé correspondent à la valeur de l'erreur à l'échelle j et à la position $x_j(i)$ sur une grille dont la résolution dépend de j (avec $i = 1, N_x(j)$).

Dans le cas horizontal, l'approche ondelette diagonale pour $\boldsymbol{B}$ consiste à calculer les variances de ces coefficients ondelettes $\hat{\varepsilon}_{j,x_j(i)}$. Comme chaque fonction ondelette contient la double information de position et d'échelle, les variances ondelettes contiennent de l'information sur la forme locale de la fonction de covariance.

De plus, il est possible de montrer (voir l'annexe A de ce chapitre) que cette approche diagonale ondelette revient à effectuer une série de moyennes locales pondérées des fonctions de covariance. Les fonctions de covariance résultantes, notées $f^x_{\boldsymbol{W}}(s)$ avec l'indice $\boldsymbol{W}$ faisant ici référence à l'approche diagonale ondelette, peuvent s'exprimer comme suit :

$$f^x_{\boldsymbol{W}}(s) = \sum_{x',s'} f^{x'}(s') \Phi^{x,s}(x', s'),$$

où $\Phi^{(x,s)}(x', s')$ est défini par

$$\sum_j \sum_{i=1}^{N_x(j)} \psi_j(x' - x_j(i))\ \psi_j(x' - x_j(i) + s')\ \psi_j(x - x_j(i))\ \psi_j(x - x_j(i) + s)$$

et peut s'interpréter comme un coefficient de pondération dans le calcul de la moyenne spatiale.

Comme attendu, $\Phi^{x,s}(x', s')$ donne plus de poids aux positions x' dans le voisinage de x, et aux valeurs de séparation s' qui sont proches de s. La figure 4.2 représente deux exemples de fonction $\Phi^{x,s}(x', s')$, pour $x = 121$ et $s = s' = 20$, et pour deux bandes ondelettes différentes.

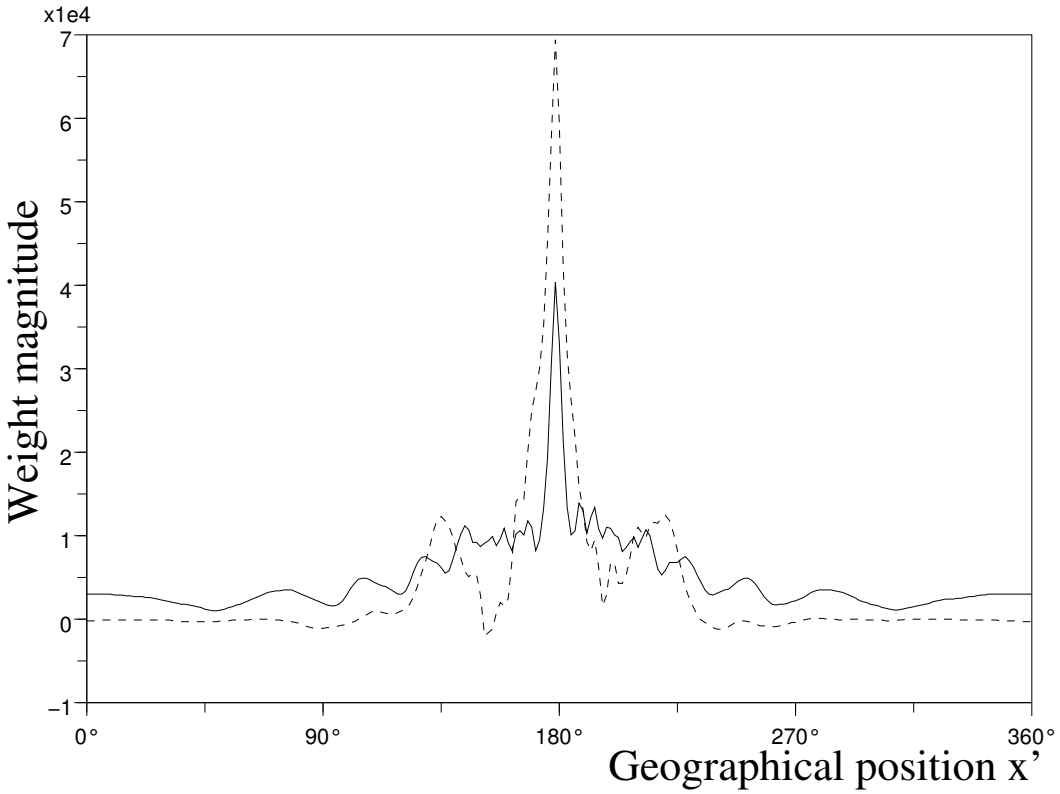


FIG. 4.2 – Représentation de $\Phi^{x,s}(x', s')$ pour $x = 121$ (correspondant à $180\ °$) et $s = s' = 20$ (correspondant à $30\ °$) et pour deux choix différents de bande ondelette, comme défini par les nombres d'onde de coupure $\{N_j\}$ (voir la section 4.2.4) : un ensemble de bandes relativement étroites $\{N_j\} = \{0, 1, 2, 3, 5, 7, 10, 15, 21, 30, 42, 63, 120\}$ (courbe continue) et un ensemble de bandes relativement larges $\{N_j\} = \{0, 4, 8, 12, 20, 28, 40, 60, 84, 120\}$ (en pointillé).

Le poids plus important donné à la position x' proche de x est illustré par les grandes valeurs de $\Phi^{x,s}(x', s')$ quand x' est proche de x.

Tout comme dans l'approche spectrale, les fonctions résultantes $f^x_{\boldsymbol{W}}(s)$ sont le résultat de moyennes spatiales. La taille de l'échantillon spatial est cependant plus petite que celle du cas spectral, parce que la fonction poids $\Phi^{x,s}(x', s')$ a des valeurs proches de zéro sauf dans une zone au voisinage direct de x (comme illustré sur la figure 4.2).

Les deux exemples de la figure 4.2. diffèrent par le choix des bandes ondelettes. Il est ainsi possible de noter que la moyenne spatiale est d'autant plus localisée que les bandes sont larges. Ceci est lié au fait que la taille des bandes détermine le compromis entre les résolutions spatiale et spectrale[3], comme nous le verrons plus loin en section 4.2.4.

Il est également à noter que le calcul, la représentation et le filtrage spatial de la matrice $\boldsymbol{B}$ restent efficaces avec une approche ondelette. En effet, il ne nécessite que le calcul de la diagonale de la matrice, qui contient les variances ondelettes. Cette approche est moins coûteuse que le calcul de la matrice complète en point de grille, puis de son filtrage par des opérateurs de moyenne spatiale.

Pour résumer, l'approche ondelette est similaire à l'approche spectrale, au sens où les fonctions locales de covariance sont moyennées spatialement. Ceci permet potentiellement de réduire le niveau de bruit d'échantillonnage, comparé à l'estimation des fonctions locales de covariance. De plus, comme la moyenne spatiale est locale plutôt que globale, il reste possible de représenter des variations géographiques des fonctions de covariance.

Ces deux caractéristiques (filtrage spatial et variations géographiques) sont étudiées expérimentalement dans deux contextes différents, dans les section 4.3 et section 4.4.

4.2.4 Isotropie des ondelettes et de la moyenne spatiale des covariances

Comme l'illustre la figure 4.1 sur le cercle et Fisher (2004) sur la sphère, les fonctions passe-bandes ψ_j sont radiales. Ceci implique que la moyenne spatiale locale correspondante des fonctions de covariance tend à être isotrope. En d'autre termes, pour une distance de séparation donnée, la moyenne locale des covariances est effectuée sur différentes directions de séparation.

D'un côté, cette isotropie apparaît comme étant un inconvénient, puisqu'elle interdit toute représentation d'anisotropie locale marquée. Cette limitation a été mise en évidence et discutée par Deckmyn et Berre (2005) avec les ondelettes de Meyer.

D'un autre côté, la moyenne spatiale dans plusieurs directions permet d'augmenter la taille de l'échantillon effectif, réduisant ainsi l'amplitude du bruit d'échantillonnage. Ainsi, si les fonctions de covariance sont approximativement isotropes, cette moyenne isotrope apparaît comme étant plutôt bénéfique.

L'équilibre entre ce pour et ce contre dépend finalement du degré d'anisotropie des fonctions de covariance réelles et de la taille de l'ensemble disponible.

Un autre point associé à cette question est qu'une partie de l'anisotropie provient des variations géographiques des écarts types d'erreur d'ébauche. Ces derniers peuvent être représentés en points de grille, tandis que les corrélations sont modélisées via l'espace des ondelettes, d'après la formulation de Fisher (2003).

4.2.5 Détails de l'approche diagonale ondelette

Le principe de moyenne locale des fonctions de covariance a été introduit dans la section précédente. Ce concept peut également être considéré pour les fonctions de corrélation, en ap-

[3] Il s'agit des contraintes naturelles associées à la relation d'incertitude de Heisenberg.

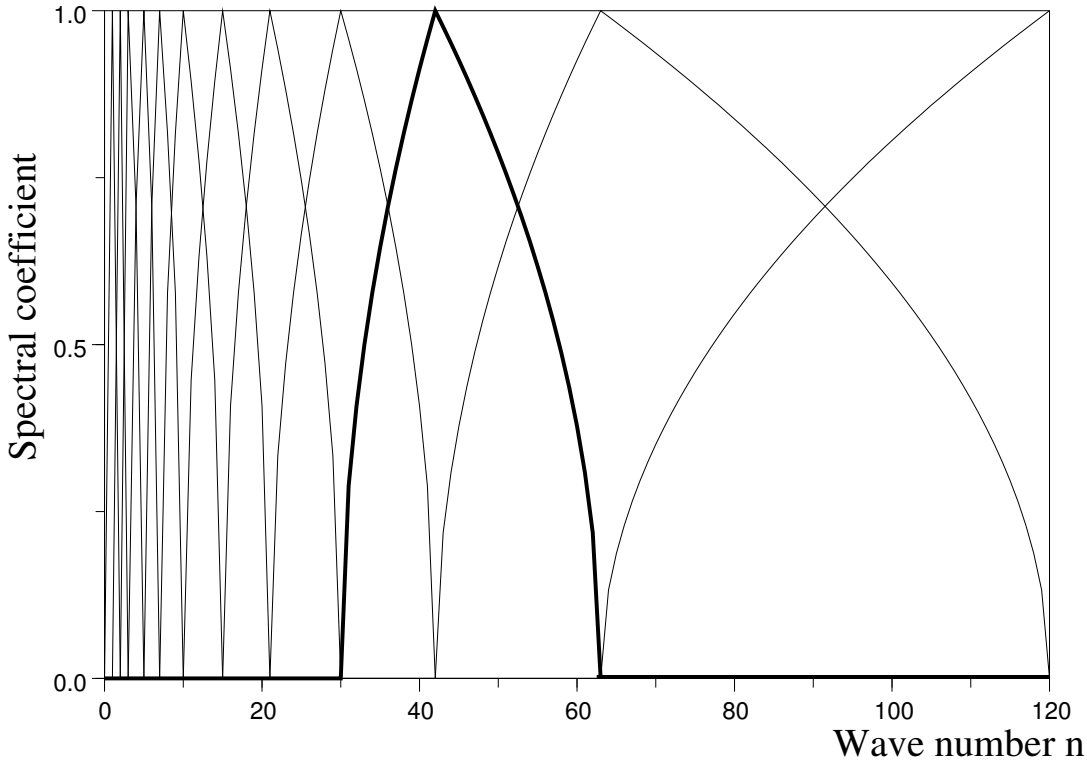


FIG. 4.3 – Coefficients spectraux des fonctions ondelettes $\psi_j(x)$ pour différentes échelles j (il y a une courbe par fonction) et pour la troncature spectrale $T = 120$. Le spectre associé à une fonction $\psi_j(x)$ particulière ($j = 10$) est représenté en gras.

pliquant le formalisme aux erreurs d'ébauche normalisées par les écarts types en points de grille, soit $\varepsilon'(x,k) = \varepsilon(x,k)/\sigma_b(x)$, avec $\sigma_b(x) = (\frac{1}{N_e}\sum_k \varepsilon(x,k)^2)^{1/2}$. C'est cette approche qui est envisagée dans la suite du chapitre, de sorte que l'on ait une formulation des covariances similaire à Fisher (2003) et Deckmyn et Berre (2005). La formulation de $\boldsymbol{B}$ est souvent déterminée par la construction de $\boldsymbol{B}^{-1/2}$, qui est détaillée dans l'annexe B de ce chapitre. Ainsi, dans le cas où l'on ne considère que les composantes horizontales, la racine carrée de la matrice de covariance d'erreur de prévision modélisée via les ondelettes, et représentée en points de grille, s'écrit

$$\boldsymbol{B}_w^{1/2} = \boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_s \boldsymbol{W}^{-1} \boldsymbol{D}_w^{1/2}, \tag{4.1}$$

avec $\boldsymbol{\Sigma}_g$ la matrice diagonale constituée des écarts types en points de grille, $\boldsymbol{\Sigma}_s$ correspondant à une multiplication par les écarts types spectraux, et $\boldsymbol{D}_w^{1/2}$ la matrice diagonale dans l'espace des ondelettes constituée des écarts types ondelettes. Les matrices $\boldsymbol{W}$ et $\boldsymbol{W}^{-1}$ correspondent respectivement à la transformation ondelette directe et à son inverse (à noter que $\boldsymbol{W}$ est une matrice rectangulaire et que $\boldsymbol{W}^{-1}$ est une matrice inverse à droite de $\boldsymbol{W}$).

La particularité des fonctions ondelettes ψ_j utilisées par Fisher (2003) sur la sphère est qu'elles sont à bande limitée et définies dans l'espace spectral de la manière suivante. Pour $(N_j)_{j\in[0,J]}$, avec $N_j < N_{j+1}$, les coefficients spectraux des fonctions ψ_j sont donnés par, pour $j \neq 0$ (n étant le nombre d'onde totale dans le cas sphérique) : $\frac{1}{\sqrt{2n+1}}\check{\psi}_{j,n} = \sqrt{\frac{n-N_{j-1}}{N_j-N_{j-1}}}$ pour $N_{j-1} \leq n < N_j$, $\sqrt{\frac{N_{j+1}-n}{N_{j+1}-N_j}}$ pour $N_j \leq n < N_{j+1}$, et 0 sinon. Pour $j = 0$, la définition est la même, excepté que la plage $N_{j-1} \leq n < N_j$ est remplacée par $0 \leq n < N_0$, pour laquelle $\frac{1}{\sqrt{2n+1}}\check{\psi}_{j,n} = 1$.

Il est possible de définir, de manière équivalente, des ondelettes sur le cercle via le spectre

de Fourier. Les coefficients spectraux de Fourier correspondants, $\check{\psi}_{j,n}$, sont représentés sur la Fig. 4.3, pour la série suivante $\{N_j\} = \{0, 1, 2, 3, 5, 7, 10, 15, 21, 30, 42, 63, 120\}$. Cette série a été choisie pour être plus proche de celle utilisée par Fisher (2003). Différentes voies pour définir une suite optimale pourront être étudiées dans le futur. En particulier, comme le discutent Fisher et Andersson (2001), c'est le choix de la largeur des bandes qui détermine le compromis entre les résolutions spectrale et point de grille. Quand les bandes sont larges, les variations spectrales sont plus petites, mais les variations géographiques sont plus grandes. Ceci fait écho à la discussion de la figure 4.2 présentée en section 4.2.3. La moyenne spatiale est plus localisée quand les bandes sont larges, ce qui améliore la représentation des variations géographiques.

On peut également mentionner que chaque champ de coefficients ondelettes $\hat{\boldsymbol{\varepsilon}}_j$ peut être représenté exactement sur une grille à plus basse résolution, qui correspond à une troncature égale à $T_j = min(N_{j+1}, T)$, T étant la troncature maximale, associée à la grille haute résolution initiale ($T = 120$ dans l'exemple ci-dessus). Ainsi, la transformation ondelette directe est appliquée de la manière suivante :

$$\hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} . \\ \hat{\boldsymbol{\varepsilon}}_j \\ . \end{pmatrix} = \boldsymbol{W}\boldsymbol{\varepsilon} = \begin{pmatrix} . \\ \boldsymbol{G}_j \check{\boldsymbol{\Psi}}_j \boldsymbol{S}_j \\ . \end{pmatrix} \boldsymbol{\varepsilon},$$

où $\check{\boldsymbol{\Psi}}_j$ est la matrice diagonale contenant les coefficients spectraux $\check{\psi}_{j,n}$, $\boldsymbol{S}_j$ désigne la transformation spectrale directe avec la troncature T_j, et $\boldsymbol{G}_j$ est la transformation inverse correspondante vers une grille à basse résolution, associée à la troncature T_j. La transformée ondelette inverse, pour sa part, est définie par $\boldsymbol{\varepsilon} = \boldsymbol{W}^{-1}\hat{\boldsymbol{\varepsilon}} = \boldsymbol{G}_J \sum_{j=1}^{J} \check{\boldsymbol{\Psi}}_j \boldsymbol{S}_j \hat{\varepsilon}_j$, où $\boldsymbol{G}_J$ est la transformation spectrale inverse à pleine résolution (puisque $T_J = T$). Un exemple de transformée ondelette d'un champ d'erreur $\hat{\boldsymbol{\varepsilon}}$ est illustré sur la figure 4.4.

4.3 Propriétés de filtrage des ondelettes dans un contexte analytique 1D

4.3.1 Matrice de covariance hétérogène analytique 1D

Un contexte analytique 1D simple est utilisé pour illustrer les propriétés de filtrage des ondelettes. Dans ce cadre, le domaine géographique est un grand cercle sur la terre, de rayon a. La coordonnée géographique est notée x, elle varie de $0\,^\circ$ à $360\,^\circ$ en terme d'angle (ou de 0 à $2\pi a$ en terme de distance). Sur ce cercle, un seul champ est considéré. Une matrice de covariance homogène est construite à partir d'une corrélation radiale gaussienne de la forme $f_H^x(s) = e^{-\frac{s^2}{2L_H^2}}$, où x est un point du cercle, s désigne la séparation, et L_H est la longueur de portée, qui est arbitrairement égale à $250\,km$.

Ainsi, une matrice de corrélation hétérogène est construite en déformant les champs à l'aide d'une c-transformation de Schmidt (Courtier et Geleyn, 1988), adaptée ici au cas du cercle, et définie par $h(x) = a\left[\pi - 2Atan\left(\frac{1}{c}tan(\frac{\pi}{2} - \frac{1}{2}\frac{x}{a})\right)\right]$; dans la suite $c = 2.4$. Cette transformation de Schmidt est utilisée sur la sphère, dans le cadre d'Arpège, pour étirer la maille du modèle, et ainsi obtenir une résolution variable. La corrélation ainsi construite est

$$f^x(s) = f_H^{h^{-1}(x)}(h^{-1}(x+s) - h^{-1}(x)).$$

La matrice de corrélation hétérogène associée est notée $\boldsymbol{B}_a$. Elle se caractérise par une hétérogénéité de ses corrélations, qui sont courtes au voisinage de $180\,^\circ$ et larges au voisinage de $0\,^\circ$.

Ceci est illustré sur le graphe en haut de la figure 4.4, qui représente la longueur de portée de corrélation $L(x)$ en un point x, d'après l'approximation de Belo Pereira et Berre (2006)

$$L^2(x) = \frac{\sigma(\varepsilon)^2(x)}{\sigma(\partial_x \varepsilon)^2(x) - (\partial_x \sigma(\varepsilon))^2(x)}, \tag{4.2}$$

où $\sigma(\varepsilon)(x)$ est l'écart type de $\varepsilon(x)$, et ∂_x correspond à la dérivation suivant la coordonnée d'espace.

Dans le cadre 1D, l'évaluation des portées calculées par la formulation 4.2 implique l'application de l'opérateur gradient, et de son adjoint, à la matrice de covariance $\boldsymbol{B}_w$. Par exemple, la matrice de covariance $\boldsymbol{B}'_{w,xx}$ de $\partial_x \varepsilon$ correspond à

$$\boldsymbol{B}'_{w,xx} = \overline{(\partial_x \varepsilon)(\partial_x \varepsilon)^*} = \partial_x \overline{\varepsilon \varepsilon^*} \partial_x^* = \partial_x \boldsymbol{B} \partial_x^*. \tag{4.3}$$

La diagonale de $\boldsymbol{B}'_{w,xx}$ permet d'obtenir les variances de $\partial_x \varepsilon$ qui sont ensuite utilisées pour le calcul de la portée dans la formule 4.2.

4.3.2 Échantillonnage de B_a et produit de Schur

Afin d'examiner quels sont les effets du bruit d'échantillonnage, des perturbations aléatoires $\boldsymbol{\varepsilon}$ ont été générées à partir de $\boldsymbol{B}_a$, d'après $\boldsymbol{\varepsilon} = \boldsymbol{B}_a^{1/2} \boldsymbol{\zeta}$, où $\boldsymbol{\zeta}$ est la réalisation d'un vecteur suivant une loi normale de moyenne nulle et de matrice de covariance égale à l'identité (Fisher et Courtier 1995). Ceci permet d'obtenir une matrice de covariance $\boldsymbol{B}_e = \frac{1}{N_e} \sum_l \boldsymbol{\varepsilon}(l) \boldsymbol{\varepsilon}(l)^T$, issue d'un ensemble dont la taille vaut N_e.

Le graphique du milieu de la figure 4.4 représente un exemple de réalisation d'une erreur de prévision ainsi générée. Il est à noter que ce champ présente des variations plus rapides au voisinage de $180\,^\circ$ (dans la région Z1) qu'au voisinage de $0\,^\circ$ (dans la région Z2), en accord avec la portée locale de la matrice analytique spécifiée (graphique du haut sur la figure 4.4). On s'attend à ce que cette variation de l'échelle soit représentée par l'approche diagonale ondelette, avec des amplitudes plus grandes pour les variances de petite échelle dans la zone Z1 que dans la zone Z2.

Ceci est conforté par le graphique en bas de la figure 4.4, qui représente les amplitudes des coefficients ondelettes de l'exemple de champ d'erreur. Comme attendu, les amplitudes des coefficients représentant les petites échelles (pour $j \geq 10$) tendent à être plus grandes dans la zone Z1 que dans la zone Z2.

La matrice de covariance $\boldsymbol{B}_e$ estimée à partir de l'ensemble peut être comparée à la matrice exacte $\boldsymbol{B}_a$, pour examiner l'effet du bruit d'échantillonnage, mais aussi à la matrice $\boldsymbol{B}_w$ modélisée à l'aide des ondelettes et définie par l'équation (4.1). Le bruit d'échantillonnage est une question particulièrement importante dans les schémas d'assimilation tels que celui du filtre de Kalman d'Ensemble (EnsKF), et rend nécessaire l'utilisation d'un filtrage par un produit de Schur (Houtekamer *et al.* , 2001). Le produit de Schur correspond à un produit terme à terme de la matrice $\boldsymbol{B}_e$ et d'une matrice $\boldsymbol{F}_{L_s}$. Ainsi, une matrice de covariance estimée à partir d'un ensemble et filtrée, que l'on notera $\boldsymbol{B}_e^{L_s}$, peut être obtenue d'après

$$\boldsymbol{B}_e^{L_s} = \boldsymbol{F}_{L_s} \circ \boldsymbol{B}_e,$$

où $\boldsymbol{F}_{L_s}$ est une matrice qui correspond à une fonction de corrélation à support compact (à prendre au sens de fini et localisé au voisinage du point d'intérêt)[4]. Dans ce chapitre, une fonc-

[4]Naturellement, cette désignation de support compact est un abus de langage. En effet, le domaine étant le cercle ou la sphère qui sont compacts au sens topologique, fournissent nécessairement des fonctions localisées en espace. Cependant, cette désignation correspond plus au fait que le support de la fonction de corrélation est non nul uniquement dans un voisinage proche du point d'intérêt, éliminant les corrélations à trop grande distance.

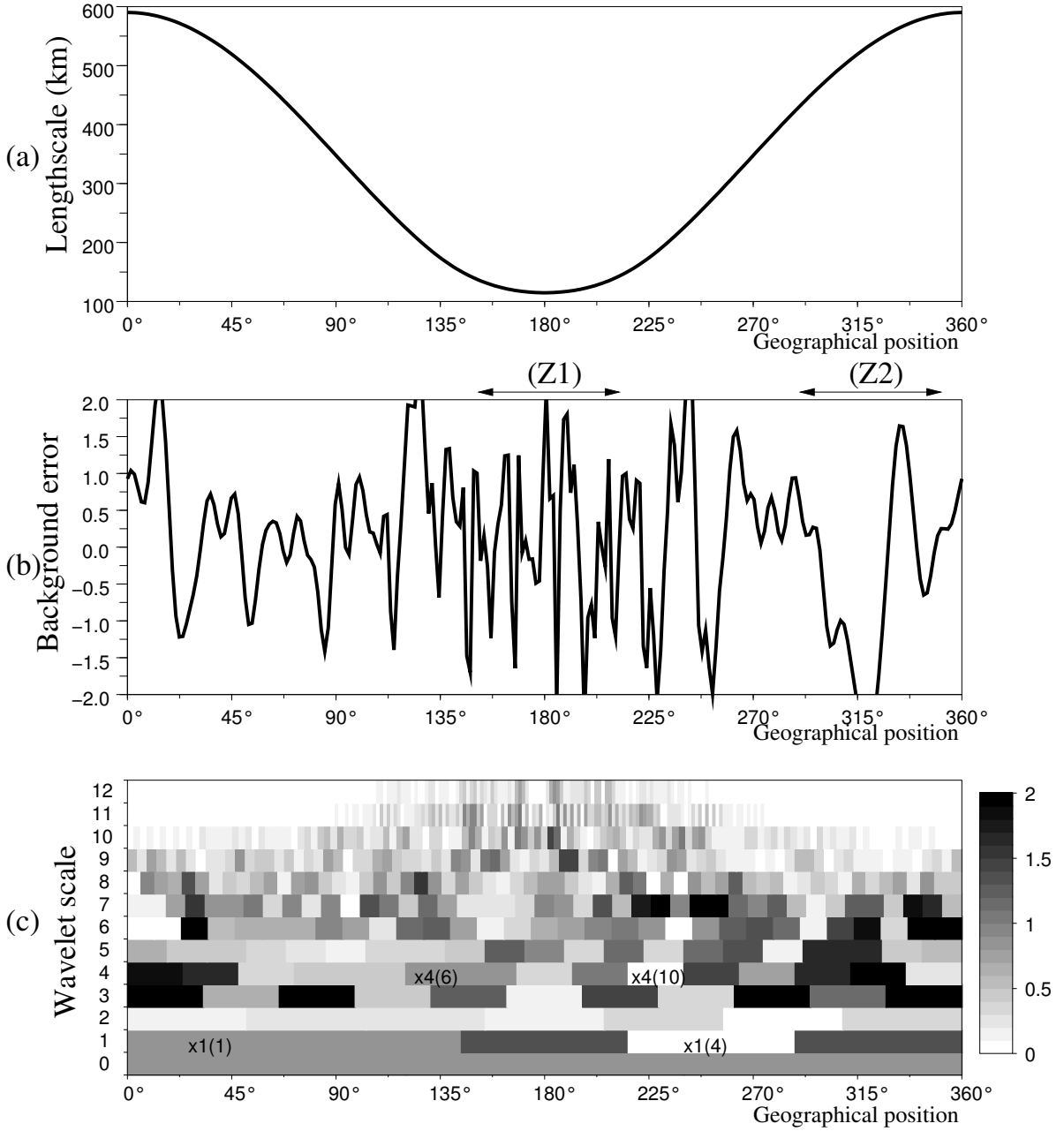


FIG. 4.4 – (a) Variations géographiques de la portée dans le contexte analytique. (b) Une réalisation de $\boldsymbol{\varepsilon}$ pour une erreur d'ébauche gaussienne associée au modèle de covariance d'ébauche analytique sur le cercle et (c) la valeur absolue des coefficients ondelettes de cette erreur : chaque courbe correspond à la valeur absolue de $\hat{\boldsymbol{\varepsilon}}_j = \boldsymbol{S}_j^{-1} \check{\boldsymbol{\Psi}}_j \boldsymbol{S}_j \boldsymbol{\varepsilon}$; pour une échelle j donnée, il y a une boîte par point de grille pour la grille à basse résolution associée. Par exemple $x_1(1)$, $x_1(4)$ pour $j = 1$ ou $x_4(6)$, $x_4(10)$ pour $j = 4$.

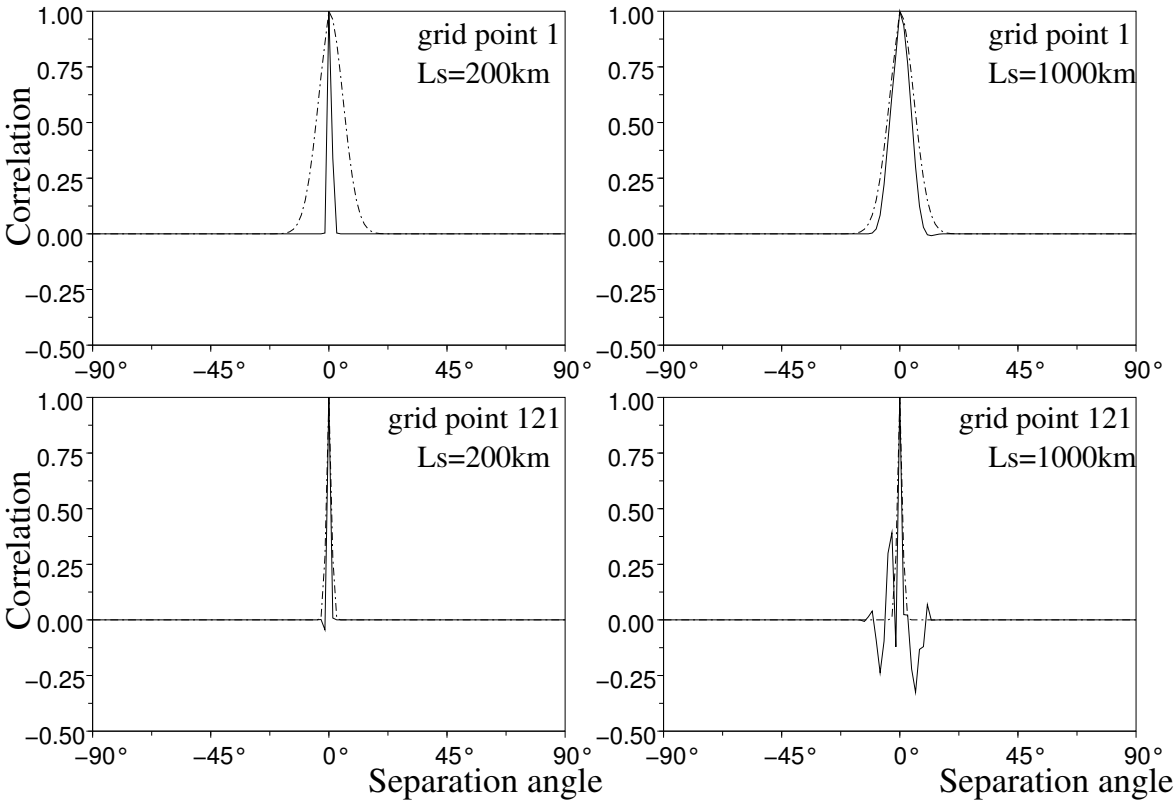


FIG. 4.5 – Fonctions de corrélation ensemblistes (courbe continue) associées aux points de grille 1 (correspondant à la région de grandes longueurs de portée) et 121 (correspondant à la région de courtes portées). Les corrélations sont directement calculées à partir d'un ensemble de 10 membres, puis filtrées à l'aide d'un produit de Schur $\boldsymbol{F}_{L_s}$ avec $L_s = 200\ km$ (graphiques de gauche) et $1000\ km$ (graphiques de droite). Les fonctions de corrélation exactes sont également représentées (courbe tiretée-pointilliée).

tion rationnelle par morceau et d'ordre de dérivation 5 est utilisée. Cette fonction correspond à l'équation (4.10) de Gaspari et Cohn (1999). Elle est utilisée de manière classique dans les méthodes ensemblistes. En particulier, elle est utilisée par Houtekamer et Mitchell (2001).

Le filtrage de Schur assure que les corrélations à longue distance sont mises à zéro. Ceci est fait au prix d'un rétrécissement des fonctions de corrélation dans les distances intermédiaires.

Ce filtre ne change pas les variances locales elles-mêmes, alors qu'elles sont également affectées par le bruit d'échantillonnage. Une normalisation par les valeurs diagonales de $\boldsymbol{B}_e^{L_s}$ est donc appliquée également, ce qui assure que l'estimation $\boldsymbol{B}_e^{L_s}$ soit une matrice de corrélation valide. Cette procédure permet de se concentrer sur l'effet du bruit d'échantillonnage sur les fonctions de corrélation.

L'effet de $\boldsymbol{F}_{L_s}$ sur les corrélations finales est sensible. Si L_s est trop courte (graphique de gauche sur la figure 4.5, correspondant à $L_s = 200\ km$), la fonction de corrélation exacte est remplacée par une fonction de corrélation excessivement homogène et étroite. Dans ce cas, les fonctions de corrélation associées à la zone de portées courtes Z1 (point de grille 121) sont bien représentées, mais celles associées à la zone de portées larges Z2 (point de grille 1) sont bien trop étroites par rapport à ce qu'elles devraient être.

D'autre part, L_s peut être augmenté (ici jusque $1000\,km$), de façon à ce que les grandes portées de la zone Z2 soient préservées. Cependant, dans la zone Z1, on voit alors apparaître des oscillations parasites à proximité du point d'intérêt et qui ne sont plus filtrées dans ce cas

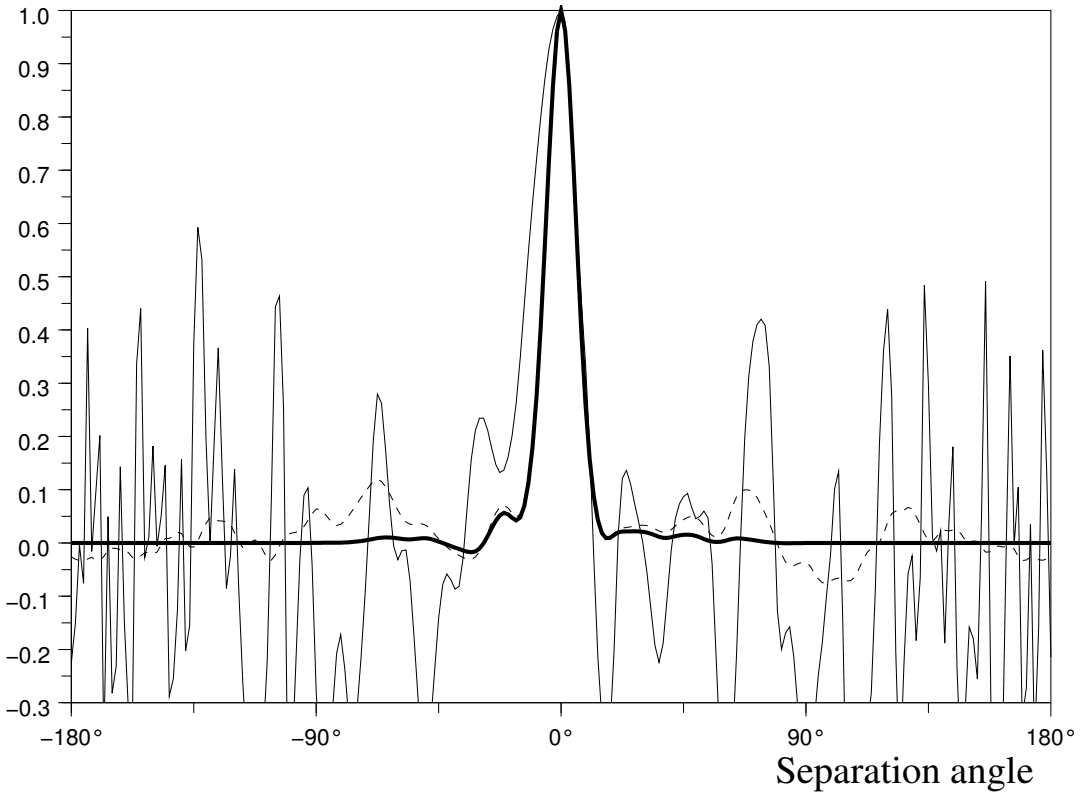


FIG. 4.6 – Fonction de corrélation ensembliste relative au point 1, et calculée à partir d'un ensemble de 10 membres : soit directement (courbe continue), soit avec l'approche ondelette diagonale sans filtre de Schur (courbe pointillée) ou soit encore avec l'approche ondelette diagonale et avec filtre de Schur tel que $L_s = 6000km$ (courbe continue en trait gras).

(Fig. 4.5, graphique de droite).

L'échelle L_s doit donc être choisie de manière judicieuse. Cette valeur optimale dépend également de la taille de l'ensemble (Houtekamer et Mitchell, 2001 ; Lorenc, 2003).

D'autre part, un filtre de Schur est aussi appliqué à la formulation ondelette pour fournir une version filtrée de cette matrice, permettant d'éliminer certaines imperfections de la formulation (en particulier on peut trouver des corrélations faiblement non nulles à l'opposé du point d'intérêt). Ceci permet d'obtenir la matrice

$$\boldsymbol{B}_w^{L_s} = \boldsymbol{F}_{L_s} \circ \boldsymbol{B}_w.$$

La nécessité relative et les effets d'un filtrage de Schur pour l'approche ondelette sont illustrés sur la figure 4.6. Il s'agit ici d'une comparaison entre les corrélations calculées à partir d'un ensemble de 10 membres, soit directement (courbe continue), soit en utilisant l'approche ondelette diagonale (courbe tiretée). La mise à zéro des corrélations de longue distance reste souhaitable pour l'approche ondelette dans cet exemple, bien que cela soit nettement moins marqué que pour l'estimation ensembliste directe. L'effet du filtre de Schur avec $L_s = 6000km$ sur la modélisation des fonctions de corrélation basées sur les ondelettes est illustré par la courbe continue en trait gras. Il met à zéro les corrélations à grande distance. La possibilité de se passer d'un filtrage de Schur pourra être explorée dans les études futures, par exemple en utilisant des bandes plus larges pour assurer une corrélation plus locale.

Dans la suite du chapitre, $\boldsymbol{B}_e^{L_s}$ et $\boldsymbol{B}_w^{L_s}$ font référence aux matrices de corrélation respectivement d'ensemble et modélisée à l'aide des ondelettes.

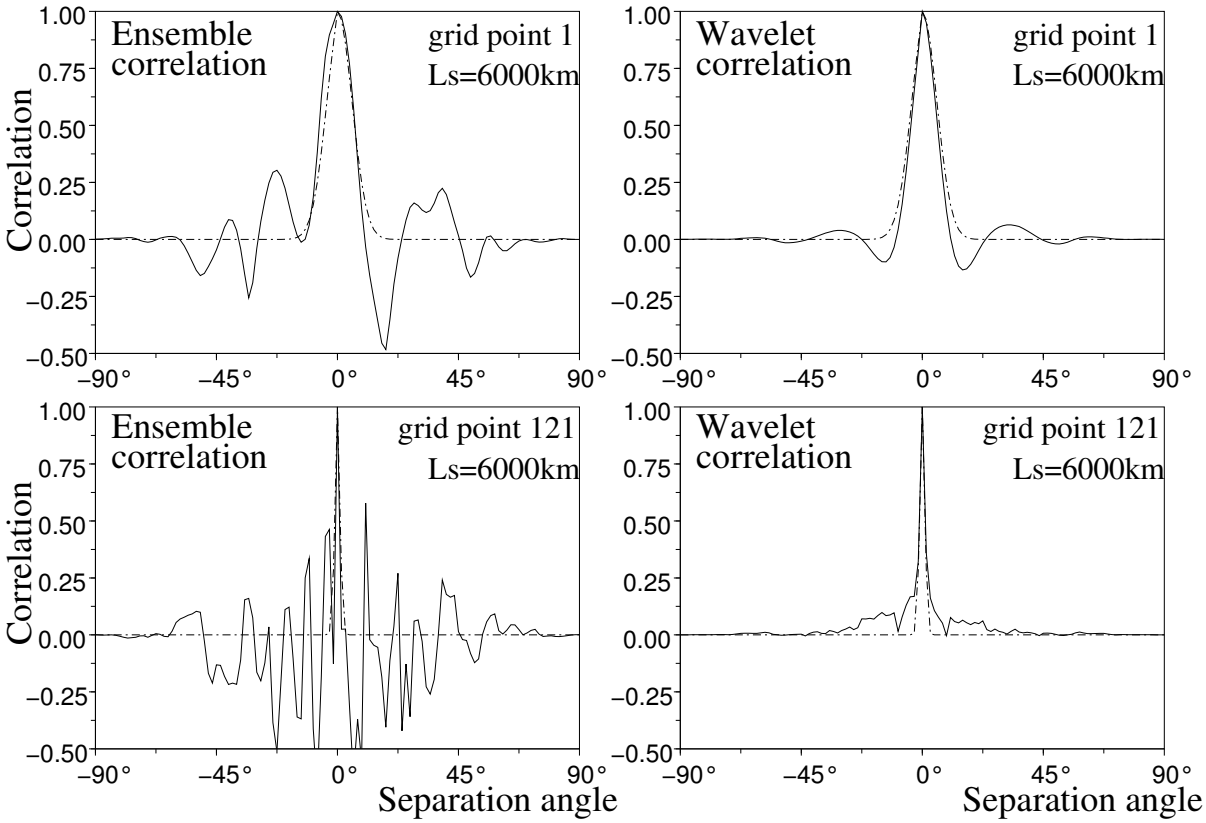


FIG. 4.7 – Fonctions de corrélation ensemblistes et en ondelette (courbe continue) relatives aux points de grille 1 (correspondant à des portées larges) et 121 (correspondant à des portées courtes). Les corrélations sont calculées à partir d'un ensemble de 10 membres et sont ensuite filtrées à l'aide d'un produit de Schur $\boldsymbol{F}_{L_s}$ en prenant $L_s = 6000\ km$. Les corrélations exactes sont également représentées (courbe tiretée-pointilliée).

4.3.3 Filtrage ondelette des fonctions de corrélation et de leurs variations

Les propriétés de filtrage ondelette peuvent être étudiées en comparant les graphiques de gauche et de droite de la figure 4.7, produites avec les paramètres $L_s = 6000km$ et $N_e = 10$ membres. Une grande valeur pour L_s a été choisie pour permettre de visualiser l'amplitude typique du bruit d'échantillonnage. Le graphique de gauche montre les corrélations ensemblistes brutes pour le point de grille 1 (graphique du dessus) et 121 (graphique du dessous). Les graphiques ondelettes correspondants sont à droite, pour ces mêmes points et ce même ensemble. Le bruit d'échantillonnage affectant les corrélations brutes est bien visible : plusieurs oscillations irréalistes de grande amplitude apparaissent au voisinage des points considérés. De tels artefacts sont moins marqués pour les corrélations ondelettes.

Ainsi, la formulation ondelette apparaît comme filtrant en partie le bruit d'échantillonnage. Cette propriété est attendue sachant que l'approche ondelette diagonale revient à moyenner localement les fonctions de covariance, ce qui revient à augmenter la taille de l'échantillon total. Une interprétation en ondelette de ce filtrage est également discutée dans l'annexe C de ce chapitre.

La figure 4.8 montre la manière dont le bruit d'échantillonnage affecte les longueurs de portée. Cette figure montre également le filtrage ondelette. Le cas $N_e = 10$ (graphique en haut à gauche) est le plus spectaculaire. En effet, l'estimation brute présente de grandes oscillations

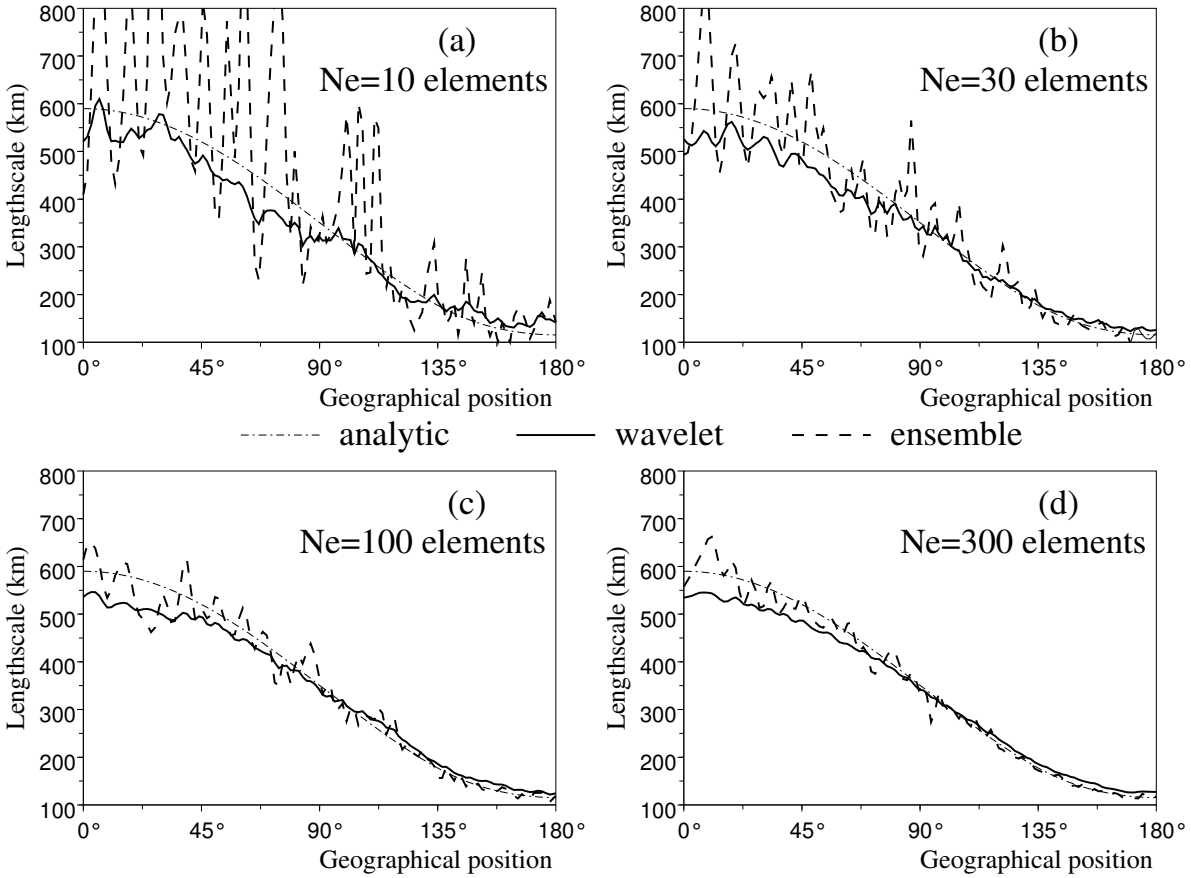


FIG. 4.8 – Variations géographiques de la longueur de portée : exactes (courbe tiretée-pointilliée), estimées à l'aide des ondelettes (courbe continue) et estimées directement à partir de l'ensemble (courbe tiretée). Les longueurs de portée sont calculées pour différentes tailles d'ensemble, $N_e = 10, 30, 100$ et 300. Pour chaque cas, les corrélations ont été filtrées à l'aide d'un produit de Schur $\boldsymbol{F}_{L_s}$ avec $L_s = 6000\ km$.

artificielles de petite échelle, et l'on remarque que les grandes valeurs de portée sont parfois très largement surestimées (*e.g.* avec $800\ km$ pour la valeur brute au voisinage de $70\ °$, tandis que la valeur théorique n'est que de $400\ km$). À l'inverse, les longueurs de portée représentées par le modèle ondelette présentent des variations géographiques plus lisses et des valeurs plus réalistes. L'approche ondelette s'avère ainsi capable de capter et représenter les variations géographiques principales (ici l'augmentation de portée entre $180\ °$ et $0\ °$) à partir d'un ensemble de petite taille ne comprenant que 10 membres.

À nouveau, la moyenne locale offerte par la formulation ondelette permet une réduction du bruit d'échantillonnage. Ces effets bénéfiques diminuent quand la taille de l'ensemble augmente.

4.3.4 Expériences d'assimilation de données

Des expériences d'assimilation sur le cercle ont également été menées. Comme dans ce contexte les vraies fonctions de corrélation sont connues, la vraie solution peut être calculée, puis comparée avec celles fournies par les modèles de corrélation (estimation et ondelette).

Dans ces expériences, l'état vrai est le champ nul (il vaut zéro en tout point). Une erreur de prévision est générée à partir de la matrice de corrélation analytique (on rappelle ici que l'on a fixé la carte d'écart type à 1), avec $\boldsymbol{B}_a^{1/2}$. Les observations sont réalisées à partir de $\boldsymbol{R}^{1/2}$, la racine carrée de la matrice de covariance d'erreur d'observations $\boldsymbol{R}$. $\boldsymbol{R}^{1/2}$ est supposée être une matrice diagonale $\sigma_o \boldsymbol{I}$, où $\boldsymbol{I}$ désigne la matrice identité (dans la suite, $\sigma_o = 0.95$: ainsi les observations sont supposées être de qualité similaire à l'ébauche). Le réseau est de densité homogène sur le domaine, avec une observation tous les cinq points de grille.

Les valeurs RMSE (Root Mean Square Error) des erreurs d'analyse sont moyennées sur le domaine[5]. Elles sont calculées pour les trois approximations de la matrice de covariance : l'estimation directe, la formulation homogène (*i.e.* l'approche diagonale spectrale) et la formulation ondelette. Ces RMSE peuvent être comparés avec la valeur RMSE correspondant aux covariances vraies. La figure 4.9 représente les différences $RMSE - RMSE(vrai)$, pour différentes échelles du filtre de Schur L_s et pour différentes tailles d'ensemble N_e. Sur cette figure, toutes les courbes présentent le même comportement (similaire à celui observé par Lorenc 2003) : les courbes sont convexes avec un minimum déterminant une valeur optimale de l'échelle du filtre de Schur L_s.

Pour l'estimation ensembliste directe (courbe tiretée), plus l'ensemble est petit, plus le RMSE est grand, et plus l'échelle L_S doit être petite. De telles dépendances de la valeur RMSE et de l'échelle optimale L_s à la taille de l'ensemble sont réduites pour l'approche ondelette (courbe continue). En outre, les résultats des ondelettes sont relativement bons, même pour des ensembles de petite taille (de l'ordre de $N_e = 10$). Ceci illustre l'impact bénéfique du filtrage ondelette.

Comparé avec l'estimation directe, un autre résultat attractif de l'approche ondelette est que la pente du RMSE est relativement faible au-delà de la valeur optimale pour L_s. Cela signifie que la qualité de l'analyse sera moins affectée par le choix d'une échelle de Schur sous-optimale, que dans le cas de l'estimation directe.

Comme attendu, les analyses produites avec l'approche ondelette sont plus proches de l'analyse optimale que celles produites par la formulation homogène. D'un autre côté, il est intéressant de noter que pour $N_e = 10$ l'approche homogène peut donner de meilleurs résultats que l'estimation directe dans le cas où l'échelle du filtre de Schur est trop large. Ceci est une autre

[5]Cette valeur RMSE correspond en pratique au calcul de $RMSE = \sqrt{1/n\, Trace(\boldsymbol{A}_{\tilde{\boldsymbol{K}}})}$, avec $\boldsymbol{A}_{\tilde{\boldsymbol{K}}}$ la matrice de covariance d'erreur d'analyse pour la matrice de gain $\tilde{\boldsymbol{K}}$.

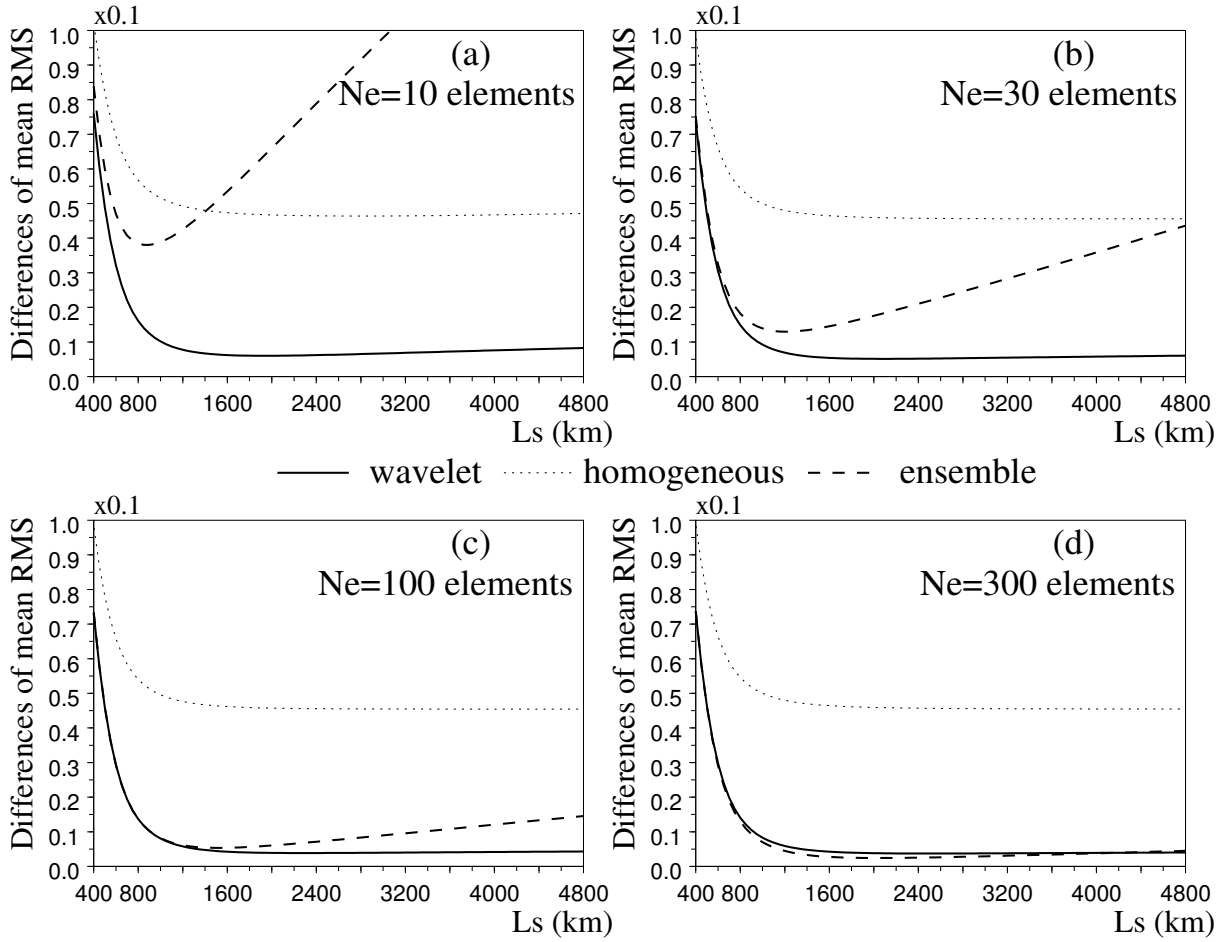


FIG. 4.9 – Différences entre la valeur RMSE moyenne de l'erreur de l'analyse sous-optimale et celle de l'analyse optimale (associée à la vérité), en fonction de l'échelle intervenant dans le produit de Schur L_s et de la taille de l'ensemble N_e (respectivement 10, 30, 100 et 300 pour les quatre quadrants). Ces différences sont représentées pour l'analyse ondelette (courbe continue), l'analyse avec la formulation homogène (courbe pointillée) et l'analyse avec l'estimation brute (courbe tiretée). Toutes ces analyses sont réalisées pour un réseau homogène avec une observation tous les cinq points.

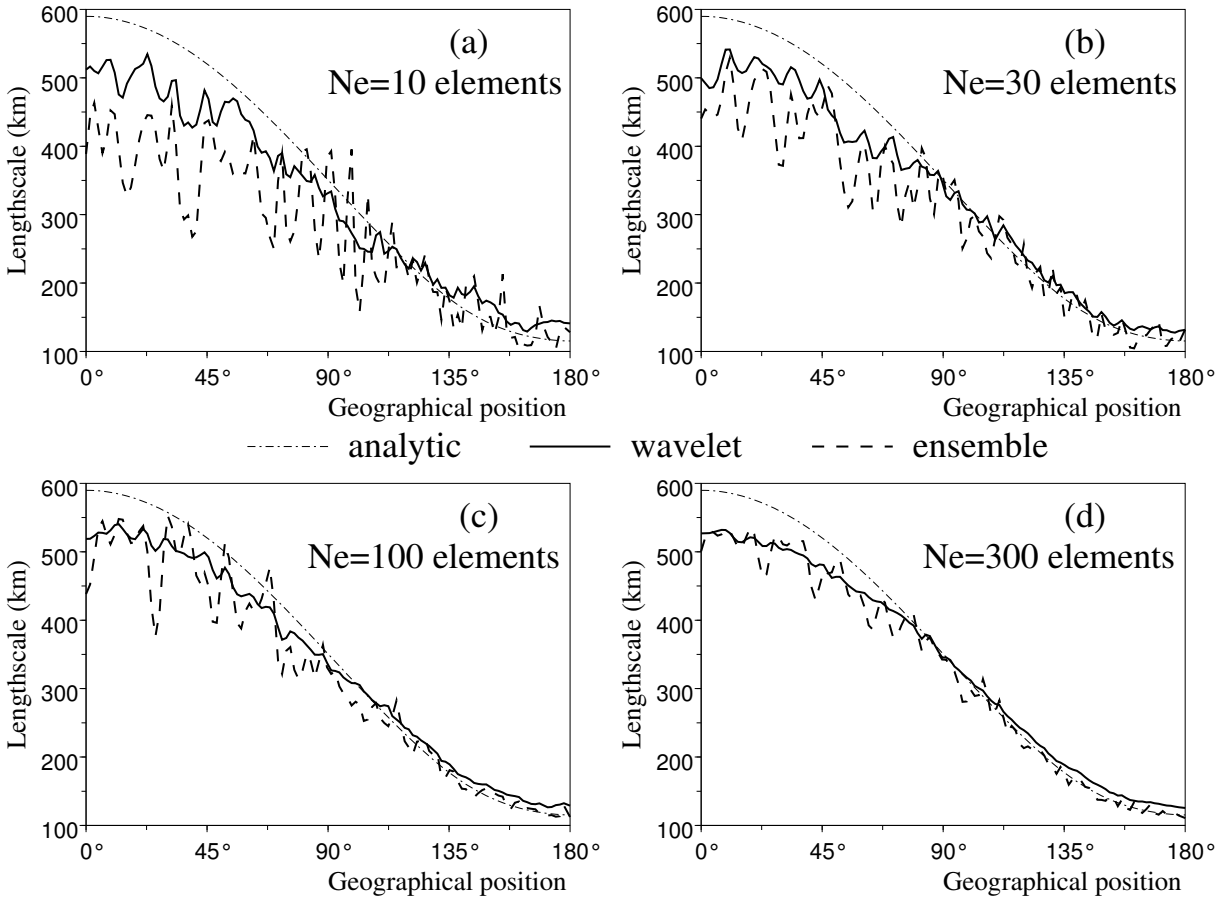


FIG. 4.10 – Même figure que 4.8, mais avec l'échelle du filtre de Schur L_s prise à sa valeur optimale, et déterminée d'après la figure 4.9. Pour les ondelettes, L_s est constante, égale à $4000\ km$. Pour l'estimation ensembliste, L_s dépend de la taille de l'ensemble N_e : (a) $(N_e = 10, L_s = 900\ km)$, (b) $(N_e = 30, L_s = 1250\ km)$, (c) $(N_e = 100, L_s = 1600\ km)$ et (d) $(N_e = 300, L_s = 2050\ km)$.

mise en évidence de l'importance du bruit d'échantillonnage associé aux ensembles de petite taille, et illustre à nouveau le potentiel bénéfique de la moyenne spatiale des fonctions de covariance dans ce cas.

Il est également possible de comparer les variations géographiques des portées quand l'échelle du filtre de Schur utilisé est optimale. Les résultats sont rapportés sur la figure 4.10. Les portées modélisées à l'aide des ondelettes restent plus précises et leurs variations sont plus lisses (à bon escient). C'est en particulièrement remarquable dans le cas des ensembles de petite taille.

Il est également intéressant de remarquer que, comparée à la figure 4.8, l'utilisation d'une échelle de Schur L_s plus petite implique un raccourcissement des portées. Ce résultat est cohérent avec le fait que plus l'échelle L_s est petite, plus la corrélation filtrée va décroître fortement (en fonction de la distance de séparation). En effet, d'après l'équation (4.2), cette variation est liée à l'augmentation de $\sigma(\partial_x \varepsilon)^2$, parce que les contributions des petites échelles sont amplifiées à la fois dans $\partial_x \varepsilon$ (comparé au spectre de ε) et quand la longueur de portée diminue.

4.4 Application à un ensemble de prévisons Arpège

4.4.1 Description de l'ensemble de prévisions Arpège sur le globe

Le modèle de prévision numérique global de Météo-France est basé sur le modèle Arpège (Courtier et Geleyn, 1998), associé à un schéma d'analyse variationnelle 4D-Var (Rabier *et al.* , 2000 ; Veersé et Thépaut, 1998). La matrice de covariance d'erreur d'ébauche est calculée en utilisant un ensemble d'assimilations perturbées (Houtekamer *et al.* , 1996 ; Fisher, 2003). Les résultats pour Arpège sont décrits en détails par Belo Pereira et Berre (2006).

Nous proposons d'illustrer ici quelques résultats typiques de la modélisation ondelette des covariances pour ce type d'ensemble de prévisions globales. Le formalisme ainsi que la liste des nombres d'ondes de coupure sont les mêmes que ceux présentés en section 4.2.4 (et Fisher, 2003). L'ensemble disponible est composé de six différences de prévisions pour chaque jour pour une période allant du 9 février au 24 mars 2002.

Dans ce cadre 2D sphérique, dont le nombre de degrés de liberté est très grand, le calcul de la longueur de portée est effectué en utilisant l'équation (4.2), avec une technique de randomisation pour diagnostiquer les portées issues de la modélisation ondelette. Dans ce dernier cas 1000 vecteurs aléatoires de ε ont été produits à partir de la racine carrée de $\boldsymbol{B}_w$ (Fisher et Courtier, 1995), et l'opérateur de dérivation ∂_x a été appliqué à ces réalisations. Les variances de $\partial_x \varepsilon$ ont ensuite été injectées dans l'équation de la portée.

Une première estimation possible pour les fonctions de covariance est de moyenner temporellement ces fonctions, de sorte que l'on diagnostique alors la "climatologie" des covariances d'erreur. Cela signifie que l'ensemble utilisé est alors constitué des six membres journaliers sur l'ensemble de la période. Ainsi, cet ensemble comprend $N_e = 264$ éléments au total.

Une seconde estimation possible consiste à étudier les covariances pour un jour donné. Ainsi le nombre d'éléments de cet ensemble est alors réduit à $N_e = 6$. Ces éléments correspondent aux différences entre les prévisions perturbées. Dans cette étude, les résultats présentés correspondent au 10 février pour le réseau de 12H UTC.

4.4.2 Carte des portées "climatologiques"

Comme le montre la figure 4.11, la méthode ensembliste permet de fournir des informations locales intéressantes sur les corrélations "climatologiques", qui apparaissent bien représentées par la formulation ondelette donnée par l'équation (4.1). Il apparaît un fort contraste terre-mer. Les longueurs de portée sont plus courtes dans l'hémisphère Nord (dense en observations) que dans l'hémisphère Sud. D'autre part, des zones de courtes portées sont bien visibles sur l'Atlantique Nord et sont associées au *rail des dépressions*, ou encore dans la zone de convergence intertropicale sur l'Afrique de l'Ouest. Les portées sont également influencées par l'orographie. Ainsi, l'Himalaya et les Andes sont bien marquées par la présence de plus courtes portées. À l'inverse, des maxima de longueur de portée sont bien marqués sur les parties tropicales des océans.

En accord avec les résultats de la section 4.3.3, ces comportements sur le globe nous confirment les avantages principaux de la formulation ondelette (comparée *e.g.* avec la formulation spectrale) : elle permet la représentation des variations géographiques des corrélations, en particulier celles qui sont associées aux processus dynamiques de l'atmosphère et aux variations de densité des observations.

On peut également noter que l'amplitude des variations géographiques est lissée sur la carte des portées en ondelette. Ceci est dû aux propriétés de filtrage de l'approche diagonale ondelette.

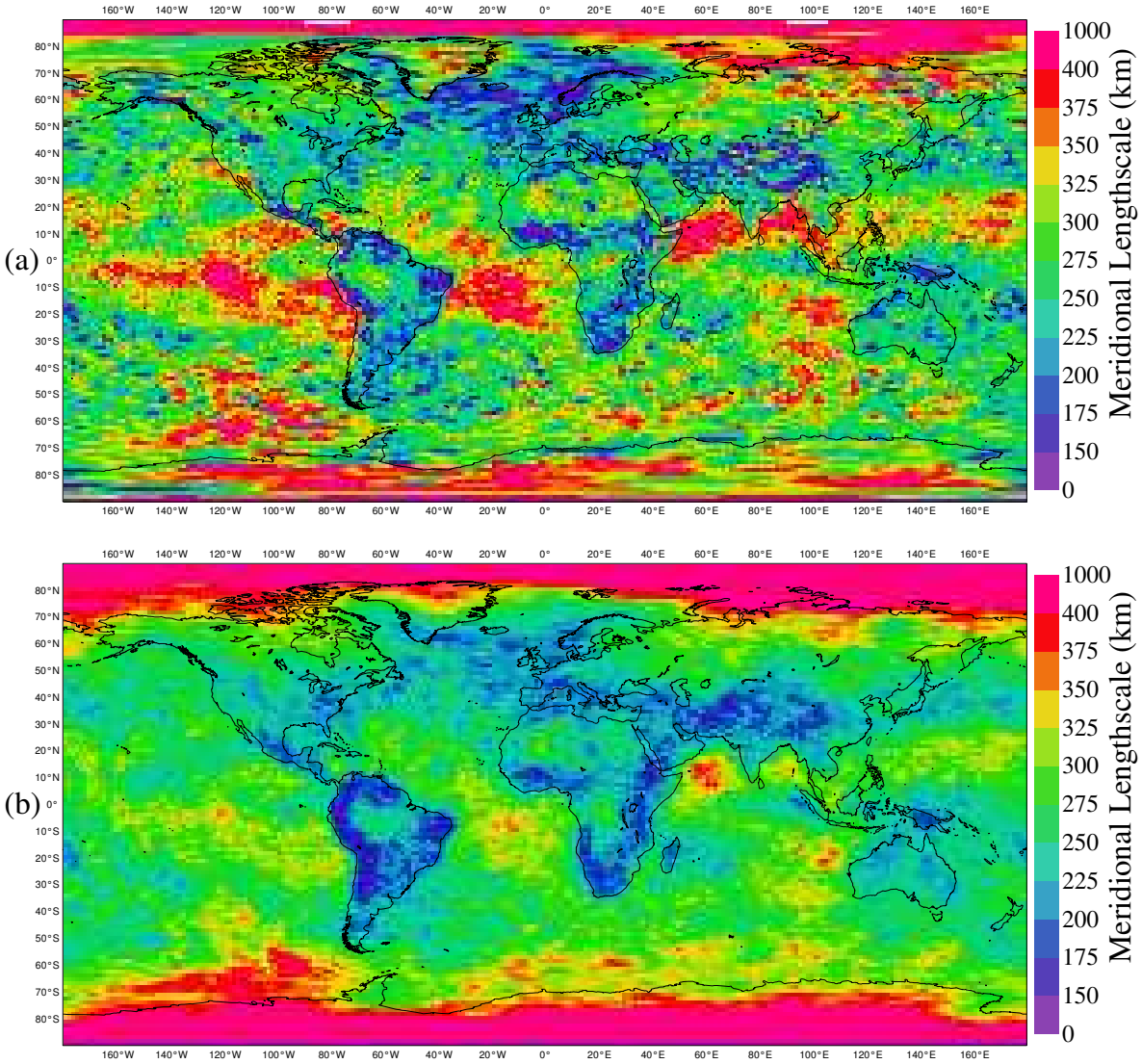


FIG. 4.11 – Longueurs de portée méridienne (en km) pour la pression de surface, moyennées sur la période de 46 jours et sur les 6 membres journaliers. (a) : les longueurs de portée brutes. (b) : les longueurs induites par la formulation ondelette.

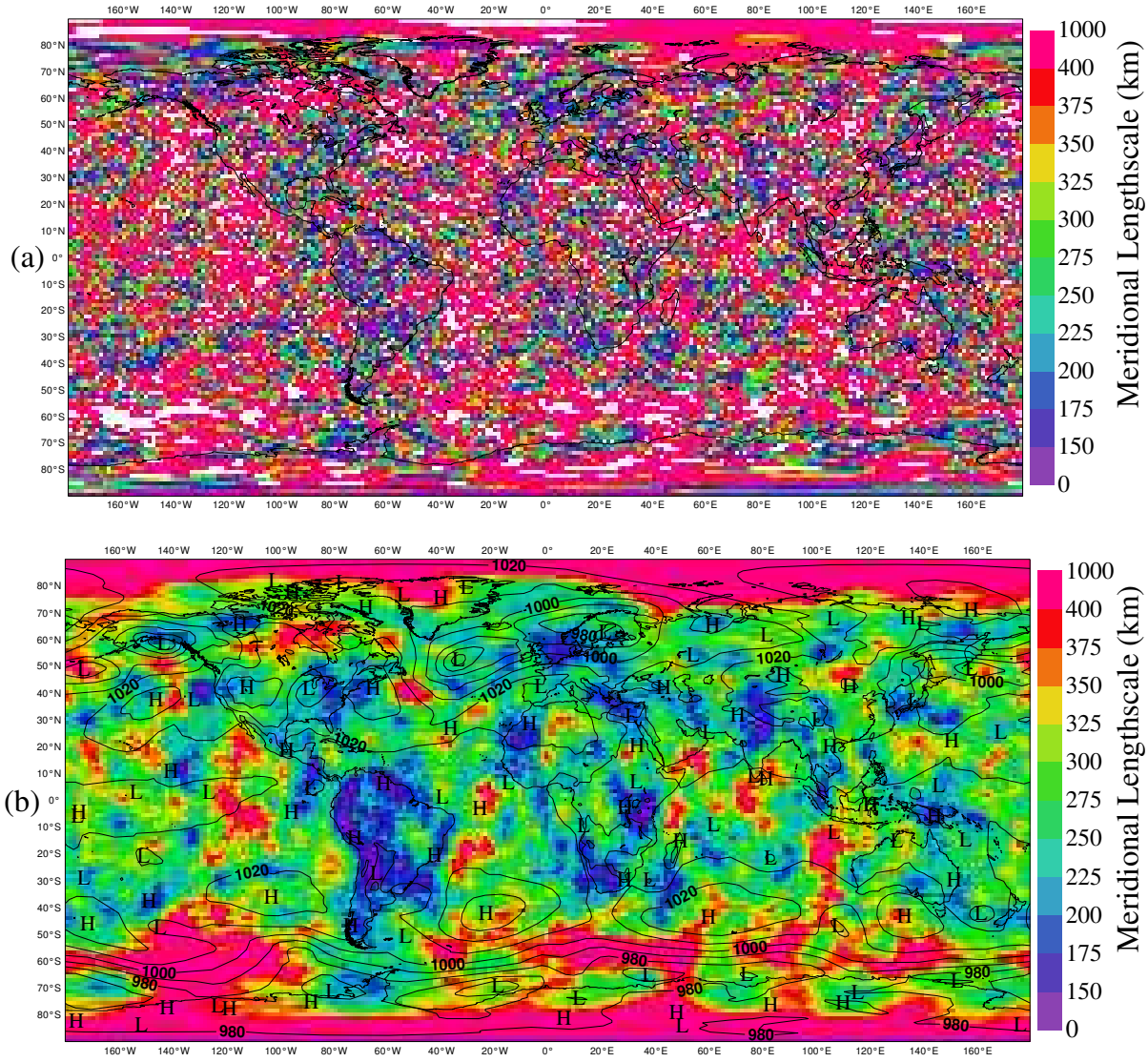


FIG. 4.12 – Longueurs de portée méridienne (en km) pour la pression de surface, valables pour le 10 février 2002 à 12H UTC, estimées à partir d'un ensemble de 6 membres. (a) : longueurs de portée brutes. (b) : longueurs de portée représentées par la formulation ondelette et superposées avec le champ de pression réduite au niveau de la mer prévu à cette date et heure.

On va maintenant montrer que ces propriétés sont encore plus importantes quand la taille de l'ensemble est réduite à $N_e = 6$.

4.4.3 Carte des portées pour un jour donné

La figure 4.12 représente les longueurs de portée pour le 10 février à 12H UTC. Les portées brutes (graphique du haut) apparaissent très bruitées, du fait de la taille réduite de l'ensemble. À l'inverse, les portées représentées par les ondelettes (graphique du bas, superposées avec le champ de pression réduite au niveau de la mer) sont relativement lisses et structurées. En particulier, on note des grandes portées dans l'océan circumpolaire austral et dans les zones tropicales de l'océan, comme dans le cas climatologique mais de manière plus prononcée. Des portées courtes sont visibles sur les continents. D'autres structures bien marquées sont à relier à la situation du jour. En particulier, on note un raccourcissement des portées dans les zones

dépressionnaires des moyennes latitudes (voir *e.g.* la dépression proche de la Scandinavie ou encore au sud de l'Argentine). Ces résultats sont en accord avec ceux décrits par Thépaut *et al.* (1995).

De telles différences entre les portées brutes et celles représentées par les ondelettes, quand la taille de l'échantillon est petite ($N_e = 6$), sont en accord avec les grandes différences de valeur sur l'estimation des portées obtenues dans le contexte 1D pour $N_e = 10$ (graphique en haut à gauche sur la figure 4.8). Ceci conforte l'idée que la formulation ondelette est capable de capter et de représenter les principales variations géographiques pertinentes, et ce même à partir d'un ensemble de petite taille, grâce à la moyenne spatiale locale.

4.5 Conclusions

Dans ce chapitre, la capacité de l'approche ondelette à restituer des variations géographiques lisses des fonctions de corrélation a été étudiée. La représentation des covariances à l'aide de l'approche diagonale dans l'espace des ondelettes revient à moyenner localement les fonctions de covariance. Grâce à cette moyenne spatiale locale, l'estimation statistique est mieux échantillonnée que dans le cas d'une estimation purement locale des fonctions de covariance. De plus, du fait que cette moyenne spatiale est locale, la représentation des variations géographiques des portées reste possible (contrairement au cas homogène). De telles propriétés de filtrage sont particulièrement attractives dans le cas de covariances d'erreur estimées à partir d'un ensemble de prévisions. Ces propriétés relatives à la formulation ondelette ont été formellement explicitées et ont été illustrées expérimentalement dans deux cadres d'étude.

Le premier cadre expérimental considéré est celui du cercle. Sur ce domaine, une formulation hétérogène a été construite, avec des variations géographiques de portée. Ainsi, la modélisation des fonctions de covariance à l'aide des ondelettes s'avère capable de représenter les fonctions locales de corrélation, ainsi que leurs variations géographiques (diagnostiquées à l'aide du calcul des portées). Il a été montré que cette représentation était plus lisse et réaliste que dans le cas d'une simple utilisation d'un filtrage par produit de Schur. Cela est particulièrement marqué avec des ensembles de petite taille (de l'ordre de 10 à 30 membres). La formulation ondelette reste compétitive jusque pour une taille d'ensemble de l'ordre de 100.

Le deuxième cadre expérimental a été donné par un ensemble de prévisions globales issues d'un modèle de prévision numérique. L'approche ondelette s'avère représenter correctement les variations géographiques "climatologiques" des portées locales, qui sont associées à la dynamique de l'atmosphère, et à l'hétérogénéité du réseau d'observations. De plus, un examen préliminaire des portées diagnostiquées pour une date donnée suggère que l'approche ondelette permet d'extraire des variations géographiques importantes, liées à la situation météorologique locale.

Ces résultats sont cohérents avec les propriétés de filtrage attendues des ondelettes, en terme de moyenne spatiale locale. Ils suggèrent que les ondelettes sont un outil prometteur pour estimer et représenter des covariances dépendantes de l'écoulement à partir d'un petit ensemble de prévisions.

4.6 Annexes

4.6.1 Annexe A : Formulation de la moyenne spatiale locale des covariances

Il est montré, dans le cadre général des "frames"[6], que l'hypothèse diagonale pour représenter les fonctions de covariance entraîne une pondération locale des fonctions de covariances initiales.

Décomposition de l'erreur d'ébauche dans une frame

Une frame (Daubechies, 1992 ; Fisher, 2004) est une famille de fonctions $\{\boldsymbol{\phi}_m, m \in \mathcal{M}\}$, où $\mathcal{M}$ est un ensemble dénombrable. Cette famille est associée à une frame duale $\left\{\tilde{\boldsymbol{\phi}}_m, m \in \mathcal{M}\right\}$, telle que le champ d'erreur $\boldsymbol{\varepsilon}$ peut s'analyser comme un ensemble de coefficients dans la frame $\hat{\varepsilon}_m$ avec

$$\hat{\varepsilon}_m = \sum_x \varepsilon(x)\tilde{\phi}^*_m(x), \tag{4.4}$$

où l'exposant * désigne l'opérateur de trans-conjugaison. Le signal analysé peut être recomposé d'après

$$\varepsilon(x) = \sum_{m\in\mathcal{M}} \hat{\varepsilon}_m \phi_m(x),$$

ou en terme vectoriel

$$\boldsymbol{\varepsilon} = \sum_{m\in\mathcal{M}} \hat{\varepsilon}_m \boldsymbol{\phi}_m.$$

Décomposition des fonctions de covariance dans une frame sous l'hypothèse diagonale

En utilisant la décomposition dans une frame, la matrice de covariance $\boldsymbol{B} = \overline{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^*}$ peut se représenter sous la forme $\boldsymbol{B} = \sum_{m,m'} \overline{\hat{\varepsilon}_m \hat{\varepsilon}^*_{m'}} \boldsymbol{\phi}_m \boldsymbol{\phi}^*_{m'}$. Sous l'hypothèse diagonale dans l'espace des coefficients de la frame, les covariances $\overline{\hat{\varepsilon}_m \hat{\varepsilon}^*_{m'}}$ sont nulles sauf pour $m = m'$. Ainsi, la matrice de covariance résultante $\boldsymbol{B}_d$ est $\boldsymbol{B}_d = \sum_m B_{mm} \boldsymbol{\phi}_m \boldsymbol{\phi}^*_m$, où $B_{mm} = \overline{\hat{\varepsilon}_m \hat{\varepsilon}^*_m}$ est la variance pour le coefficient m de la frame. Ainsi, la fonction de covariance f^x_d relative au point x a l'expression suivante :

$$f^x_d(s) = \sum_m B_{mm} \phi_m(x) \phi^*_m(x+s),$$

où s est la valeur de séparation (en points de grille). À partir de (4.4), les coefficients B_{mm} peuvent se réécrire

$$\begin{aligned} B_{mm} &= \sum_{x',s'} \overline{\varepsilon(x')\varepsilon(x'+s')^*}\ \tilde{\phi}^*_m(x')\tilde{\phi}_m(x'+s') \\ &= \sum_{x',s'} f^{x'}(s')\ \tilde{\phi}^*_m(x')\tilde{\phi}_m(x'+s'), \end{aligned}$$

[6] Il n'y a pas de traduction très claire pour ce terme. Yves Meyer parle de repère oblique, cependant un repère sous-entend la notion de base et donc de famille génératrice libre. Or les frames peuvent ne pas être une famille libre, rendant le terme de "repère" sans doute peu approprié ici. Stéphane Mallat ne le traduit pas, et conserve cette dénomination.

où $f^{x'}(s') = \overline{\varepsilon(x')\varepsilon(x'+s')^*}$ est la fonction de covariance relative au point x' de la matrice $\boldsymbol{B}$ complète initiale, d'où l'expression résultante pour $f_d^x(s)$:

$$f_d^x(s) = \sum_{x',s'} f^{x'}(s') \, (\sum_m \tilde{\phi}_m^*(x')\tilde{\phi}_m(x'+s')\phi_m(x)\phi_m^*(x+s)),$$

qui peut également se mettre sous la forme

$$f_d^x(s) = \sum_{x',s'} f^{x'}(s')\Phi^{x,s}(x',s'),$$

avec

$$\Phi^{x,s}(x',s') = \sum_m \tilde{\phi}_m^*(x')\tilde{\phi}_m(x'+s')\phi_m(x)\phi_m^*(x+s).$$

La fonction de covariance résultante $f_d^x(s)$ peut donc s'interpréter comme une moyenne spatiale pondérée de la fonction de covariance locale initiale $f^{x'}(s')$, où les poids $\Phi^{x,s}(x',s')$, relatifs aux paires $\{x,s\}$, sont des fonctions de la position x' et de la séparation s'.

Approche diagonale dans la base de Fourier

Soit N_g le nombre de points de grille sur le cercle, et $e_m(x) = exp(im\frac{2\pi x}{N_g})$. Alors, la famille

$$\left\{ \phi_m = \frac{1}{\sqrt{N_g}} e_m, m \in [0, N_g - 1] \right\}$$

est une frame dont la duale est simplement $\tilde{\phi}_m = \phi_m$. Dans ce cas, les coefficients de pondération sont donnés par

$$\begin{aligned}
\Phi^{x,s}(x',s') &= \frac{1}{N_g^2} \sum_m e_m^*(x') e_m(x'+s') e_m(x) e_m^*(x+s) \\
&= \frac{1}{N_g^2} \sum_m e_m(s'-s) \\
&= \frac{1}{N_g} \delta(s'-s),
\end{aligned}$$

avec $\delta(s'-s) = 1$ quand $s' = s$ et $\delta(s'-s) = 0$ ailleurs. Ceci signifie qu'un poids nul est donné à toutes les valeurs $f^{x'}(s')$ telles que $s' \neq s$ et qu'un poids uniforme $\frac{1}{N_g}$ est donné à toutes les valeurs $f^{x'}(s')$ telles que $s' = s$, quelle que soit la position de x'. En d'autres termes, chaque fonction de covariance résultante f_d^x correspond en fait à la moyenne spatiale uniforme (sur tout le domaine) des fonctions de covariance $f^{x'}$.

Approche diagonale dans l'espace des ondelettes

Dans le cas des ondelettes, $\mathcal{M}$ est un ensemble d'indices $m = (j, \{x_j(i), i = 1, N_x(j)\})$, où j désigne une échelle et $x_j(i)$ une position sur une grille associée à j. Ceci définit à nouveau une frame telle que, $\phi_m(x) = \tilde{\phi}_m(x) = \psi_j(x - x_j(i))$, où les fonctions ψ_j sont les ondelettes de Fisher. Dans ce cas, les coefficients de pondération $\Phi^{(x,s)}(x',s')$ sont donnés par

$$\sum_j \sum_{i=1}^{N_x(j)} \psi_j(x' - x_j(i)) \; \psi_j(x' - x_j(i) + s') \; \psi_j(x - x_j(i)) \; \psi_j(x - x_j(i) + s).$$

À la différence du cas spectral, les fonctions poids résultantes $\Phi^{(x,s)}(x', s')$ varient avec la position x'. Comme on pouvait s'y attendre, les tests numériques indiquent que les poids $\Phi^{(x,s)}(x', s')$ tendent à être maximum pour les positions x' proches de x et pour des valeurs de séparations s' proches de s.

En d'autres termes, les fonctions de covariance résultantes f_d^x peuvent être vues comme des moyennes spatiales locales des fonctions de covariance $f^{x'}$.

4.6.2 Annexe B : Construction de $B_w^{-1/2}$ et de $B_w^{1/2}$

La formulation de $\boldsymbol{B}_w^{1/2}$ est donnée par la construction de $\boldsymbol{B}_w^{-1/2}$. Cette dernière matrice est conçue comme un opérateur qui transforme l'erreur d'ébauche $\boldsymbol{\varepsilon}$ en une nouvelle variable , dont la matrice de covariance est proche de la matrice identité (voir *e.g.* Deckmyn et Berre, 2005 ; ou Gustafsson *et al.* , 2001) :

$$\boldsymbol{B}_w^{-1/2} = \boldsymbol{D}_w^{-1/2} \boldsymbol{W} \boldsymbol{\Sigma}_s^{-1} \boldsymbol{\Sigma}_g^{-1},$$

où $\boldsymbol{\Sigma}_g$ est la matrice diagonale des écarts types en points de grille de $\boldsymbol{\varepsilon}$, $\boldsymbol{\Sigma}_s^{-1} = \boldsymbol{S}^{-1} \boldsymbol{D}_s^{-1/2} \boldsymbol{S}$ correspond à une normalisation par les écarts types spectraux de $\check{\boldsymbol{\varepsilon}}' = \boldsymbol{S} \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\varepsilon}$ ($\boldsymbol{S}$ désignant ici la transformée spectrale, et $\boldsymbol{D}_s$ est la matrice diagonale correspondant aux variances dans l'espace spectral), et $\boldsymbol{D}_w^{1/2}$ est la matrice diagonale constituée des écarts types ondelettes de $\hat{\boldsymbol{\varepsilon}}'' = \boldsymbol{W} \check{\boldsymbol{\varepsilon}}'$. Les matrices $\boldsymbol{W}$ et $\boldsymbol{W}^{-1}$ correspondent respectivement aux transformations ondelettes directe et inverse.

Ainsi, l'expression de $\boldsymbol{B}_w^{1/2}$ est alors donnée par $\boldsymbol{B}_w^{1/2} = (\boldsymbol{B}_w^{-1/2})^{-1} = \boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_s \boldsymbol{W}^{-1} \boldsymbol{D}_w^{1/2}$.

4.6.3 Annexe C : Illustration dans l'espace des ondelettes des propriétés de filtrage

Comme mentionné aux sections 4.2 et 4.3.3, les propriétés de filtrage des ondelettes sont attendues du fait que l'approche diagonale ondelette revient à moyenner localement les fonctions de covariance. Ceci peut s'interpréter comme une vision point de grille de la propriété de filtrage, dans le sens où les fonctions de covariance point de grille sont moyennées dans l'espace des points de grille. Dans cette section, deux visions complémentaires, pour le filtrage ondelette, sont discutées : il s'agit du point de vue ondelette et spectral.

La matrice de covariance dans l'espace des ondelettes représente les covariances entre différentes échelles et pour différentes localisations géographiques. Un "point de vue ondelette" des propriétés de filtrage en jeu est donné par le fait que l'approche diagonale ondelette implique la mise à zéro des termes non diagonaux. Ces termes correspondent aux corrélations entre les modes ondelettes à différentes positions (*e.g.* pour une échelle j donnée). L'examen formel de l'équation associée (non détaillée ici) suggère le résultat suivant : négliger les corrélations ondelettes hors-diagonale évite aux petites échelles distantes de contribuer (de manière erronée) à la fonction de covariance locale (pour une position de référence donnée).

Cette interprétation est validée expérimentalement, comme l'illustre la figure 4.13. Cette figure représente l'amplitude (en valeur absolue) des coefficients ondelettes $\boldsymbol{W} f^0$ de la fonction de covariance relative à la position géographique $0\,°$, notée f^0. Dans le cas ensembliste avec termes non-diagonaux (graphique en haut à droite), l'amplitude des petites échelles même éloignées (*e.g.* pour $j \geq 10$ proche de $180\,°$) peut être anormalement grande (ceci étant dû au bruit d'échantillonnage), tandis que la valeur exacte (graphique en haut à gauche) présente des amplitudes proches de zéro. A l'inverse, dans le cas de l'approche diagonale ondelette, ces valeurs sont effectivement proches de zéro, et ce, même avec un ensemble de petite taille ($N_e = 10$). Ce

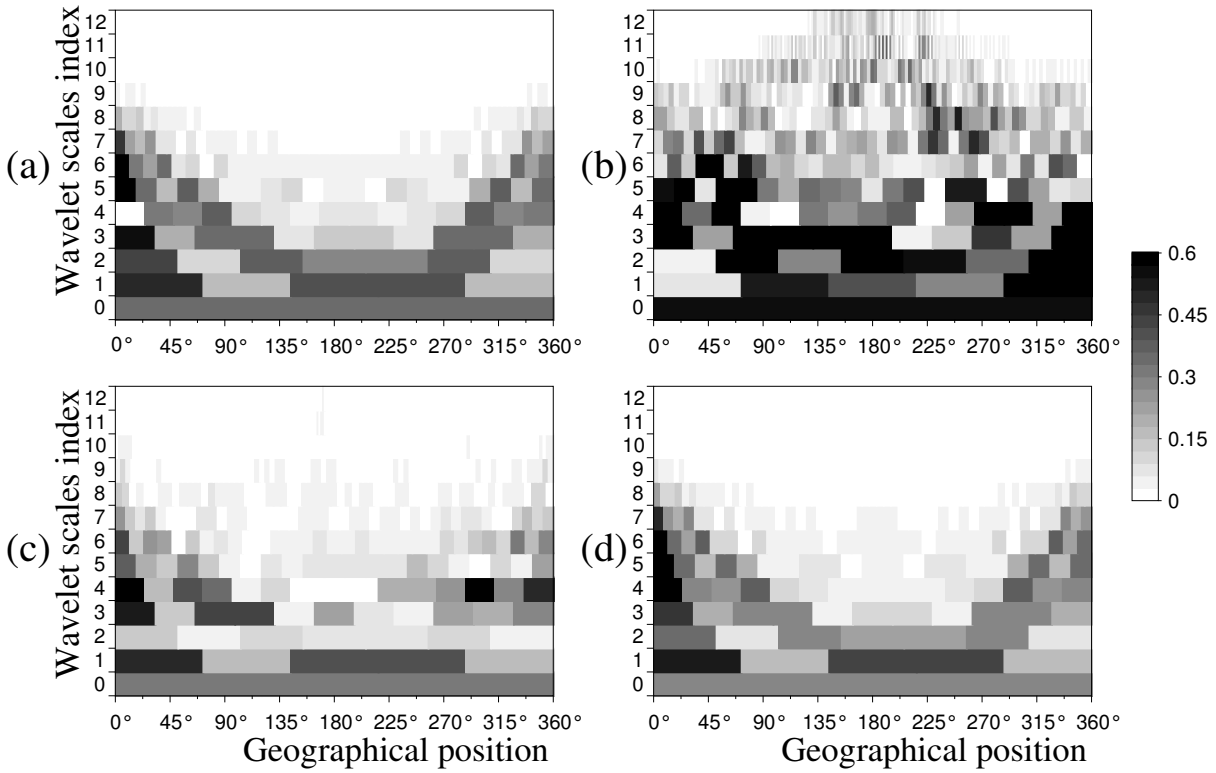


FIG. 4.13 – Amplitude des coefficients ondelettes pour la fonction de covariance relative à la position géographique $0\,°$. (a) transformée ondelette de cette fonction de covariance, (b) approche ensembliste estimée à partir de 10 membres et pour la matrice de covariance non filtrée, (c) approche diagonale ondelette estimée à partir de 10 membres, et (d) approche diagonale ondelette pour un ensemble de taille infinie. L'indice des échelles pour la formulation ondelette est noté j : les grandes valeurs de j correspondent aux petites échelles ondelettes.

résultat est en accord avec la mise à zéro implicite des corrélations entre les modes ondelettes positionnés au voisinage de $0\,^\circ$ et ceux positionnés au voisinage de $180\,^\circ$ par exemple.

Finalement, il est également intéressant de mentionner qu'un "point de vue spectral" des propriétés de filtrage des ondelettes peut également être considéré. Il n'est pas détaillé ici par souci de concision.De façon résumée, il s'agit de noter que la mise à zéro des corrélations croisées entre des échelles ondelettes différentes revient à annuler des contributions de petite échelle (et bruitées) aux variations géographiques des covariances. Ceci est semblable à l'effet attendu d'une moyenne spatiale locale en point de grille.

Chapitre 5

Estimation de la longueur de portée des fonctions de corrélation d'erreur d'ébauche et de leurs statistiques d'échantillonnage

Traduction d'un article publié : **Pannekoucke O., Berre L. and Desroziers G., 2008.** ***Background error correlation length-scale estimates and their sampling statistics*****, QJRMS, 134, 497–508. .**

Situation du papier par rapport à la problématique du filtrage ondelette du bruit d'échantillonnage

Ce chapitre porte sur les propriétés statistiques de l'estimation des portées de corrélation. Vis-à-vis de la problématique du filtrage en ondelette du bruit d'échantillonnage, la partie la plus importante de ce chapitre correspond à la section 5.4. En particulier, la section 5.4.4 permet de mettre en évidence, dans un cadre académique, le fait que la structure spatiale du bruit d'échantillonnage est plutôt de petite échelle (et ce même si le signal étudié est de grande échelle). Ce résultat conforte ainsi la pertinence de la moyenne spatiale réalisée par les ondelettes, dont l'efficacité est illustrée dans la section 5.4.5.

Les lecteurs plus particulièrement intéressés par les ondelettes pourront donc se concentrer sur ces sections. Les autres parties de ce chapitre ne portent pas spécifiquement sur les ondelettes, mais elles permettent d'approfondir des aspects statistiques de l'estimation des portées (présentation et évaluation de différentes formules de portées, amplitudes respectives du biais et de l'écart type de l'estimation de la portée à partir d'un ensemble).

RÉSUMÉ

Ce chapitre présente différentes formules pour estimer les longueurs de portée des fonctions de corrélation, ainsi qu'une évaluation de leur performance dans les applications pratiques. En particulier, deux nouvelles formules, simples d'utilisation, sont introduites. Elles ne nécessitent que le calcul de la corrélation en un point pour une direction donnée. Un cadre analytique 1D permet de montrer que ces formules conduisent toutes à des valeurs de portée réalistes, et qu'elles sont bien capables de représenter les variations géographiques de ce paramètre.

L'estimation des longueurs de portée à partir d'un ensemble fini est également abordée. Tandis qu'un biais positif apparaît dans l'estimation à partir d'un ensemble de petite taille, il est montré que l'écart type joue un rôle

prépondérant dans l'erreur d'estimation. La structure spatiale de l'erreur d'échantillonnage est diagnostiquée, et les effets de techniques de filtrage spatial sur le biais et l'écart type sont illustrés.

Pour finir, un ensemble de prévisions perturbées Arpège est utilisé, permettant de montrer les applications du calcul de la portée dans un cas réel.

MOT CLÉS: Assimilation de données, diagnostic, approximation de la longueur de portée, ensemble d'assimilations, bruit d'échantillonnage.

5.1 Introduction

Afin de déterminer de manière optimale une condition initiale pour les modèles de prévision numérique, les schémas d'assimilation modernes sont basés sur une combinaison optimale des observations et d'une ébauche. Cette ébauche correspond généralement à la dernière prévision disponible à courte échéance. Cette analyse dérive des méthodes de la théorie de l'estimation optimale. Dans ce cadre, les deux sources d'information sont mises en relation à l'aide de leur matrice de covariance d'erreur respective. Ces matrices de covariance d'erreur déterminent le poids à apporter à chaque information contribuant à l'analyse. Cependant, la spécification de ces statistiques reste un problème majeur pour les systèmes d'assimilation en opérationnel.

La connaissance de la forme de la fonction de corrélation est particulièrement importante. En effet, la fonction de corrélation joue un rôle majeur dans le filtrage et la propagation spatiale de l'information observée (Daley, 1991).

C'est pourquoi le diagnostic des longueurs de portée des corrélations d'erreur d'ébauche est souvent utilisé. Cette longueur permet de décrire la forme de la fonction de corrélation. Il est possible de définir une échelle caractéristique de différentes façons (toute l'information sur la fonction de structure est alors condensée en une simple information d'échelle). La définition adoptée classiquement en assimilation est celle donnée par Daley (1991). Elle est basée sur la courbure de la fonction de corrélation à l'origine. Typiquement, plus la longueur de portée est courte, plus la décroissance de la corrélation en fonction de la distance est rapide.

Comme le montrent un certain nombre d'auteurs (Hollingsworth, 1987 ; Bouttier, 1993 ; Rabier *et al.* , 1998 ; Ingleby, 2001 ; Belo Pereira et Berre, 2006 ; Deckmyn et Berre, 2005), ce diagnostic de portée renseigne également sur la manière dont la dynamique de l'atmosphère et la densité du réseau d'observation agissent sur la structure spatiale de l'erreur. Ainsi, il est intéressant de pouvoir diagnostiquer et interpréter ces échelles en différentes positions géographiques. Par ailleurs ce diagnostic local est en liaison directe avec les efforts de recherche dans le domaine de la représentation de l'hétérogénéité et de l'anisotropie locale (*e.g.* Fisher, 2003 ; Buehner, 2005).

De plus, il sera montré ici que les longueurs de portée locale peuvent être approximées par différentes formules. Un premier objectif de ce chapitre est donc d'évaluer la capacité de ces différentes approches à diagnostiquer les variations géographiques des portées locales.

En outre, avec un ensemble de prévisions, il est possible de calculer les covariances d'erreur d'ébauche "du jour". Cependant, les ensembles disponibles étant de taille finie, l'estimation est entachée d'un bruit d'échantillonnage, qui détériore l'estimation des covariances (d'autant plus que la taille de l'ensemble est petite).

En ce qui concerne les corrélations, la question du bruit d'échantillonnage a été généralement étudiée vis-à-vis des valeurs de corrélation à longue distance, identifiées comme étant particulièrement bruitées (Houtekamer et Mitchell, 2001). Cependant, relativement peu d'études se sont intéressées à la représentation locale des portées. Ainsi, un deuxième objectif du chapitre est l'étude du bruit d'échantillonnage affectant les longueurs de portée.

La structure du chapitre est la suivante. Dans la section 5.2, différentes formules pour le

calcul des portées sont déduites de la définition de Daley. Les résultats expérimentaux sont présentés en section 5.3, dans un cadre analytique 1D. La section 5.4 montre la sensibilité de l'estimation de portée à la taille de l'ensemble, ainsi que la structure spatiale du bruit d'échantillonnage. La section 5.5 présente la comparaison du calcul de portée issue de deux formules différentes, dans le cas de la sphère, et en utilisant un ensemble de prévisions globales perturbées Arpège.

5.2 Formules pour la longueur de portée

Un des problèmes majeurs des schémas d'assimilation de données opérationnels est de mieux ajuster la matrice de covariance d'erreur d'ébauche $\boldsymbol{B} = \mathbb{E}(\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_b^T)$, où $\boldsymbol{\varepsilon}_b$ est une erreur de prévision supposée sans biais. Dans l'objectif de mieux connaître la forme de la fonction de corrélation au voisinage de l'origine, le diagnostic de longueur est souvent introduit.

La longueur de portée est formellement définie, en assimilation de données, d'après Daley (1991). Sa définition est similaire à celle de la micro-échelle turbulente de Taylor. Dans cette section, un rappel de la définition de Daley est complétée par des formules qui s'en déduisent.

5.2.1 Formule de Daley

Pour une fonction de corrélation ρ lisse et isotrope à l'origine, la longueur de portée de Daley est donnée par

$$L_D = \sqrt{-\frac{1}{\Delta\rho(0)}}, \tag{5.1}$$

en 1D et $L_D = \sqrt{-\frac{2}{\Delta\rho(0)}}$ en 2D. Cette longueur est proportionnelle à la micro-échelle turbulente de Taylor qui est définie de manière similaire. Cette formule est obtenue à partir de la décomposition de Taylor de la fonction de corrélation à l'origine :

$$\rho(\delta x) \approx \rho(0) + \frac{\delta x^2}{2}\frac{d^2\rho}{dx^2}(0) = 1 - \frac{\delta x^2}{2L_D^2}. \tag{5.2}$$

L'hypothèse d'isotropie est nécessaire afin d'assurer la continuité de la dérivée seconde en $x = 0$, *i.e.* $\frac{d^2\rho}{dx^2}(0^-) = \frac{d^2\rho}{dx^2}(0^+)$. Une interprétation géométrique de cette longueur est qu'elle correspond à la longueur pour laquelle la parabole osculatrice à l'origine vaut 0.5. Une illustration graphique de cette définition est donnée sur la figure 5.1, où une fonction de corrélation (courbe continue) et sa parabole osculatrice à l'origine (courbe pointillée) sont représentées. La longueur de portée déduite de cette interprétation géométrique est $L_D = 250\,km$, pour cette fonction de corrélation particulière.

De plus la longueur de portée est également reliée à la courbure de la fonction de corrélation à l'origine. En effet, le rayon de courbure de la fonction de corrélation à la distance r est défini par $R(r) = \frac{\left(1+\left(\frac{d\rho}{dx}(r)\right)^2\right)^{3/2}}{\frac{d^2\rho}{dx^2}(r)}$. Ainsi, à l'origine, $\frac{d\rho}{dx}(0) = 0$ entraîne $R(0) = \frac{1}{\frac{d^2\rho}{dx^2}(0)} = -L_D^2$.

Il est à noter que la portée de Daley ne renseigne pas sur l'anisotropie de la fonction de corrélation. De plus, elle nécessite la connaissance de la dérivée seconde de la fonction de corrélation à l'origine. Le calcul de cette dérivée seconde est coûteux, car il nécessite idéalement la connaissance de la fonction de corrélation dans sa totalité. Dans les paragraphes suivants des approximations de cette formule sont décrites.

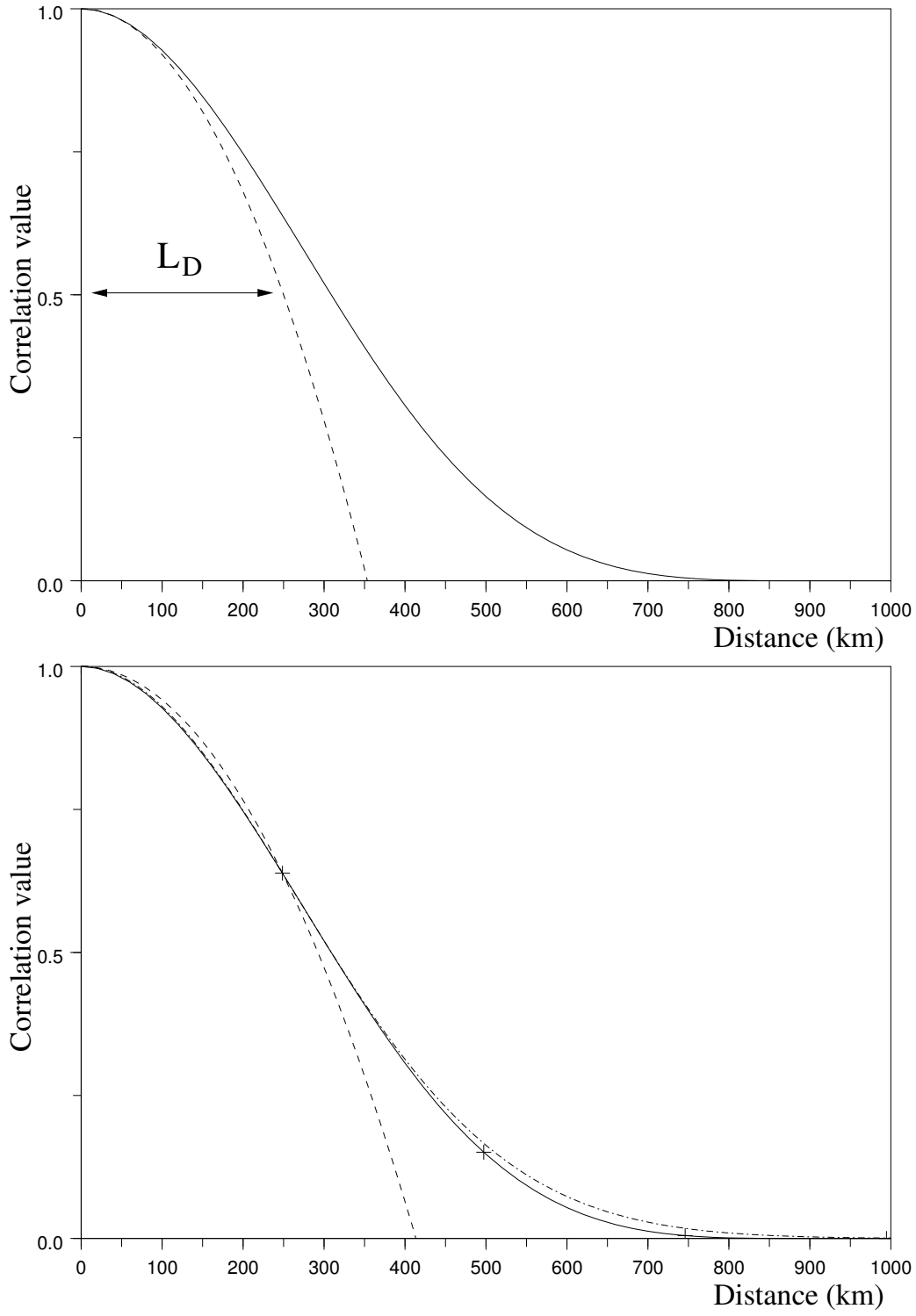


FIG. 5.1 – En haut : fonction de corrélation (courbe continue) de Gaspari et Cohn (voir la section 5.3.1) et sa parabole osculatrice à l'origine (courbe tiretée). En bas : approximations à l'origine de la fonction de corrélation de type parabolique (courbe continue) et gaussienne (courbe tiretée-pointilliée), à partir de la valeur de la corrélation à la distance $\delta x = 248km$, sur une grille régulière (croix).

5.2.2 Formule de Belo Pereira-Berre

Belo Pereira et Berre (2006) (noté dans la suite B&B) ont proposé une formule relativement peu coûteuse pour le calcul de la portée. Cette formule est basée sur l'approximation de l'écart type de la dérivée de l'erreur donnée par $\sigma^2(\partial_x \varepsilon_b(x)) = (\partial_x \sigma(\varepsilon_b(x)))^2 - \sigma^2(\varepsilon_b(x))\partial_x^2\rho(0)$, où $\partial_x = \frac{\partial}{\partial x}$ désigne la dérivation le long de la coordonnée. En revenant à la définition de la portée de Daley, il vient que

$$L_{B\&B} = \sqrt{\frac{\sigma(\varepsilon_b(x))^2}{\sigma(\partial_x \varepsilon_b(x))^2 - (\partial_x \sigma(\varepsilon_b(x)))^2}}, \tag{5.3}$$

avec $\sigma\left(\varepsilon_b(x)\right)$ l'écart type de $\varepsilon_b(x)$. Cette définition nécessite le calcul de l'écart type de l'erreur, son gradient, mais également le calcul de l'écart type du gradient. Dans le cas d'un domaine périodique (cercle, tore ou sphère), le calcul du gradient peut être effectué soit en points de grille, soit en passant par l'espace spectral associé.

5.2.3 Formules basées sur l'approximation par une parabole et par une gaussienne

Comme le suggère l'équation (5.2), une discrétisation du laplacien apparaissant dans l'équation (5.1) conduit à une expression simple de la portée

$$L_{Pb} = \frac{|\delta x|}{\sqrt{2\left(1 - \rho(\delta x)\right)}}. \tag{5.4}$$

Cette longueur de portée est appelée, dans la suite, *longueur de portée basée sur la parabole* ou *Pb*. Elle est basée sur l'approximation de la fonction de corrélation par une fonction parabolique, comme le représente la figure 5.1. L'exemple présenté sur cette figure suggère que pour une distance (assez grande) de séparation, la parabole peut décroître moins rapidement (entre l'origine et la distance δx choisie) que la vraie fonction de corrélation. Ceci suggère que la qualité de l'approximation de la portée basée sur la parabole peut dépendre de la qualité de l'approximation de la fonction de corrélation et de la distance de séparation considérée δx.

Afin d'étudier la sensibilité de l'approximation de la forme de la corrélation, il est intéressant de considérer un autre modèle analytique pour la fonction de corrélation ρ à proximité de l'origine. En approximant la corrélation à l'origine par une fonction gaussienne, on obtient : $\rho(\delta x) = exp(-\frac{\delta x^2}{2\,L_D^2})$. En inversant cette équation, l'expression de la portée, associée à la valeur de la corrélation à la distance δx, donne

$$L_{Gb} = \frac{|\delta x|}{\sqrt{-2\ \ln\ \rho(\delta x)}}. \tag{5.5}$$

Cette longueur est appelée, dans la suite, *longueur de portée basée sur l'approximation gaussienne* ou *Gb*. L'approximation du calcul de la portée est simple à implémenter dans un contexte réel et demeure peu coûteuse. Le graphique du haut de la figure 5.1 illustre l'approximation gaussienne à l'origine pour la fonction de corrélation discrétisée.

Il est à noter que quand la fonction de corrélation est proche de 1, les deux approximations de la portée basées sur la parabole et la gaussienne sont égales. En effet, avec $\eta = 1 - \rho$, le développement de Taylor est donné par $L_{Pb} = L_{Gb} = \frac{\delta x}{\sqrt{2\eta}}$.

5.2.4 Longueur de portée directionnelle

Les formules (5.4) et (5.5) peuvent être définies le long d'une direction orientée, de la manière suivante. Soit $\boldsymbol{\delta x}$ un déplacement dans une direction orientée par le vecteur unitaire $\boldsymbol{u} = \frac{\boldsymbol{\delta x}}{|\boldsymbol{\delta x}|}$ sur le domaine (cercle, plan, sphère 2D, sphère 2D + verticale, *etc*). Alors, la longueur de portée vectorielle basée sur la parabole est définie par

$$L_{Pb,\boldsymbol{u}} = \frac{|\boldsymbol{\delta x}|}{\sqrt{2\,(1-\rho(\boldsymbol{\delta x}))}}\,\boldsymbol{u}, \tag{5.6}$$

et la longueur de portée vectorielle basée sur la gaussienne est

$$L_{Gb,\boldsymbol{u}} = \frac{|\boldsymbol{\delta x}|}{\sqrt{-2\ \ln\,\rho(\boldsymbol{\delta x})}}\,\boldsymbol{u}, \tag{5.7}$$

avec $|\boldsymbol{\delta x}|$ l'amplitude du déplacement. Ainsi, ce diagnostic offre une caractérisation de la forme de la fonction de corrélation dans les différentes directions de l'espace. De manière similaire, les formules (5.1) et (5.3) peuvent être définies de manière directionnelle pour une fonction de corrélation anisotrope. Pour l'équation (5.1), il suffit de remplacer $\Delta\rho(0)$ par $\frac{\partial^2\rho}{\partial\boldsymbol{u}^2}(0^+) = \lim_{t\to 0^+} 2\,\{\rho(t\,\boldsymbol{u}) - 1\}\,t^{-2}$, qui correspond à la dérivée seconde, calculée suivant le vecteur $\boldsymbol{u}$, de la fonction de corrélation anisotrope. Pour l'équation (5.3), la portée directionnelle est obtenue en calculant le gradient $\partial_{\boldsymbol{u}}\varepsilon_b$ et $\partial_{\boldsymbol{u}}\sigma(\varepsilon_b)$, où $\partial_{\boldsymbol{u}}$ désigne la dérivation suivant le vecteur $\boldsymbol{u}$.

Il doit être mentionné que ces longueurs de portée peuvent être calculées pour différents types de domaine, qu'ils soient bornés ou non. Ainsi, de telles formulations sont adaptées dans les applications en océanographie ou dans les modèles à aire limitée, aussi bien que pour les modèles globaux de circulation générale.

Dans le cas particulier d'un domaine 1D et pour le cas de la portée basée sur une parabole, la portée directionnelle se traduit par une portée à gauche, notée $L_{Pb}(-\delta x)$, et une portée à droite, notée $L_{Pb}(+\delta x)$. Naturellement, une définition similaire est valable pour la portée basée sur la gaussienne. Dans la suite, la portée à gauche est désignée à l'aide de l'indice $^-$ et la portée à droite par l'indice $^+$. À noter que le ratio $\frac{L^+}{L^-}$ est un indicateur de l'anisotropie : l'anisotropie est caractérisée par une valeur différente de 1. Si le ratio est supérieur à 1 alors la fonction de corrélation accuse une asymétrie à droite.

5.2.5 Autres formules approximant la portée

Cette manière de construire des approximations pour la portée est très générale. Ainsi, il existe d'autres définitions qui sont formellement déduites de relations du type $\rho(\delta x) = f(\delta x, L_D)$. La contrainte principale est que ces formules soient effectivement inversibles par rapport à L_D. Il en résulte que la portée est alors déduite de la valeur de corrélation, associée à une grille, d'après $L_D = f^{-1}\,(\delta x, \rho(\delta x))$. Le choix d'une telle relation entre la corrélation et la portée peut être recherché d'après la forme de la fonction de corrélation, permettant ainsi d'obtenir une meilleure approximation de la portée. Ainsi, ce choix peut dépendre du champ physique considéré ou de la formulation utilisée pour modéliser les corrélations dans le schéma d'assimilation.

De plus, il est à noter que de tels développements sont applicables pour le diagnostic de fonctions de structure de forme spatiale plus complexe en 1D, 2D et 3D.

Dans la suite, le cas du cercle 1D et de la sphère 2D sont considérés pour illustrer la théorie.

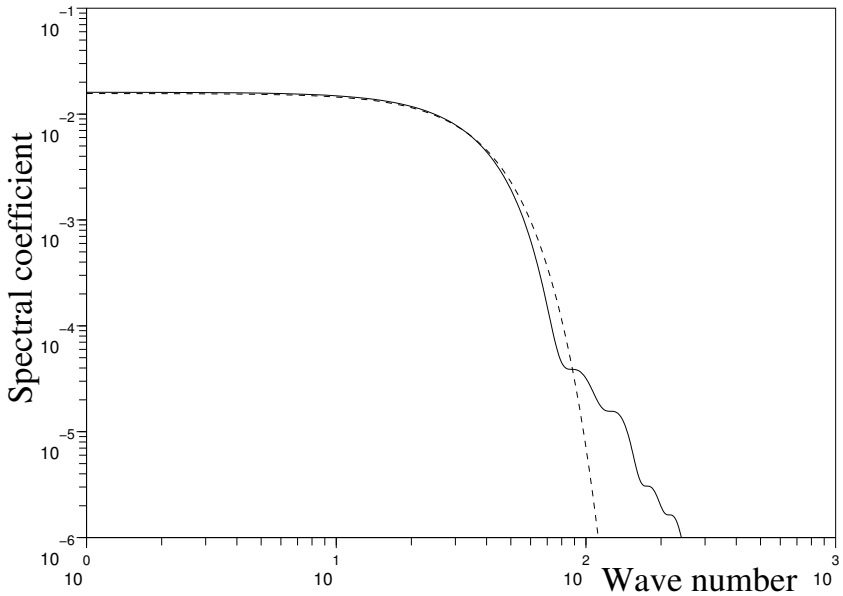


FIG. 5.2 – Spectre des corrélations de Gaspari et Cohn (courbe continue) et gaussienne (courbe tiretée).

5.3 Application dans un contexte 1D hétérogène

5.3.1 Un cadre 1D simple

Ce cadre analytique est celui décrit au chapitre 4. Il est rapidement rappelé pour fixer les notations. Cet exemple analytique permet d'évaluer la qualité des différentes formulations de la portée explorées dans ce chapitre.

Dans ce contexte, le domaine est un cercle équatorial de rayon a, et de coordonnée $\frac{x}{a}$ correspondant à un angle variant de $0\,^\circ$ à $360\,^\circ$. A nouveau, un seul champ est considéré sur ce domaine. Un tenseur de corrélation homogène gaussien est donné par $B_h(x,y) = e^{-\frac{(x-y)^2}{2L_H^2}}$, où x et y sont deux points sur le cercle, et L_H est la longueur de portée, qui est ici arbitrairement fixée à $L_H = 250km$. De plus, une corrélation homogène non gaussienne est également considérée. Elle correspond à la corrélation introduite par Gaspari and Cohn (1999 Eq 4.10) $C_h(x,y) = \rho_L(x-y)$ avec

$$\rho_L(r) = \begin{cases} -\frac{1}{4}\left(\frac{r}{L}\right)^5 + \frac{1}{2}\left(\frac{r}{L}\right)^4 + \frac{5}{8}\left(\frac{r}{L}\right)^3 - \frac{5}{3}\left(\frac{r}{L}\right)^2 + 1 & ,0 \leq r \leq L, \\ \frac{1}{12}\left(\frac{r}{L}\right)^5 - \frac{1}{2}\left(\frac{r}{L}\right)^4 + \frac{5}{8}\left(\frac{r}{L}\right)^3 + \frac{5}{3}\left(\frac{r}{L}\right)^2 - 5\left(\frac{r}{L}\right) + 4 - \frac{2}{3}\left(\frac{L}{r}\right) & ,L \leq r \leq 2L, \\ 0 & ,2L \leq r, \end{cases}$$

et $L = \sqrt{0.3}L_H$ de sorte que la portée théorique est la même que pour le cas gaussien. Les spectres de ces deux corrélations sont représentés en figure 5.2.

Une corrélation hétérogène est calculée à partir d'un étirement par la transformation de Schmidt (Courtier et Geleyn 1988), adaptée au cas du cercle et définie par

$$h(x) = a\left[\pi - 2Atan\left(\frac{1}{c}tan(\frac{\pi}{2} - \frac{1}{2}\frac{x}{a})\right)\right],$$

avec $c = 2.4$. Cette transformation est inversible d'inverse h^{-1}. Ainsi, la formulation hétérogène est donnée par le tenseur

$$B(x,y) = B_h\left(h^{-1}(x), h^{-1}(y)\right), \tag{5.8}$$

dont les fonctions de corrélation sont relativement étroites au voisinage de $180\,^\circ$, et étalées au voisinage de $0\,^\circ$.

Les versions discrétisées de ces tenseurs de corrélation sur une grille correspondent aux matrices de covariance et dépendent de la résolution de la grille. Pour une troncature T donnée, le nombre de points de grille est $N_g = 2T + 1$ et la résolution homogène associée est $\delta x = \frac{2\pi a}{N_g}$. Dans la suite, les exemples sont illustrés pour $T = 120$.

5.3.2 Calcul de la portée dans un cas hétérogène

Pour cette expérience numérique, le terme $\sigma(\partial_x \varepsilon_b)^2$ apparaissant dans la portée de B&B est calculé de la manière suivante. Pour un point d'indice i, ce terme s'exprime ainsi :

$$\sigma(\partial_x \varepsilon_b)_i^2 = \boldsymbol{\delta}_i^T \boldsymbol{D}\boldsymbol{B}\boldsymbol{D}^*\boldsymbol{\delta}_i,$$

où $\boldsymbol{D}$ désigne l'opérateur de dérivation (l'adjoint de cet opérateur n'est autre que son opposé : $\boldsymbol{D}^* = -\boldsymbol{D}$) construit dans l'espace de Fourier, et $\boldsymbol{\delta}_i$ le dirac dont la valeur est nulle sauf pour la position i où il vaut 1. De même, la portée de Daley est calculée directement en exprimant le laplacien à l'origine, ce laplacien étant lui aussi calculé en passant par l'espace de Fourier.

Dans le cadre 1D, les différentes portées sont représentées sur la figure 5.3 pour les corrélations hétérogènes basées sur le tenseur de corrélation gaussienne et celles basées sur celui de G&C. Les longueurs de portée basées sur la parabole et la gaussienne sont calculées en prenant la valeur moyenne de leurs portées à gauche et à droite $\frac{1}{2}\left(L^+ + L^-\right)$. La portée de Daley calculée numériquement est prise pour référence.

Dans la première expérience, représentée sur le graphique en haut sur la figure 5.3, le tenseur analytique utilisé est le tenseur hétérogène déduit du tenseur de corrélation gaussienne. Il apparaît que toutes les formules de portée sont capables de représenter les variations géographiques des fonctions de corrélation sur le domaine : des portées larges au voisinage de $0\,^\circ$, et courtes au voisinage de $180\,^\circ$. Les différences entre ces résultats restent faibles. Le plus gros écart est rencontré pour la portée basée sur la parabole.

Dans la deuxième expérience, représentée en bas de la figure 5.3, le tenseur analytique hétérogène utilisé correspond à la corrélation de G&C. À nouveau, les formules conduisent à des portées dont les valeurs sont similaires et réalistes.

Dans les deux expériences de la figure 5.3, les portées basées sur la parabole sont quelque peu moins précises que celles basées sur la gaussienne. Ceci indique que ce diagnostic de la portée est légèrement sensible à l'approximation sous-jacente des fonctions de corrélation (comme mentionné au paragraphe 5.2.3). Par exemple, dans la deuxième expérience, l'approximation gaussienne de la corrélation de G&C apparaît meilleure que l'approximation parabolique.

Finalement, il ressort de ces résultats que toutes les formulations conduisent à des valeurs de portée similaires et réalistes, et que les variations géographiques sont bien représentées.

5.4 Statistiques d'échantillonnage des portées

5.4.1 Influence de la taille de l'ensemble dans le cas 1D

En pratique, les portées sont estimées à partir d'un ensemble de taille finie (*e.g.* Belo Pereira et Berre 2006). La Figure 5.4 représente les effets de la distribution d'échantillonnage sur l'estimation de la portée L_D pour des ensembles de petite taille. Dans cette expérience, le tenseur de corrélation hétérogène est issu du tenseur de corrélation gaussienne sur le cercle. La portée

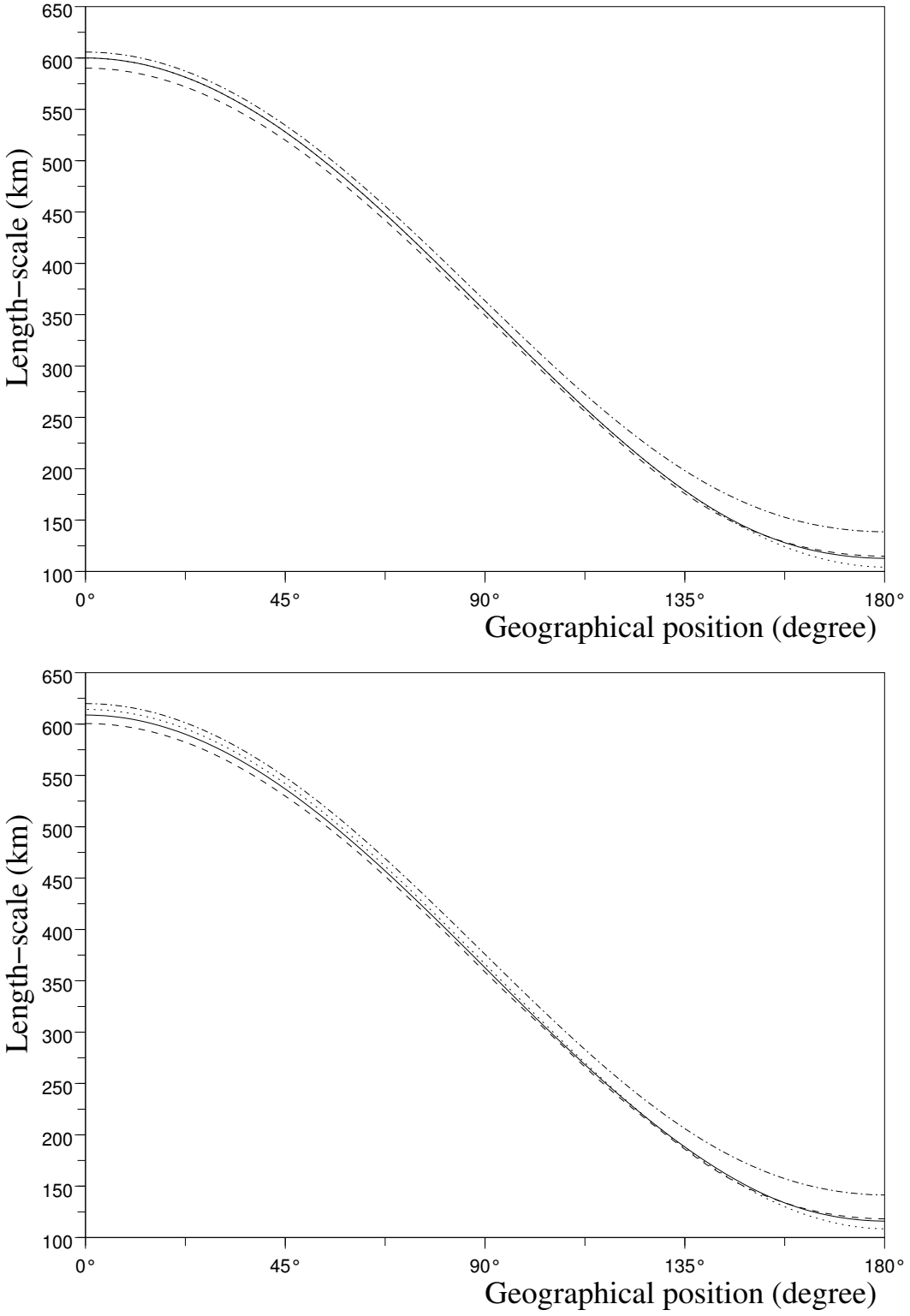


FIG. 5.3 – Longueur de portée en chaque point du domaine, calculée d'après différentes formules, et pour les tenseurs de corrélation hétérogènes issus des corrélations : gaussienne (graphique du haut) et G&C (graphique du bas). Le problème a été discrétisé sur le cercle à troncature $T120$. Daley (courbe continue), Belo Pereira-Berre (courbe tiretée), portée moyenne basée sur la parabole (courbe tiretée-pointilliée) et basée sur la gaussienne (courbe pointillée).

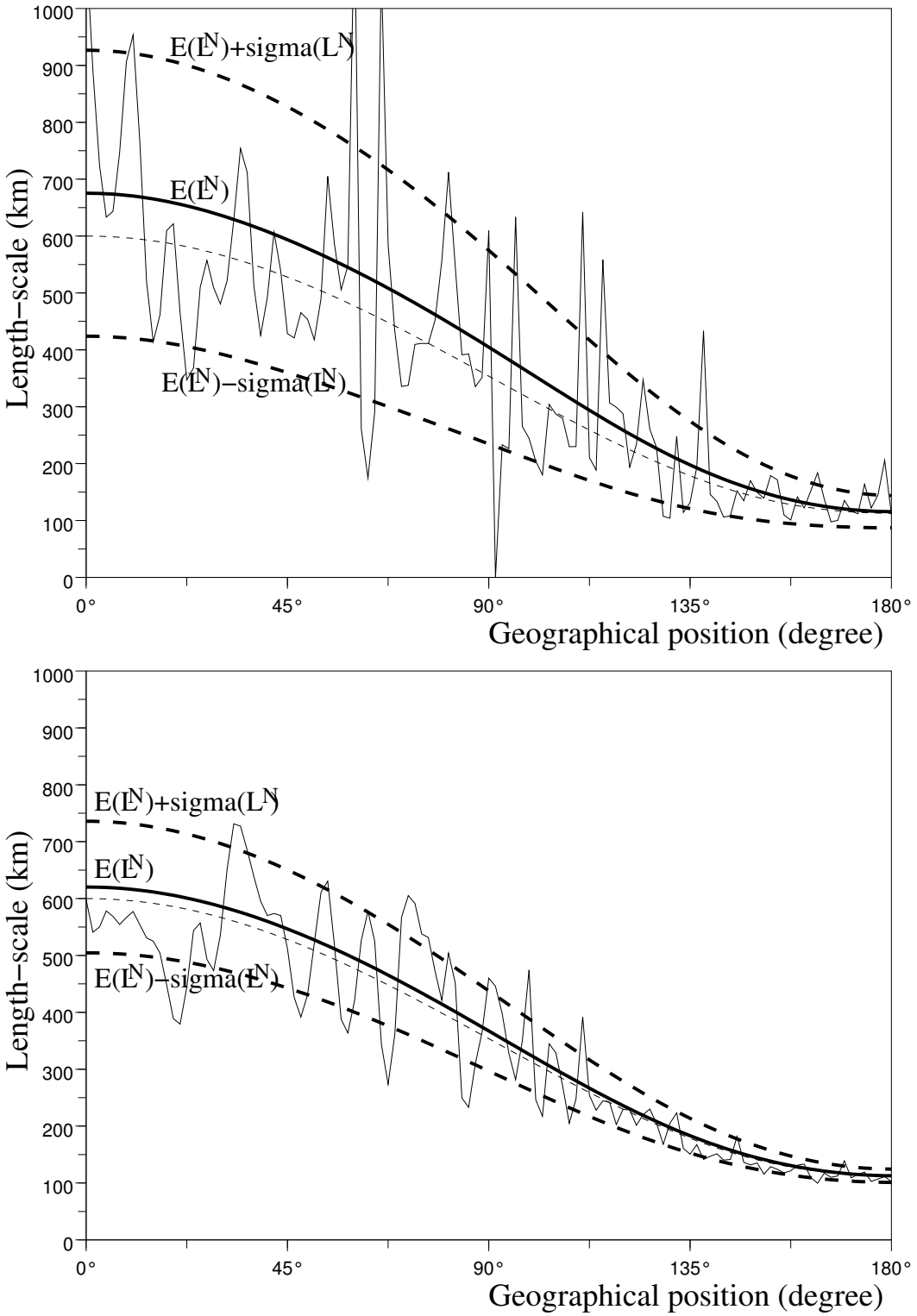


FIG. 5.4 – Sensibilité de l'estimation de la portée en fonction de la taille de l'ensemble. La carte des portées exactes (courbe tiretée en trait fin) est comparée aux portées estimées (courbe continue en trait fin) pour $N = 10$ (graphique du haut) et $N = 30$ (graphique du bas). L'espérance $\mathbb{E}(L^N)$ (courbe continue en gras) illustre l'existence d'un biais et la plage $\mathbb{E}(L^N) \pm \sigma(L^N)$ (courbe tiretée en gras) (avec $\sigma(L^N)$ l'écart type de L^N) représente la zone caractéristique de répartition des valeurs prises par la portée estimée.

vraie L_D (courbe tiretée) est comparée aux portées estimées (courbe continue en trait fin) à partir de 10 membres (graphique du haut) et 30 membres (graphique du bas). Les variations de portée sont bruitées par des oscillations à hautes fréquences.

Chaque échantillon de N membres conduit à une estimation particulière du champ de portée L^N, qui peut être considérée comme une série de variables aléatoires. Il est alors intéressant de déterminer l'espérance $\mathbb{E}(L^N)$ de ces variables aléatoires, ainsi que leurs autres caractéristiques statistiques (écart type $\sigma(L^N)$, distribution d'échantillonnage, *etc*)

L'espérance mathématique $\mathbb{E}$ est définie numériquement pour un champ aléatoire α comme $\mathbb{E}(\alpha) \approx \frac{1}{N_s}\sum_k \alpha_k$ où α_k sont N_s réalisations indépendantes de α. Par exemple, en désignant par L^N le champ aléatoire de portées estimé à partir de N membres, $\mathbb{E}(L^N) \approx \frac{1}{N_s}\sum_k L^N_k$, où L^N_k est la k^{ime} longueur de portée estimée à partir du k^{ime} échantillon de N membres. Dans la suite, N_s est pris comme étant très grand de manière à stabiliser les résultats.

La figure 5.4 montre que les longueurs de portée estimées sont biaisées puisque leur espérance $\mathbb{E}(L^N)$ ne correspond pas à la vérité. Pour 10 membres, $\mathbb{E}(L^{10}) \neq L$, avec un biais maximum au voisinage de $0\,^\circ$ et atteignant une valeur de $75km$ (12% d'erreur). De plus, l'écart type montre une dispersion importante des portées estimées, de l'ordre de 40% pour 10 membres (resp. 20% pour 30 membres).

Afin de mieux comprendre la manière dont la taille finie de l'ensemble influence l'estimation, la distribution d'échantillonnage peut être calculée expérimentalement. Dans le cas particulier des portées Pb et Gb, la distribution de l'échantillon peut également être déduite analytiquement, à partir de la distribution de la corrélation ρ^N entre deux points séparés d'une distance δx. Ceci est décrit en annexe.

5.4.2 Distribution d'échantillonnage pour la portée Gb

La distribution des fréquences expérimentales des portées calculées est représentée sur la figure 5.5 pour $N = 25$. Cette figure montre que la distribution d'échantillonnage est asymétrique avec un coefficient d'asymétrie ("skewness" en anglais) positif[1] et l'existence d'un biais $b^N_{Gb} = \mathbb{E}(L^N_{Gb}) - L^\infty_{Gb}$. Ces résultats expérimentaux sont en accord avec les études analytiques telles que montrées en annexe. Une asymétrie positive implique que des valeurs de portée éventuellement grandes sont fréquemment rencontrées pour de tels ensembles de petite taille.

La courbe continue sur la figure 5.6 représente le pourcentage d'erreur relative associé au biais $\frac{\mathbb{E}(L^N_{Gb})-L^\infty_{Gb}}{L^\infty_{Gb}}$ pour une discrétisation donnée δx. Dans ce cas, le cercle est discrétisé à la troncature $T120$ ($\delta x \approx 166km$) et $L^\infty_{Gb} = 250km$. Ainsi l'erreur est grande pour des ensembles de petite taille, et faible pour des ensembles de grande taille. La convergence en fonction de la taille de l'ensemble est relativement rapide et varie en $\mathcal{O}(N^{-1})$. Pour 10 membres, l'erreur relative associée au biais est de 10%.

Cependant, on montre maintenant que l'écart type a un effet majeur, par rapport à celui du biais, sur l'erreur d'estimation. La figure 5.6 montre le rapport $\frac{\sigma^N_{L_{Gb}}}{L^\infty_{Gb}}$ où

$$\sigma^N_{L_{Gb}} = \sqrt{\mathbb{E}\left\{\left(L^N_{Gb} - \mathbb{E}(L^N_{Gb})\right)^2\right\}}.$$

[1] On rappelle que le coeffcient d'asymétrie désigne un facteur de forme pour la distribution de probabilité d'une variable aléatoire X, défini par $skew = \frac{\mu_3}{\mu_2^{3/2}}$, où $\mu_p = \mathbb{E}\left\{[X - \mathbb{E}(x)]^p\right\}$ est le moment centré d'ordre p. Cette quantité caractérise l'écart entre la moyenne et le mode de la distribution, normalisé par la dispersion, soit formellement $skew \equiv \frac{moyenne-mode}{dispersion}$. Ainsi, l'asymétrie est positive quand la queue de distribution est étalée à droite.

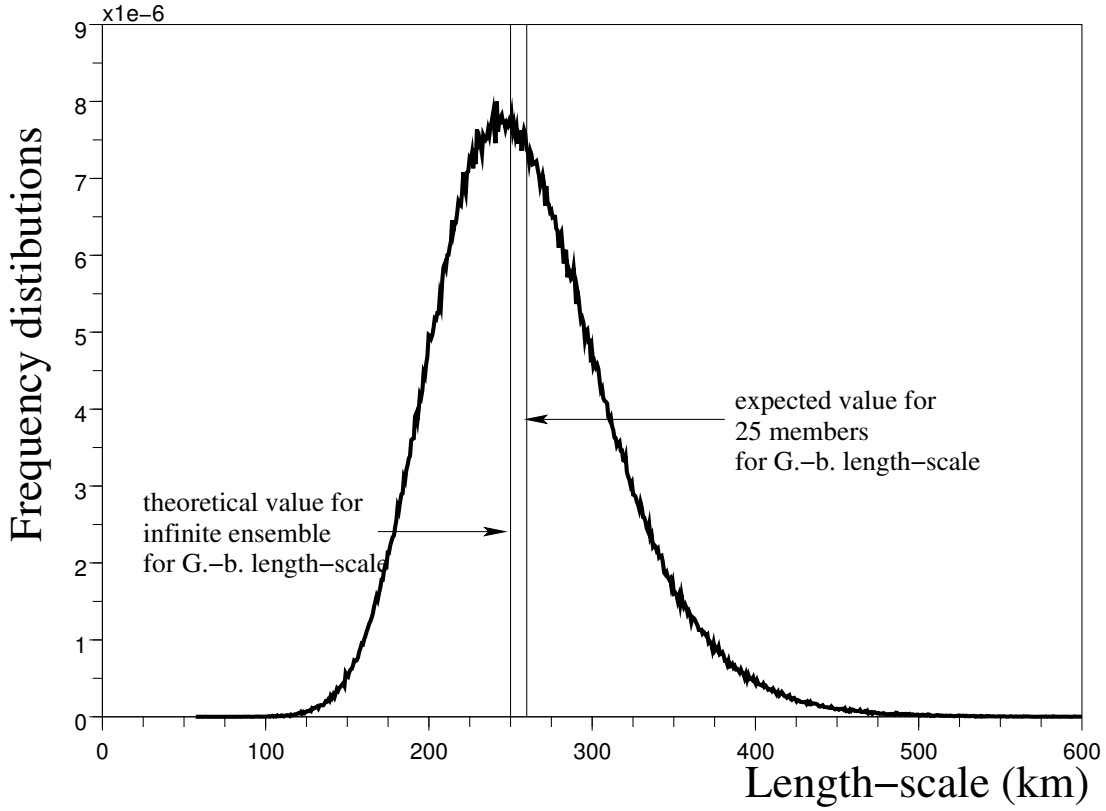


FIG. 5.5 – Distribution d'échantillonnage pour la portée Gb avec $\delta x = 166\,km$ et $N = 25$. La portée théorique de $L_H = 250\,km$ est surestimée par l'espérance de la portée estimée.

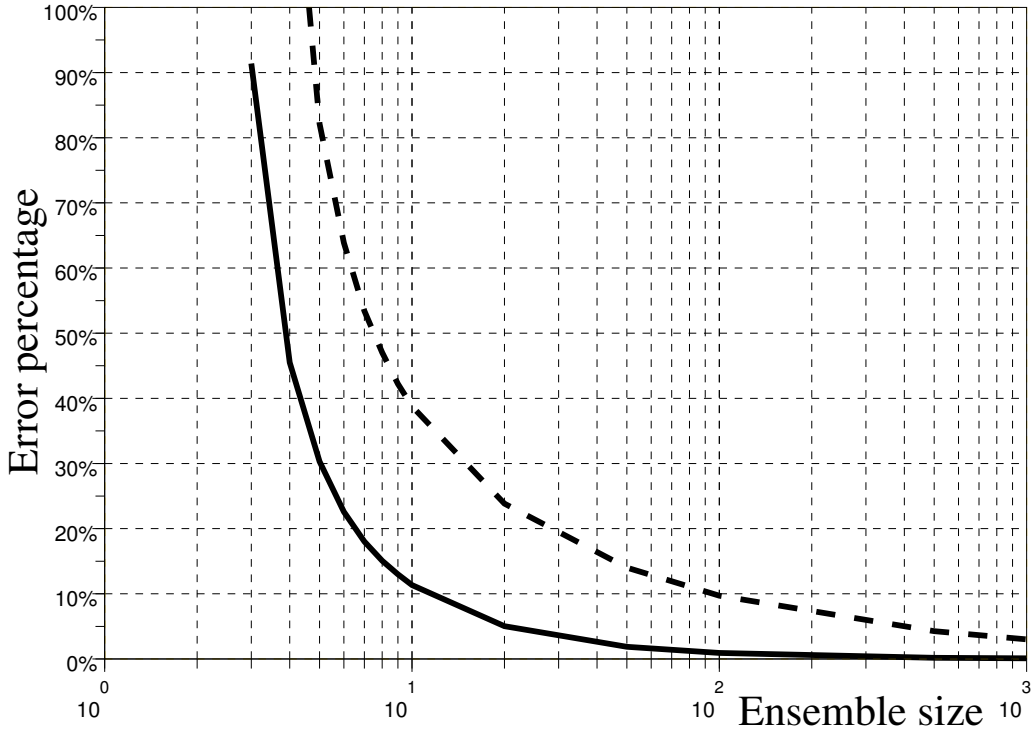


FIG. 5.6 – Convergence de l'erreur de portée en fonction de la taille de l'ensemble : le biais (courbe continue) et l'écart type (courbe tiretée). Ces deux quantités sont normalisées par la portée vraie (voir le texte pour les détails).

Comme attendu (voir annexe), un comportement en $\mathcal{O}(N^{-1/2})$ est observé. Pour 10 membres, le rapport est de 40%. Ceci illustre l'effet majoritaire de la contribution de l'erreur associée à l'écart type, par rapport à celui introduit par le biais, sur l'estimation de la portée.

5.4.3 Comparaison avec d'autres formules de calcul de la portée

Déterminer la distribution d'échantillonnage analytique pour les portée de Daley et B&B n'est pas facile, parce que ces portées dépendent de la forme de la fonction, et non plus seulement d'une corrélation. Cependant, des expériences numériques montrent que ces portées ont un comportement similaire à celle de Pb et Gb. La figure 5.7 illustre les distributions d'échantillonnage pour les portées issues des quatre formules : Daley, B&B, Pb et Gb. Ces portées sont estimées à partir d'un ensemble de 10 membres. Dans ce cas, le tenseur de covariance utilisé correspond au tenseur homogène gaussien sur le cercle discrétisé à la troncature $T120$ et avec la portée L_H (donnée plus haut). Il apparaît que les distributions d'échantillonnage sont proches les unes des autres. En particulier, les portées de Daley et B&B présentent elles aussi un biais.

5.4.4 Structure spatiale du bruit d'échantillonnage

Comme illustré sur la figure 5.4, les variations spatiales des portées estimées semblent être aléatoires et décorrélées spatialement, comparées aux variations de la portée exacte. Ceci suggère que le champ de portée estimé est entaché d'un bruit d'échantillonnage, dont l'amplitude est relativement grande (comparée à celles du champ de portée exact) dans les petites échelles.

Afin d'étudier cette question, différents spectres d'énergie ont été calculés : celui de la carte

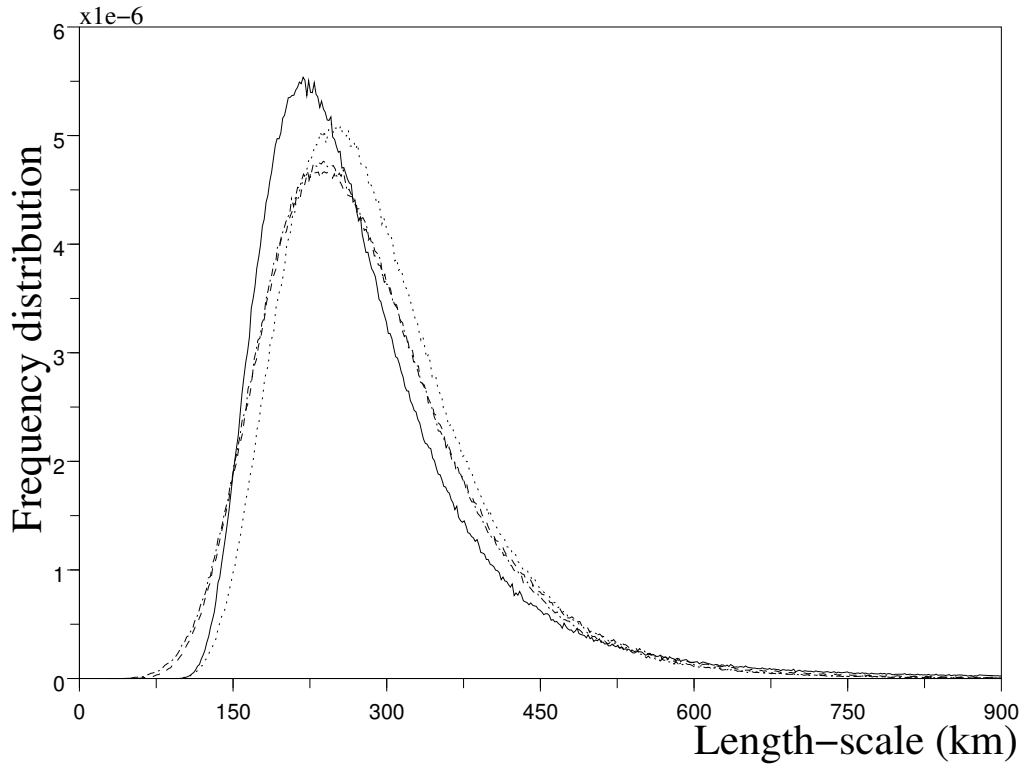


FIG. 5.7 – Comparaison de la distribution d'échantillonnage pour la portée, estimée à partir d'un ensemble de 10 membres. Le tenseur de corrélation considéré correspond au tenseur homogène gaussien, de portée $L_H = 250\,km$, sur le cercle discrétisé à la troncature $T = 120$. Les portées estimées sont : Daley (courbe continue), Gb (courbe tiretée-pointilliée), Pb (courbe pointillée) et Belo Pereira& Berre (courbe tiretée). À noter que les portées Gb et B&B sont quasiment superposées.

de portée exacte, celui de la carte de portée estimée à partir d'un ensemble et enfin le spectre de l'erreur d'estimation correspondante. Les résultats sont montrés sur la figure 5.8. Comme attendu, d'après la figure 5.4, la carte de portée estimée à partir d'un ensemble contient plus d'énergie dans les petites échelles que la carte exacte. Ceci correspond à des contributions artificielles du bruit d'échantillonnage, dont le spectre d'énergie est proche d'un bruit blanc.

Ces résultats indiquent que les techniques de filtrage spatial, basées sur les approches spectrale ou ondelette, seraient utiles pour diminuer cette part de bruit. C'est ce qui est abordé dans le prochain paragraphe.

5.4.5 Réduction du bruit d'échantillonnage à l'aide d'un filtrage spatial

La modélisation des covariances d'erreur d'ébauche est souvent basée sur l'approche diagonale dans l'espace spectral (Courtier *et al.* , 1998). Plus récemment, Fisher (2003) a également proposé une modélisation basée sur l'approche diagonale ondelette. Comme discuté au chapitre 4 (Pannekoucke *et al.* , 2007), l'utilisation de ces techniques revient à moyenner spatialement les fonctions locales de corrélation. Cela permet potentiellement une diminution du bruit d'échantillonnage.

Dans l'approche spectrale, cette moyenne spatiale est globale, au sens où elle correspond à une moyenne uniforme sur l'ensemble du domaine. Dans le cas des ondelettes, l'approche diagonale correspond à une moyenne spatiale plus localisée. Ceci signifie que les ondelettes augmentent la taille effective de l'échantillon, en introduisant un échantillonnage spatial local (à multiplier par la taille de l'ensemble d'échantillonnage initial), tout en permettant de représenter les variations géographiques de la portée.

L'efficacité de cette approche de filtrage en ondelettes a été illustré par Pannekoucke *et al.* (2007) dans un contexte 1D hétérogène. Ici, on se focalise sur un cas homogène 1D, de manière à illustrer les effets du filtrage spatial sur le biais et l'écart type affectant l'estimation de la portée.

Ce cas 1D correspond au tenseur de corrélation homogène gaussien, sur le cercle discrétisé $T120$ (comprenant $N_g = 241$ points de grille) avec une portée théorique égale à $L_H = 250km$. Une estimation de la matrice de corrélation, avec un ensemble de N membres, conduit à une matrice de corrélation hétérogène. Pour un point k, la portée Gb est calculée comme la moyenne $L_e^N = (L^+ + L^-)/2$. La corrélation modélisée à partir de l'hypothèse diagonale en spectral est homogène, et correspond à la moyenne sur l'ensemble du domaine des N_g fonctions de corrélation estimées. La portée Gb correspondante est alors $L_{ds} = \left(L(\bar{\rho}^+) + L(\bar{\rho}^-)\right)/2$ où $\bar{\rho}^+ = \frac{1}{N_g}\sum_k \rho_k^+$ et $\bar{\rho}^- = \frac{1}{N_g}\sum_k \rho_k^-$.

Différentes variables aléatoires sont également introduites : L_{ds}^N est la portée résultant de l'hypothèse diagonale spectrale pour un ensemble de N membres ; L_{dw}^N est la portée équivalente mais pour le cas ondelette. Le graphique en haut de la figure 5.9 représente les erreurs relatives $\mathbb{E}(L_e^N)/L_H - 1$ (courbe continue), $\mathbb{E}(L_{ds}^N)/L_H - 1$ (courbe tiretée-pointilliée) et $\mathbb{E}(L_{dw}^N)/L_H - 1$ (courbe tiretée). Ces quantités représentent le biais normalisé par L_H, pour les différentes estimations. Ces erreurs sont représentées en fonction du nombre de membres dans l'ensemble avec $N \in [6, 200]$. Comme déjà montré en section 5.4.2, $\mathbb{E}(L_e^N)/L_H - 1$ converge en $\mathcal{O}(N^{-1})$, tandis que $\mathbb{E}(L_{ds}^N)/L_H - 1$ est proche de zéro partout. $\mathbb{E}(L_{dw}^N)/L_H - 1$ est faible, même s'il reste différent de zéro, même pour de grands ensembles. Ce défaut correspond à une caractéristique connue des portées restituées dans la modélisation par l'hypothèse diagonale ondelette : les portées peuvent être sousestimées ou surestimées avec une erreur de moins de 10% (Pannekoucke *et al.* , 2007).

Pour apprécier la précision de l'estimation, les rapports σ_e^N/L_H, σ_{ds}^N/L_H et σ_{dw}^N/L_H sont

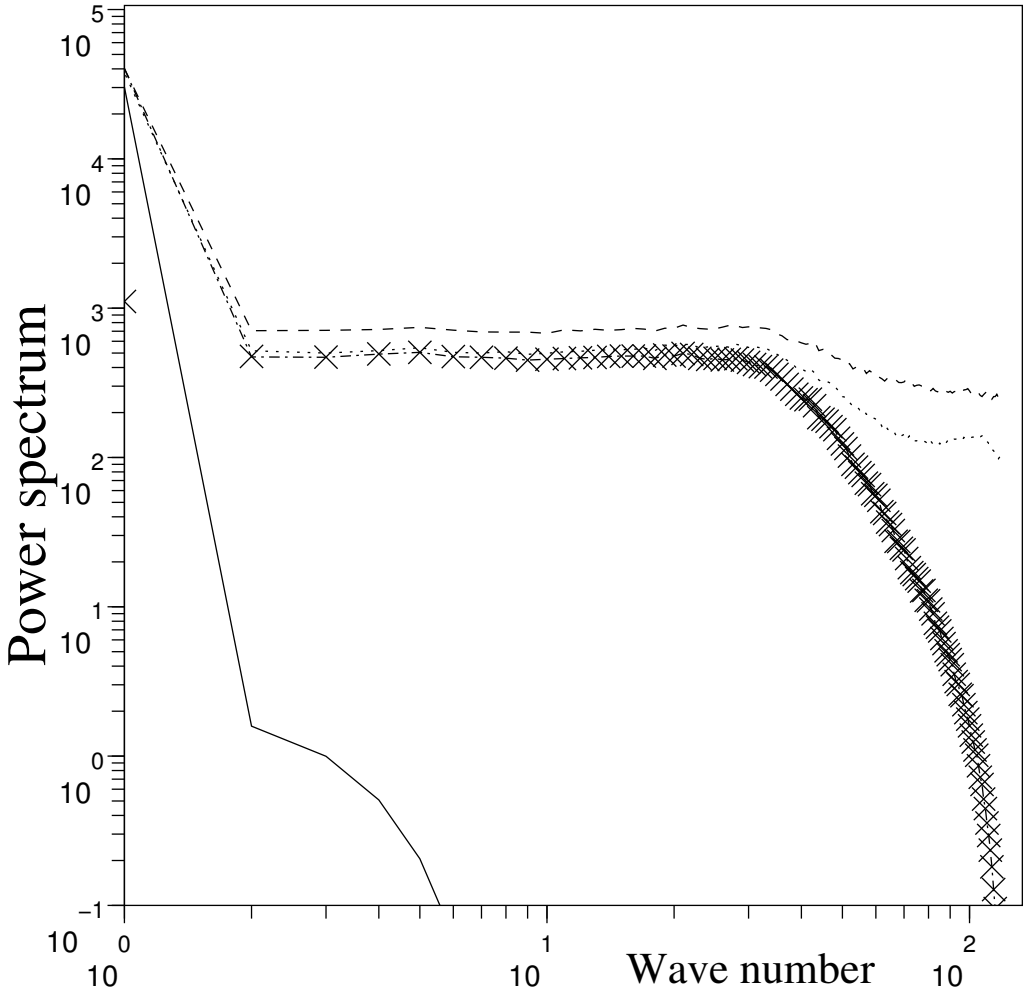


FIG. 5.8 – Spectre d'énergie des cartes de portée exactes (courbe continue) et estimées à partir d'un ensemble de 10 membres : Daley (courbe tiretée), B&B (courbe pointillée) et Gb (courbe tiretée-pointilliée). Pb étant similaire à Gb, ce résultat n'est pas montré. Le spectre d'énergie de l'estimation de l'erreur pour Gb est représenté par des croix. Ce spectre s'avère superposé au spectre d'énergie de la carte Gb sauf pour le nombre d'onde 1. À noter également que le spectre de la portée exacte (courbe continue) reflète la prédominance du nombre d'onde 1 en accord avec la figure (5.3).

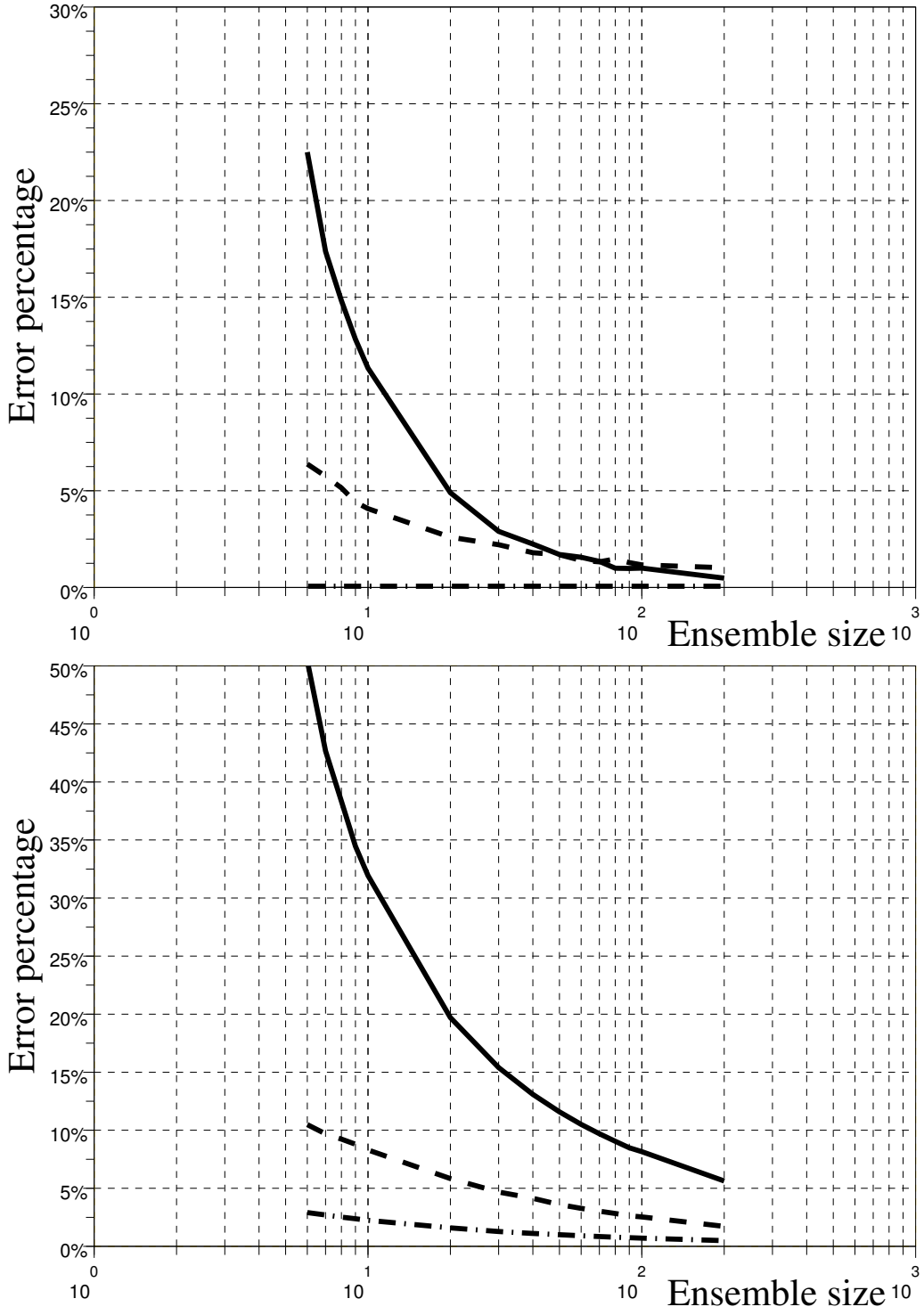


FIG. 5.9 – Comparaison de la convergence de l'erreur d'estimation de la portée en fonction de la taille de l'ensemble : pour les portées directement estimées (courbe continue), pour les portées résultantes de l'hypothèse diagonale dans l'espace des ondelettes (courbe tiretée) et dans l'espace spectral (courbe tiretée-pointilliée). En haut : le biais normalisé par L_H. En bas : l'écart type normalisé par L_H. (voir le texte pour les détails.)

également représentés (graphique du bas sur la figure 5.9). Ces rapports représentent l'écart type d'erreur normalisé par la portée L_H. À nouveau, σ_e^N/L_H converge vers zéro en $\mathcal{O}(N^{-1/2})$, ce qui est un résultat attendu (cf annexe). σ_{ds}^N/L_H converge à la même vitesse mais avec un facteur multiplicatif de l'ordre de $1/17 \approx 1/\sqrt{N_g}$: pour $N = 6$, $\sigma_e^N/L_H \approx 50\%$, tandis que $\sigma_{ds}^N/L_H \approx 3\%$. La convergence de σ_{dw}^N/L_H est encadrée par les deux convergences précédentes : la vitesse de convergence évolue comme l'estimation ensembliste, mais avec un facteur multiplicatif proche de $1/5$. En effet, pour $N = 6$, $\sigma_e^N/L_H \approx 50\%$ tandis que $\sigma_{dw}^N/L_H \approx 10\%$.

Ces résultats illustrent donc notamment la capacité de la formulation ondelette à représenter les valeurs de portée avec une plus grande précision (ici d'un facteur 5) qu'avec l'estimation ensembliste brute.

5.5 Application à un ensemble de prévisions Arpège

Le système opérationnel utilisé à Météo-France est fondé sur le modèle Arpège (Courtier et Geleyn, 1988), et sur un schéma d'assimilation variationnel 4D-Var (Rabier *et al.* , 2000 ; Veersé et Thépaut, 1998). La matrice de covariance d'erreur d'ébauche est calculée en utilisant un ensemble d'assimilations perturbées (Houtekamer *et al.* , 1996 ; Fisher 2003). Les détails des résultats pour cet ensemble Arpège sont décrits dans Belo Pereira et Berre (2006).

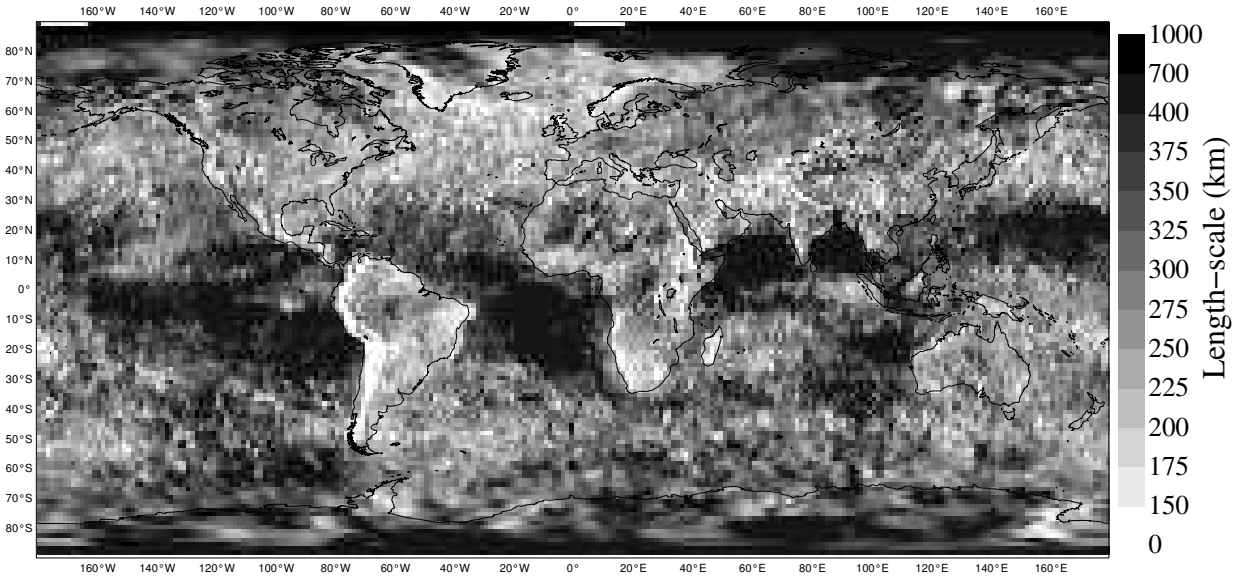
L'ensemble disponible comprend 6 différences de prévisions pour chaque jour sur la période allant du 9 février au 24 mars 2002. Les covariances moyennes sont calculées sur la période dans son ensemble, c'est-à-dire sur les 49 jours. La figure 5.10 montre les résultats obtenus avec les portées zonales calculées à partir des formules de B&B (graphique du haut) et de Gb (graphique du bas), pour le logarithme de la pression de surface. Comme dans le paragraphe précédent, le gradient zonal intervenant dans la formule de B&B est calculé dans l'espace spectral. Ces formules, approximant la portée, présentent des variations géographiques marquées. On peut noter, par exemple, le contraste terre-mer ou encore l'influence de l'orographie, avec *e.g.* des portées larges au dessus des zones tropicales sur les océans et des portées courtes au dessus des Andes. En fait, il n'y a que peu de différences entres les deux formulations de la portée. Ainsi, ces faibles différences confortent l'idée que les portées basées sur l'approximation gaussienne, fournissent une approximation réaliste des portées exactes.

5.6 Conclusions

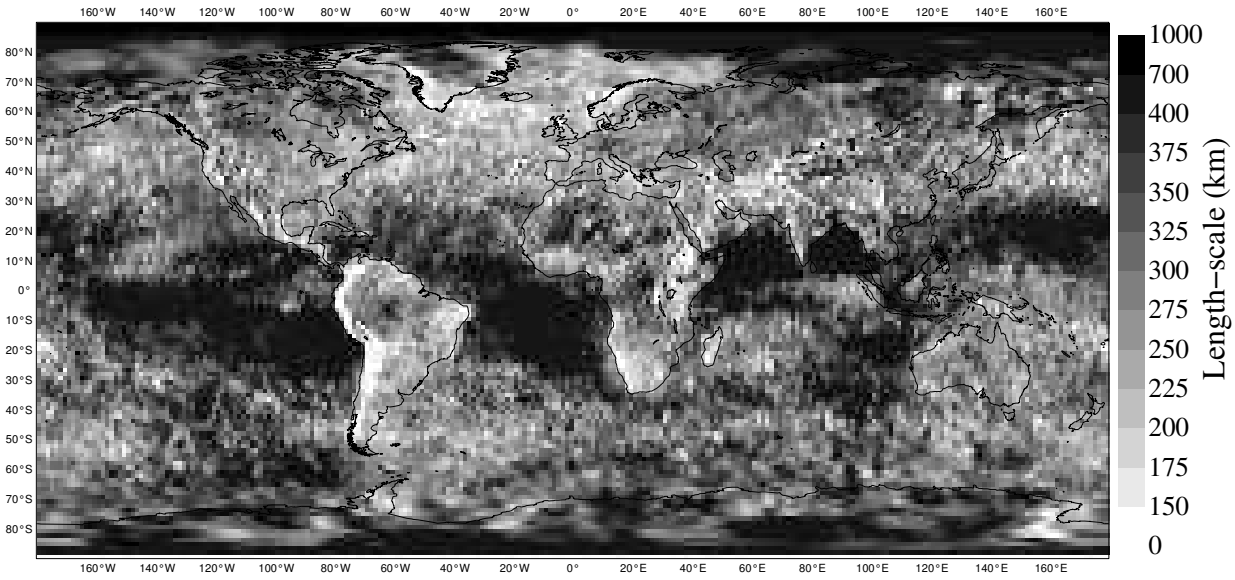
Dans ce chapitre, des approximations de la longueur de portée, définie par Daley, ont été exposées et étudiées. En particulier, une estimation peu coûteuse basée sur l'hypothèse gaussienne a été proposée. D'une part, un contexte 1D a permis de montrer que cette estimation est capable de restituer des valeurs de portée, ainsi que des variations géographiques réalistes.

D'autre part, une étude de la distribution d'échantillonnage a été menée, à la fois analytiquement et de manière expérimentale. Ainsi, il a été montré que l'estimation des portées à partir d'un ensemble (de taille N) est biaisée, avec un biais positif indiquant que l'estimation d'une portée conduit en moyenne à une portée qui lui est supérieure. De plus, ce biais tend vers zéro en $\mathcal{O}(N^{-1})$. Il a de plus été montré que l'effet de ce biais est petit par rapport à l'écart type de l'erreur d'échantillonnage. Ce dernier converge également vers zéro, mais avec une vitesse en $\mathcal{O}(N^{-1/2})$.

En outre, l'examen de la variation géographique des portées et de leur spectre d'énergie indique que le bruit d'échantillonnage tend à être décorrélé spatialement (à la manière d'un bruit blanc). Cela conforte l'idée qu'une technique de moyenne spatiale, telle que celle basée sur les



(a) Belo Pereira−Berre zonal length−scale



(b) Gaussian−based zonal length−scale

FIG. 5.10 – Cartes des portées zonales pour le logarithme de la pression de surface, calculées avec la formule de Belo Pereira-Berre (a), et avec la formule basée sur la gaussienne (b). Cette carte correspond à une moyenne climatologique, calculée sur une période de 49 jours.

ondelettes, peut être intéressante à considérer pour filtrer spatialement le bruit d'échantillonnage.

Finalement, la formule de portée de Belo Pereira et Berre a été comparée avec celle issue de l'approximation gaussienne, dans un cadre 2D sphérique à partir d'un ensemble de prévisions Arpège. Les portées ainsi diagnostiquées et leur variations géographiques sont similaires. Ainsi, l'approximation de la forme de la fonction de corrélation par une gaussienne apparaît comme étant raisonnable pour estimer la valeur de la portée.

5.7 Annexe

5.7.1 Approximation de la distribution d'échantillonnage de la corrélation

La distribution normale d'une paire d'erreur d'ébauche corrélée $\boldsymbol{\varepsilon}_b = (\varepsilon_1, \varepsilon_2)$ s'écrit

$$f_b(\boldsymbol{\varepsilon}_b) = \frac{1}{2\pi|\boldsymbol{B}|^{1/2}} exp\left(-1/2||\boldsymbol{\varepsilon}_b||^2_{\boldsymbol{B}^{-1}}\right),$$

où $\boldsymbol{B} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$ est la matrice de covariance d'erreur d'ébauche, de corrélation ρ, et dont le déterminant est $|\boldsymbol{B}|$. À partir d'un ensemble de N échantillons $(\varepsilon_1^1, \varepsilon_2^1), \ldots (\varepsilon_1^N, \varepsilon_1^N)$ les variances etimées sont données par $S_1^2 = \frac{1}{N}\sum_k {\varepsilon_1^k}^2$, et $S_2^2 = \frac{1}{N}\sum_k {\varepsilon_2^k}^2$. La corrélation estimée est donnée par $C = \frac{1}{N}\sum_k \varepsilon_1^k\varepsilon_2^k/(S_1S_2)$. Naturellement, S_1^2, S_2^2 et C sont des variables aléatoires.

Fisher (1953) donne l'expression de la distribution d'échantillonnage de C (Kendall *et al.* , 1998 ; Hotteling, 1953). Cependant, cette formule est en fait trop complexe, et des approximations de la distribution doivent être utilisées. Pour un ensemble de grande taille, cette distribution est proche d'une gaussienne (théorème central limite). Pour des ensembles de petite taille, cette distribution est loin d'être gaussienne. En particulier elle a une asymétrie ("skewness") négative, dont l'amplitude augmente avec la valeur de corrélation. Quand N n'est pas trop petit, typiquement, $N \geq 25$, Fisher a proposé une transformation adaptée pour laquelle la variable transformée converge plus rapidement vers une loi normale. Ainsi, la variable aléatoire $Z = tanh^{-1}C$ suit, avec une bonne approximation, la distribution gaussienne :

$$Z \sim \mathcal{N}\left(\mu_Z(\rho, N), \sigma_Z^2(\rho, N)\right), \tag{5.9}$$

avec la moyenne $\mu_Z(\rho, N) = \zeta + \frac{\rho}{2(N-1)} + \frac{\rho(5+\rho^2)}{8(N-1)^2} + \frac{\rho(11+2\rho^2+3\rho^4)}{16(N-1)^3} + \mathcal{O}\left(N^{-4}\right)$ où $\zeta = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ et l'écart type $\sigma_Z(\rho, N)^2 = \frac{1}{N-1} + \frac{4-\rho^2}{2(N-1)^2} + \frac{22-6\rho^2-3\rho^4}{6(N-1)^3} + \mathcal{O}\left(N^{-4}\right)$. Un simple changement de variable conduit à la distribution d'échantillonnage de la corrélation

$$f_C(c) = \frac{1}{(1-c^2)\sigma_Z(\rho, N)\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{[Z(c) - \mu_Z(\rho, N)]^2}{\sigma_Z(\rho, N)^2}\right\}, \tag{5.10}$$

où par définition $P(C \in [c, c+dc]) = f_C(c)dc$, avec P la mesure de probabilité associée à la variable C.

Le graphique en haut de la figure 5.11 représente cette distribution d'échantillonnage (courbe continue en trait gras) pour $N = 25$ membres et $\rho = exp(-\delta x^2/2L_H^2) \approx 0.8$. Cette distribution

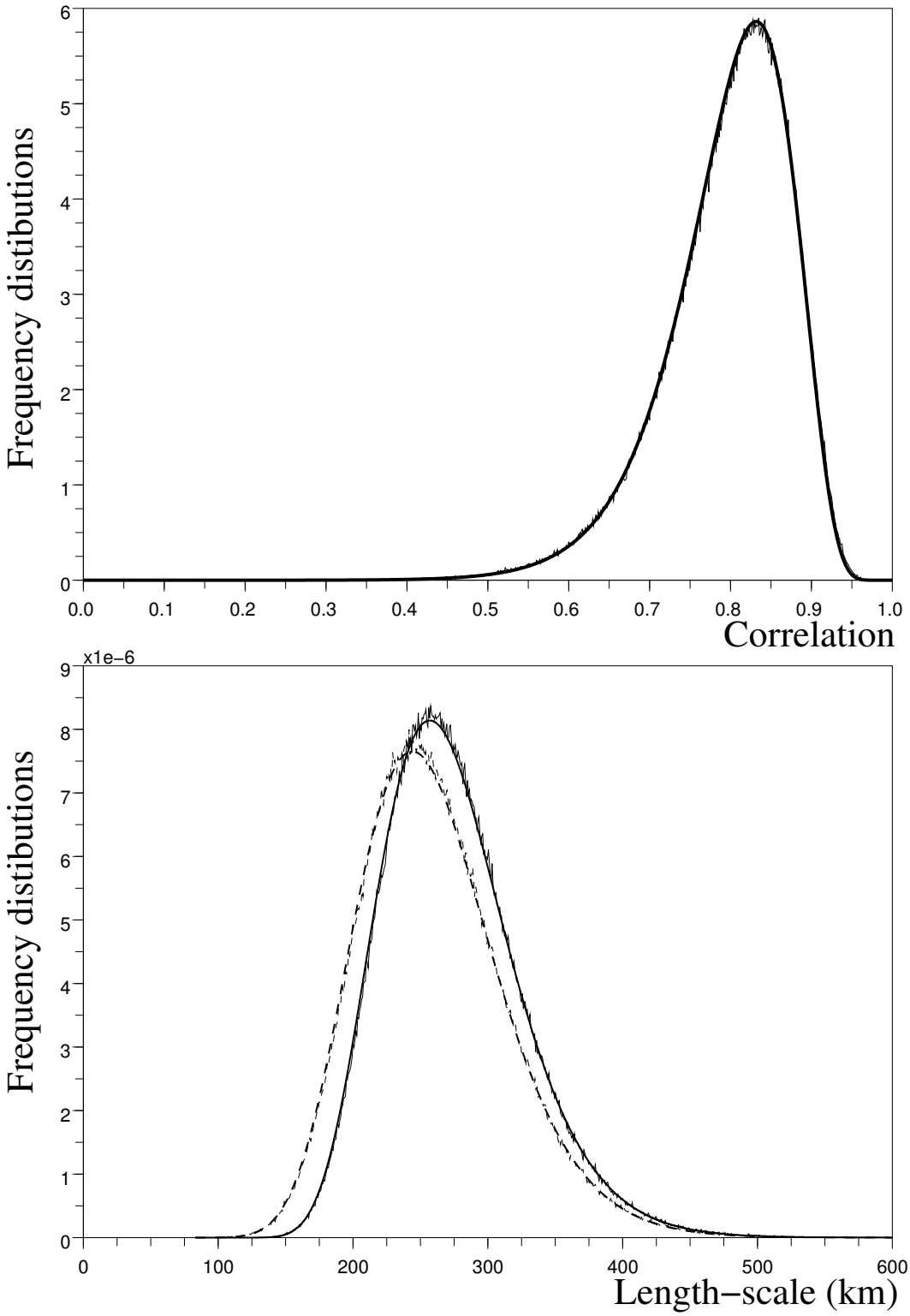


FIG. 5.11 – Comparaison entre l'approximation et l'estimation de la distribution d'échantillonnage résultant d'une estimation à partir d'un ensemble de 25 membres. En haut : distribution d'échantillonnage pour la corrélation, avec $\rho \approx 0.8$, analytique approchée (courbe continue en trait gras) et estimée expérimentalement (courbe continue en trait fin). En bas : distribution d'échantillonnage pour les deux approximations Pb et Gb, pour $\delta x = 166\,km$. La distribution analytique approchée pour la Pb (resp. Gb) est la courbe continue en trait gras (resp. la courbe tiretée en trait gras), tandis que que la distribution estimée expérimentalement est la courbe continue en trait fin (resp. la courbe tiretée en trait fin).

est comparée à l'estimation expérimentale de la distribution de fréquence (courbe continue en trait fin). Il apparaît que l'approximation de Fisher est de bonne qualité.

Comme le suggère l'équation de μ_Z et de σ_Z^2, le biais et l'écart type de Z convergent vers zéro en $\mathcal{O}(N^{-1})$ et $\mathcal{O}(N^{-1/2})$ respectivement. Cette vitesse de convergence est similaire à celle de la corrélation C.

5.7.2 Approximation de la distribution d'échantillonnage pour la portée Gb

L'application de la transformation de Fisher à la longueur de portée conduit à l'approximation de la distribution d'échantillonnage de la portée. Le calcul est présenté pour la portée Gb, sachant qu'il est similaire dans le cas Pb.

En fait, pour la portée Gb, les corrélations doivent être strictement positives. Ainsi la distribution d'échantillonnage de la corrélation doit être restreinte aux valeurs de corrélation strictement positives. Soit $\chi_{(0,1]}$ la fonction caractéristique définie sur $[-1, 1]$; cette fonction est égale à un sur $(0, 1]$ et nulle ailleurs. La variable aléatoire associée à la corrélation positive est alors $C^+ = \chi_{(0,1]}C$. Sa distribution d'échantillonnage est $f_{C^+}(c) = \Lambda(\rho, N)^{-1} f_C(c)$, $c > 0$, où $\Lambda(\rho, N) = \int_0^1 f_C(c)dc$ est un facteur de normalisation. Cette normalisation peut être approchée par $\Lambda(\rho, N) \approx 1 - \frac{\Gamma(N)}{\Gamma(N+1/2)\sqrt{2\pi}} \frac{(1-\rho^2)^{N/2}}{\rho}$ (Hotteling, 1953). Le changement de variable $C^+ = \exp\left(-\frac{\delta x^2}{2L_{Gb}^{N\ 2}}\right)$, avec L_{Gb}^N l'estimateur de la portée Gb calculé à partir de N membres, conduit à la distribution d'échantillonnage

$$f_{L_{Gb}^N}(l) = \frac{\Lambda(\rho, N)^{-1}\delta x^2}{2l^3 \sinh(\frac{\delta x^2}{2l^2})\sigma_Z(\rho, N)\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{[Z(\rho(l)) - \mu_Z(\rho, N)]^2}{\sigma_Z(\rho, N)^2}\right\}, \tag{5.11}$$

avec $\rho(l) = \exp\left(-\frac{\delta x^2}{l^2}\right)$.

Le graphique du bas sur la figure 5.11 représente la distribution expérimentale de fréquence et l'approximation analytique de la distribution d'échantillonnage pour les portées Pb et Gb. Ces résultats sont obtenus avec une valeur de corrélation de $\rho \approx 0.8$, utilisée dans la section précédente. Il apparaît que l'approximation analytique de la distribution d'échantillonnage est en accord avec la distribution de fréquence expérimentale, et ce avec une bonne précision.

Dans le cas d'un ensemble de taille infinie, la valeur de portée attendue est $L_{Gb}^\infty = \frac{\delta x}{\sqrt{-2\,ln(\rho))}}$. Dans le cas d'une estimation à partir de N membres, la distribution d'échantillonnage résultante est asymétrique, avec une asymétrie positive, et L_{Gb}^N est biaisée, avec un biais positif : $b_{Gb}^N = \mathbb{E}(L_{Gb}^N) - L_{Gb}^\infty > 0$. A nouveau, il peut être déduit de l'approximation analytique de distribution d'échantillonnage que, pour N grand, le biais converge vers zéro en $\mathcal{O}(N^{-1})$, et que l'écart type d'erreur d'échantillonnage converge également vers zéro mais en $\mathcal{O}(N^{-1/2})$.

Ces résultats sont également valables pour les autres formules de la portée sous l'hypothèse que l'erreur d'ébauche suit une loi gaussienne.

Chapitre 6

La structure spatiale et la dynamique des portées des corrélations d'erreur via un filtrage en ondelettes

RÉSUMÉ

Les ondelettes peuvent être utilisées pour représenter les corrélations locales des erreurs d'ébauche, estimées à partir d'un ensemble de taille finie. Leurs premières applications opérationnelles sont basées sur des corrélations qui ne varient pas temporellement, et qui sont moyennées sur plusieurs semaines typiquement. Les ondelettes sont utilisées ici pour diagnostiquer les variations spatiales et journalières des portées des corrélations.

On montre que les portées modélisées par les ondelettes sont relativement robustes lorsque les statistiques en ondelette sont moyennées sur un ensemble de six membres et sur une période de 24 heures (avec cinq réseaux successifs d'analyse). Cette robustesse peut être reliée à la structure spatiale du bruit d'échantillonnage et aux propriétés de filtrage des ondelettes.

De plus, l'évolution temporelle des cartes de portée est examinée sur quelques jours. Il apparaît que la dynamique des portées des corrélations est relativement complexe, et qu'elle est liée en partie à la situation météorologique. Ces résultats diagnostiques soutiennent l'idée d'utiliser les ondelettes pour extraire et filtrer des corrélations dépendantes de l'écoulement, à partir d'un ensemble d'assimilations.

6.1 Introduction

Dans le contexte de l'assimilation de données, les corrélations spatiales des erreurs d'ébauche influencent directement la façon dont l'information observée est filtrée et propagée spatialement. En particulier, la portée d'une fonction locale de corrélation (Daley 1991) permet de décrire la forme de cette fonction, qui peut être plus ou moins étendue spatialement. Par ailleurs, ces corrélations dépendent en principe de la situation météorologique. Cette dépendance reste cependant assez mal connue en pratique, en plus d'être difficile à représenter dans les schémas d'assimilation.

Des approches basées sur des modèles linéaires tangents ont notamment été proposées par le passé, pour diagnostiquer les variations des corrélations en fonction de l'écoulement. D'une part, Thépaut *et al.* (1995) ont utilisé des expériences d'assimilation 4D-Var avec une seule observation, pour examiner la forme des écarts types et des corrélations sur une situation barocline. Les résultats indiquent notamment que la portée des corrélations est deux fois plus courte près de la dépression étudiée. D'autre part, Bouttier (1993) a utilisé une méthode basée sur une évolution linéaire tangente de covariances initiales empiriques, pour diagnostiquer les variations spatio-temporelles des covariances. Cette approche a été complétée (Bouttier 1994) par une approche reposant sur un filtre de Kalman étendu. Ces deux dernières études ont mis en

évidence là aussi une influence importante de la situation météorologique sur la portée locale des corrélations.

Plus récemment, une méthode basée sur un ensemble d'assimilations perturbées a été proposée par Houtekamer et al (1996), qui présente par ailleurs des liens importants avec le filtre de Kalman d'ensemble (Evensen 1994). Cette approche permet notamment de représenter l'effet de l'analyse et des non linéarités du modèle atmosphérique, au niveau de l'évolution temporelle des erreurs du système d'assimilation. Cependant, les variations spatio-temporelles induites pour les portées des corrélations ont été d'une part peu documentées. Un diagnostic peu coûteux des portées locales sera donc ici mis en oeuvre pour étudier ces variations avec ce type d'ensemble.

D'autre part, les estimations de corrélation issues de l'ensemble sont affectées par un bruit d'échantillonnage, qui provient de la taille finie de l'ensemble. Il a néanmoins été montré par Pannekoucke *et al.* (2007), dans un cadre académique, que les ondelettes présentent des propriétés de filtrage potentiellement intéressantes pour réduire l'amplitude de ce bruit.

L'objet de ce papier est donc d'une part d'étudier ces propriétés de filtrage dans le cadre d'un ensemble d'assimilations réel (sections 6.3 et 6.4). Il s'agit d'autre part de mettre à profit ces propriétés pour documenter les variations spatio-temporelles des portées des corrélations (section 6.5).

6.2 Données ensemblistes, estimation des portées et ondelettes

6.2.1 Ensemble d'assimilations Arpège

L'ensemble utilisé dans cette étude est semblable à celui décrit dans Belo Pereira et Berre (2006) et Berre *et al.* (2007). Il comprend six expériences d'assimilation perturbées et indépendantes, couvrant la période du 18 janvier au 17 février 2005. Les assimilations ont été réalisées avec la version 3D-FGAT de l'assimilation Arpège, avec une géométrie de type T359 L46 c1.0 pour les prévisions. Des analyses sont calculées toutes les six heures (00H UTC, 06H UTC, 12H UTC and 18H UTC). Pour chaque membre de l'ensemble, une prévision à six heures d'échéance est lancée à partir de chacune de ces quatre analyses journalières, pour servir d'ébauche au réseau d'analyse suivant.

Un ensemble de six membres supplémentaires a par ailleurs été calculé pour les cinq réseaux allant du 18 janvier à 00H UTC au 19 janvier à 00H UTC. Cela permet donc de disposer d'un ensemble total de 12 membres pour ces dates là.

On notera ${\boldsymbol{x}^b}_{t,k}$ l'ébauche valable à l'instant t, et associée au membre k. Cela correspond à une prévision à six heures d'échéance, tronquée en $T179$. Les erreurs d'ébauche sont estimées par l'écart à la moyenne $\overline{{\boldsymbol{x}^b}_t}$ des N membres de l'ensemble : $\varepsilon^b_{t,k} = {\boldsymbol{x}^b}_{t,k} - \overline{{\boldsymbol{x}^b}_t}$, avec $\overline{{\boldsymbol{x}^b}_t} = \frac{1}{N}\sum_{k=1}^{N} {\boldsymbol{x}^b}_{t,k}$.

6.2.2 Diagnostic des portées des corrélations

Les portées locales L des corrélations ont été calculées à l'aide de la formule de Belo Pereira et Berre (2006) :

$$L^2(x) = \frac{\sigma^2\left(\varepsilon^b(x)\right)}{\sigma^2\left(\partial_x\varepsilon^b(x)\right) - \left\{\partial_x\sigma(\varepsilon^b(x))\right\}^2}, \tag{6.1}$$

pour la portée dans la direction zonale x. La portée dans la direction méridienne y est définie d'une façon similaire. Les gradients sont calculés avec une méthode spectrale.

6.2.3 Les ondelettes et leurs propriétés de filtrage

Les estimations des statistiques (telles que les portées locales), issues d'un ensemble de taille finie, sont affectées par un bruit d'échantillonnage. Pour réduire ce bruit, une approche en ondelettes peut être envisagée.

Les ondelettes correspondent à une approche intermédiaire entre les représentations en points de grille et en spectral (Fisher 2003). Une ondelette contient en effet à la fois une information de position et une information d'échelle. Il a été montré (Pannekoucke *et al.* , 2007) de façon formelle, et dans un cadre expérimental académique, que les ondelettes permettent de filtrer spatialement le bruit d'échantillonnage (au travers d'une moyenne spatiale locale des fonctions de corrélation).

6.3 Variations de robustesse avec ou sans les ondelettes et une moyenne temporelle

Les applications opérationnelles actuelles des ondelettes sont basées sur des moyennes temporelles sur plusieurs semaines (Fisher 2003). Le but de cette section est de montrer qu'une estimation des cartes de portées, basée sur les ondelettes et une moyenne journalière, permet d'accéder à une information robuste.

6.3.1 Estimation de la robustesse des estimations

Compte tenu du fait qu'une moyenne temporelle est habituellement utilisée pour estimer les statistiques en ondelettes, on considère ici l'estimation des portées à l'aide d'un ensemble de six membres, et avec une moyenne temporelle sur un nombre N_t de réseaux, qui vaut soit 1, 3 ou 5. Ces trois estimations correspondent donc respectivement à des valeurs instantanées ($N_t = 1$), moyennées sur une période de 12 heures ($N_t = 3$), et moyennées sur une période de 24 heures ($N_t = 5$).

Pour estimer la robustesse des portées estimées, on calcule la corrélation entre les cartes issues de deux ensembles indépendants et avec une taille d'échantillon temporel N_t identique. Ce diagnostic pour les cartes des portées est semblable à ce qui est présenté par Berre *et al.* (2007) pour les cartes des écarts types. L'idée est que si l'erreur d'échantillonnage est forte, la corrélation entre les deux cartes d'erreur doit être proche de 0. Inversement, si l'erreur d'échantillonnage est faible, la corrélation entre les deux cartes d'erreur doit être proche de 1.

On définit $L_i(\theta, \phi)$ comme étant la carte des portées estimées à partir de l'ensemble i (=1 ou 2). (θ, ϕ) correspond aux coordonnées de longitude et de latitude. La moyenne spatiale sur l'ensemble du globe est définie par $< \cdot >= \frac{1}{4\pi} \int_{S^2} \cdot \, cos\phi \, d\phi d\theta$. Ainsi, la valeur moyenne de L_i est donnée par

$$< L_i >= \frac{1}{4\pi} \int_{S^2} L_i(\theta, \phi) \, cos\phi \, d\phi d\theta, \tag{6.2}$$

tandis que son écart type est

$$\sigma_i = \sqrt{\left\langle (L_i - < L_i >)^2 \right\rangle}. \tag{6.3}$$

La corrélation $\rho_{1,2}$ entre les deux cartes de portée est donnée par la formule suivante :

$$\rho_{1,2} = \frac{\left\langle (L_1 - < L_1 >) \, (L_2 - < L_2 >) \right\rangle}{\sigma_1 \sigma_2}, \tag{6.4}$$

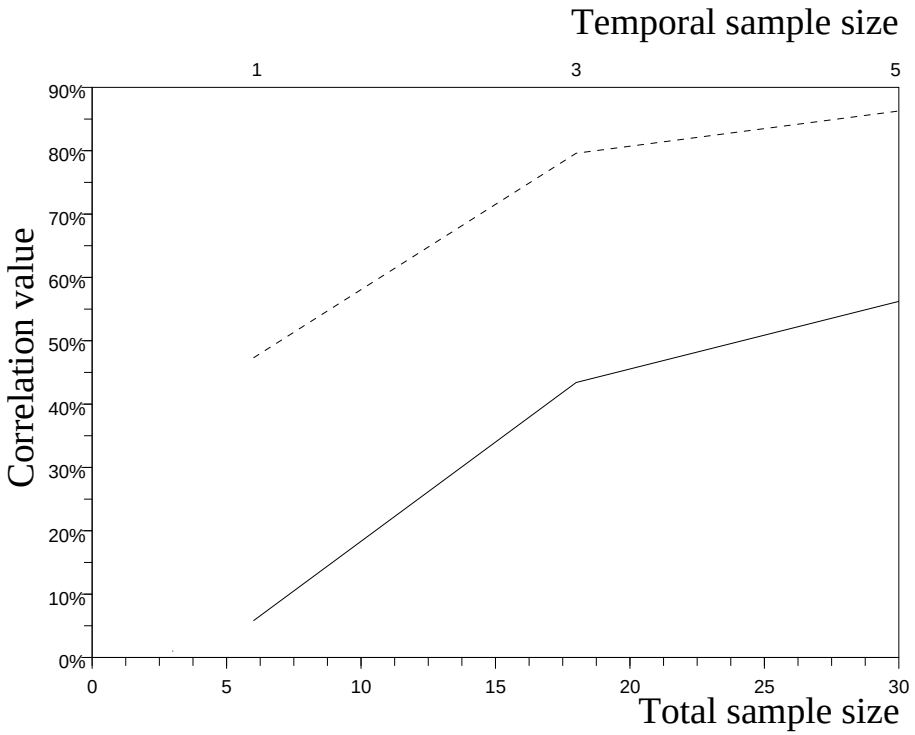


FIG. 6.1 – Corrélation (en moyenne sur la période) entre les cartes de portée, calculées de façon brute (courbe pleine) ou avec les ondelettes (courbe tiretée), en fonction de la taille N_t de l'échantillon temporel (N_t= 1, 3 ou 5), et à l'aide d'un ensemble de six membres.

Cette mesure de la robustesse des cartes est calculée en moyenne sur l'ensemble de la période.

6.3.2 Résultats numériques

Les variations obtenues pour cette corrélation sont illustrées sur la figure 6.1. Un premier résultat attendu correspond à l'augmentation de la robustesse des portées lorsque l'on augmente la taille de l'échantillon temporel. Cela reflète en effet l'augmentation de la taille de l'échantillon statistique.

Par ailleurs, un deuxième résultat frappant est que la corrélation des portées journalières (N_t=5) estimées avec les ondelettes s'élève à plus de 85%. Cela constitue un résultat intéressant, dans la mesure où l'on aurait pu penser a contrario qu'il faille envisager une moyenne sur plusieurs semaines pour atteindre ce niveau de robustesse.

Lorsque les portées sont estimées de façon brute (i.e. sans les ondelettes), la robustesse de ces portées journalières n'est elle que de l'ordre de 55%. L'augmentation de robustesse apportée par les ondelettes paraît en fait cohérente avec leurs propriétés de filtrage : on s'attend à ce que le bruit d'échantillonnage soit filtré spatialement par les ondelettes, conduisant ainsi à des estimations plus robustes.

Dans la suite de cette étude, on se propose de documenter ces propriétés de filtrage d'une part, puis de les exploiter pour étudier la dynamique spatio-temporelle des portées journalières.

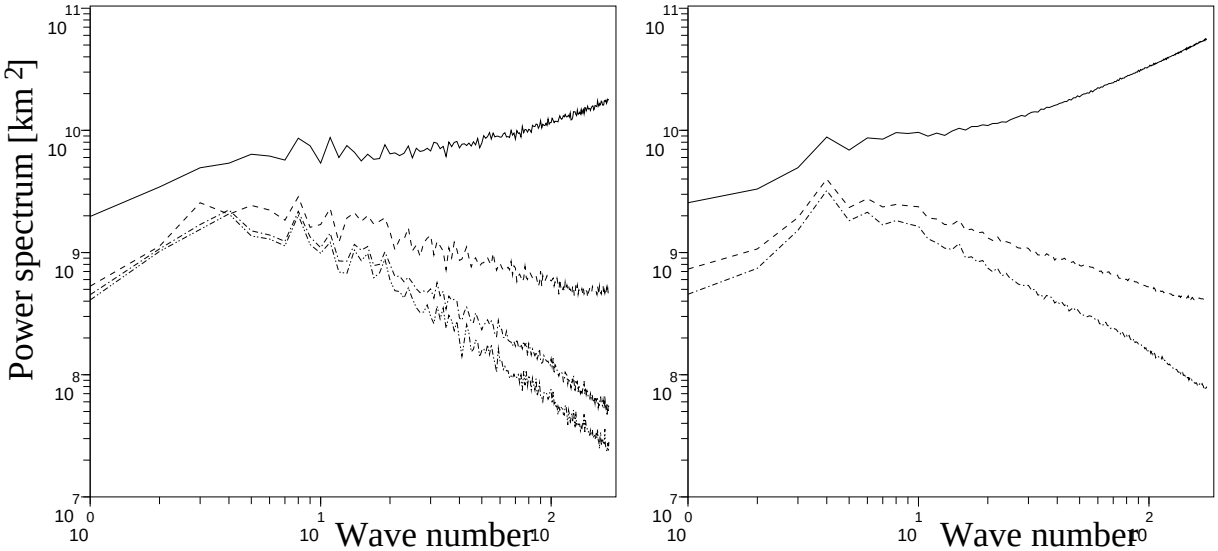


FIG. 6.2 – Spectres d'énergie des cartes des portées journalières (correspondant à une moyenne temporelle sur 5 réseaux successifs, centrée sur le réseau de 12H), relatives au 18 janvier 2005 (panel de gauche), et en moyenne sur l'ensemble de la période (panel de droite), pour différentes tailles d'ensemble : $N = 1$ (courbe pleine), $N = 3$ (courbe tiretée), $N = 6$ (courbe tiretée et pointillée), et $N = 12$ (courbe tiretée et bipointillée, pour le panel de gauche seulement).

6.4 Structure spatiale du bruit et filtrage en ondelette

Les ondelettes présentent des propriétés de filtrage spatial qui sont potentiellement bénéfiques pour l'estimation des corrélations (Pannekoucke *et al.* , 2007). Un tel filtrage spatial est en effet a priori pertinent, lorsque le bruit d'échantillonnage est d'assez petite échelle par rapport au signal à extraire (Berre *et al.* , 2007).

L'objet de cette section est donc de mettre en évidence la structure spatiale du bruit d'échantillonnage, ainsi que le type de filtrage apporté par les ondelettes. Compte tenu de la section précédente et de la figure 6.1, cette problématique est appliquée au cas des portées journalières (obtenues par un moyennage des statistiques sur les cinq réseaux successifs répartis sur une période de 24 heures).

6.4.1 Spectres d'énergie des cartes en fonction de la taille de l'ensemble

Pour examiner la structure spatiale du bruit d'échantillonnage, une première idée est de tracer l'évolution des spectres d'énergie (des cartes de portée) en fonction de la taille N de l'ensemble.

Pour un champ de portées L, le spectre d'énergie est défini par

$$E\left(L_i\right)(n) = \sum_{m=-n}^{n} L_{in}^{m} (L_{in}^{m})^*, \tag{6.5}$$

où L_n^m est le coefficient spectral (en harmoniques sphériques) du champ L. L'astérique $()^*$ est l'opérateur de conjugaison des nombres complexes.

L'évolution des spectres en fonction de N est représentée sur la figure 6.2. Il apparaît d'une part que l'énergie tend à diminuer lorsque la taille de l'ensemble augmente, et d'autre part que

cette baisse d'énergie est plus forte dans les petites échelles que dans les grandes échelles. Cela est notamment observable sur une journée prise au hasard (panel de gauche). Par ailleurs, ce résultat est relativement robuste statistiquement, dans la mesure où il apparaît également de façon très similaire en moyenne sur l'ensemble de la période (panel de droite).

Pour interpréter ces résultats, on peut proposer le modèle suivant. Le champ des portées brutes L_i d'un ensemble i donné peut être formellement décomposé comme la somme d'un signal exact $L_\star$ et d'un bruit d'échantillonnage L_i^e :

$$L_i = L_\star + L_i^e,$$

ce qui conduit à, pour l'énergie de ce champ :

$$E(L_i) = E(L_\star) + E(L_i^e),$$

si l'on néglige le terme croisé égal à $2L_\star(L_i^e)^*$. En effet, à un instant donné (par exemple) et pour un nombre d'onde n donné, ce terme croisé peut être vu comme la contribution de n à la covariance spatiale entre $L_\star$ et L_i^e, moyennée sur l'ensemble du globe (e.g. Courtier *et al.* . 1998). Or on peut considérer que ces deux champs sont décorrélés spatialement, dans la mesure où le bruit L_i^e est un processus aléatoire, déterminé essentiellement par les valeurs des perturbations d'observation, plutôt que par les variations du signal.

Par ailleurs, lorsque la taille de l'ensemble augmente, on s'attend à ce que l'amplitude $E(L_i^e)$ du bruit diminue (tandis que celle du signal exact reste inchangée par définition), conduisant ainsi à une diminution de l'énergie $E(L_i)$ des portées estimées. Cette interprétation est conforme à l'évolution observée sur la figure 6.2.

En outre, le fait que la diminution d'énergie soit plus importante dans les petites échelles que dans les grandes échelles suggère que l'amplitude du bruit est relativement forte dans les petites échelles.

L'objet de la section suivante est de mettre en évidence cette caractéristique d'une façon plus directe.

6.4.2 Amplitudes absolue et relative du bruit d'échantillonnage

Pour estimer l'amplitude du bruit d'échantillonnage, on peut utiliser la différence entre les estimations L_1 et L_2 de deux ensembles indépendants et de même taille (cf Berre *et al.* (2007) pour une approche semblable sur les écarts types) :

$$L_1 - L_2 = (L_\star + L_1^e) - (L_\star + L_2^e) = L_1^e - L_2^e.$$

Sous l'hypothèse que les deux bruits L_1^e et L_2^e sont deux processus aléatoires décorrélés, cela conduit à $E(L_1 - L_2) = E(L_1^e) + E(L_2^e) = 2E(L^e)$, sachant que $E(L_1^e) = E(L_2^e)$ (dans la mesure où les deux ensembles indépendants ont la même taille) peut être noté $E(L^e)$. Cela signifie que l'énergie du bruit peut être estimée par la moitié de l'énergie de $L_1 - L_2$:

$$E(L^e) = \frac{1}{2}E(L_1 - L_2)$$

Les amplitudes respectives (en énergie) des portées brutes (courbe pleine) et du bruit associé (courbe tiretée) sont représentées sur la figure 6.3. Il apparaît que le niveau relatif de bruit (en proportion de l'amplitude des portées brutes) est plus important dans les petites échelles que dans les grandes échelles.

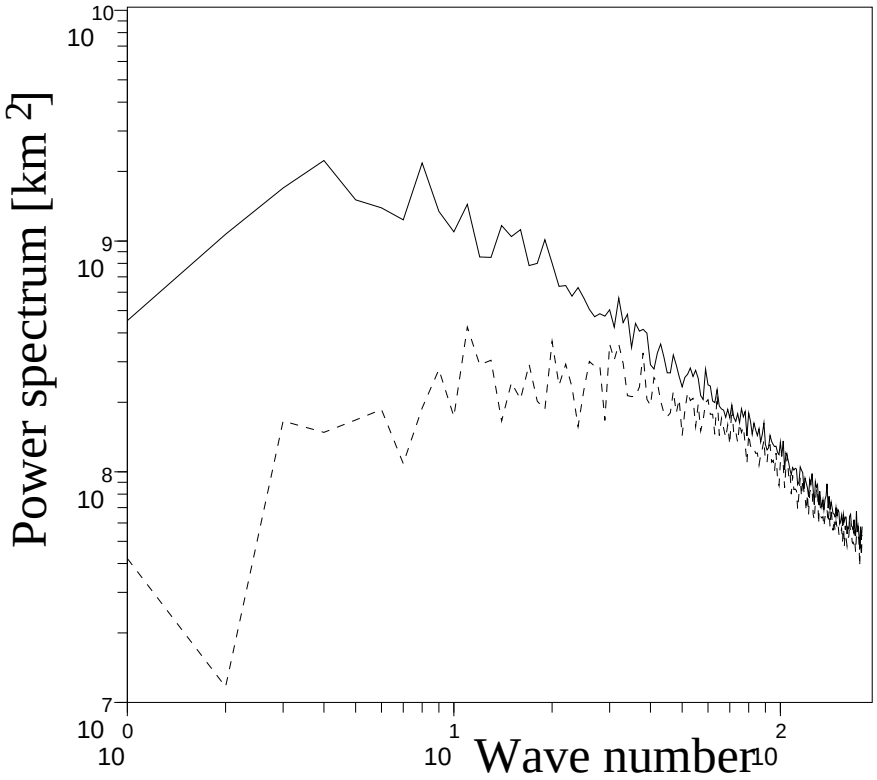


FIG. 6.3 – Spectres d'énergie des cartes des portées journalières brutes (courbe pleine), et du bruit d'échantillonnage (courbe tiretée), pour le 18 janvier 2005, avec $N = 6$.

Cette prédominance du bruit dans les petites échelles est cohérente avec la figure 6.2, qui indique une forte sensibilité des portées brutes à la taille N de l'ensemble, au niveau des petites échelles. Ces résultats suggèrent ainsi qu'un filtrage spatial des portées serait pertinent, afin de réduire l'amplitude du bruit de petite échelle.

6.4.3 Effets du filtrage en ondelette sur les spectres des portées

Compte tenu du niveau élevé de bruit dans les petites échelles, il serait souhaitable que ces petites structures soient filtrées par les ondelettes. Pour voir si ce type de filtrage a lieu, les spectres respectifs des portées brutes et des portées en ondelette ont été représentés sur la figure 6.4. Il apparaît que l'amplitude des petites échelles est fortement atténuée par les ondelettes, en accord avec les effets de filtrage attendus.

En outre, il s'avère que l'amplitude des grandes échelles est également réduite par les ondelettes. Cette atténuation dans les grandes échelles semble en fait pertinente, dans la mesure où les grandes structures de portée sont elles aussi affectées par du bruit d'échantillonnage, même si cela est moins marqué que pour les petites échelles.

Cela est effectivement suggéré d'une part par la figure 6.3, qui indique que le bruit a une amplitude significative dans les grandes échelles. Pour confirmer cela, on peut également tracer la corrélation spectrale entre les cartes des portées brutes (figure 6.5) :

$$\rho(n) = \frac{\sum_{m=-n}^{n} Cov(L_{1_n}^m, L_{2_n}^m)}{\sqrt{P(L_1)(n)P(L_2)(n)}}, \tag{6.6}$$

avec

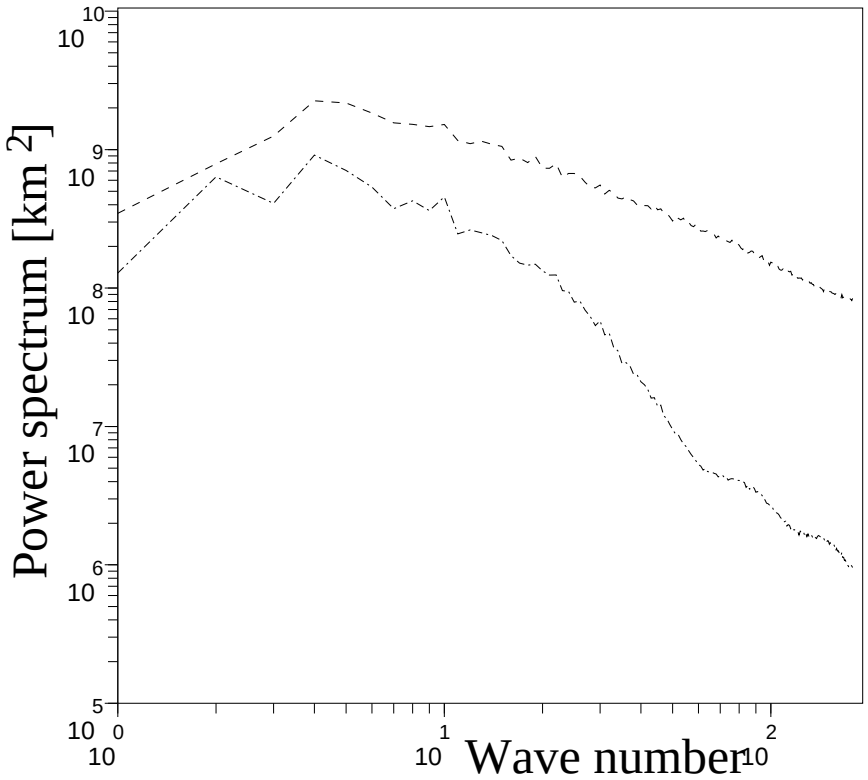


FIG. 6.4 – Spectres d'énergie des cartes des portées journalières, avec $N = 6$: pour les portées brutes (courbe tiretée), et pour les portées en ondelettes (courbe tiretée-pointillée).

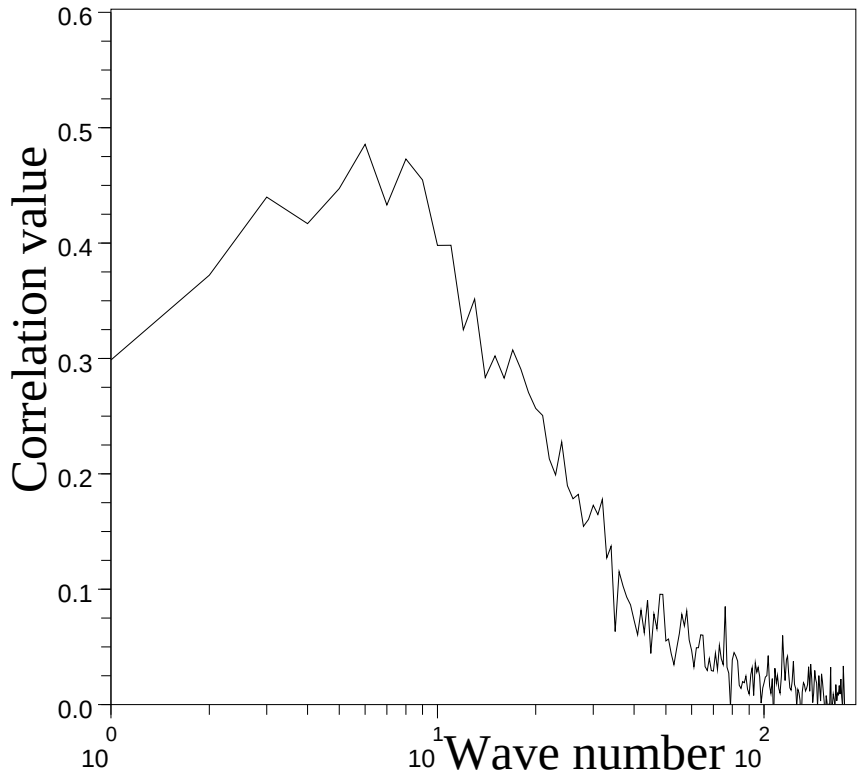


FIG. 6.5 – Corrélation ρ entre les coefficients spectraux des portées brutes, issues de deux ensembles indépendants et de même taille ($N = 6$).

$$Cov(L_{1_n}^m, L_{2_n}^m) = \frac{1}{N_t} \sum_t \left(L_{1_n}^m(t) - \overline{L_{1_n}^m}\right) \left(L_{2_n}^m(t) - \overline{L_{2_n}^m}\right)^*, \tag{6.7}$$

$\overline{L_{i_n}^m} = \frac{1}{N_t} \sum_t L_{i_n}^m(t)$ (il s'agit de la moyenne temporelle du coefficient spectral), et

$$P(L_i)(n) = \sum_{m=-n}^{n} Cov(L_{i_n}^m, L_{i_n}^m).$$

Cette corrélation est en effet une autre façon de mesurer le niveau relatif de bruit, en fonction de l'échelle horizontale (Berre *et al.* , 2007). La figure 6.5 indique ainsi d'une part que le niveau de bruit est relativement élevé dans les petites échelles (conduisant à des corrélations spectrales proches de zéro). D'autre part, il apparaît que la corrélation spectrale est de l'ordre de 30 à 50% dans les grandes échelles.

Ces résultats suggèrent que l'atténuation partielle des grandes structures de portée par les ondelettes est pertinente, même si cette atténuation reste nettement moins forte (à juste titre) que pour les petites échelles.

Le filtrage opéré par les ondelettes semble donc justifié, et cela explique la bonne robustesse induite pour les cartes de portée. Compte tenu de ces résultats, il apparaît intéressant d'examiner la dynamique spatio-temporelle des portées à l'aide des ondelettes.

6.5 Dynamique spatio-temporelle des portées

6.5.1 Situation météorologique et évolution

Une étude de cas a été menée pour examiner les liens entre la dynamique des portées et celle de l'écoulement. L'étude couvre la période du 20 janvier à 06H UTC au 23 janvier à 00H UTC. Les figures 6.6 et 6.7 représentent (sous forme d'isolignes) le champ de pression réduite au niveau de la mer, issu des prévisions opérationnelles à six heures d'échéance du modèle Arpège.

Le domaine considéré couvre une partie de l'Amérique du Nord, de l'Atlantique Nord et de l'Europe. Au cours de cette période, trois centres de basse pression (notés respectivement L1, L2 et L3) apparaissent et évoluent, ainsi que trois centres de haute pression (H1, H2 et H3).

L1 correspond à une dépression qui se développe au nord-est des Etats-Unis, et qui est présente dès le 20 janvier à 06H UTC. L1 se déplace dans la direction du nord-est, vers la mer du Labrador. Cette dépression s'atténue à partir du 22 janvier à 18H UTC, pour complètement disparaître le 23 janvier.

L'anticyclone H1 est situé au nord-est des Açores, et il tend à s'affaiblir au cours de la période. L'anticyclone H2 se trouve sur le Canada, et il est relativement étendu. Ce centre se propage vers la côte orientale des Etats-Unis tout en s'atténuant, et il disparaît le 22 janvier.

Le 21 janvier à 06H UTC, H3 apparaît sous la forme d'un maximum local de pression de surface à l'ouest du Canada. H3 se renforce ensuite, tout en se propageant vers les Etats-Unis, pour former un grand centre anticyclonique qui s'étend de l'Oklahoma jusqu'au Labrador.

La dépression L2 apparaît le 21 janvier à 18H UTC au nord des Bermudes. Elle résulte du renforcement du thalweg associé à L1. Enfin, le 22 janvier à 12H UTC, la dépression L3 se développe au sud des Grands Lacs, et elle se déplace vers l'est.

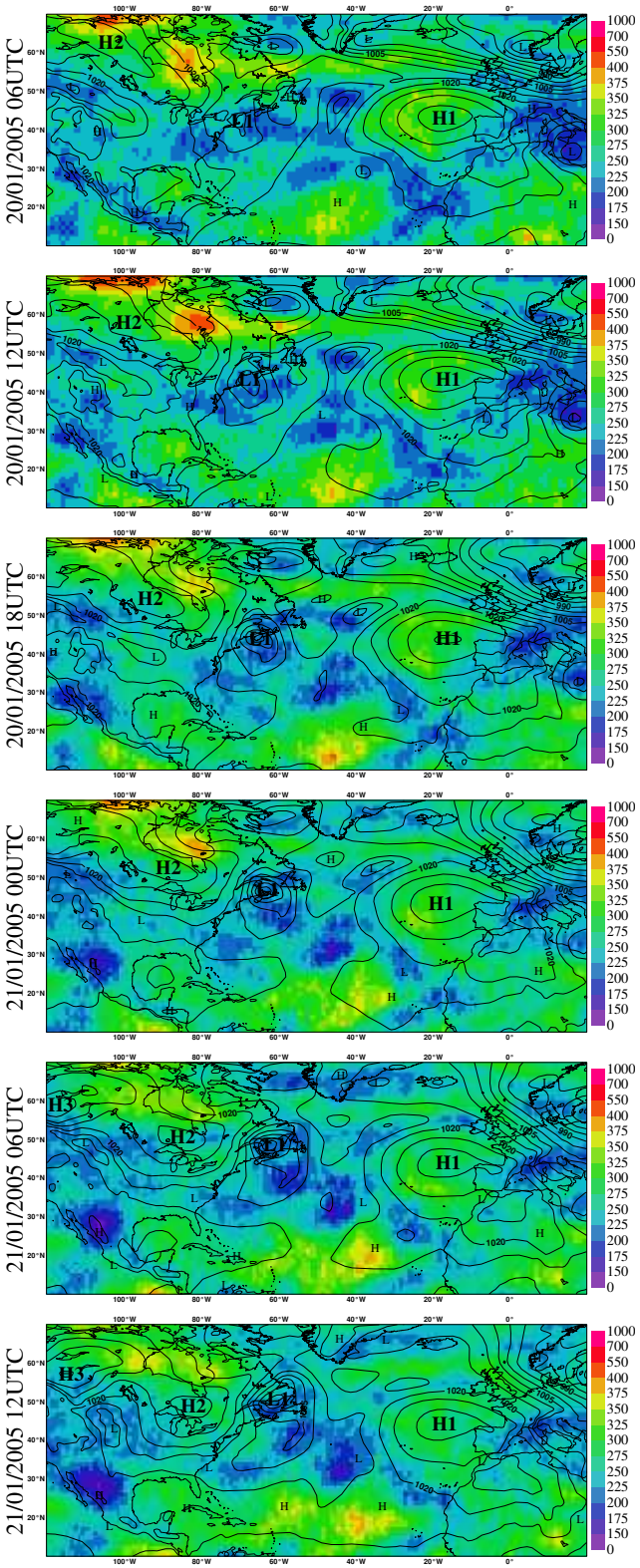


FIG. 6.6 – Evolution du champ des portées de la pression de surface (en couleur, avec un filtrage en ondelettes), et de la situation météorologique (pression réduite au niveau de la mer, avec des isolignes écartées de 5 hPa), du 20/01/2005 à 06H UTC au 21/01/2005 à 12H UTC.

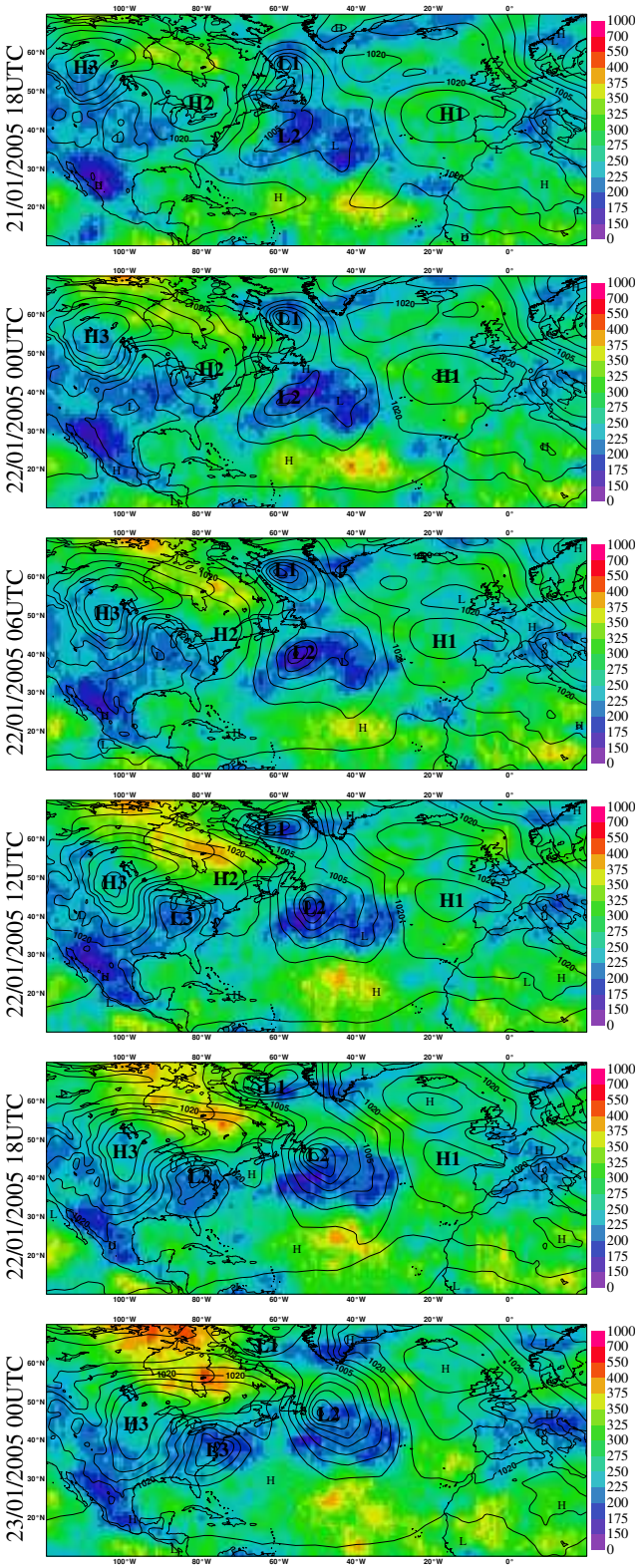


FIG. 6.7 – Evolution du champ des portées de la pression de surface (en couleur, avec un filtrage en ondelettes), et de la situation météorologique (pression réduite au niveau de la mer, avec des isolignes écartées de 5 hPa), du 21/01/2005 à 18H UTC au 23/01/2005 à 00H UTC.

6.5.2 Liens entre la portée locale et l'écoulement

Les portées obtenues avec le filtrage en ondelettes et un ensemble de six membres sont également représentées sur les figures 6.6 et 6.7. Les portées les plus courtes sont en bleu, tandis que les portées les plus longues sont en rouge. L'étude de cas présentée permet de mettre en évidence certaines liaisons entre la portée locale et la situation météorologique.

On peut noter en particulier la présence de portées courtes à proximité des dépressions. Par exemple, la dépression L1 est associée à une région de courtes portées (de l'ordre de 175 à $225\,km$), qui se déplace de façon cohérente avec le système dépressionnaire vers le nord-est. Il apparaît que les portées tendent à décroître au sud-est du minimum de pression de L1 (avec des valeurs de l'ordre de 150 à $175\,km$). Cette évolution correspond à l'émergence et au creusement de la dépression L2. On peut ainsi noter, par exemple le 22/01 à 06H UTC, que chacune des dépressions L1 et L2 est associée à une région spécifique de courtes portées. De façon similaire, l'apparition et la propagation de L3 peut être reliée à une diminution locale des portées.

De manière relativement symétrique, on peut noter que les valeurs de portée les plus grandes se trouvent dans les régions anticycloniques, telles que les systèmes H2 au Canada (avec des valeurs supérieures à $400\,km$ localement) et (dans une moindre mesure) H1 près de l'Europe.

6.5.3 Autres types de variations de portée

De façon plus générale, l'examen des figures 6.6 et 6.7 indique aussi que la dynamique des portées est relativement riche et complexe. En particulier, malgré les connexions visibles que l'on a décrites précédemment, les variations spatio-temporelles des portées ne semblent pas se réduire à une relation simple et systématique avec la situation météorologique locale.

Par exemple, un minimum local quasi-stationnaire de portées apparaît au niveau du Mexique. Ces courtes portées semblent liées à des effets orographiques, plutôt qu'à la présence d'un système dépressionnaire intense.

Un autre exemple correspond aux variations de portée au niveau de la dépression L2, le 23 janvier à 00H UTC. On peut noter que la partie méridionale est associée à des portées assez courtes, tandis que des portées plus grandes apparaissent dans la partie septentrionale. Cela illustre le fait que les variations de portée ne sont pas simplement dues à des variations du champ concerné (ici la pression de surface), mais qu'elles résultent plutôt d'un ensemble complexe de facteurs.

Ces résultats sont en partie attendus du reste, dans la mesure où la dynamique des erreurs d'ébauche résulte notamment de l'évolution des erreurs d'analyse, et en particulier de la densité des observations.

Ces propriétés dynamiques soutiennent l'idée de recourir à un ensemble d'assimilations en temps réel, pour simuler pleinement la dynamique des portées, plutôt que d'utiliser, par exemple, une paramétrisation simplifiée en fonction de la valeur du champ de l'ébauche.

6.6 Conclusions

Ce chapitre a permis de montrer que les ondelettes sont un outil pertinent et robuste pour la modélisation des variations spatiales et journalières des corrélations, dans le cadre du modèle global Arpège.

Sachant que les corrélations sont habituellement calculées sous la forme d'une moyenne sur plusieurs semaines, la sensibilité des cartes de portée à la taille de l'échantillon temporel a d'abord été étudiée. Il apparaît d'une part que les ondelettes permettent d'estimer les portées

du jour avec une robustesse plus grande que l'estimation ensembliste directe. D'autre part, une moyenne temporelle sur une période de 24 heures (associée à cinq réseaux successifs d'analyse) et sur un ensemble de six membres suffit pour que les ondelettes fournissent des cartes de portée relativement robustes (à savoir avec une cohérence supérieure à 80%).

Dans une deuxième partie, l'étude approfondie des spectres d'énergie des cartes de portée permet d'expliquer cette bonne robustesse des corrélations en ondelette. Il apparaît d'une part que les petites échelles sont davantage affectées par le bruit d'échantillonnage que les grandes échelles. D'autre part, l'évolution des spectres d'énergie en fonction de la taille de l'ensemble indique, au niveau du signal recherché, que l'amplitude des structures de grande échelle tend à prédominer par rapport à celle des structures de petite échelle. Ces deux caractéristiques expliquent l'efficacité des ondelettes : comme elles reposent sur des moyennes spatiales locales des statistiques, elles permettent d'atténuer le bruit de petite échelle et d'extraire le signal de (plus ou moins) grande échelle.

Compte tenu de ces résultats, les ondelettes ont ensuite été utilisées pour étudier la dynamique spatio-temporelle des portées sur une période de trois jours consécutifs. Cela a permis de mettre en évidence la richesse et la complexité de la dynamique des portées. Celle-ci reflète en partie l'influence de l'écoulement, avec par exemple des portées plus courtes près de certaines dépressions. La relation entre les portées et la situation météorologique locale s'avère cependant relativement complexe, du fait notamment de l'influence d'autres facteurs comme la densité des observations. La dynamique des portées ne semble donc pas pouvoir se réduire à une simple dépendance à l'écoulement décrite par l'ébauche. Il semble approprié de recourir plutôt à un ensemble d'assimilations pour décrire l'évolution des portées dans toute sa richesse, et de s'appuyer sur les propriétés de filtrage des ondelettes pour filtrer le bruit d'échantillonnage et extraire l'information utile fournie par l'ensemble.

Il reste à préciser la stratégie de mise en oeuvre opérationnelle des ondelettes dans le cadre d'un ensemble d'assimilations en temps réel. Les résultats de ce chapitre suggèrent qu'un échantillon de l'ordre de 30 éléments suffit pour avoir des corrélations robustes en ondelette. Si les moyens de calcul le permettent, un ensemble de 30 membres pourrait donc permettre d'avoir une représentation des corrélations caractéristiques de la situation en cours. Si l'ensemble disponible contient moins de membres (par exemple six), des stratégies basées sur des moyennes temporelles pourront être envisagées (par exemple avec une moyenne pondérée des statistiques sur les cinq derniers réseaux).

Une autre perspective naturelle concerne la représentation de la dynamique des corrélations verticales, qui est également accessible en principe avec les ondelettes.

Chapitre 7

Conclusions et perspectives

Ce travail de thèse s'est concentré sur l'utilisation des ondelettes, pour la représentation des variations spatio-temporelles des corrélations d'erreur d'ébauche. Différents résultats ont été obtenus.

Dans le chapitre 4, la capacité de l'approche ondelette à restituer des variations géographiques lisses des fonctions de corrélation a été étudiée. La représentation des covariances à l'aide de l'approche diagonale dans l'espace des ondelettes revient à moyenner localement les fonctions de covariance. Grâce à cette moyenne spatiale locale, l'estimation statistique est mieux échantillonnée que dans le cas d'une estimation purement locale des fonctions de covariance. De plus, du fait que cette moyenne spatiale est locale, la représentation des variations géographiques des portées reste possible (contrairement au cas homogène). De telles propriétés de filtrage sont particulièrement attractives dans le cas des covariances d'erreur estimées à partir d'un ensemble de prévisions. Ces propriétés relatives à la formulation ondelette ont été formellement explicitées et ont été illustrées expérimentalement dans deux cadres d'étude.

Le premier cadre expérimental considéré est celui du cercle. Sur ce domaine, une formulation hétérogène a été construite, permettant des variations géographiques de portées. Ainsi, la modélisation des fonctions de covariance à l'aide des ondelettes s'est avérée capable de représenter les fonctions locales de corrélation et leurs variations géographiques (diagnostiquées à l'aide du calcul des portées). Il a été montré que cette représentation était plus lisse et réaliste que dans le cas d'une simple utilisation d'un filtrage par produit de Schur. Cela est particulièrement marqué avec des ensembles de petite taille (de l'ordre de 10 à 30 membres). La formulation ondelette reste compétitive jusque une taille d'ensemble de l'ordre de 100.

Le deuxième cadre expérimental a été donné par un ensemble de prévisions globales issues d'un modèle de prévision numérique. L'approche ondelette s'avère représenter correctement les variations géographiques climatologiques des portées locales, qui sont associées à la dynamique de l'atmosphère et à l'hétérogénéité du réseau d'observations. De plus, un examen préliminaire des portées diagnostiquées pour une date donnée suggère que l'approche ondelette permet d'extraire des variations géographiques importantes, liées à la situation météorologique locale.

Ces résultats sont cohérents avec les propriétés de filtrage attendues des ondelettes, en terme de moyenne spatiale locale. Ils suggèrent que les ondelettes sont un outil prometteur pour estimer et représenter des covariances dépendantes de l'écoulement à partir d'un petit ensemble de prévisions.

Dans le chapitre 5, des approximations de la longueur de portée, définie par Daley, ont été

exposées et étudiées. En particulier, une estimation économique basée sur l'hypothèse gaussienne a été proposée. Un contexte 1D a permis de montrer que cette formule est capable de restituer des valeurs de portée et des variations géographiques réalistes.

D'autre part, une étude de la distribution d'échantillonnage de la portée a été menée, à la fois analytiquement et de manière expérimentale. Ainsi, il a été montré que l'estimation des portées à partir d'un ensemble (de taille N) est biaisée, avec un biais de signe positif (correspondant à une surestimation de la portée). De plus, ce biais tend vers zéro en $\mathcal{O}(N^{-1})$. Il a de plus été montré que ce biais est petit par rapport à l'écart type de l'erreur d'échantillonnage. Ce dernier converge également vers zéro, mais avec une vitesse en $\mathcal{O}(N^{-1/2})$.

En outre, l'examen de la variation géographique des portées et de leur spectre d'énergie indique que le bruit d'échantillonnage tend à être décorrélé spatialement (à la manière d'un bruit blanc). Cela conforte l'idée qu'une technique de moyenne spatiale, telle que celle basée sur les ondelettes, peut être intéressante à considérer pour filtrer spatialement le bruit d'échantillonnage.

Finalement, la formule de calcul de la portée de Belo Pereira et Berre a été comparée avec celle issue de l'approximation gaussienne, dans un cadre 2D sphérique à partir d'un ensemble de prévisions Arpège. Les portées ainsi diagnostiquées et leur variations géographiques sont similaires. Ainsi, l'approximation de la fonction de corrélation par une gaussienne apparaît comme étant raisonnable pour estimer la valeur de la portée.

Pour finir, le chapitre 6 a permis de montrer que les ondelettes sont un outil pertinent et robuste pour la modélisation des variations spatiales et journalières des corrélations, dans le cadre du modèle global Arpège.

Sachant que les corrélations sont habituellement calculées sous la forme d'une moyenne sur plusieurs semaines, la sensibilité des cartes de portée à la taille de l'échantillon temporel a d'abord été étudiée. Il apparaît d'une part que les ondelettes permettent d'estimer les portées du jour avec une robustesse plus grande que l'estimation ensembliste directe. D'autre part, une moyenne temporelle sur une période de 24 heures (associée à cinq réseaux successifs d'analyse) et sur un ensemble de six membres suffit pour que les ondelettes fournissent des cartes de portée relativement robustes (à savoir avec une cohérence supérieure à 80%).

Dans une deuxième partie, l'étude approfondie des spectres d'énergie des cartes de portée a permis d'expliquer cette bonne robustesse des corrélations en ondelette. Il apparaît d'une part que les petites échelles sont davantage affectées par le bruit d'échantillonnage que les grandes échelles. D'autre part, l'évolution des spectres d'énergie en fonction de la taille de l'ensemble indique, au niveau du signal recherché, que l'amplitude des structures de grande échelle tend à prédominer par rapport à celle des structures de petite échelle. Ces deux caractéristiques expliquent l'efficacité des ondelettes : comme elles reposent sur des moyennes spatiales locales des statistiques, elles permettent d'atténuer le bruit de petite échelle et d'extraire le signal de (plus ou moins) grande échelle.

Compte tenu de ces résultats, les ondelettes ont été utilisées pour étudier la dynamique spatio-temporelle des portées sur une période de trois jours consécutifs. Cela a permis de mettre en évidence la richesse et la complexité de la dynamique des portées. Celle-ci reflète en partie l'influence de l'écoulement, avec par exemple des portées plus courtes près de certaines dépressions. La relation entre les portées et la situation météorologique locale s'avère cependant relativement complexe, du fait notamment de l'influence d'autres facteurs comme la densité des observations. La dynamique des portées ne semble donc pas pouvoir se réduire à une simple dépendance à l'écoulement décrite par l'ébauche. Il semble approprié de recourir plutôt à un ensemble d'assimilations pour décrire l'évolution des portées dans toute sa richesse, et

de s'appuyer sur les propriétés des ondelettes pour filtrer le bruit d'échantillonnage et extraire l'information utile fournie par l'ensemble.

Il reste à préciser la stratégie de mise en oeuvre opérationnelle des ondelettes dans le cadre d'un ensemble d'assimilations en temps réel. Les résultats de ce chapitre suggèrent qu'un échantillon de l'ordre de 30 éléments suffit pour avoir des corrélations robustes en ondelette. Si les moyens de calcul le permettent, un ensemble de 30 membres pourrait donc permettre d'avoir une représentation des corrélations caractéristiques de la situation en cours. Si l'ensemble disponible contient moins de membres (par exemple six), des stratégies basées sur des moyennes temporelles pourront être envisagées (par exemple avec une moyenne pondérée des statistiques sur les cinq derniers réseaux).

Les perspectives ouvertes par ces travaux concernent notamment la représentation de la dynamique des corrélations verticales, qui est également accessible en principe avec les ondelettes. En effet, dans ce cadre les corrélations verticales sont alors non plus spécifiées en fonction du nombre d'onde, mais en fonction de l'échelle et de la position géographique. Il est alors possible de régionaliser ces dépendances verticales, ce qui permet d'envisager de les faire évoluer dans le temps. Une manière d'y parvenir est de procéder d'une façon similaire aux corrélations horizontales, *i.e.* en utilisant l'information d'un ensemble d'analyses perturbées.

Des études d'impact sur la qualité des analyses et des prévisions pourront être menées pour évaluer les effets de cette modélisation de la dynamique des corrélations. Cela peut notamment passer par des études de cas, portant sur des situations météorologiquement intenses (tempêtes, cyclones tropicaux, convection,...).

Une autre extension naturelle porte sur les modélisations des covariances croisées et des opérateurs de balance associés. Par ailleurs, compte tenu du caractère isotrope des ondelettes utilisées dans cette thèse, des travaux sur des ondelettes plus anisotropes constituent un prolongement naturel des recherches effectuées. Différent travaux pourront permettre d'approfondir une telle recherche. Ainsi, il est possible de s'inspirer de l'analyse à l'aide de fonctions polarisées. Une autre voie possible est d'envisager des transformations ondelettes décrites en point de grille plutôt qu'en spectral. Naturellement, ces approches devront prendre en compte les contraintes de temps associées à l'opérationnel, ainsi que celles plus directes de l'implémentation dans un schéma d'assimilation déjà existant.

Les études présentées ici ont porté sur des ondelettes sphériques en vue d'une assimilation couvrant tout le globe. Des applications analogues sont naturellement envisagées pour une assimilation régionale à échelle fine, compte tenu de l'existence d'ondelettes adaptées à la géométrie des modèles à aire limitée. Il existe pour ces applications des ondelettes plus adaptées (ondelettes orthogonales, paquets d'ondelettes, ondelettes complexes,...) dont le potentiel n'a pas encore été exploité. De nouveaux outils seront aussi à rechercher dans d'autres disciplines, par exemple le traitement des images 2D.

Enfin la meilleure connaissance des caractéristiques statistiques de la portée permet également d'envisager d'autres utilisations que celles présentées ici. En particulier, cette connaissance peut servir à mieux estimer la portée réelle, puis à utiliser cette information pour estimer de manière objective les paramètres intervenant dans d'autres modèles de fonctions de covariance. Par exemple, dans la modélisation des covariances d'erreur d'ébauche basée sur un opérateur de diffusion, il est possible de déduire le tenseur local de diffusion à partir des diagnostics de portée locale. Ce type de prolongement a déjà commencé à être étudié : des collaborations avec des chercheurs du CERFACS ont permis de diagnostiquer les portées dans le schéma d'assimilation variationnelle d'OPA-Var. L'estimation du tenseur local de diffusion a

également été mise en oeuvre pour le schéma d'assimilation MOCAGE-PALM (Pannekoucke et Massart, 2008).

Références

Auger L. and Tangborn A. 2004. *A wavelet-based reduced rank kalman filter for assimilation of stratospheric chemical tracer observations*. *Monthly Weather Review*, **132**, 1220–1237.

Belo Pereira M. and Berre L. 2006. *The use of an Ensemble approach to study the Background Error Covariances in a Global NWP model*. *Mon. Wea. Rev.*, **134**, 2466–2489

Bergé P., Pomeau I. and Vidal C. 1997. *L'ordre dans le chaos*. *Hermann*, p353.

Berre L., Pannekoucke O., Desroziers G., Stefanescu S., Chapnik B. and Raynaud L. 2007. *A variational assimilation ensemble and the spatial filtering of its error covariances : increase of sample size by local spatial averaging*. Proceedings of Workshop on Flow-dependent Aspects of Data Assimilation, 11-13 June 2007, 151–168. ECMWF, Research Department, Shinfield Park, Reading, England.

Bouttier F. 1993. *The dynamics of error covariances in a barotropic model*. *Tellus*, **45A**, 408–423.

Bouttier F. 1994.*A dynamical estimation of error covariances in an assimilation system*. *Mon. Wea. Rev.*, **122**, 2376–2390.

Bouttier F. 1994b. *Sur la prévision de la qualité des prévisions météorologiques*. PhD thesis, Université Paul Sabatier, Toulouse, 240pp.

Buehner M. and Charron M. 2006. *Spectral and spatial localization of background error correlations for data assimilation*. Submitted to *Q.J.R. Meteorol. Soc.*

Burke-Hubbard B. 2000. *Ondes et ondelettes. La Saga d'un outil mathématique*. *Pour la science*. pp236

Chapnik B., Desroziers G., Rabier F. and Talagrand O. 2006. *Diagnosis and tuning of observational error statistics in a quasi-operational data assimilation setting*. *Q. J. R. Meteorol. Soc.*, **132**, 543–565.

Coifman R. and Wickerhauser W. 1992. *Entropy-based algorithms for best-basis selection*. *IEEE Trans. Inform. Theory*, **38**, 713–718.

Courtier P. 1997. *Dual formulation of four-dimensional variational data assimilation.Quart. J.R. Meteor. Soc.*, **123**, 2449–2461.

Courtier P. and Geleyn J.F. 1988. *A global numerical weather prediction model with variable resolution : Application to the shallow-water equations.Q.J.R. Meteorol. Soc.*, **114**, 1321–1346.

Courtier P., Andersson E., Heckley W., Pailleux J., Vasiljević D., Hamrud M., Hollingsworth A., Rabier F. and Fisher M. 1998. *The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I : Formulation*. *Q.J.R. Meteorol. Soc.*, **124**, 1783–1807.

Daley, R. 1991. *Atmospheric Data Analysis*. Cambridge University Press, 471pp.

Daubechies I. 1992. *Ten lectures on Wavelets. CBMS-NSF regional conference series in applied mathematics, SIAM, 357pp.*

Deckmyn A. and Berre L. 2005. *A wavelet approach to representing background error covariances in a LAM.* Mon. Wea. Rev.*, **133**, 1279-1294*

Dee D. 1995. On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.*, **123**, 1128–1145.

Derber J. and Bouttier F. 1999. *A reformulation of the background error covariance in the ECMWF global data assimilation system.* Tellus*, **51A**, 195–221*

Desroziers G. 1997. *A Coordinate Change for Data Assimilation in Spherical Geometry of Frontal Structures*. *Mon. Wea. Rev.*, **125**, 3030–3038.

Desroziers G. and Ivanov S. 2001. *Diagnosis and adaptive tuning of information error parameters in a variational assimilation*. *Q. J. R. Meteorol. Soc.*, **127**, 1433–1452.

Donoho D. and Johnstone I. 1994. *Ideal spatial adaptation by wavelet shrinkage*. *Biometrika*, **81**, 425–455.

Donoho D., Mallat S., von Sachs R. and Samuelides S. 2003. *Signal and covariance estimation with macrotiles*. *IEEE trans. signal process*, **52**, 614–627.

Evensen G. 1994. *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error.* J. Geophys. Res.*, **99** (C5) 10 143–10 162*

Farge M. 1992. *Wavelet transforms and their applications to turbulence*. *Annual Review of Fluid Mechanics*, **24**, 395–457.

Fisher M. and Courtier P. 1995. *Estimating the covariance matrices of analysis and forecast error in variational data assimilation. ECMWF Technical Memorandum, **220**, 29pp.*

Fisher M. and Anderson E. 2001. *Developments in 4D-Var and Kalman Filtering*. ECMWF Technical Memorandum, **347**, 38pp.

Fisher M. 2003. *Background error covariance modelling. Processing of the ECMWF Seminar on "Recent developments in data assimilation for atmosphere and ocean", Reading, 8–12 September 2003, 45–63.*

Fisher M. 2004. *Generalized frames on the sphere, with application to the background error covariance modelling. Processing of the ECMWF Seminar on "Developments in Numerical Methods for Atmospheric and Ocean Modelling", Reading, 6–10 September 2004, 87–101.*

Freeden W. and Windheuser U. 1996. *Spherical wavelet transform and its discretization*. *Advanced in Computational Mathematics*, **5**, 51–94.

Freeden W. and Schreiner S. 1998. *Orthogonal and non-orthogonal multiresolution analysis, scale discrete and exact fully discrete wavelet transform on the sphere*. *Constructive Approximation*, **14**, 493–515.

Gaspari G. and Cohn S. 1999. *Construction of correlation functions in two and three dimensions.* Q.J.R. Meteorol. Soc.*, **125**, 723–757.*

Gustafsson N., Berre L., Hörnquist S., Huang X-Y., Lindskog M., Navasques B., Mogensen KS. and Thorsteinsson S. 2001. *Three-dimensional variational data assimilation for a limited area model.* Tellus*, **53A**, 425–446.*

Hollingsworth A. 1987. *Short- and medium-range numerical weather prediction*. Collection of papers presented at the WMO/IUGG symposium, Tokyo, 4–8 August 1986.

Holschneider M. 1990. *Wavelet analysis on the circle*. *Journal of Mathematical Physics*, **31**, 39–44.

Houtekamer P.L., Lefaivre L., Derome J., Ritchie H. and Mitchell H.L. 1996. *A system simulation approach to ensemble prediction.* Mon. Wea. Rev.*, **124**, 1225–1242.*

Houtekamer P.L. and Mitchell H.L. 2001. *A sequential ensemble Kalman filter for Atmospheric Data Assimilation.* Mon. Wea. Rev.*, **129**, 123–137.*

Houtekamer P. and Mitchell H. 2005. *Ensemble Kalman filtering*. *Q. J. R. Meteorol. Soc.*, **131**, 3269–3289.

Ingleby B. 2001. *The statistical structure of forecast errors and its representation in The Met.*

Office Global Model. *Q.J.R. Meteorol. Soc.*, **124**, 1783–1807.

Kalman R. 1960. *A New Approach to Linear Filtering and Prediction Problems*, Transactions of the ASME–Journal of Basic Engineering.**82** : 35–45.

Lönnberg P. 1988. *Developpements in the ECMWF analysis scheme*. Proc. ECMWF Seminar on "Data assimilation and the use of satellite data", Reading, 5–9 September 1988, 75–120.

Lorenc A. 1988. *Optimal nonlinear objective analysis. Quart. J.R. Meteor. Soc.*, **114**, 205–240.

Lorenc A. 2003. *The potential of ensemble Kalman filter for NWP – a comparison with 4D-Var*. *Q.J.R. Meteorol. Soc.*, **129**, 3183–3203.

Lopez P. 2002. *Implementation and validation of a new prognostic large-scale cloud and precipitation scheme for climate and data-assimilation purposes*. *Q.J.R. Meteorol. Soc.*, **128**, 229–257.

Mallat S. 1989. *A Theory for Multiresolution Signal Decomposition : The Wavelet Representation*. *IEEE Transactions on Pattern Analysis and Machine*. **11**, 674–693.

Mallat S., Papanicolaou G. and Zhang Z. 1998 *Adaptive covariance estimation of locally stationary processes*. *Annals of Statistics*, **26**, 1–47.

Mallat S. 2001. *Une exploration des signaux en ondelettes*. *Édition de l'École Polytechnique*. pp636.

Manneville P. 2004. *Instabilités, chaos et turbulence*. *Édition École Polytechnique*, p352.

Pannekoucke O, Berre L. and Desroziers G. 2007. *Filtering properties of wavelets for the local background error correlations*. *Q.J.R. Meteorol. Soc.*, **133**, 363–379.

Pannekoucke O, Berre L. and Desroziers G. 2008. *Background error correlation length-scale estimates and their sampling statistics*. Submitted to *Quarterly Journal of the Royal Meteorological Society*.

Pannekoucke O., Berre L. and Desroziers G. 2007c. *The spatial structure and dynamics of error correlation length-scales with a wavelet filtering approach*. To be submited.

Pannekoucke O. and Massart S. 2008. *Estimation of the local diffusion tensor and normalization for heterogeneous correlation modeling using a diffusion equation*. *Quarterly Journal of the Royal Meteorological Society* Accepted with minor revisions.

Parrish D. and Derber J. 1992. *The national meteorological center spectral statistical interpolation analysis system*. *Mon. Wea. Rev.*, **120**, 1747–1763.

Rabier F., Jarvinen H., Klinker E., Mahfouf J.F. and Simmons A. 2000. *The ECMWF operational implementation of four-dimensional variational assimilation. I : Experimental results with simplified physics*. *Q.J.R. Meteorol. Soc.*, **126**, 1148–1170.

Talagrand O. 2002. *Data assimilation for the Earth system*. *Kluwer Academic Publisher*, 2002.

Tangborn A. 2004. *Wavelet approximation of error covariance propagation in data assimilation*. *Tellus A*,**56**, 16–28.

Tarantola A. 2005. *Inverse problem theory and model parameter estimation*. *SIAM*, 352p.

Thépaut J-N., Courtier P., Belaud G. and Lemaître G. 1995. *Dynamical structure functions in a four-dimensional variational assimilation : a case study*. *Q.J.R. Meteorol. Soc.*, **122**, 535–561.

Torrésani B. 1995. *Analyse continue par ondelettes*. *EDP Sciences*, p239.

Veersé F. and Thépaut J-N. 1998. *Multiple-truncation incremental approach for four-dimensional variational data assimilation*. *Q.J.R. Meteorol. Soc.*, **124**, 1889–1908.

Weaver A. and Courtier P. 2001. *Correlation modelling on the sphere using a generalized diffusion equation*. *Quart. J. Roy. Meteor. Soc.*, **127**, 1815–1846.

McWilliams J.C. 1984. *The emergence of isolated coherent vorticies in turbulent flow*. *J. Fluid. Mech.*, **146**, 21–43.

Annexes

Annexe A

Version originale du chapitre 4 (Pannekoucke *et al.* , 2007)

QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY
Q. J. R. Meteorol. Soc. **133**: 363–377 (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/qj.133

Filtering properties of wavelets for the local background error correlations

O. Pannekoucke*, L. Berre and G. Desroziers
GAME/CNRM (Météo-France, CNRS), Toulouse, France

Abstract:
Background error covariances can be estimated from an ensemble of forecast differences. The finite size of the ensemble induces a sampling noise in the calculated statistics. It is shown formally that a wavelet diagonal approach amounts to locally averaging the correlations, and its ability to spatially filter this sampling noise is thus investigated experimentally.
This is first studied in a simple analytical one dimensional framework. The capacity of a wavelet diagonal approach to model the scale variations over the domain is illustrated. Moreover, the sampling noise appears to be better filtered than when only using a Schur filter, in particular for small ensembles.
The filtering properties are then illustrated for an ensemble of Météo-France Arpège forecasts. This is done both for the "time-averaged correlations", and for the "correlations of the day". It is shown that the wavelets are able to extract some length scale variations that are related to the meteorological situation.

KEY WORDS Spherical wavelet; wavelet on the circle; background error covariance; assimilation ensemble; sampling noise; Ensemble Kalman Filter.

Received 20 December 2005; Revised 22 November 2006; Accepted 26 November 2006

1 Introduction

Most data assimilation schemes seek to provide an optimal combination of observations and of a background given by a short-term forecast. The optimal analysis is basically derived from statistical estimation theory. In such a theory, the two sets of information are associated with covariance matrices corresponding to their respective errors. The error covariance matrices determine the respective weights given to each piece of information in the analysis. However, the correct specification of those statistics remains a major challenge in data assimilation systems.

The estimation of background error covariances is a particularly difficult problem since in operational practice the background state is a vector of dimension $10^5 - 10^7$. In that case, it is not only intractable to handle such a huge corresponding error covariance matrix, but it is also impossible to specify it exactly, since there is a lack of available statistical information (Dee 1995).

To overcome these difficulties, a statistical model for the background error covariances has to be defined. Such a model often relies on the hypothesis that the background error correlations are homogeneous and isotropic (Gaspari and Cohn 1999). This assumption is equivalent to considering that the background error correlation matrix is diagonal in spectral space (Courtier et al 1998) and thus facilitates the representation of background error statistics.

One technique for specifying the background error covariance matrix is to use an ensemble of assimilations, obtained by a perturbation of observations and of the background (Houtekamer et al. 1996). This procedure has recently been applied at ECMWF and Météo-France for specifying the stationary component of background error covariances (Fisher 2003; Belo Pereira and Berre 2006). In that case, the covariances are computed over several weeks and the hypothesis of homogeneity is often assumed. Such an approach is also partly related to the Ensemble Kalman Filter (EnKF), originally proposed by Evensen (1994), where flow-dependent covariances are calculated from the ensemble.

The hypotheses of homogeneity and isotropy, are known to be rather crude representations for the "real" error structures. Lönnberg (1988) suggested that horizontal and vertical correlations vary geographically: horizontal scales tend to be broader in the tropics than at high latitudes because of atmospheric dynamics (Ingleby 2001). Bouttier (1993, 1994) also showed that correlation scales depend on the meteorological situation and on data density.

Using an ensemble of assimilations, Belo Pereira and Berre (2006) have shown such heterogeneities and

*Correspondence to: Météo-France CNRM/GMAP, 42 av. G. Coriolis, 31057 Toulouse Cedex France. e-mail: olivier.pannekoucke@meteo.fr

anisotropies in the background error correlations by using a new economical algorithm for estimating the correlation length scales.

The EnKF is one approach for obtaining heterogeneous and flow-dependent background error correlations. However, due to the relatively small sampling size, covariances are noisy and have to be filtered by an additional treatment. A Schur product (Houtekamer and Mitchell 2001) is generally applied to the raw statistics. The use of the Schur product in EnKF has been discussed by Lorenc (2003).

From a different point of view, Fisher (2003) has recently introduced the idea to use wavelets on the sphere to improve the representation of background error correlations, and in particular to allow some heterogeneity in the description of those errors. Such a formulation is now operationally applied at ECMWF to represent stationary but heterogeneous correlations, with for example larger correlations in the Tropics. A similar approach has been considered by Deckmyn and Berre (2005) for a limited area model.

The wavelet representation of background error covariances implemented at ECMWF has been obtained by using an ensemble of analyses over several weeks, which is expected to provide valid statistics. Moreover, there is scope to combine the use of wavelets and ensembles in order to obtain at the same time heterogeneous and flow-dependent background error correlations.

The aim of the paper is to show that wavelets provide an effective tool to allow some variability in the correlations, but also to filter the noise due to the small size of an ensemble. These properties of the wavelet formulation are in particular investigated by using the diagnostic of the local correlation length scale proposed by Belo Pereira and Berre (2006). It may be also mentioned that some analogous filtering effects are under investigation at the Meteorological Service of Canada (Buehner and Charron 2006), through spectral and spatial localization.

The structure of the paper is the following. In section 2, we explain that a wavelet diagonal approach amounts to applying a local spatial averaging of covariance functions. This allows the sample size to be increased and to preserve the representation of geographical variations. The ability of wavelets to extract useful information from a small ensemble is first discussed in section 3 for a toy analysis problem on a circle. Section 4 shows the application of the same wavelet representation in 2D on the sphere, with actual background errors provided by an ensemble of forecasts from the Météo-France Arpège system. Results are produced both over a long period, and on a single date with only a few members. Conclusions are given in section 5.

2 Wavelet spatial averaging of covariance functions

2.1 Local covariance functions

For the sake of simplicity, we will consider a 1D cyclic domain in this section. Derivations are also valid in a 2D cyclic domain, by considering horizontal positions as vectors with two components (x and y). They can be also generalized easily to 2D spherical domains and to 3D contexts (by including the appropriate metrical terms in the formulae).

The separation s between two positions x and x" is defined as the difference $s = x" - x$. Note that s can be positive or negative: the absolute value $|s|$ is the separation distance, while the sign of s corresponds to the orientation of the separation. (In a 2D cyclic context, the separation direction is given by the argument of s, when seeing s as a complex number: $s = s_x + is_y = |s|\, exp(i\, arg(s))$, where s_x and s_y are the x and y components of s).

The local error covariance function $f^x(s)$ at a reference point x is often calculated from an ensemble of N_e forecast differences ε, according to the following equation (in a horizontal context) :

$$f^x(s) = \overline{\varepsilon(x)\varepsilon(x+s)} = \frac{1}{N_e}\sum_k \varepsilon(x,k)\varepsilon(x+s,k),$$

where $\varepsilon(x,k)$ is a forecast error realization at position x, s is the separation value between the two considered points, k is the ensemble member index, and the overline is the ensemble average. This ensemble average is thus calculated over N_e members of the ensemble.

The sampling size is thus equal to N_e only. A first problem is that the finite size of the sample induces a noise, which often renders necessary the use of e.g. a Schur filter (Houtekamer and Mitchell 2001). A second important problem is that the sample covariance matrix will be rank deficient if the sample size is too small. This will limit the analysis increments to lie in a low-dimensional subspace in the analysis. In the remainder of the study, we will focus on the first problem (sampling noise).

2.2 Spectral diagonal approach: a global spatial averaging

It is possible to represent the background error covariance matrix $\boldsymbol{B}$ in spectral space (e.g. Courtier et al 1998). Under the assumption of horizontal homogeneity, the spectral covariance matrix is simply a diagonal matrix (or block-diagonal in a three-dimensional context, with vertical covariances in the diagonal blocks). The diagonal of $\boldsymbol{B}$ contains the variances of the error spectral coefficients.

Equivalently, it can be shown (see appendix A) that this spectral diagonal approach amounts to calculating $\frac{1}{N_g}\sum_{x'} f^{x'}(s)$, which is the global spatial average of the covariance functions (N_g is the number of gridpoints in the domain). The resulting local covariance functions, noted $f^x_{\boldsymbol{S}}(s)$, where the subscript $\boldsymbol{S}$ refers to the spectral diagonal approach, are all equal to this global spatial average:

$$f^x_{\boldsymbol{S}}(s) = \frac{1}{N_g}\sum_{x'} f^{x'}(s),$$

or equivalently

$$f^x_{\boldsymbol{S}}(s) = \frac{1}{N_g\, N_e}\sum_{x'}\sum_k \varepsilon(x',k)\varepsilon(x'+s,k).$$

DOI: 10.1002/qj

The globally averaged covariance function is thus estimated as an average over a number of pairs of error realizations, which is equal to $\mathcal{N} = N_g N_e$ (instead of $\mathcal{N} = N_e$ when estimating the local covariance functions).

Such a huge increase of sample size has an obvious counterpart: the *global spatial average* does not allow one to represent any geographical variations. Therefore, one may wonder if it is possible to consider a *local spatial average*, in order both to increase the sampling size and to represent some geographical variations.

2.3 Wavelet diagonal approach: a set of local spatial averages

The direct and inverse wavelet transforms, which are used by Fisher (2003), are defined as follows: $\hat{\boldsymbol{\varepsilon}}_j = \boldsymbol{\varepsilon} \otimes \boldsymbol{\psi}_j$ and $\boldsymbol{\varepsilon} = \sum_j \hat{\boldsymbol{\varepsilon}}_j \otimes \boldsymbol{\psi}_j$, where $\otimes$ is the convolution product in physical space for both expressions, and $\boldsymbol{\psi}_j$ are radial band-pass functions for different scales j (see Courtier et al 1998 for the properties of the convolution with radial functions on the sphere). It may be mentioned that such functions $\boldsymbol{\psi}_j$ are not orthogonal, and different from traditional wavelets, which instead are orthogonal with respect to integer dilation and translation on a regular lattice.

The functions $\hat{\boldsymbol{\varepsilon}}_j$ and $\boldsymbol{\psi}_j$ can be represented in grid-point space. Some examples of wavelet functions $\boldsymbol{\psi}_j$ are shown in a one dimensional framework (Fig. 1) and can be compared with spectral functions. Note that each wavelet function has both a specific position and a specific scale. The associated coefficients $\hat{\varepsilon}_{j,x_j(i)}$ of the transformed error field correspond to the error values at scale j and at position $x_j(i)$ on a grid whose resolution depends on j (with $i = 1, N_x(j)$).

In a horizontal framework, the wavelet diagonal approach for $\boldsymbol{B}$ consists in calculating variances of these wavelet coefficients $\hat{\varepsilon}_{j,x_j(i)}$. As each wavelet function contains information both on position and scale, the wavelet variances contain information on the local shape of the covariance functions.

Moreover, it can be shown (see appendix A) that this wavelet diagonal approach amounts to computing a set of weighted local spatial averages of the covariance functions. The resulting local covariance functions, noted $f^x_{\boldsymbol{W}}(s)$ with the subscript $\boldsymbol{W}$ refering to the wavelet diagonal approach, can be expressed as follows:

$$f^x_{\boldsymbol{W}}(s) = \sum_{x',s'} f^{x'}(s') \Phi^{x,s}(x',s'),$$

where $\Phi^{(x,s)}(x',s')$ is defined by

$$\sum_j \sum_{i=1}^{N_x(j)} \psi_j(x' - x_j(i))\, \psi_j(x' - x_j(i) + s')\, \psi_j(x - x_j(i))\, \psi_j(x - x_j(i) + s)$$

and can be seen as a weighting coefficient in the calculation of the spatial average.

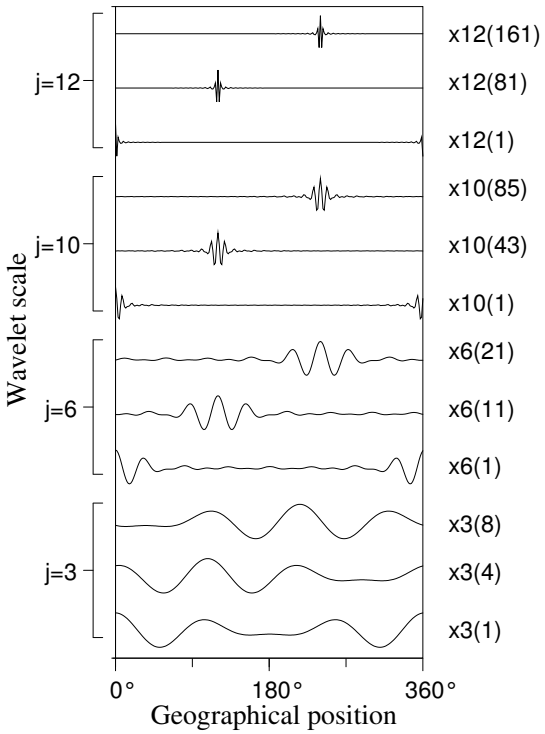


Figure 1. Some wavelet functions $\psi_j(x - x_j(i))$ for different scales j and different points $x_j(i)$.

As expected, $\Phi^{x,s}(x',s')$ will give more weight to positions x' in the neighbourhood of x, and to separation values s' that are close to s. Figure (2) represents two examples of function $\Phi^{x,s}(x',s')$, for $x = 121$ and $s = s' = 20$, and for two choices of wavelet bands. The larger weight given to position x' close to x is illustrated by the larger values of $\Phi^{x,s}(x',s')$ when x' is near x.

As in the spectral approach, the implied functions $f^x_{\boldsymbol{W}}(s)$ are the result of a spatial average. The spatial sample size is nevertheless likely to be smaller than in the spectral case, because the weighting functions $\Phi^{x,s}(x',s')$ have values close to zero except in the neighbourhood of x (as illustrated in fig. 2).

The two examples in fig. (2) differ by the choice of wavelet bandwidths. It may thus be noticed that the spatial average will tend to be more localized when the bandwidth is larger. This is related to the fact that the bandwidth determines the trade-off between spatial and spectral resolutions, as will be further discussed in section 2(e).

It may also be mentioned that the calculation, representation and spatial filtering of the matrix $\boldsymbol{B}$ remain efficient with a wavelet diagonal approach. It only requires the calculation of a diagonal matrix, which contains the wavelet variances. This is much cheaper than calculating a full gridpoint covariance matrix, and then applying a spatial averaging operator.

To summarize, the wavelet approach is similar to the spectral approach, in the sense that the local covariance functions are averaged spatially. This potentially allows one to reduce the level of sampling noise, compared to the

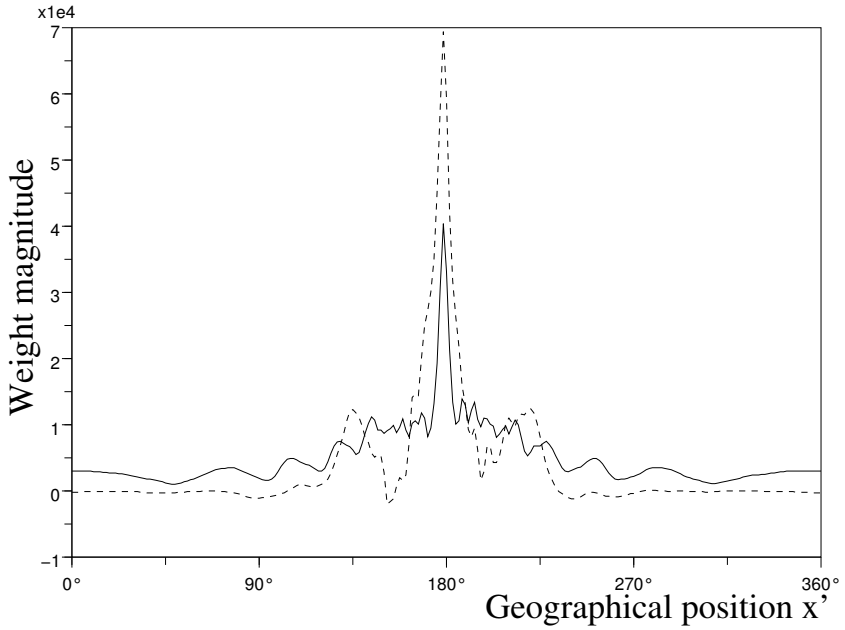


Figure 2. Representation of $\Phi^{x,s}(x',s')$ for $x = 121$ (corresponding to 180 °) and $s = s' = 20$ (corresponding to 30 °) and for two different choices of wavelet bands, as defined by the cutoff wave numbers $\{N_j\}$ (see section 2(e)) : a set of relatively tight bands $\{N_j\} = \{0, 1, 2, 3, 5, 7, 10, 15, 21, 30, 42, 63, 120\}$ (solid line) and a set of relatively wide bands $\{N_j\} = \{0, 4, 8, 12, 20, 28, 40, 60, 84, 120\}$ (dashed line).

estimation of the local covariance functions. Moreover, as the spatial averaging is local instead of global, it remains possible to represent some geographical variations of the covariance functions.

These two features (spatial filtering and geographical variations) will be studied experimentally in two different frameworks, in sections 3 and 4 respectively.

2.4 Isotropy of wavelets and of covariance averaging

As illustrated in fig. (1) on the circle and in Fisher (2004) on the sphere, the band-pass functions ψ_j are radial. This implies that the corresponding local averaging of covariances tends to be isotropic. In other words, covariances along different separation directions are averaged together, for a given separation distance.

On the one hand, this may be seen as a drawback, as it prohibits the representation of possible anisotropies. This limitation has been illustrated and discussed by Deckmyn and Berre (2005) with Meyer wavelets. On the other hand, averaging over several directions allows the sample size to be further increased, which can reduce the amplitude of sampling error. In other words, if the covariances are nearly isotropic, this isotropic averaging may rather be beneficial.

The balance between these pros and cons will thus depend on the degree of anisotropy of the actual covariances and on the available ensemble size. Another related point is that part of the covariance anisotropies arises from the geographical variations of the background error standard deviations. The latter can be represented in gridpoint space, while modelling correlations in wavelet space, as in Fisher (2003).

2.5 Details of the wavelet diagonal approach

The concept of local averaging of covariance functions has been introduced in the previous section. This concept can also be considered for correlation functions, by applying the equations to the background error normalized by the gridpoint standard deviations, namely $\varepsilon'(x,k) = \varepsilon(x,k)/\sigma_b(x)$, with $\sigma_b(x) = (\frac{1}{N_e}\sum_k \varepsilon(x,k)^2)^{1/2}$. This will be the approach used in the remainder of the paper, in order to have a similar covariance formulation as Fisher (2003) and Deckmyn and Berre (2005). The formulation of $\boldsymbol{B}$ is often determined by the design of $\boldsymbol{B}^{-1/2}$, which is detailed in appendix B. Thus, in a horizontal context, the square root of the wavelet-modelled gridpoint covariance matrix can be written

$$\boldsymbol{B}_w^{1/2} = \boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_s \boldsymbol{W}^{-1} \boldsymbol{D}_w^{1/2}, \tag{1}$$

where $\boldsymbol{\Sigma}_g$ is a diagonal matrix of gridpoint standard deviations, $\boldsymbol{\Sigma}_s$ corresponds to a multiplication by spectral standard deviations, and $\boldsymbol{D}_w^{1/2}$ is a diagonal matrix of wavelet standard deviations. Matrices $\boldsymbol{W}$ and $\boldsymbol{W}^{-1}$ correspond respectively to the direct and inverse wavelet transforms (note that $\boldsymbol{W}$ is a rectangular matrix, and that $\boldsymbol{W}^{-1}$ is the left inverse of $\boldsymbol{W}$).

The particular wavelet functions ψ_j introduced by Fisher (2003) on the sphere are band-limited and defined in spectral space as follows. For $(N_j)_{j\in[0,J]}$, with $N_j < N_{j+1}$, the spectral coefficients of functions ψ_j are given by, for $j \neq 0$ (n being the total wave number in the spherical case): $\frac{1}{\sqrt{2n+1}}\check{\psi}_{j,n} = \sqrt{\frac{n-N_{j-1}}{N_j-N_{j-1}}}$ for $N_{j-1} \leq n < N_j$, $\sqrt{\frac{N_{j+1}-n}{N_{j+1}-N_j}}$ for $N_j \leq n < N_{j+1}$, and 0 otherwise. For $j = 0$, the definition is the same, except that the range

DOI: 10.1002/qj

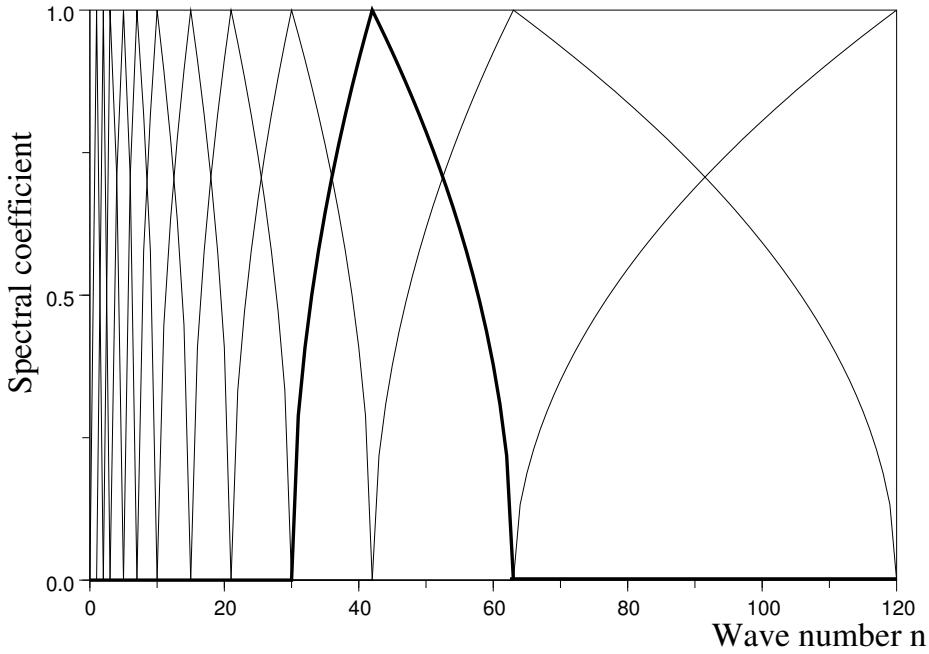


Figure 3. Spectral coefficients of wavelet functions $\psi_j(x)$ for different scales j (there is one curve for each function) and truncation $T = 120$. The spectrum associated with a particular $\psi_j(x)$ function ($j = 10$) is shown in bold.

$N_{j-1} \leq n < N_j$ is replaced by $0 \leq n < N_0$, for which $\frac{1}{\sqrt{2n+1}}\check{\psi}_{j,n} = 1$.

It is possible to define equivalent wavelets in a one dimensional Fourier space. The corresponding Fourier spectral coefficients $\check{\psi}_{j,n}$ are represented in Fig. 3 , for the following set $\{N_j\} = \{0, 1, 2, 3, 5, 7, 10, 15, 21, 30, 42, 63, 120\}$. This set has been chosen in order to be similar to Fisher (2003). Ways to define an optimized choice for this set may be explored in future studies. In particular, as discussed by Fisher and Andersson (2001), it is the choice of the bandwidths which determines the trade-off between spectral and spatial resolution. When the bands are broader, the spectral variations are smaller, but the geographical variations are allowed to be larger. This is connected to the discussion of Figure 2 in section 2(c). The spatial averaging is more localized when the bands are broader, which allows more geographical variations to be represented.

It may also be mentioned that each wavelet field $\hat{\boldsymbol{\varepsilon}}_j$ can be represented exactly on a low-resolution grid, which corresponds to a truncation equal to $T_j = min(N_{j+1}, T)$, T being the maximum truncation, associated to the original full resolution grid ($T = 120$ in the example above). The direct wavelet transform is thus applied as follows:

$$\hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} . \\ \hat{\boldsymbol{\varepsilon}}_j \\ . \end{pmatrix} = \boldsymbol{W}\boldsymbol{\varepsilon} = \begin{pmatrix} . \\ \boldsymbol{G}_j\check{\boldsymbol{\Psi}}_j\boldsymbol{S}_j \\ . \end{pmatrix} \boldsymbol{\varepsilon},$$

where $\check{\boldsymbol{\Psi}}_j$ is the diagonal matrix containing the $\check{\psi}_{j,n}$ spectral coefficients, $\boldsymbol{S}_j$ is the spectral transform associated with truncation T_j, and $\boldsymbol{G}_j$ is the corresponding inverse transform onto a grid that corresponds to truncation T_j. The inverse wavelet transform is conversely defined by $\boldsymbol{\varepsilon} = \boldsymbol{W}^{-1}\hat{\boldsymbol{\varepsilon}} = \boldsymbol{G}_J \sum_{j=1}^{J} \check{\boldsymbol{\Psi}}_j\boldsymbol{S}_j\hat{\varepsilon}_j$, where $\boldsymbol{G}_J$ is the full resolution inverse spectral transform (since $T_J = T$). A representation of an example of wavelet field $\hat{\boldsymbol{\varepsilon}}$ will be shown in figure 4.

3 Wavelet filtering properties in a 1D analytical framework

3.1 A simple 1D analytical case with varying length scales

A simple 1D analytical framework has been considered to highlight the filtering properties of wavelets. In this framework, the geographical domain is supposed to be an earth great circle of radius a, and the coordinate x is the geographical position varying from $0\,^\circ$ to $360\,^\circ$ in terms of angle (or 0 to $2\pi a$ in terms of distance). On this circle, only one field is considered. A homogeneous covariance matrix is obtained from a radial correlation function $f_H^x(s) = e^{-\frac{s^2}{2L_H^2}}$, where x is a point on the circle, s is a separation value, and L_H is the length scale, which is here arbitrarily set equal to 250 km.

Then, a heterogeneous correlation is computed using a c-stretching Schmidt transformation (Courtier and Geleyn 1988), adapted to the circle and defined by $h(x) = a\left[\pi - 2Atan\left(\frac{1}{c}tan(\frac{\pi}{2} - \frac{1}{2}\frac{x}{a})\right)\right]$ with $c = 2.4$ (the Schmidt transformation is used for a different purpose in the Arpège global stretched model to obtain a variable resolution). The resulting correlation function is:

$$f^x(s) = f_H^{h^{-1}(x)}(h^{-1}(x+s) - h^{-1}(x)).$$

The associated matrix obtained, noted $\boldsymbol{B}_a$, is characterized by heterogeneous correlations which are relatively

DOI: 10.1002/qj

sharp around $180\,^\circ$ and broad around $0\,^\circ$. This is illustrated in the top panel of Fig. 4, which represents the local correlation length scales $L(x)$ at point x, approximated by (Belo Pereira and Berre 2006)

$$L^2(x) = \frac{\sigma(\varepsilon)^2(x)}{\sigma(\partial_x \varepsilon)^2(x) - (\partial_x \sigma(\varepsilon))^2(x)}, \qquad (2)$$

where $\sigma(\varepsilon)(x)$ is the standard deviation of $\varepsilon(x)$, and ∂_x is the derivative along the coordinate.

In the 1D framework, evaluating the length scales that are implied by the formulation (1) involves application of the gradient operator and its adjoint to the covariance matrix $\boldsymbol{B}_w$. For instance, the covariance matrix $\boldsymbol{B}'_{w,xx}$ of $\partial_x \varepsilon$ corresponds to

$$\boldsymbol{B}'_{w,xx} = \overline{(\partial_x \varepsilon)(\partial_x \varepsilon)^*} = \partial_x \overline{\varepsilon \varepsilon^*} \partial_x^* = \partial_x \boldsymbol{B} \partial_x^* \qquad (3)$$

The diagonal of $\boldsymbol{B}'_{w,xx}$ then provides the variances of $\partial_x \varepsilon$ that are used in the length scale equation.

3.2 Randomization of $\mathbf{B}_a$ and Schur product

In order to examine the effects of the sampling noise, random perturbations $\boldsymbol{\varepsilon}$ have been generated from $\boldsymbol{B}_a$: $\boldsymbol{\varepsilon} = \boldsymbol{B}_a^{1/2} \boldsymbol{\zeta}$, where $\boldsymbol{\zeta}$ is a random sample of a normal distribution with zero mean and the identity as a covariance matrix (Fisher and Courtier 1995). This allows one to obtain an ensemble covariance matrix $\boldsymbol{B}_e = \frac{1}{N_e} \sum_k \boldsymbol{\varepsilon}(k) \boldsymbol{\varepsilon}(k)^T$, where N_e is the ensemble size.

The middle panel of Fig. 4 shows an example of random error realization. One may notice that this field presents shorter variations near $180\,^\circ$ (Z1 region) than near $0\,^\circ$ (Z2 region), in accordance with the local length scales of the specified covariance function (upper panel of Fig. 4). These varying length scales are expected to be captured by the wavelet diagonal approach, with a larger amplitude of small scale wavelet variances in Z1 than in Z2.

This is supported by the bottom panel of Figure 4, which shows the amplitudes of the wavelet coefficients of the error field example. As expected, the amplitudes of small scale coefficients (for $j \geq 10$) tend to larger in Z1 than in Z2.

The ensemble covariance matrix $\boldsymbol{B}_e$ can be compared to the exact covariance matrix $\boldsymbol{B}_a$ to examine the effects of the sampling noise, and also to the corresponding wavelet ensemble covariance matrix $\boldsymbol{B}_w$ defined by equation (1). Sampling noise is a particularly important issue in assimilation schemes such as the Ensemble Kalman filter, and makes the use of a Schur product appropriate (Houtekamer et al 2001). The Schur product corresponds to an element-wise product with a matrix $\boldsymbol{F}_{L_s}$. A filtered ensemble covariance matrix $\boldsymbol{B}_e^{L_s}$ can be obtained with

$$\boldsymbol{B}_e^{L_s} = \boldsymbol{F}_{L_s} \circ \boldsymbol{B}_e,$$

where $\boldsymbol{F}_{L_s}$ is the matrix which corresponds to a compactly-supported correlation function. In this paper, a fifth-order piecewise rational function (Gaspari and Cohn 1999, Eq. (4.10)) is used, as in Houtekamer and Mitchell (2001).

The Schur filter ensures that the long distance correlations are set to zero, at the expense of an artificial sharpening of the correlation functions in the intermediate distances.

This filter does not change the local variances themselves, while they are also affected by sampling noise. A normalization by the diagonal values of $\boldsymbol{B}_e^{L_s}$ is therefore applied in addition, to ensure that $\boldsymbol{B}_e^{L_s}$ becomes a valid correlation matrix. This procedure allows us to concentrate on the sampling noise effects on the correlation functions.

The impact of $\boldsymbol{F}_{L_s}$ on the final correlation is sensitive. If L_s is too short (left panels of Fig. 5, which correspond to $L_s = 200\ km$), the exact correlation functions are replaced by an excessively homogeneous and sharp correlation function. In this case, the correlation related to the short length scale area Z1 (grid-point 121) is well represented, but the correlation function in the large length scale area Z2 (grid-point 1) is shorter than it should be.

Therefore, L_s may be increased (here to 1000 km), in order to preserve the large length scale in the Z2 area. However, in the Z1 area, some spurious oscillations in the short distances are no longer filtered in this case (Fig. 5, right panels).

The scale L_s has thus to be chosen judiciously. This optimal value depends also on the ensemble size (Houtekamer and Mitchell 2001, Lorenc 2003).

A Schur filter is also applied to provide a filtered version of the wavelet ensemble covariance matrix:

$$\boldsymbol{B}_w^{L_s} = \boldsymbol{F}_{L_s} \circ \boldsymbol{B}_w.$$

The relative necessity of this Schur filter for the wavelet approach and its effect are illustrated in Figure 6. It is a comparison between correlations calculated from a 10 member ensemble, either directly (solid line) or based on a wavelet diagonal approach (dashed line). Zeroing spurious long distance correlations remains desirable for the wavelet approach in this example, although this is much less marked than in the direct estimation. The effect of a Schur filter with $L_s = 6000$ km on the wavelet-based correlations is illustrated (bold solid line). It sets large distance correlations to zero. The possibility to avoid applying this Schur filter may be explored in future studies, for instance by using bandwidths that are broad enough to make the correlations sufficiently local.

In the remainder of the paper, $\boldsymbol{B}_e^{L_s}$ and $\boldsymbol{B}_w^{L_s}$ will be simply refered to as the ensemble and wavelet correlation matrices respectively.

3.3 Wavelet filtering of the correlation functions and of their variations

The wavelet filtering properties can be examined by comparing the left and right panels of Figure 7, which have been produced with $L_s = 6000$ km and $N_e = 10$ elements. A large value of L_s is chosen to illustrate the typical amplitude of sampling noise. The left panels show the raw

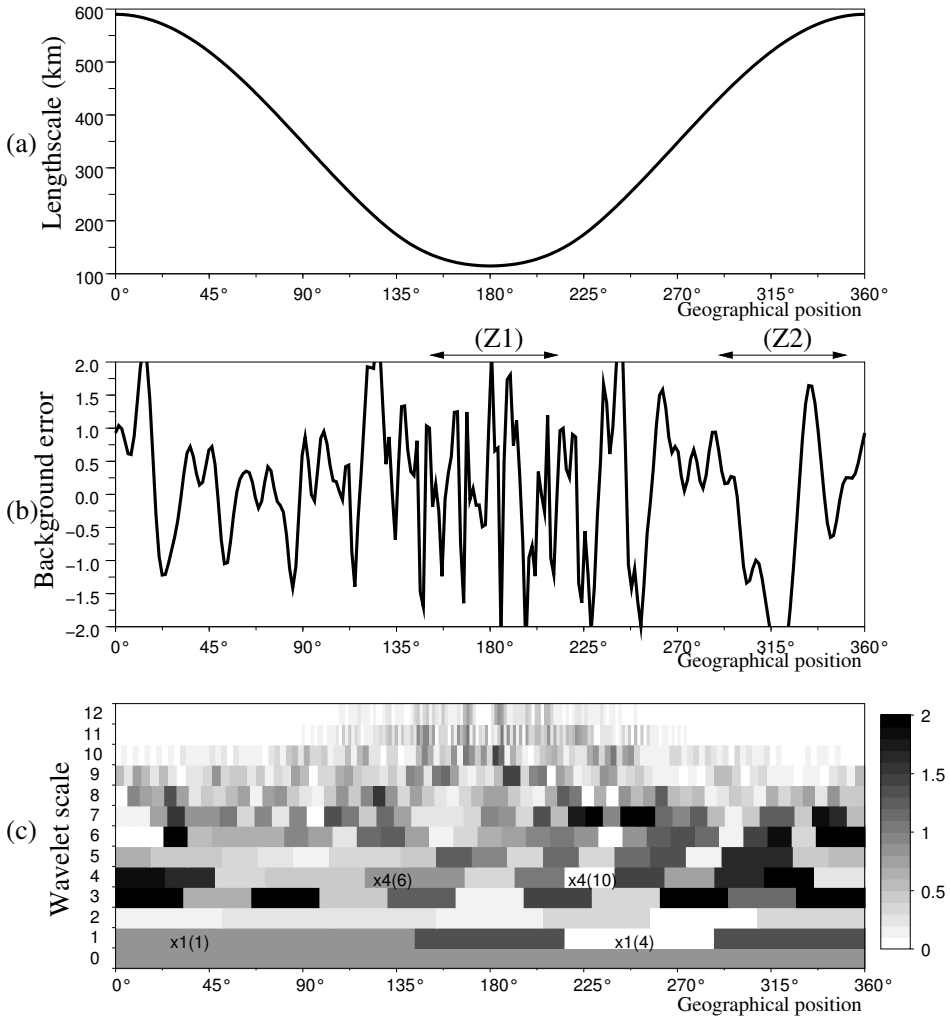


Figure 4. (a) Geographical variations of the length scale in the analytical framework. (b) Sample $\boldsymbol{\varepsilon}$ of a Gaussian background error associated with the analytical background covariance model on the circle and (c) the absolute value of its wavelet coefficients : each row is the absolute value of vector $\hat{\boldsymbol{\varepsilon}}_j = \boldsymbol{S}_j^{-1} \check{\boldsymbol{\Psi}}_j \boldsymbol{S}_j \boldsymbol{\varepsilon}$; for a given scale j there is one box per subgrid-point, such as $x_1(1), x_1(4)$ for $j = 1$ or $x_4(6), x_4(10)$ for $j = 4$.

ensemble correlations at gridpoints 1 (top panel) and 121 (bottom panel). The corresponding right panels are for the wavelet-based correlations at these gridpoints. The sampling noise is quite visible in the ensemble correlations, with many large and spurious oscillations. Such artefacts are much less marked in the wavelet correlations.

The wavelet formulation thus appears to partly filter the sampling noise. This is expected knowing that the wavelet diagonal assumption amounts to locally averaging the covariance functions, which implies an increase of the total size of the sample. A wavelet illustration of these filtering properties is also discussed in appendix C.

Fig. 8 illustrates the effects on the length scale variations of the sampling noise and of the wavelet filtering. The case $N_e = 10$ (top left panel) is the most spectacular. The raw length scales exhibit some large and spurious small scale oscillations, and the largest length scale values are often much exagerated (e.g. with 800 km raw values around 70 °, while the exact value is around 400 km). In contrast, the wavelet implied length scales have relatively smooth variations and accurate values. The wavelet approach thus appears to be able to capture and represent the main geographical variation of interest (increase of length scale towards 0 °) from a small ensemble with only 10 members.

Again the wavelet local averaging reduces the effects of sampling noise. These beneficial effects decrease when the size of the ensemble increases.

3.4 Data assimilation experiments

Assimilation experiments on the circle have also been performed. As the true correlations are known, the true solution can be computed, and can be compared with that of the correlation model.

In these experiments, the true state is taken as being zero everywhere. The background error is generated from the true correlation matrix (remember that the standard deviation is equal to 1) with $\boldsymbol{B}_a^{1/2}$, while observations are generated with $\boldsymbol{R}^{1/2}$ (which is the square-root of the observation error covariance matrix $\boldsymbol{R}$). $\boldsymbol{R}^{1/2}$ is assumed to be a diagonal matrix $\sigma_o \boldsymbol{I}$ where $\boldsymbol{I}$ denotes the identity matrix (thereafter $\sigma_o = 0.95$: the observations are

DOI: 10.1002/qj

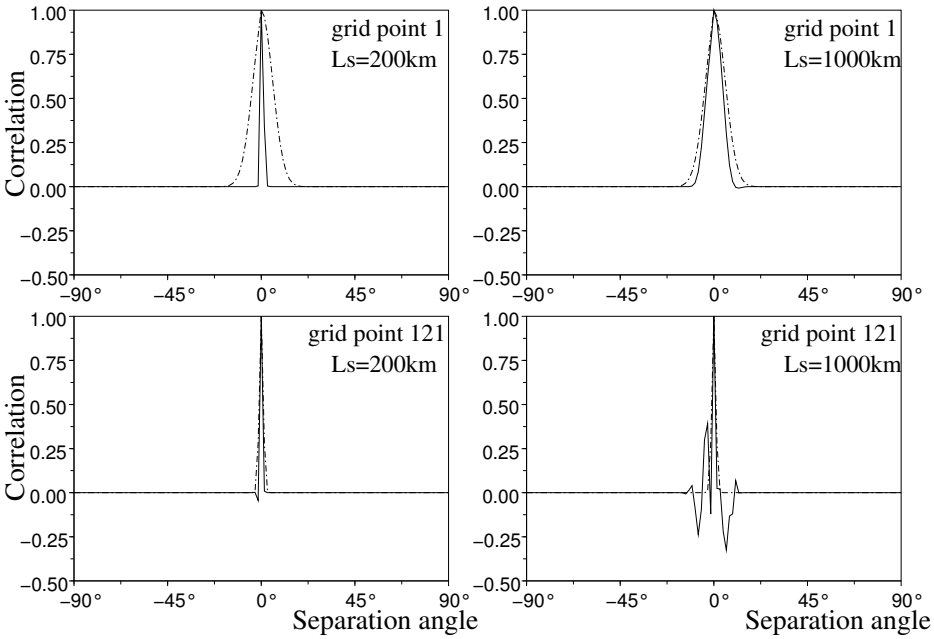


Figure 5. Ensemble correlations (solid line) related to grid points 1 (corresponding to the large length scale area) and 121 (corresponding to the short length scale area). Correlations are directly computed from a 10 element ensemble, and then filtered with a Schur filter $\boldsymbol{F}_{L_s}$ with $L_s = 200\ km$ (left panels) and $1000\ km$ (right panels). Exact correlations are also shown (chain-dotted line).

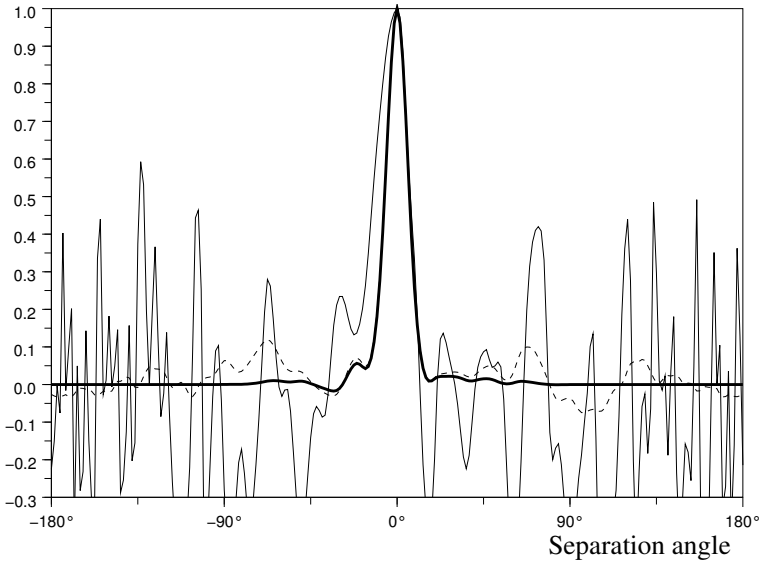


Figure 6. Ensemble correlations related to gridpoint 1, calculated from a 10 element ensemble: directly (solid line), with a wavelet diagonal approach and no Schur filter (dashed line), with a wavelet diagonal approach and a Schur filter such as $L_s = 6000km$ (bold solid line).

assumed to have a similar quality as the background). There is one observation every five gridpoints.

Root Mean Square (RMS) errors (averaged over the domain) of analysis can be calculated for three different covariance approximations: the direct ensemble estimation, the homogeneous formulation (i.e. the spectral diagonal approach), and the wavelet formulation. These RMS can then be compared with the RMS corresponding to the true covariances. Fig. 9 corresponds to the difference $RMS - RMS(true)$, for different Schur scales L_s and for different ensemble sizes N_e. In this figure, every curve has the same behaviour (similar to Lorenc 2003): the shape is convex with a minimum that determines an optimal L_s value.

For the direct ensemble estimation (dashed line), the smaller the ensemble is, the larger the RMS is, and the shorter the scale L_s has to be. Such dependencies of the RMS and optimal L_s on the ensemble size are much smaller for the wavelet approach (full line). Moreover, the wavelet results are relatively good even for a small ensemble size such as $N_e = 10$. This illustrates the beneficial impact of wavelet filtering.

Compared with the direct ensemble estimation, another attractive result of the wavelet approach is that the RMS slope is small beyond the optimal choice for L_s. This means that the analysis quality will be less affected

DOI: 10.1002/qj

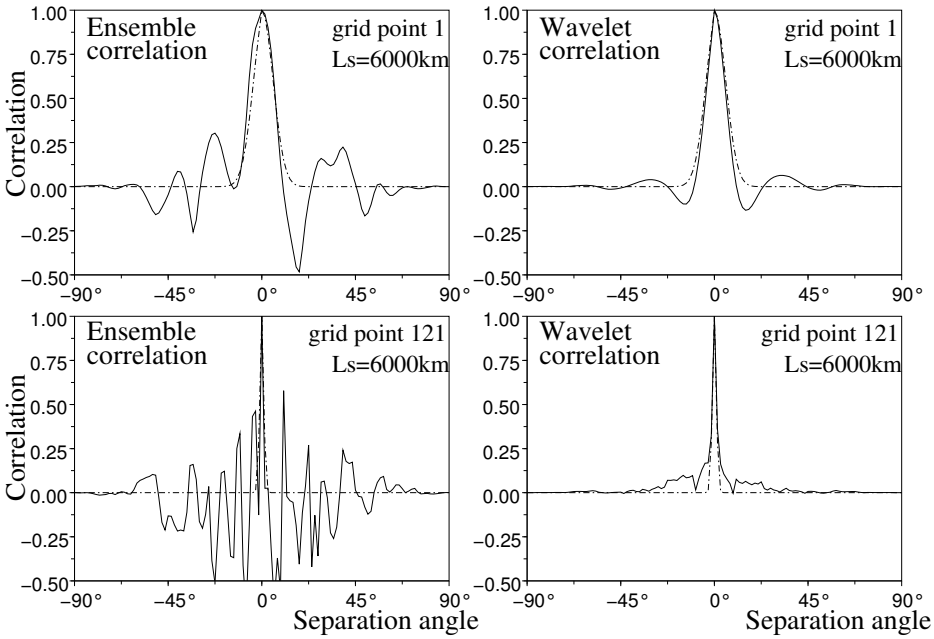


Figure 7. Ensemble and wavelet correlations (solid line) related to grid points 1 (corresponding to the large length scale area) and 121 (corresponding to the short length scale area). Correlations are computed from a 10 element ensemble and then filtered with a Schur filter $\boldsymbol{F}_{L_s}$ with $L_s = 6000\ km$. Exact correlations are also shown (chain-dotted line).

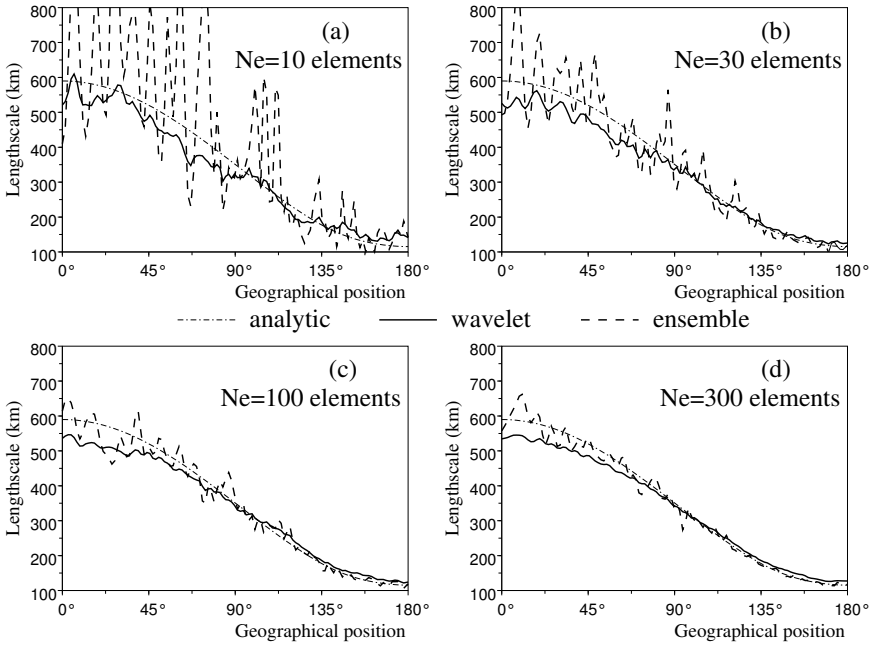


Figure 8. Geographical variation of the length scale: exact (chain-dotted line), wavelet estimated (solid line) and ensemble estimated (dashed line). Length scales are computed for different ensemble sizes, namely $N_e = 10, 30, 100$ and 300. For each case, the wavelet and ensemble correlation estimations have been filtered using a Schur filter $\boldsymbol{F}_{L_s}$ with $L_s = 6000\ km$.

by a suboptimal Schur scale than in the direct ensemble estimation.

As expected, analyses produced with the wavelet approach are closer to optimality than those produced with the homogeneous formulation. On the other hand, it is interesting to notice that for $N_e = 10$, the homogeneous approach can give better results than a direct ensemble estimation, if the Schur length is too large. This is another illustration of the importance of sampling noise in small ensembles, and of the potential benefit of spatially averaging the ensemble covariances in this case.

It is also possible to compare geographical variations the lengthscales, when the optimal Schur lengths are used. The results are illustrated in Fig. 10. The wavelet implied length scale values still appear to be more accurate, and their variations are smoother. This is particularly noticeable for small ensembles.

It may be mentioned also that, compared to Figure

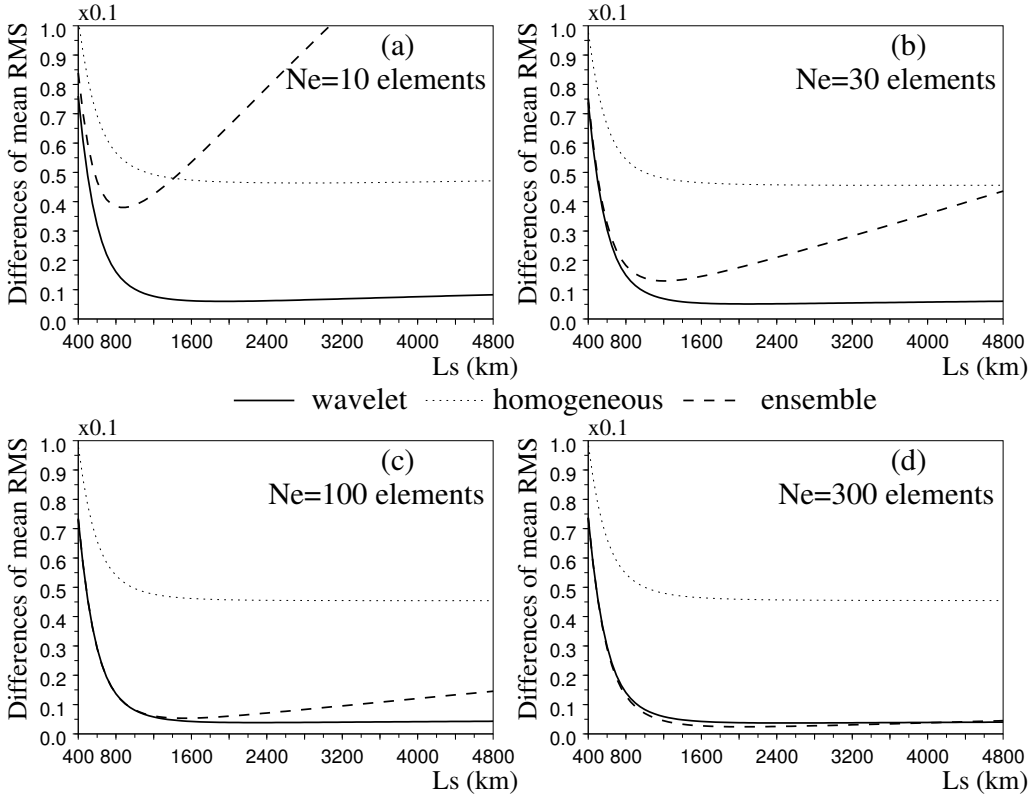


Figure 9. Differences between analysis RMS error and true analysis RMS error, as a function of the scale of the Schur filter L_s and of the size N_e of the ensemble (respectively 10, 30, 100 and 300 for the four different panels). Differences are shown in each panel for wavelet analysis (solid line), analysis with homogeneous formulation (dotted line) and analysis with raw ensemble covariances (dashed line). All analyses are performed with one observation every 5 gridpoints.

(8), the use of shorter Schur lengths L_s implies a shortening of the length scales. This is consistent with the fact that the smaller L_s is, the faster the filtered correlation functions will decrease (as a function of separation distance). With respect to Eq. (2), this is linked with an increase of $\sigma(\partial_x \varepsilon)^2$, because small scale contributions are emphasized both in $\partial_x \varepsilon$ (compared to ε) and when decreasing the correlation length scale.

4 Application to an ensemble of Arpège forecasts

4.1 The ensemble data set of global Arpège forecasts

The Météo-France operational NWP system is based on the Arpège model (Courtier and Geleyn 1988), and on a 4D-Var assimilation scheme (Rabier et al 2000, Veersé and Thépaut 1998). The background error covariance matrix is calculated by using an ensemble of perturbed assimilation runs (Houtekamer et al 1996, Fisher 2003). The detailed results for Arpège are described in Belo Pereira and Berre (2006).

We propose to illustrate some typical results of the wavelet covariance modelling on this kind of NWP ensemble data. The formalism and the cutoff wave numbers are the same as those mentioned in section 2 (e) (and Fisher 2003). The available ensemble consists in a set of 6 forecast differences for each day of the period from 9 February to 24 March 2002.

In this 2D high dimensional framework, local length scales are calculated by using equation (2), with a specific randomization technique for the wavelet-implied length scales. In the latter case, a set of 1000 random vectors ε has been generated from the square root of $\boldsymbol{B}_w$ (Fisher and Courtier 1995), and the gradient operator ∂_x has been applied to these vectors ε. The variances of $\partial_x \varepsilon$ are then used in the length scale equation.

A first possible option is to temporally average the spatial covariances, in order to examine the "climatology" of the error covariances. The total sample is made of $N_e = 264$ elements in this case.

A second option is to study the covariances for a particular date. The total sample is reduced to only $N_e = 6$ elements in this case, which correspond to differences between six perturbed forecasts, which are valid on 10 February at 12 UTC.

4.2 Local "climatological" length scales

As shown in Fig. 11, the ensemble method provides some interesting local information on the "climatological" correlations, which appear to be well represented by the wavelet formulation of equation (1). The length scales are smaller in the (data rich) Northern Hemisphere than in the (data poor) Southern Hemisphere. Some local length scale minima can be identified in the storm track region of the Northern Atlantic, and near the Inter Tropical

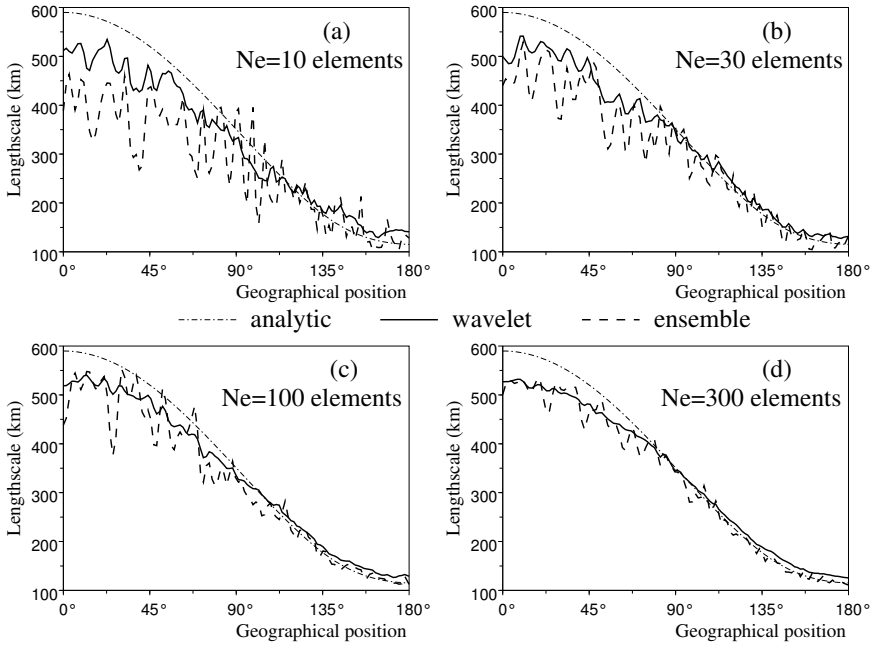


Figure 10. Same as figure 8, but with optimal Schur lengths L_s, chosen according to Fig. 9. For the wavelets, L_s is constant and equal to 4000 km. For the ensemble estimation, L_s depends on the size N_e : (a) $(N_e = 10, L_s = 900\ km)$, (b) $(N_e = 30, L_s = 1250\ km)$, (c) $(N_e = 100, L_s = 1600\ km)$ and (d) $(N_e = 300, L_s = 2050\ km)$.

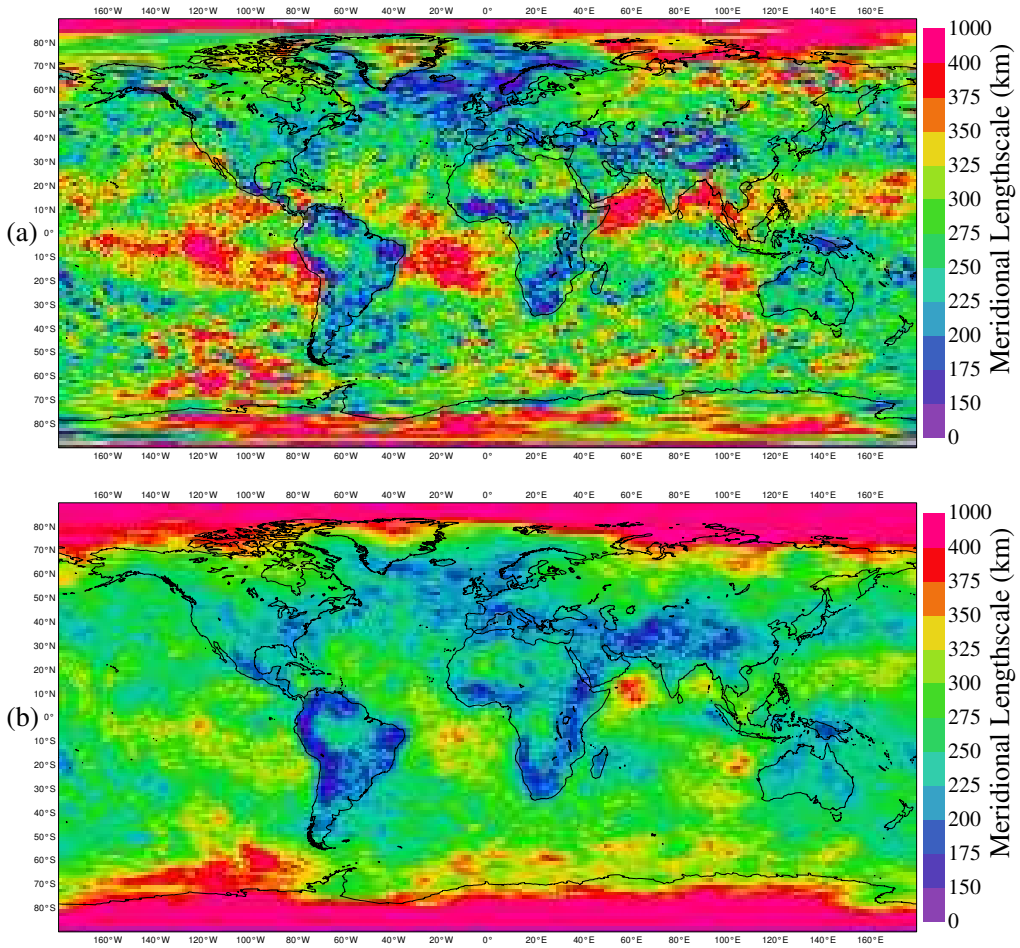


Figure 11. Meridional length scales (in km) for surface pressure, averaged over the 46 day period and the 6 ensemble members. (a): raw length scales. (b): wavelet implied length scales.

Convergence Zone (ITCZ) area in Western Africa. Length scales are also smaller over orographic regions such as the Himalaya and the Andes, and in the Southern part of Africa. In contrast, some length scale maxima are visible over tropical oceans.

In accordance with the results in section 33.3, these behaviours on the whole globe confirm one of the main advantages of a wavelet covariance formulation (compared e.g. with a spectral formulation): it allows some correlation variations to be represented, such as those that are induced by atmospheric processes and by data density contrasts.

It can be also noticed that the extreme variations tend to be smoothed in the wavelet map. This is due to the filtering properties of the wavelet diagonal approach. These filtering properties will now be shown to be even more important when the ensemble size is reduced to $N_e = 6$.

4.3 Local length scales of a particular date

Fig. 12 shows the corresponding length scales for a particular date, the 10 February at 12 UTC. The raw length scales (top panel) appear to be very noisy, due to the small size of the ensemble. In contrast, the wavelet implied length scale map (bottom panel, superposed with the background field of sea level pressure) is relatively smooth and well structured. Large values appear clearly in the southern circumpolar ocean and over tropical oceans, as in the climatological case but in a more pronounced way. Small length scales are visible over land. Some other structures related to the local weather situation can also be identified, such as small values in the vicinity of some mid-latitude lows over sea (see e.g. the lows near Scandinavia and south of Argentina). This is consistent with results described by Thépaut et al (1995) for instance.

Such strong differences between the raw and wavelet implied length scales, when the sampling size is small ($N_e = 6$), are in agreement with the large length scale differences found in the 1D framework with $N_e = 10$ (top left panel of Fig. 8). They support the idea that the wavelet formulation is able to capture and to represent the main relevant length scale variations, from a small ensemble of forecasts, thanks to local spatial averaging.

5 Conclusions

In this paper, the ability of a wavelet diagonal approach to ensure a smooth representation of geographical variations of the correlation functions was studied. Representing the covariances by a diagonal matrix in wavelet space amounts to locally averaging the covariance functions. Due to this spatial averaging, the statistical estimate is more sampled than when estimating the purely local covariance functions. Moreover, as this spatial averaging is local, the representation of geographical variations remains possible. Such filtering properties look particularly attractive when estimating error covariances from an ensemble of forecasts. These aspects of wavelet covariance modelling were made explicit formally, and they were illustrated experimentally in two different frameworks.

The first experimental framework is a simple 1D context with varying length scales. The wavelet covariance model was shown to be able to represent the local correlation functions and their length scale variations, in a smoother and more realistic way than when only using a Schur filter. This is particularly noticeable when using small ensembles with e.g. 10 or 30 members. The wavelet formulation remains competitive with up to 100 members.

The second experimental framework consists in an ensemble of global NWP forecasts. The wavelet approach appears to represent well the "climatological" variations of the local length scales, which are induced by atmospheric processes and by data density contrasts. Moreover, a preliminary examination of the length scales of a particular day suggests that the wavelet approach allows some important length scale variations to be extracted, which are connected with the local weather situation.

These results are consistent with the expected filtering properties of wavelets in terms of local spatial averaging. They suggest that wavelets may be a promising tool to estimate and represent flow-dependent covariances from a small ensemble of forecasts.

Acknowledgement

The authors would like to thank Mike Fisher, Anthony Weaver and Andrew Tangborn for fruitful discussions. We also thank Bernard Chapnik, Claude Fischer and Florence Rabier, for their careful reading of the manuscript.

A Appendix A : Expressions of covariance spatial averaging

It is shown, in the general layout of frames, that the diagonal covariance assumption leads to a weighted spatial averaging of covariances.

A.1 Background error expansion in a frame

A frame (Daubechies 1992, Fisher 2004) is a family of functions $\{\boldsymbol{\phi}_m, m \in \mathcal{M}\}$, where $\mathcal{M}$ is a countable set. This family is associated to a dual frame $\left\{\tilde{\boldsymbol{\phi}}_m, m \in \mathcal{M}\right\}$, so that the error field $\boldsymbol{\varepsilon}$ can be analyzed as a set of frame coefficients $\hat{\varepsilon}_m$ with

$$\hat{\varepsilon}_m = \sum_x \varepsilon(x) \tilde{\phi}^*_m(x), \tag{4}$$

where the exponent $*$ stands for the conjugate transpose operator. The signal is recomposed according to

$$\varepsilon(x) = \sum_{m \in \mathcal{M}} \hat{\varepsilon}_m \phi_m(x),$$

or in vector form:

$$\boldsymbol{\varepsilon} = \sum_{m \in \mathcal{M}} \hat{\varepsilon}_m \boldsymbol{\phi}_m.$$

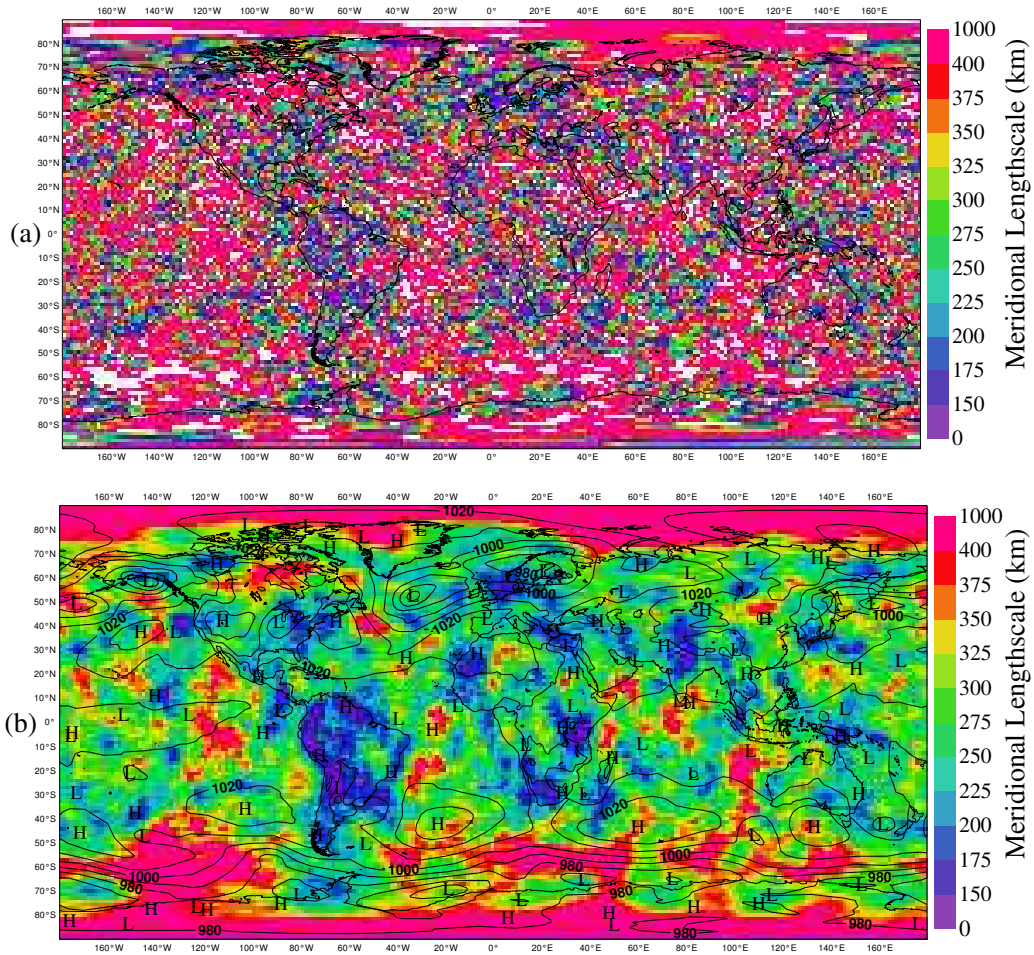


Figure 12. Meridional length scales (in km) for surface pressure on 10 February 2002 at 12 UTC, computed from 6 ensemble members. (a): raw length scales. (b): wavelet implied length scales, superposed with the background field of sea level pressure.

A.2 Expansion of covariances in a frame under the diagonal assumption

Using the previous frame decomposition, the covariance matrix $\boldsymbol{B} = \overline{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^*}$ can be expanded as $\boldsymbol{B} = \sum_{m,m'} \overline{\hat{\varepsilon}_m \hat{\varepsilon}^*_{m'}} \boldsymbol{\phi}_m \boldsymbol{\phi}^*_{m'}$. When using the diagonal assumption in the frame space, covariances $\overline{\hat{\varepsilon}_m \hat{\varepsilon}^*_{m'}}$ are zero except if $m = m'$. Then, the resulting covariance matrix $\boldsymbol{B}_d$ is $\boldsymbol{B}_d = \sum_m B_{mm} \boldsymbol{\phi}_m \boldsymbol{\phi}^*_m$, where $B_{mm} = \overline{\hat{\varepsilon}_m \hat{\varepsilon}^*_m}$ is the variance for the frame coefficient m. Thus, the associated covariance function f^x_d at position x has the following expression:

$$f^x_d(s) = \sum_m B_{mm} \phi_m(x) \phi^*_m(x+s),$$

where s is a separation value (in gridpoint space). From (4), coefficients B_{mm} can be rewritten

$$\begin{aligned} B_{mm} &= \sum_{x',s'} \overline{\varepsilon(x')\varepsilon(x'+s')^*}\ \tilde{\phi}^*_m(x') \tilde{\phi}_m(x'+s') \\ &= \sum_{x',s'} f^{x'}(s')\ \tilde{\phi}^*_m(x') \tilde{\phi}_m(x'+s'), \end{aligned}$$

where $f^{x'}(s') = \overline{\varepsilon(x')\varepsilon(x'+s')^*}$ is the local covariance function at position x' for the full covariance matrix $\boldsymbol{B}$, and then the resulting expression for $f^x_d(s)$ is

$$f^x_d(s) = \sum_{x',s'} f^{x'}(s') \left(\sum_m \tilde{\phi}^*_m(x') \tilde{\phi}_m(x'+s') \phi_m(x) \phi^*_m(x+s)\right),$$

which can be in turn rewritten as

$$f^x_d(s) = \sum_{x',s'} f^{x'}(s') \Phi^{x,s}(x',s'),$$

with

$$\Phi^{x,s}(x',s') = \sum_m \tilde{\phi}^*_m(x') \tilde{\phi}_m(x'+s') \phi_m(x) \phi^*_m(x+s).$$

The implied covariance function $f^x_d(s)$ may thus be seen as a weighted spatial average of the local covariance functions $f^{x'}(s')$, where the weights $\Phi^{x,s}(x',s')$, for a pair $\{x,s\}$, are functions of position x' and separation s'.

A.3 Diagonal assumption in Fourier space

Let N_g be the number of grid points on a circle, and $e_m(x) = exp(im\frac{2\pi x}{N_g})$. Then, the family

DOI: 10.1002/qj

$\left\{\phi_m = \frac{1}{\sqrt{N_g}} e_m, m \in [0, N_g - 1]\right\}$ is a frame whose dual is simply $\tilde{\phi}_m = \phi_m$. In this case, the weighting coefficients are given by

$$\begin{aligned}\Phi^{x,s}(x', s') &= \frac{1}{N_g^2} \sum_m e_m^*(x') e_m(x' + s') e_m(x) e_m^*(x + s) \\ &= \frac{1}{N_g^2} \sum_m e_m(s' - s) \\ &= \frac{1}{N_g} \delta(s' - s),\end{aligned}$$

where $\delta(s' - s) = 1$ when $s' = s$ and $\delta(s' - s) = 0$ elsewhere. This means that a zero weight is given to all values $f^{x'}(s')$ such as $s' \neq s$ and that a uniform weight $\frac{1}{N_g}$ is given to all values $f^{x'}(s')$ such as $s' = s$, whatever the position x' is. In other words, each implied covariance function f_d^x is simply a global spatial average of the local covariance functions $f^{x'}$.

A.4 Diagonal assumption in wavelet space

For the wavelet case, $\mathcal{M}$ is a set of $m = (j, \{x_j(i), i = 1, N_x(j)\})$, where j is a scale and $x_j(i)$ a position on a subgrid associated to j. This defines a frame such that, $\phi_m(x) = \tilde{\phi}_m(x) = \psi_j(x - x_j(i))$, where functions ψ_j are Fisher's wavelets. In this case, the weighting coefficients $\Phi^{(x,s)}(x', s')$ are given by

$$\sum_j \sum_{i=1}^{N_x(j)} \psi_j(x' - x_j(i))\, \psi_j(x' - x_j(i) + s') \\ \psi_j(x - x_j(i))\, \psi_j(x - x_j(i) + s).$$

In contrast with the spectral case, the resulting weights $\Phi^{(x,s)}(x', s')$ will vary with the position x'. As expected, numerical tests indicate that the weights $\Phi^{(x,s)}(x', s')$ tend to be maximum for positions x' close to x and for separation values s' that are also close to s. In other words, the implied covariance function f_d^x may be seen as a local spatial average of the covariance functions $f^{x'}$.

B Appendix B: Design of $\mathbf{B}_w^{-1/2}$ and of $\mathbf{B}_w^{1/2}$

The formulation of $\boldsymbol{B}_w^{1/2}$ is determined by the design of $\boldsymbol{B}_w^{-1/2}$. The latter matrix is conceived as an operator that transforms the background error variable $\boldsymbol{\varepsilon}$ into a transformed variable, whose covariance matrix is close to an identity matrix (see e.g. Deckmyn and Berre (2005) or Gustafsson et al (2001)):

$$\boldsymbol{B}_w^{-1/2} = \boldsymbol{D}_w^{-1/2} \boldsymbol{W} \boldsymbol{\Sigma}_s^{-1} \boldsymbol{\Sigma}_g^{-1},$$

where $\boldsymbol{\Sigma}_g$ is a diagonal matrix of gridpoint standard deviations of $\boldsymbol{\varepsilon}$, $\boldsymbol{\Sigma}_s^{-1} = \boldsymbol{S}^{-1} \boldsymbol{D}_s^{-1/2} \boldsymbol{S}$ corresponds to a normalisation by spectral standard deviations of $\check{\boldsymbol{\varepsilon}}' = \boldsymbol{S} \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\varepsilon}$ ($\boldsymbol{S}$ being the spectral transform, and $\boldsymbol{D}_s$ is the corresponding diagonal matrix of variances in spectral space), and $\boldsymbol{D}_w^{1/2}$ is the diagonal matrix of wavelet standard deviations of $\hat{\boldsymbol{\varepsilon}}'' = \boldsymbol{W} \check{\boldsymbol{\varepsilon}}'$. Matrices $\boldsymbol{W}$ and $\boldsymbol{W}^{-1}$ correspond respectively to the direct and inverse wavelet transforms.

The associated expression of $\boldsymbol{B}_w^{1/2}$ is then $\boldsymbol{B}_w^{1/2} = (\boldsymbol{B}_w^{-1/2})^{-1} = \boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_s \boldsymbol{W}^{-1} \boldsymbol{D}_w^{1/2}$.

C Appendix C: A wavelet illustration of the filtering properties

As mentioned in sections 2 and 3.3, the filtering properties of wavelets can be expected since the wavelet diagonal assumption amounts to locally averaging the covariance functions. This may be considered as a "gridpoint vision" of the filtering properties, in the sense that gridpoint covariance functions are seen as being averaged in gridpoint space. In this section, we will evoke two other complementary visions of the filtering properties (in wavelet and spectral spaces respectively).

Note that the full covariance matrix in wavelet space represents the covariances between different scales at different locations. A "wavelet vision" of the filtering properties at play is thus to note in particular that the wavelet diagonal assumption implies zeroing off-diagonal terms, which correspond to correlations between wavelet modes at different positions (e.g. for a given scale j). The formal examination of the associated equations (not detailed here for the sake of conciseness) suggests the following result: neglecting these wavelet off-diagonal correlations prevents small scale distant modes from (spuriously) contributing to the local covariance function (at a given reference position).

This interpretation is supported experimentally, as illustrated in Fig. 13. This figure shows the amplitude (absolute value) of the wavelet coefficients $\boldsymbol{W} f^0$ of the local covariance function at geographical position $0\,^\circ$, namely f^0. In the ensemble off-diagonal case (top right panel), the amplitude of the small scale distant wavelet modes (e.g. for $j \geq 10$ near $180\,^\circ$) can be spuriously large (due to sampling noise), while the exact values (top left panel) tend to be close to zero. In contrast, the wavelet diagonal approach produces realistic small values, even when a small ensemble is used ($N_e = 10$). This is consistent with the implicit zeroing of correlations between wavelet modes positioned around $0\,^\circ$ and those positioned around $180\,^\circ$ for instance.

Finally, it may be also mentioned that a "spectral vision" of the wavelet filtering properties can be considered as well. It will not be detailed here for the sake of conciseness. Briefly summarized, it consists in noticing that zeroing cross-correlations between different wavelet scales amounts to zeroing some small scale (noisy) contributions to the geographical variations of covariances. This is similar to the expected effect of a local spatial averaging in gridpoint space.

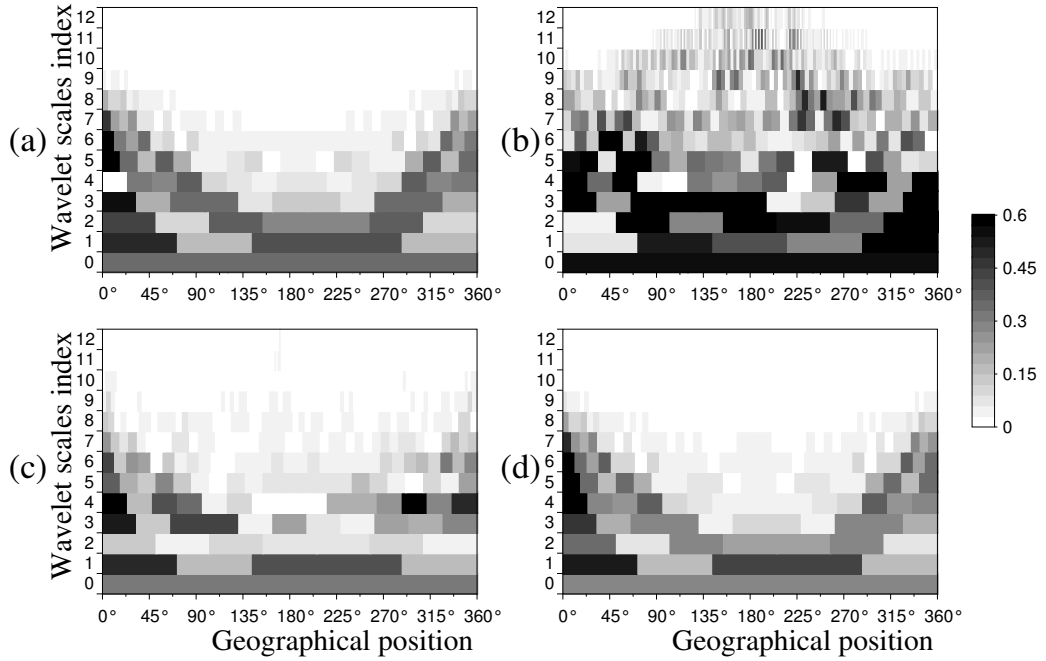


Figure 13. Amplitude of the wavelet coefficients of the local covariance function at geographical position 0 °. (a) exact covariance, (b) ensemble approach with 10 members and a full covariance matrix (in gridpoint or wavelet space), (c) ensemble wavelet diagonal approach with 10 members, and (d) analytical wavelet diagonal approach (equivalent to (c) but with an infinite number of members). The wavelet scale index is j: large j values correspond to small scale wavelets.

References

Belo Pereira M. and Berre L. 2006. *The use of an Ensemble approach to study the Background Error Covariances in a Global NWP model*. *Mon. Wea. Rev.*, **134**, 2466–2489.

Bouttier F. 1993. *The dynamics of error covariances in a barotropic model*. *Tellus*, **45A**, 408–423.

Bouttier F. 1994. *A dynamical estimation of error covariances in an assimilation system. Mon. Wea. Rev.*, **122**, 2376–2390.

Buehner M. and Charron M. 2007. *Spectral and spatial localization of background error correlations for data assimilation.* To appears in *Q.J.R. Meteorol. Soc.*

Courtier P. and Geleyn J.F. 1988. *A global numerical weather prediction model with variable resolution : Application to the shallow-water equations*. *Q.J.R. Meteorol. Soc.*, **114**, 1321–1346.

Courtier P., Andersson E., Heckley W., Pailleux J., Vasiljević D., Hamrud M., Hollingsworth A., Rabier F. and Fisher M. 1998. *The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation*. *Quart. J. Roy. Meteor. Soc.*, **124**, 1783–1807.

Daley R. 1991. *Atmospheric Data Analysis*. Cambridge University Press. p471.

Deckmyn A. and Berre L. 2005. *A wavelet approach to representing background error covariances in a LAM*. *Mon. Wea. Rev.*, **133**, 1279-1294.

Dee D. 1995. *On-line estimation of error covariance parameters for atmospheric data assimilation. Mon. Wea. Rev.*, **123**, 1128–1145.

Derber J. and Bouttier F. 1999. *A reformulation of the background error covariance in the ECMWF global data assimilation system. Tellus*, **51A**, 195–221.

Daubechies I. 1992. *Ten lectures on Wavelets.* CBMS-NSF regional conference series in applied mathematics, SIAM, 357pp.

Evensen G. 1994. *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error. J. Geophys. Res.*, **99** (C5) 10 143–10 162.

Fisher M. and Courtier P. 1995. *Estimating the covariance matrices of analysis and forecast error in variational data assimilation.* ECMWF Technical Memorandum, **220**, 29pp.

Fisher M. 2003. *Background error covariance modelling*. Processing of the ECMWF Seminar on "Recent developments in data assimilation for atmosphere and ocean", Reading, 8–12 September 2003, 45–63.

Fisher M. 2004. *Generalized frames on the sphere, with application to the background error covariance modelling.* Processing of the ECMWF Seminar on "Developments in Numerical Methods for Atmospheric and Ocean Modelling", Reading, 6–10 September 2004, 87–101.

Gaspari G. and Cohn S. 1999. *Construction of correlation functions in two and three dimensions*. *Q.J.R. Meteorol. Soc.*, **125**, 723–757.

Gustafsson N., Berre L., Hörnquist S., Huang X-Y., Lindskog M., Navasques B., Mogensen KS. and Thorsteinsson S. 2001. *Three-dimensional variational data assimilation for a limited area model. Tellus*, **53A**, 425–446.

Houtekamer P.L., Lefaivre L., Derome J., Ritchie H. and Mitchell H.L. 1996. *A system simulation approach to ensemble prediction*. *Mon. Wea. Rev.*, **124**, 1225–1242.

Houtekamer P.L. and Mitchell H.L. 2001. *A sequential ensemble Kalman filter for Atmospheric Data Assimilation*. *Mon. Wea. Rev.*, **129**, 123–137.

Ingleby B. 2001. *The statistical structure of forecast errors and its representation in The Met. Office Global Model*. *Q.J.R. Meteorol. Soc.*, **124**, 1783–1807.

Lönnberg P. 1988. *Developpements in the ECMWF analysis scheme*. Proc. ECMWF Seminar on "Data assimilation and the use of satellite data", Reading, 5–9 September 1988, 75–120.

Lorenc A. 2003. *The potential of ensemble Kalman filter for NWP – a comparison with 4D-Var. Q.J.R. Meteorol. Soc.*, **129**, 3183–3203.

Rabier F., Jarvinen H., Klinker E., Mahfouf J.F. and Simmons A. 2000. *The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. Q.J.R. Meteorol. Soc.*, **126**, 1148–1170.

Thépaut J-N., Courtier P., Belaud G. and Lemaître G. 1995 *Dynamical structure functions in a four-dimensional variational assimilation : a case study. Q.J.R. Meteorol. Soc.*, **122**, 535–561.

Veersé F. and Thépaut J-N. 1998. *Multiple-truncation incremental approach for four-dimensional variational data assimilation*. *Q.J.R. Meteorol. Soc.*, **124**, 1889–1908.

Annexe B

Version originale du chapitre 5 (Pannekoucke *et al.* , 2008)

QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY
Q. J. R. Meteorol. Soc. **134**: 497–511 (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/qj.212

Background error correlation length-scale estimates and their sampling statistics

O. Pannekoucke*, L. Berre and G. Desroziers
GAME/CNRM (Météo-France, CNRS), Toulouse, France

Abstract: This article presents different formulae to estimate correlation length-scales, and an evaluation of their qualities for practical diagnostic applications. In particular, two new and simple formulae are introduced, which only require the computation of correlation with a single point for a given direction. It is then shown in a 1D heterogeneous context that all formulations lead to similar realistic length-scale values, and that they represent geographical variations rather well.
The estimation of length-scales within a finite ensemble is also studied. While a positive bias occurs when the ensemble size is too small, the standard deviation of the length-scale estimation is shown to be the main influence on the estimation error. The spatial structure of sampling noise is then diagnosed, and effects of spatial filtering techniques on the bias and standard deviation are illustrated.
Finally, an ensemble of perturbed forecasts from a global NWP model is used, showing a real application example.

WARNING : This is a preprint of an article accepted for publication in QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY ref: *Q. J. R. Meteorol. Soc.* **134**: 497–508 (2008) see the website for final version http://www.interscience.wiley.com/

KEY WORDS Data assimilation; Diagnosis; Length-scale approximation; ensemble; sampling noise.

Received 28 June 2007; Revised 19 December 2007; Accepted 19 December 2007

1 Introduction

In order to objectively determine initial conditions for numerical weather prediction, modern data assimilation schemes rely on specified error statistics to obtain an approximately optimal combination of observations and a background given by a short-range forecast. This near-optimal analysis is derived from statistical estimation theory. In this framework, the two sets of information are associated with covariance matrices corresponding to their respective errors. The error covariance matrices determine the respective weights given to each piece of information in the analysis. However, the correct specification of those statistics remains a major challenge in data assimilation systems.

The structure of correlation functions is particularly important, as it determines how the observed information is filtered and propagated spatially. Typically, when the background error structure is large scale, the correlation functions are relatively wide. This implies that small scale observed details tend to be filtered out in the analysis step, and that locally observed information is propagated on large spatial distances (Daley, 1991). Diagnostics of the length-scale of background error correlation functions are often used as an approximate indicator of the degree of spatial smoothing. Following the classical definition of a differential length-scale by Daley (1991, p110), the length-scale diagnosis describes the curvature of the correlation functions near their origin. Thus, the smaller the length-scale is, the faster the correlation decreases with distance.

As illustrated by several authors (Hollingsworth, 1987; Bouttier, 1993; Rabier *et al.*, 1998; Ingleby 2001, Belo Pereira and Berre, 2006; Deckmyn and Berre, 2005), this length-scale diagnosis also gives information about atmospheric dynamic (Ingleby, 2001) and data density effects (Bouttier, 1993) on the background error spatial structures. Therefore, it is attractive to be able to diagnose and to interpret length-scales at different locations. This is all the more important as ongoing research is devoted to the representation of existing heterogeneities and anisotropies (e.g. Fisher, 2003; Buehner, 2005).

In this paper, the local length-scales have been approximated by different formulae. A first purpose of the paper is thus to evaluate the ability of these various formulae to diagnose the geographical variations of the local length-scales.

In addition, with the availability of forecast ensembles, it is possible to calculate flow-dependent background error covariances "of the day" (Kalnay, 2002). However, the finite size of the ensemble induces a sampling noise, which is detrimental for the covariance estimation.

*Correspondence to: Météo-France CNRM/GMAP, 42 av. G. Coriolis, 31057 Toulouse Cedex France. e-mail: olivier.pannekoucke@meteo.fr

Regarding correlations, the sampling noise has been studied mostly with respect to long distance correlation values, which were identified as particularly noisy (Houtekamer and Mitchell, 2001). By contrast, relatively little is known about the level of noise in the estimated local length-scales. A second purpose of the current paper is thus to study this sampling noise in length-scale estimates.

The current paper deals with the length-scale as a diagnostic of exisiting correlation estimates. The focus is not on modelling correlation functions on the basis of estimated length-scales, althaugh this is another potential application in the future.

The structure of the paper is as follows. In section 2, different formulae for length-scale are derived from Daley's definition. Experimental results are illustrated in section 3, in a simple 1D analytical framework. Section 4 shows the sensitivity of length-scale estimation to the ensemble size and the spatial structure of sampling noise. Section 5 presents the comparison of two length-scale formulae, in the spherical case, using an ensemble of perturbed forecasts from the French NWP model Arpège.

2 Length-scale formulae

One of the main issues for a data assimilation system is to better specify the background error covariance matrix $\boldsymbol{B} = \mathbb{E}(\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_b^T)$, where $\boldsymbol{\varepsilon}_b$ is the forecast error assumed unbiased. In order to characterize the curvature of the correlation functions near their origin, length-scale diagnosis is often introduced.

The differential length-scale is defined in data assimilation following Daley (1991, p110). The definition is similar to the turbulent microscale. In this section, the Daley length-scale is reviewed, and formulae are derived to approximate it.

The decomposition of covariances into standard deviations and correlations is common e.g. in variational schemes. This is appropriate if standard deviations and correlations do not vary much on scales smaller than the correlation length-scale.

2.1 Daley formula

For a smooth and isotropic correlation function ρ at the origin, the Daley length-scale is given by

$$L_D = \sqrt{-\frac{1}{\nabla^2 \rho(0)}}, \tag{1}$$

in one dimension and $L_D = \sqrt{-\frac{2}{\nabla^2 \rho(0)}}$ in two dimensions. This length-scale is proportional to the turbulent (or Taylor) microscale which is similarly defined. This formula is obtained from a Taylor expansion of the correlation at the origin $\rho(0)$:

$$\rho(\delta x) \approx \rho(0) + \frac{\delta x^2}{2}\frac{d^2\rho}{dx^2}(0) = 1 - \frac{\delta x^2}{2L_D^2}. \tag{2}$$

The isotropic assumption is required in order to ensure the continuity of the second order derivative at 0, *i.e.* $\frac{d^2\rho}{dx^2}(0^-) = \frac{d^2\rho}{dx^2}(0^+)$. A geometrical interpretation of this definition of length-scale is given as the scale for which the tangential parabola at the origin is equal to 0.5. This is illustrated in the top panel of Fig. 1, where a correlation function (solid line) and its tangential parabola at the origin (dashed line) are represented. The length-scale deduced from the above geometrical interpretation is $L_D = 250\,km$, for this particular correlation function. The length-scale is also related to the curvature of the correlation function, at the origin. The radius of curvature of the correlation function at the distance r is defined by $R(r) = \frac{\left(1+\left(\frac{d\rho}{dx}(r)\right)^2\right)^{3/2}}{\frac{d^2\rho}{dx^2}(r)}$. At the origin, $\frac{d\rho}{dx}(0) = 0$ leading to $R(0) = \frac{1}{\frac{d^2\rho}{dx^2}(0)} = -L_D^2$.

Note that the Daley length-scale does not give information about the correlation anisotropy. Moreover, it requires the knowledge of the second order derivative of the correlation function. The calculation of this second order derivative can be rather costly, as ideally it should involve the calculation of the whole correlation function. The next subsections will thus describe convenient approximations of this formula.

2.2 Belo Pereira-Berre formula

Belo Pereira and Berre (2006) (hereafter noted B&B) have proposed a relatively costless formula for the computation of length-scale. Under local differentiability and local homogeneity assumptions, the variance of the spatial derivative of the forecast error can be approximated by $\left(\sigma(\partial_x \varepsilon_b(x))\right)^2 = \left(\partial_x \sigma(\varepsilon_b(x))\right)^2 - \left(\sigma(\varepsilon_b(x))\right)^2 \partial_x^2 \rho(0)$, where $\partial_x = \frac{\partial}{\partial x}$ is the derivative along the coordinate. From the Daley length-scale definition, it follows:

$$L_{B\&B} = \sqrt{\frac{\left(\sigma(\varepsilon_b(x))\right)^2}{\left(\sigma(\partial_x \varepsilon_b(x))\right)^2 - \left(\partial_x \sigma(\varepsilon_b(x))\right)^2}}, \tag{3}$$

where $\sigma\left(\varepsilon_b(x)\right)$ is the standard deviation of $\varepsilon_b(x)$. This formula method requires the computation of forecast error standard deviation, its gradient and also the standard deviation of the gradient of forecast error. In the case of a periodic domain, the computation of the gradient can be done either in grid-point space or in spectral space.

2.3 Parabola-based and Gaussian-based formula

As suggested by equation (2), a direct discretization of the Laplacian appearing in Eq. (1) leads to a simple expression of the length-scale

$$L_{Pb} = \frac{\delta x}{\sqrt{2\left(1 - \rho(\delta x)\right)}}. \tag{4}$$

This length-scale is called hereafter the parabola-based length-scale (Pb). It is based on the approximation of the correlation function by a parabolic function, as represented in Fig. 1. As suggested by the example shown in

DOI: 10.1002/qj

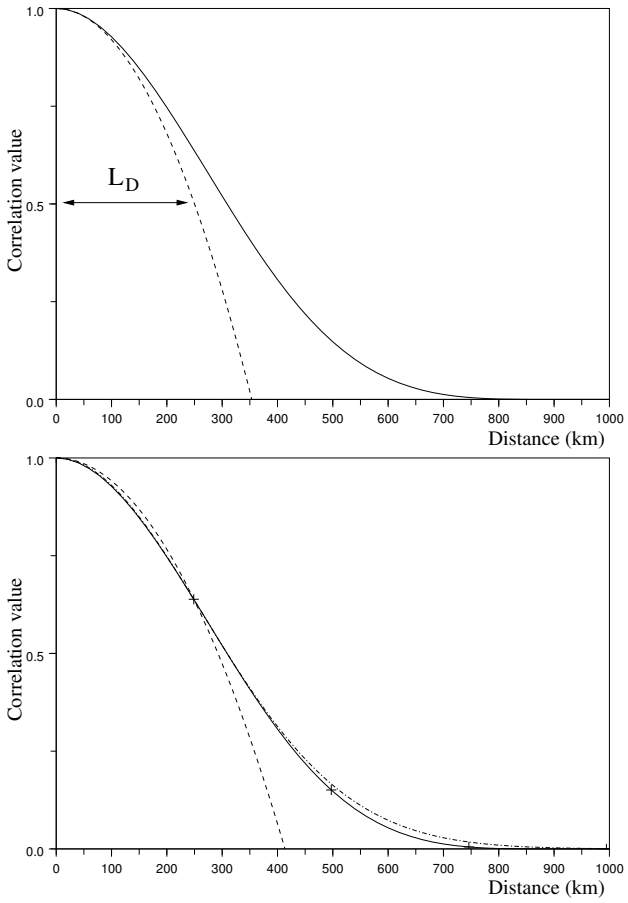


Figure 1. Top panel : Gaspari and Cohn (see section 3.1) correlation function (solid line) and its tangential parabola (dashed line). Bottom panel : Parabolic (dashed line) and Gaussian (dash-dotted line) approximations at the origin of the correlation function (solid line), determined by its value for $\delta x = 248km$, on a regular grid (crosses).

the bottom panel of Fig. 1, for some separation distances (those smaller thant the chosen distance δx), the parabolic function may decrease less quickly (from the origin to the chosen distance δx) than the true correlation function.

This suggests that the quality of the parabolic length-scale approximation may depend on the quality of the correlation function approximation and on the considered separation distance δx. Experiments indicate that the sensitivity to the choice of δx is relatively small, and that using a small value for δx provides a somewhat more accurate estimate of the length-scale. In this paper, δx corresponds to the resolution of the grid (*i.e.* the smallest possible δx).

In order to study this sensitivity to the correlation shape approximation, it is thus interesting to consider another analytical model of the correlation function ρ near the origin. By approximating the correlation at the origin by a Gaussian, the following equation is obtained: $\rho(\delta x) = exp(-\frac{\delta x^2}{2\,L_D^2})$. Inverting this equation to extract the length-scale formulation, associated to correlation at distance δx, brings

$$L_{Gb} = \frac{\delta x}{\sqrt{-2\ \ln\ \rho(\delta x)}}. \tag{5}$$

This length-scale is called hereafter the Gaussian-based length-scale (Gb). This approximation of the length-scale computation is easy to implement in real applications and costless. The bottom panel of Fig. 1 illustrates the Gaussian approximation at the origin of the discretized correlation function.

Note that when the correlation is close to one, then both Parabola-based and Gaussian-based length-scales are equal. Let $\eta = 1 - \rho$, then a Taylor expansion leads to $L_{Pb} = L_{Gb} = \frac{\delta x}{\sqrt{2\eta}}$.

2.4 Directional length-scale

Formulae (4) and (5) can be defined along an arbitrary direction as follows. Let $\boldsymbol{\delta x}$ be the displacement in a direction $\boldsymbol{u} = \frac{\boldsymbol{\delta x}}{|\boldsymbol{\delta x}|}$ of the domain (circle, plane, 2D-sphere, 3D-sphere,...). Then the vectorial parabola-based and Gaussian-based length-scale are thus defined by replacing δx by $\boldsymbol{\delta x}$ in equation (4) and (5). Thus it offers a characterization of the correlation for different directions.

Similarly, formula (1) and (3) can be defined directionally for an anisotropic correlation function. For equation (1), it consists in replacing $\Delta\rho(0)$ by $\frac{\partial^2\rho}{\partial \boldsymbol{u}^2}(0^+) = \lim_{t\to 0^+} 2\left\{\rho(t\,\boldsymbol{u}) - 1\right\} t^{-2}$, which is the second order derivative, calculated in the oriented direction $\boldsymbol{u}$, of the anisotropic correlation function. For equation (3), the directional length-scale is obtained by calculating the gradients $\partial_{\boldsymbol{u}}\varepsilon_b$ and $\partial_{\boldsymbol{u}}\sigma(\varepsilon_b)$, where $\partial_{\boldsymbol{u}}$ is the derivation along $\boldsymbol{u}$.

It should be noted that these length-scales can be calculated whether the domain is bounded or not. Thus such formulations are suitable in oceanography or for a limited area model, as well as for a global meteorological model.

In the particular case of a 1D domain, one can define a directional parabola-based left length-scale as $L_{Pb}(-\delta x)$ and a right length-scale as $L_{Pb}(+\delta x)$. A similar definition is given for the directional Gaussian-based length-scale. Thereafter, the left directional length-scale is designed by a superscript $^-$ and the right one by the superscript $^+$. Note that the ratio $\frac{L^+}{L^-}$ is an indicator of anisotropy.

2.5 Other length-scale formulae

The length-scale can be approximated in other ways, by considering various analytical expressions for correlation

$$\rho(\delta x) = f(\delta x, L_D). \tag{6}$$

The main constraint is that the formula for f has to be invertible. Thus the length-scale can be deduced from the correlation value, associated to a particular grid, with $L_D = f^{-1}\left(\delta x, \rho(\delta x)\right)$. The choice of a particular relation

between correlation and length-scale may arise from estimated correlation functions. It might depend on the physical field, or on the model used to represent the correlation function in the system.

For instance, if a SOAR function (Daley, 1991, p117) is a good model to approximate the correlation function, then one has to invert Eq. (6), with $f(\delta x, L) = (1 + \frac{\delta x}{L})e^{-\frac{\delta x}{L}}$. This inversion can be achieved by using a Newton algorithm to resolve $F(L) = 0$, with $F(L) = \rho(\delta x) - f(\delta x, L)$ where δx and $\rho(\delta x)$ are given.

Moreover, it can be noticed that such development may be applied on more complex diagnosis in 1D, 2D and 3D.

In the following, the 1D circle and the 2D sphere will be considered in order to illustrate the theory.

3 Application in a 1D analytical heterogeneous framework

3.1 A simple 1D analytical framework

Following Pannekoucke *et al.* (2007), a simple 1D analytical framework is considered to evaluate the quality of the various formulations of length-scale explored in this paper. In this framework, the geographical domain is supposed to be the equatorial circle of radius a, and the coordinate $\frac{x}{a}$ is the angle of the geographical position, varying from $0\,^\circ$ to $360\,^\circ$. On this circle, only one field is considered. A homogeneous Gaussian correlation tensor, is produced following $B_h(x,y) = e^{-\frac{(x-y)^2}{2L_H^2}}$, where x and y are two points on the circle, and L_H is the length-scale, which is here arbitrarily set equal to $L_H = 250km$. Moreover, two non-Gaussian homogeneous correlation tensors have been also defined. The first one is based on Gaspari and Cohn (1999 Eq 4.10) as $C_h(x,y) = \rho_L(x-y)$ with

$$\rho_L(r) = \begin{cases} -\frac{1}{4}\left(\frac{r}{L}\right)^5 + \frac{1}{2}\left(\frac{r}{L}\right)^4 + \\ \quad \frac{5}{8}\left(\frac{r}{L}\right)^3 - \frac{5}{3}\left(\frac{r}{L}\right)^2 + 1, 0 \le r \le L, \\ \frac{1}{12}\left(\frac{r}{L}\right)^5 - \frac{1}{2}\left(\frac{r}{L}\right)^4 + \frac{5}{8}\left(\frac{r}{L}\right)^3 + \frac{5}{3}\left(\frac{r}{L}\right)^2 \\ \quad - 5\left(\frac{r}{L}\right) + 4 - \frac{2}{3}\left(\frac{L}{r}\right), L \le r \le 2L, \\ 0, 2L \le r, \end{cases}$$

and $L = \sqrt{0.3}L_H$ in order to obtain the same theoretical length-scale as in the Gaussian case. The second non-Gaussian homogeneous tensor is similarly defined with the Second Order Auto Regressive (SOAR) correlation (Daley, 1991 p117) $\rho(r) = (1 + \frac{r}{L_H})e^{-\frac{r}{L_H}}$. The spectra on the circle of these three correlations are represented in figure 2.

Then, a heterogeneous correlation is computed using a c-stretching Schmidt transformation (Courtier and Geleyn 1988), adapted to the circle and defined by $h(x) = a\left[\pi - 2Atan\left(\frac{1}{c}tan(\frac{\pi}{2} - \frac{1}{2}\frac{x}{a})\right)\right]$ with $c = 2.4$ (the Schmidt transformation is used for a different purpose in the Arpège global stretched model to obtain a variable resolution). Its inverse is denoted by h^{-1}. The resulting heterogeneous correlation tensor is

$$B(x,y) = B_h\left(h^{-1}(x), h^{-1}(y)\right), \tag{7}$$

which provides correlation functions that are relatively sharp around $180\,^\circ$, and broad around $0\,^\circ$.

A discretized version of these correlation tensors on a given grid leads to covariance matrices, that depend on the resolution of the grid. For a given truncation T, the number of grid points is $N_g = 2T + 1$ and the homogeneous associated resolution is then $\delta x = \frac{2\pi a}{N_g}$. In this paper, we will use $T = 120$ as an example. In this experimental framework, an ensemble of generated background errors is constructed following the method described by Fisher and Courtier (1995) : $\varepsilon_b = \boldsymbol{B}^{1/2}\boldsymbol{\zeta}$, where $\boldsymbol{\zeta}$ is a Gaussian random realization with covariance matrix $\boldsymbol{I}$ and mean equal to zero.

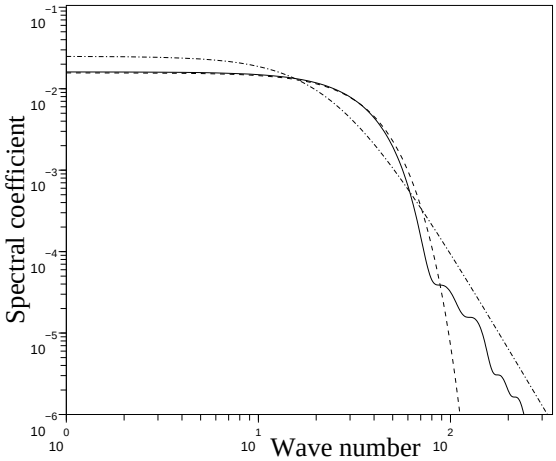


Figure 2. Spectrum of the Gaspari and Cohn correlation (solid line), of the Gaussian correlation (dashed line), and of a SOAR correlation (dash-dotted line).

3.2 Computation of length-scales in a heterogeneous case

For this numerical test, the term $(\sigma(\partial_x \varepsilon_b))^2$ that appears in the B&B length-scale is formally computed as follows. At a point index i, the term is

$$(\sigma(\partial_x \varepsilon_b)_i)^2 = \boldsymbol{\delta}_i^T \boldsymbol{D}\boldsymbol{B}\boldsymbol{D}^*\boldsymbol{\delta}_i,$$

with $\boldsymbol{D}$ the differential operator constructed in Fourier space, and $\boldsymbol{\delta}_i$ the Dirac vector whose value is set to one at index i and set to zero otherwise. In a similar way, the Daley length-scale is computed by a direct computation of the Laplacian at the origin for each correlation function. The Laplacian is computed in Fourier space.

In the 1D framework, the various formulations of length-scale are represented in figure 3 for the heterogeneous G&C-based correlation tensor and for the heterogeneous SOAR-based tensor. The parabola-based and Gaussian-based length-scales are computed as the mean

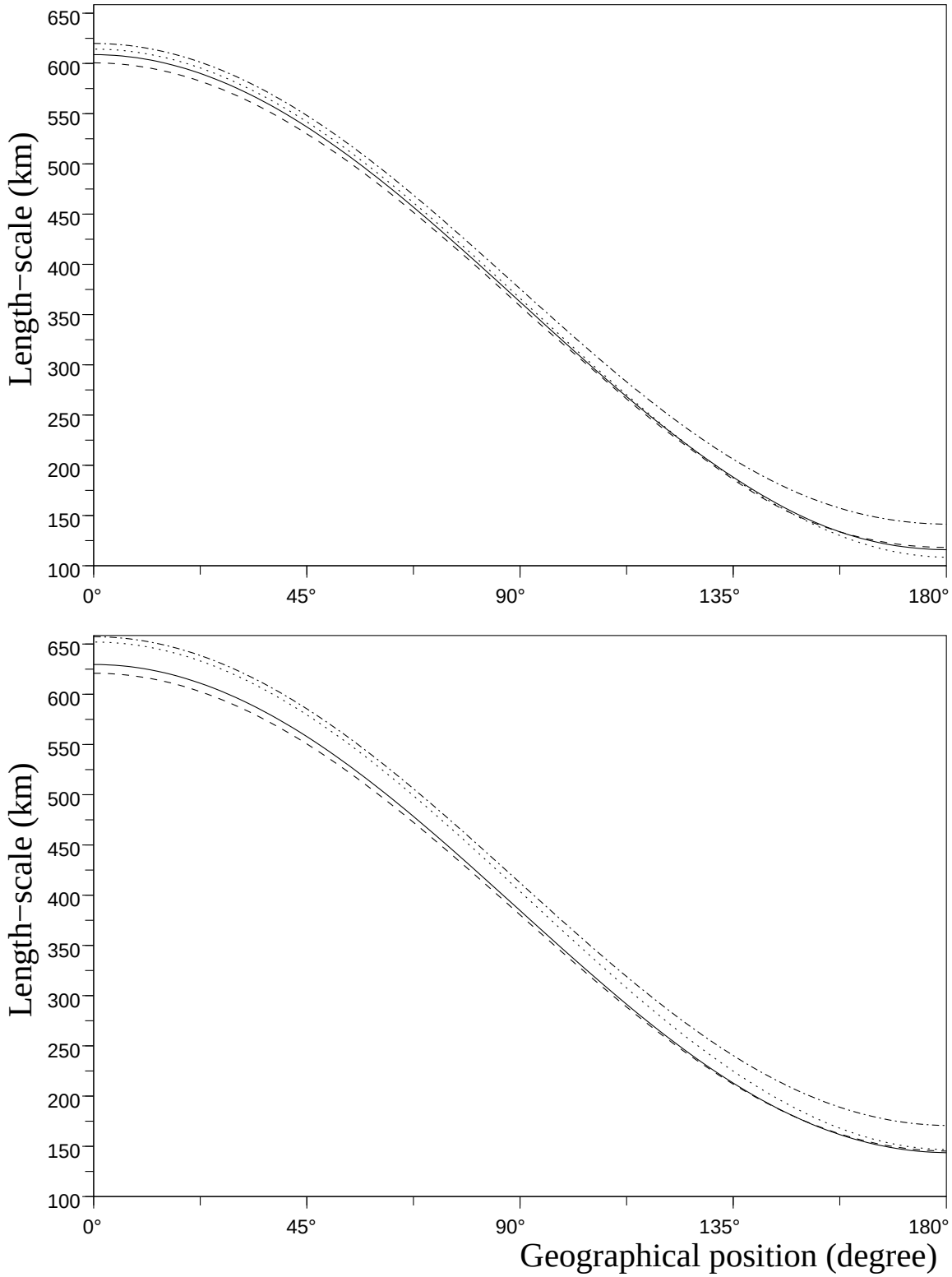


Figure 3. Local length-scales, computed with different formulae, for both G&C correlation tensor (top panel) and SOAR correlation tensor (bottom panel), discretized on the $T120$ circle. Daley (solid line), Belo Pereira-Berre (dashed line), mean parabola-based length-scales (dash-dotted line) and mean Gaussian-based length-scales (dotted line).

value $\frac{1}{2}(L^+ + L^-)$. The Daley length-scale is considered as being the numerical truth and thus the reference.

In the first experiment, represented in the top panel of Fig. 3, the analytical tensor is the heterogeneous G&C-based one. It appears that each length-scale formula is able to represent the geographical variations of the correlation structure : large length-scales near $0\,^\circ$, and small ones near

DOI: 10.1002/qj

180 °. The differences between the various formulations are small. The largest discrepancy is encountered for the parabola-based length-scale.

In the second experiment, represented in the bottom panel of Fig. 3, the analytical tensor is the heterogeneous SOAR-based one. Again, all formulations lead to similar realistic length-scale values.

In the two experiments of Fig. 3 , the parabola-based results are somewhat less accurate than the Gaussian-based length-scales. Moreover, in the bottom panel of Figure 2, the B&B formula is more accurate than the Gaussian-based formula. This indicates that the length-scale diagnosis is slightly sensitive to the underlying correlation function approximation (as mentioned in section 2.3).

Finally, it can be concluded that all formulations lead to similar length-scale values, and the geographical variations are thus well represented in these simulations.

4 Length-scale sampling statistics

4.1 Ensemble size effects in the circle framework

In practical applications, length-scales are usually estimated from a finite ensemble (*e.g.* Belo Pereira and Berre 2006). Figure 4 represents the sampling effect on the estimation of the length-scale L_D for small ensembles. In this experiment, the correlation tensor is the heterogeneous Gaussian tensor on the circle. The true length-scale L_D (dashed line) is compared to estimated length-scales (thin solid line) from 10 members (top) and from 30 members (bottom). Length-scale variations are noised by high frequency variations.

Each sample of N members leads to a particular field of length-scale estimates L^N, which can be considered as a set of random variables. It is thus interesting to know the expectation $\mathbb{E}(L^N)$ of these random variables, and their other statistical characteristics (standard deviation $\sigma(L^N)$, sampling distribution, *etc*).

The expectation function $\mathbb{E}$ is introduced as follows. It is numericaly defined, for a field α, as $\mathbb{E}(\alpha) \approx \frac{1}{N_s}\sum_k \alpha_k$ where α_k are N_s independent realizations of α. For instance, if L^N denotes the length-scale estimated from N members, $\mathbb{E}(L^N) \approx \frac{1}{N_s}\sum_k L^N_k$, where L^N_k is the k^{th} length-scale map estimated from the k^{th} sample of N members. Thereafter, N_s is large and is arbitrarly fixed in order to ensure stable statistics.

Figure 4 shows that the estimated length-scale $\mathbb{E}(L^N)$ is biased : for 10 members, $\mathbb{E}(L^{10}) \neq L$, and the length-scale bias near 0 ° is around $75km$ (12% of total). Moreover, the standard deviation illustrates an even larger distortion of the estimated length-scale, namely by 40% for 10 members (resp. 20% for 30 members).

In order to better understand how the finite size of the ensemble influences the estimation, the sampling distribution of the length-scale can be computed experimentally. In the particular case of Parabola-based and Gaussian-based length-scales, the sampling distribution can also be deduced analytically, from the sampling distribution of the correlation ρ^N between two points separated by a distance δx. This is shown in the appendix.

4.2 Gaussian-based length-scale sampling distribution

The experimental frequency distribution is represented in figure 5 for $N = 25$. It shows that the sampling distribution is positively skewed, and the existence of a bias $b^N_{Gb} = \mathbb{E}(L^N_{Gb}) - L^\infty_{Gb}$. These experimental results are consistent with analytical studies, as shown in the appendix. Positive skewness implies that large length-scale values are often encountered with such ensemble sizes.

The full line in Fig. 6 represents the relative error percentage associated to the bias $\frac{\mathbb{E}(L^N_{Gb})-L^\infty_{Gb}}{L^\infty_{Gb}}$ for a given discretization δx. In that case, the $T120$ discretized circle is considered ($\delta x \approx 166km$) and $L^\infty_{Gb} = 250km$. The error is large for a small ensemble and tiny for a large one. The convergence to the infinite-ensemble value is relatively fast, as it is in $\mathcal{O}(N^{-1})$. For 10 members, the bias ratio is 10%.

However the standard deviation is larger as shown now. Figure 6 shows the ratio $\frac{\sigma^N_{L_{Gb}}}{L^\infty_{Gb}}$ where $\sigma^N_{L_{Gb}} = \sqrt{\mathbb{E}\left\{\left(L^N_{Gb} - \mathbb{E}(L^N_{Gb})\right)^2\right\}}$. The behaviour in $\mathcal{O}(N^{-1/2})$ is observed as expected (see the appendix). For 10 members, the ratio is 40%. This illustrates the predominance of the error standard deviation over the error bias in the length-scale estimation.

4.3 Comparison with other length-scale formulae

Trying to find the sampling distribution of Daley or B&B length-scale analytically is not easy, because it depends on the shape of the function and not only on one correlation. However, numerical experiments indicate a similar behavior to that of the Pb and Gb cases. Figure 7 shows the sample distribution of length-scale for the four formulations: Daley, B&B, Pb and Gb. These length-scales are estimated from a 10 member ensemble. The true correlation tensor used here is a homogeneous Gaussian correlation tensor over the $T120$ discetized circle with length-scale L_H. It appears that the sampling distributions are similar to each other. In particular, both Daley and B&B length-scales present some bias.

4.4 The spatial structure of sampling noise

As illustrated in Fig. 4, the spatial variations of the estimated length-scales tend to be random and spatially uncorrelated, compared to variations of the exact length-scales. This suggests that the estimated length-scale field is affected by a sampling noise, whose amplitude is relatively large in the small scales (compared to the exact length-scale field).

In order to explore this issue, energy spectra have been calculated for geographical maps of the exact length-scales, of the ensemble-estimated length-scales and of the corresponding estimation errors. The results are shown in Fig. 8. As expected from Fig. 4, the ensemble-estimated

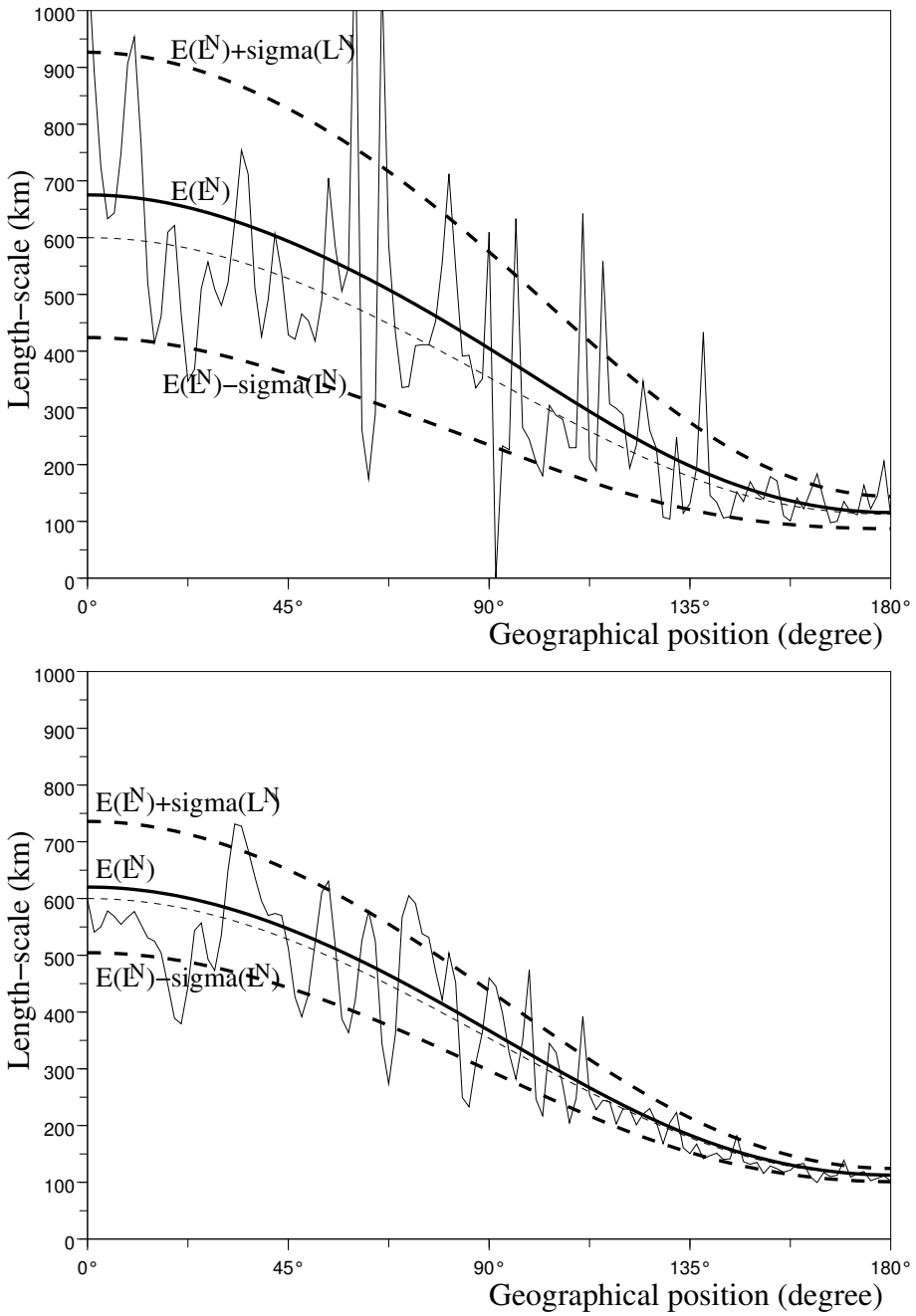


Figure 4. Sensitivity of length-scale estimation to the ensemble size. The true length-scale map (thin dashed line) is compared with the estimated length-scale (thin solid line) for $N = 10$ members (top panel) and $N = 30$ members (bottom panel). The curve of the expectation $\mathbb{E}(L^N)$ (bold solid line) illustrates the existence of a bias, and the usual range $\mathbb{E}(L^N) \pm \sigma(L^N)$ (bold dashed lines, where $\sigma(L^N)$ is the standard deviation of L^N) offers a representation of the expected range of values reached by the estimated length-scales.

length-scale maps spuriously contain much more small scale energy than the exact length-scale map. This corresponds to the artificial contribution of sampling noise, whose energy spectrum is close to a white noise.

These results indicate that spatial filtering techniques based on spectral or wavelet techniques may be worth considering. This is illustrated in the next subsection.

4.5 Sampling noise reduction through spatial filtering

Background error correlation modeling is often based on a spectral diagonal approach (Courtier *et al.*, 1998). More recently, Fisher (2003) has also defined an error correlation modeling with a wavelet diagonal approach. As discussed in Pannekoucke *et al.* (2007), using these techniques amounts to spatially averaging the local correlation functions.

In the spectral diagonal approach, this spatial averaging is global, in the sense that it is calculated as a uniform

DOI: 10.1002/qj

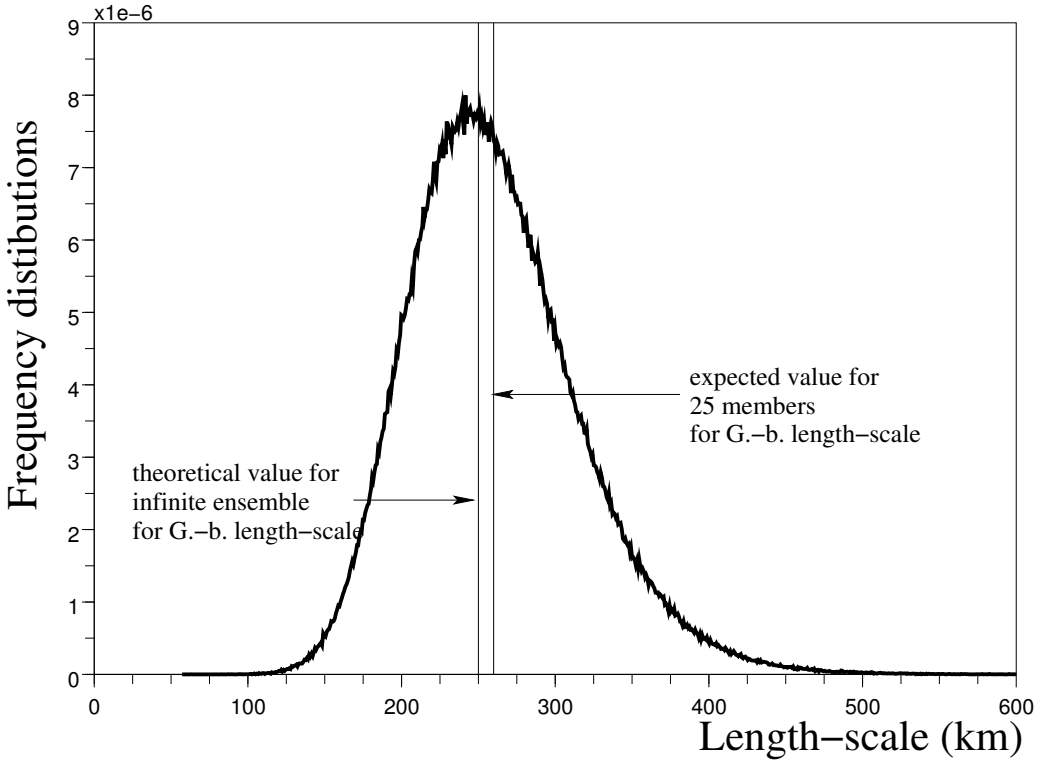


Figure 5. Sampling distribution of the Gaussian-based length-scale for $\delta x = 166\,km$ and 25 members. The theoretical length-scale $L_H = 250\,km$ is over-estimated by the expected length-scale from the finite ensemble.

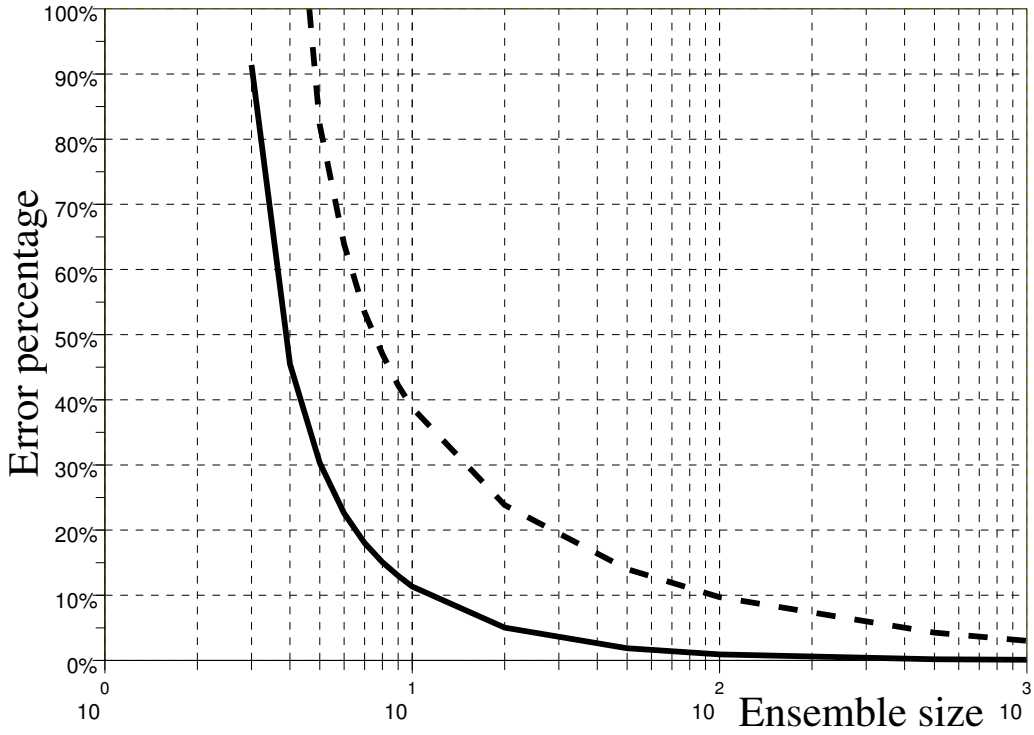


Figure 6. Convergence of length-scale error with the ensemble size: the bias (solid line) and the standard deviation (dashed line), both normalized by the true length-scale. (See text for details.)

average over the whole domain. In the wavelet diagonal approach, this spatial averaging is rather local. This means that wavelets allow the size of the statistical sample to be increased, by introducing a local spatial sample (multiplied by the ensemble sample), while keeping the

DOI: 10.1002/qj

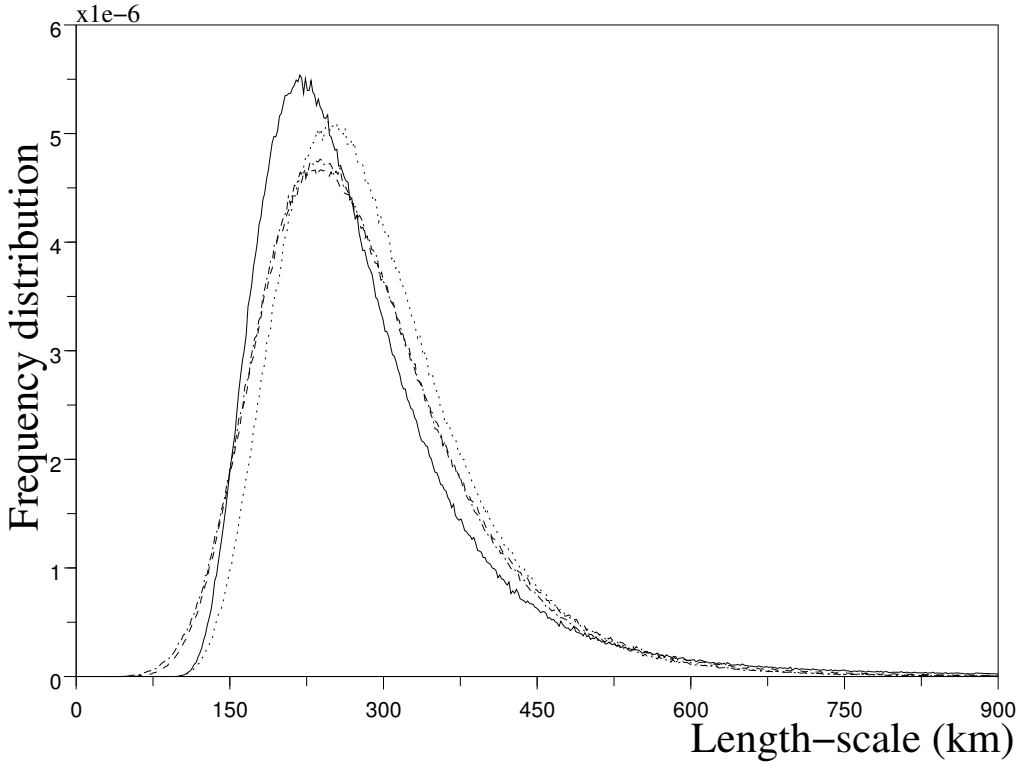


Figure 7. Comparison of sampling distributions of the length-scale, estimated from a 10 member ensemble. The correlation tensor is set equal to the Gaussian homogeneous tensor associated to the length-scale $L_H = 250\,km$ and truncation $T = 120$. Estimated length-scales : Daley (solid line), Gaussian-based (dash-doted line), parabola-based (dotted line) and Belo Pereira& Berre (dashed line). Note that Gb is almost superposed with B&B.

possbility to represent geographical variations.

The efficiency of this wavelet filtering approach has been illustrated by Pannekoucke *et al.* (2007) in a 1D heterogeneous case. Here, we will focus on a 1D homogeneous case, in order to illustrate the effect of spatial filtering on the bias and standard deviation of the length-scale error.

This 1D case corresponds to a homogeneous Gaussian correlation tensor, discretized on a T120 circle (associated to $N_g = 241$ grid points), with a theoretical length-scale equal to $L_H = 250km$. An estimation of the correlation matrix, with an ensemble of N members, leads to a heterogeneous covariance matrix. At a given point k, the Gaussian-based length-scale at this point is the mean length-scale $L_e^N = (L^+ + L^-)/2$. The correlation modelled with the diagonal assumption in spectral space is homogeneous, and corresponds to the average of the N_g estimated correlation functions. The resulting Gaussian based length-scale is thus $L_{ds} = \left(L(\bar{\rho}^+) + L(\bar{\rho}^-)\right)/2$ where $\bar{\rho}^+ = \frac{1}{N_g}\sum_k \rho_k^+$ and $\bar{\rho}^- = \frac{1}{N_g}\sum_k \rho_k^-$.

Different random variables are also introduced: L_{ds}^N is the length-scale resulting from a spectral diagonal assumption estimated with an ensemble of N members ; L_{dw}^N is the corresponding length-scale resulting from a wavelet diagonal assumption. Relative errors $\mathbb{E}(L_e^N)/L_H - 1$ (solid line), $\mathbb{E}(L_{ds}^N)/L_H - 1$ (dash-dotted line) and $\mathbb{E}(L_{dw}^N)/L_H - 1$ (dash line) are represented on the top panel of Fig. 9 for ensemble size $N \in [6, 200]$. This error corresponds to the bias normalized by the length-scale L_H. As shown in section 4.2, $\mathbb{E}(L_e^N)/L_H - 1$ converges as $\mathcal{O}(N^{-1})$, while $\mathbb{E}(L_{ds}^N)/L_H - 1$ is close to zero everywhere. $\mathbb{E}(L_{dw}^N)/L_H - 1$ is small, although it remains different from zero even for a large ensemble. This is due to a known defect of the wavelet diagonal assumption : length-scale can be under or over-estimated with as much as 10% error (Pannekoucke *et al.*, 2007).

Then to appreciate the accuracy of the estimation, the ratios σ_e^N/L_H, σ_{ds}^N/L_H and σ_{dw}^N/L_H are also represented (bottom panel of Fig. 9). These ratios represent the error standard deviation normalized by the length-scale L_H. Again, σ_e^N/L_H converges to zero as $\mathcal{O}(N^{-1/2})$. σ_{ds}^N/L_H converges at the same rate but with a factor close to $1/17 \approx 1/\sqrt{N_g}$: for $N = 6$, $\sigma_e^N/L_H \approx 50\%$, while $\sigma_{ds}^N/L_H \approx 3\%$. The convergence of σ_{dw}^N/L_H is between these two extreme convergences : it has the same rate as the ensemble, but with a factor close to $1/5$: for $N = 6$, $\sigma_e^N/L_H \approx 50\%$ while $\sigma_{dw}^N/L_H \approx 10\%$.

These results illustrate the property of *e.g.* a wavelet formulation to represent the length-scale values with a better accuracy (here a factor 5) than the direct ensemble estimation.

DOI: 10.1002/qj

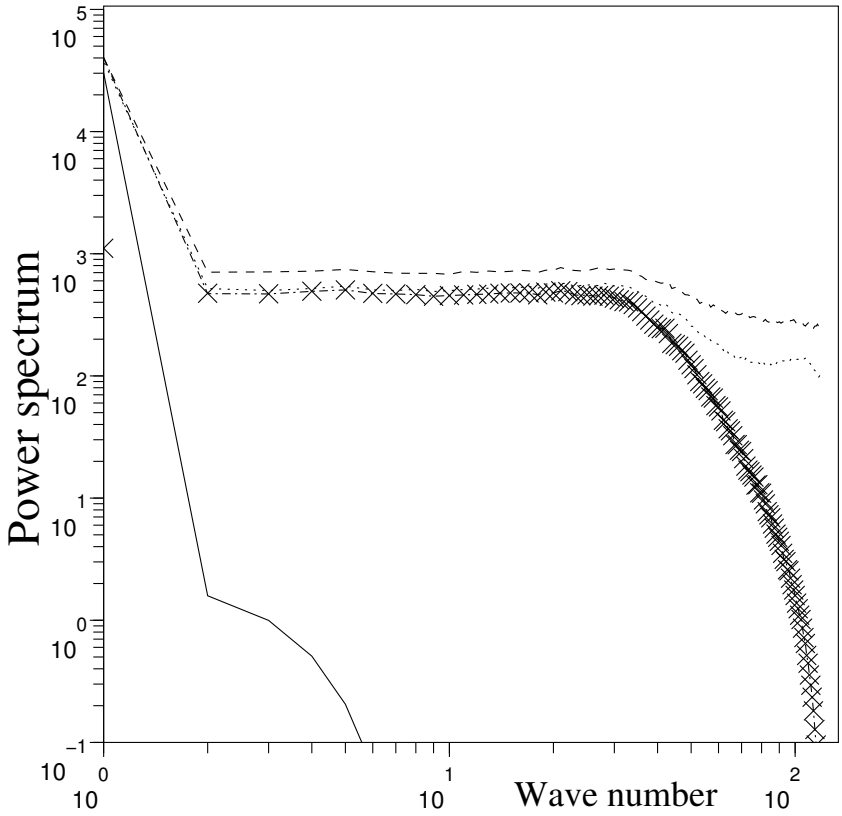


Figure 8. Energy spectra for geographical maps of the exact length-scales (solid line) and of the ensemble estimated length-scales, with 10 members: Daley (dashed line), B&B (dotted line) and Gb (dash-dotted line). Pb (not shown) is similar to Gb. The energy spectrum of the Gb estimation errors is represented by crosses, which are almost superposed with the Gb map energy, except for wavenumber 1. Note also that the spectrum of the exact length-scales (solid line) reflects the predominance of wave number 1, in accordance with figure (3).

5 Application to an ensemble of NWP forecasts

An application to an ensemble of NWP forecasts has been studied, by using an operational non-stretched version of the Arpège model (Courtier and Geleyn, 1988), whose assimilation system is a 4D-Var scheme (Rabier *et al.*, 2000; Veersé and Thépaut, 1998). The background error covariance matrix is calculated by using an ensemble of perturbed assimilation runs (Houtekamer *et al.*, 1996, Fisher 2003). The detailed results for this Arpège ensemble are described in Belo Pereira and Berre (2006).

The available ensemble consists in a set of 6 forecast differences for each day of the period 9 February to 24 March 2002, and time-averaged covariances are calculated over this 49-day period. Figure 10 presents the results obtained with the B&B zonal length-scale (top panel) and with the Gaussian-based zonal length-scale (bottom panel) for the logarithm of surface pressure. As in the previous subsection, the zonal gradient in the B&B length-scale is computed in spectral space. Each formulation represents well the land-sea contrast, and the influence of the orography, with *e.g.* larger values over tropical oceans and smaller values near the Andes. Actually there are only slight differences between the two formulations of length-scale. This supports the idea that the Gaussian-shape assumption near the origin is acceptable, leading to realistic length-scale values.

It may be mentioned that such maps of length-scales provide a full vision of geographical variations in the curvature of correlation functions. Such geographical variations can thus be examined with more details than when only plotting correlation functions at a few selected points on the globe (as *e.g.* in Baker *et al.*(1987)). On the other hand, it should be reminded that length-scales give information about the correlation curvature near the origin only, while full correlation functions provide information about all separation distances. These two diagnostics are thus to be seen as complementary.

6 Conclusion

Some approximations of the theoretical Daley (1991, p110) length-scale have been discussed in this paper. In particular, an economical estimation based on a Gaussian assumption has been investigated. Firstly, it has been shown in a 1D heterogeneous context that the different length-scale formulae provide similar realistic length-scale values and variations.

Secondly, a study of the sampling distribution of the estimated length-scales has been carried out, both analytically and experimentally. It has been shown that the estimated length-scales are affected by a positive bias when the ensemble size N is small. This bias converges towards zero in $\mathcal{O}(N^{-1})$. This bias has been shown to be

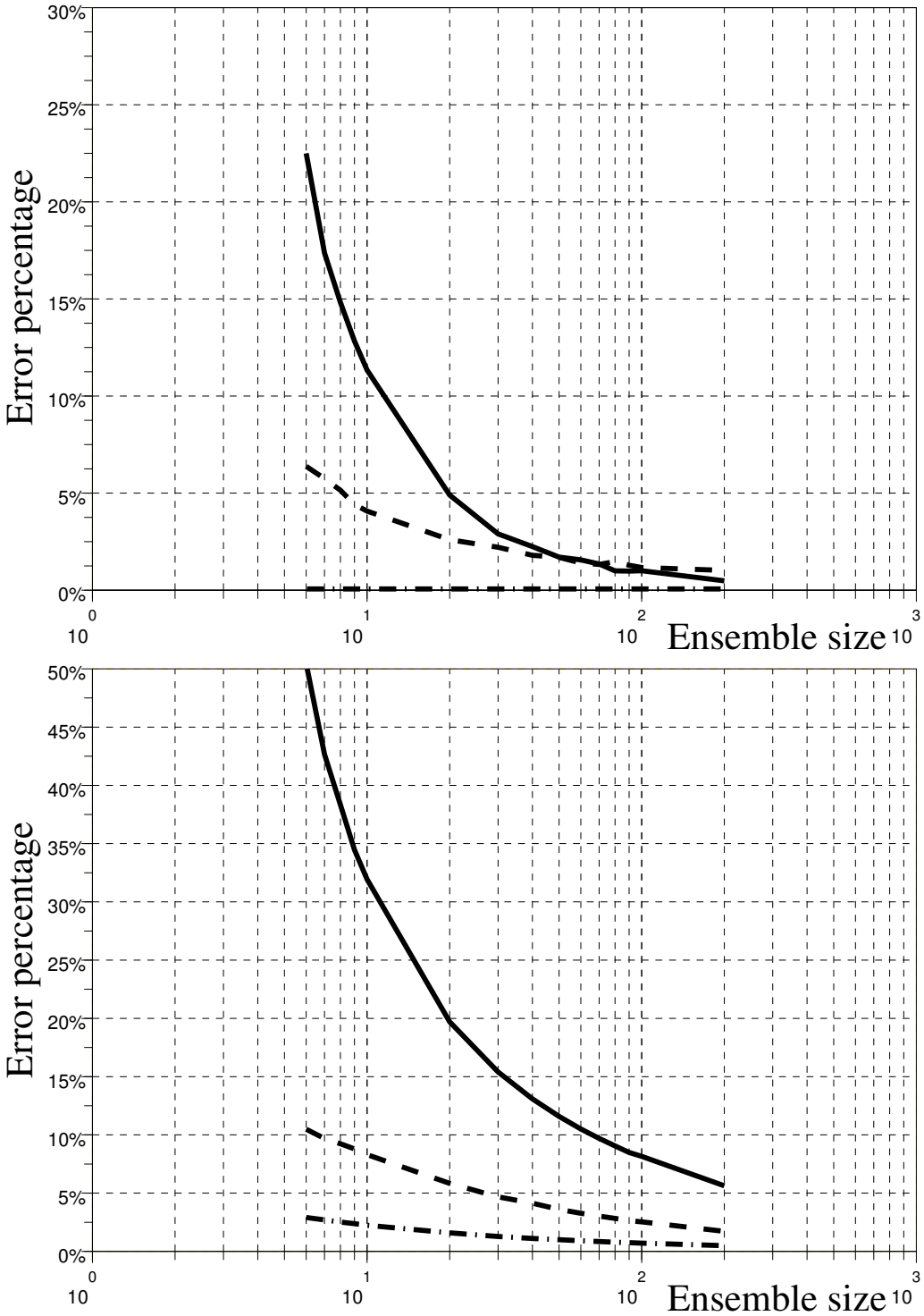


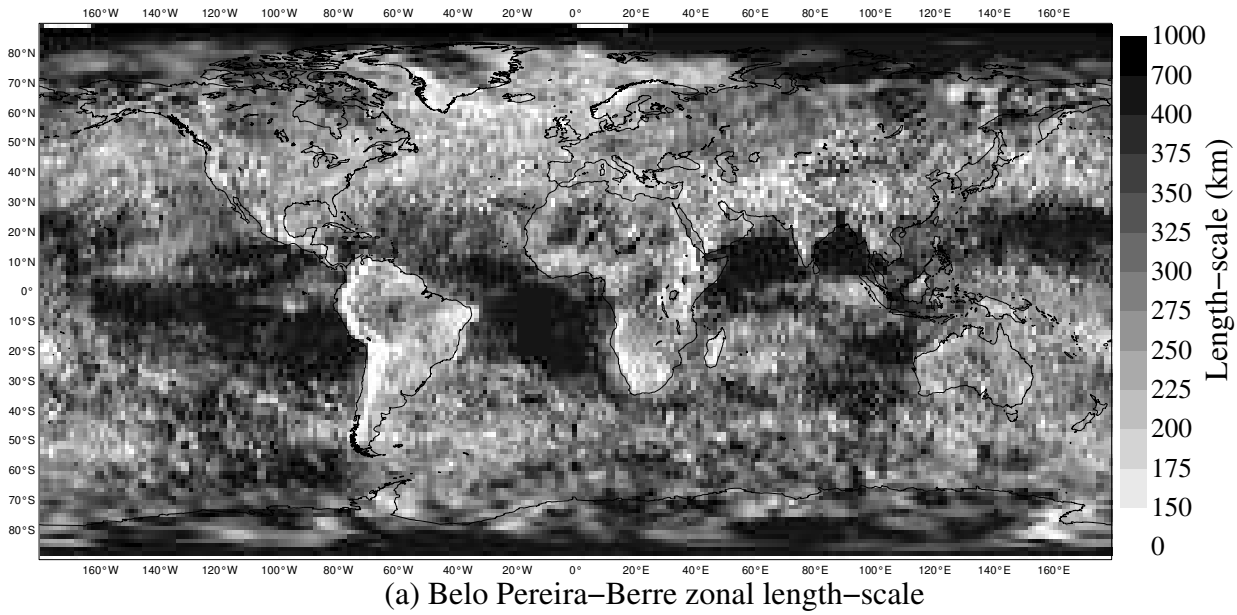
Figure 9. Comparison of the convergence of length-scale error with the ensemble size : directly estimated (solid line), resulting from a diagonal asumption in wavelet space (dashed line) and in spectral space (dash-dotted line). Top panel : bias normalized by L_H. Bottom panel : standard deviation normalized by L_H. (See text for details.)

smaller than the estimation error standard deviation. The latter converges towards zero in $\mathcal{O}(N^{-1/2})$.

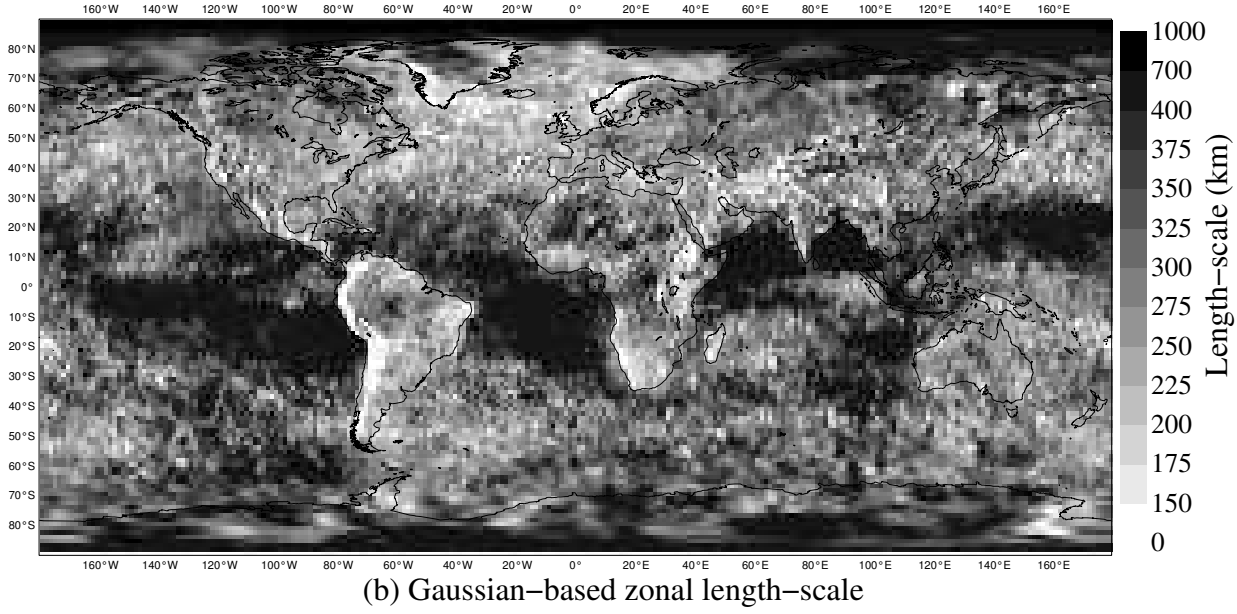
In addition, the examination of length-scale geographical variations and of their energy spectrum indicates that the sampling noise tends to be uncorrelated spatially (typically like white noise). This suggest that local space averaging techniques, such as those based on wavelets, are worth considering in order to spatially filter sampling noise.

Finally, the Belo Pereira and Berre formula has been compared to the Gaussian-Based length-scale on a 2D spherical example from a NWP ensemble data set. Length-scale values and variations appear to be similar according to the two formulae. This indicates that the

DOI: 10.1002/qj



(a) Belo Pereira–Berre zonal length–scale



(b) Gaussian–based zonal length–scale

Figure 10. Zonal length-scale of the surface pressure logarithm, numerically computed with Belo Pereira-Berre (a) and Gaussian-based (b) formulae. This is a 'climatological' average, calculated over 49 days.

assumption that the shape of the correlation function is Gaussian is reasonable in order to estimate the length-scale (defined by Daley (1991)).

The possibility to calibrate a correlation model from length-scale estimates is another potential application of the considered formulae in this paper, even if this issue has not been investiagated here. A relatively obvious limitation is that the knowledge of the length-scales may not be sufficient to determine an accurate model of the whole correlation functions, e.g. because length-scales characterize the curvature of correlation functions near their origin only. With this perspective in mind, correlation modelling based *e.g.* on wavelets may be more attractive than modelling based *e.g.* on a Gaussian approximation and on a specification of length-scales only.

7 Appendix

7.1 Approximation of the correlation sampling distribution

The Normal distribution of a correlated background error pair $\boldsymbol{\varepsilon}_b = (\varepsilon_1, \varepsilon_2)$ may be written as $f_b(\boldsymbol{\varepsilon}_b) = \frac{1}{2\pi|\boldsymbol{B}|^{1/2}} exp\left(-1/2||\boldsymbol{\varepsilon}_b||^2_{\boldsymbol{B}^{-1}}\right)$, where the covariance matrix is $\boldsymbol{B} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$, whose determinant is $|\boldsymbol{B}|$ and whose correlation is ρ. From N sample values $(\varepsilon_1^1, \varepsilon_2^1) \ldots (\varepsilon_1^N, \varepsilon_1^N)$, the corresponding sample variances are $S_1^2 = \frac{1}{N}\sum_k {\varepsilon_1^k}^2$, $S_2^2 = \frac{1}{N}\sum_k {\varepsilon_2^k}^2$, and the sample correlation is $C = \frac{1}{N}\sum_k \varepsilon_1^k\varepsilon_2^k/(S_1S_2)$. Of course, S_1^2, S_2^2 and C are random variables.

Fisher (1953) has expressed the sampling distribution of C (Kendall *et al.*, 1998, Hotteling, 1953). But this formulation is in fact too complex, and some approximations of this sampling distribution must be used. For a large ensemble, the distribution is close to Gaussian. For a small ensemble, the distribution is not Gaussian, and its skewness increases with the correlation value. When N is not too small, typically $N \geq 25$ members, Fisher has proposed a suitable transformation, where the convergence to Gaussianity of the new variable is accelerated. Then the random variable $Z = tanh^{-1}C$ follows, with a good approximation, a Gaussian distribution:

$$Z \sim \mathcal{N}\left(\mu_Z(\rho,N), \sigma_Z^2(\rho,N)\right), \tag{8}$$

with the mean $\mu_Z(\rho,N) = \zeta + \frac{\rho}{2(N-1)} + \frac{\rho(5+\rho^2)}{8(N-1)^2} + \frac{\rho(11+2\rho^2+3\rho^4)}{16(N-1)^3} + \mathcal{O}\left(N^{-4}\right)$ where $\zeta = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ and the standard deviation $\sigma_Z(\rho,N)^2 = \frac{1}{N-1} + \frac{4-\rho^2}{2(N-1)^2} + \frac{22-6\rho^2-3\rho^4}{6(N-1)^3} + \mathcal{O}\left(N^{-4}\right)$. A simple change of variable leads to the sampling distribution of correlation

$$f_C(c) = \frac{1}{(1-c^2)\sigma_Z(\rho,N)\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{\left[Z(c)-\mu_Z(\rho,N)\right]^2}{\sigma_Z(\rho,N)^2}\right\}, \tag{9}$$

where by definition $P(C \in [c, c+dc]) = f_C(c)dc$, where P is the probability measure. The top panel of figure 11 shows this sampling distribution (bold solid line) for $N = 25$ members and $\rho = exp(-\delta x^2/2L_H^2) \approx 0.8$. An experimental frequency distribution (solid line) is obtained numerically for $N = 25$. It illustrates the accuracy of the Fisher approximation.

As suggested by the above equations of μ_Z and of σ_Z^2, the bias and standard deviation of Z converge towards zero in $\mathcal{O}(N^{-1})$ and in $\mathcal{O}(N^{-1/2})$ respectively. The rates of convergence are similar for the correlation C.

7.2 Approximation of the Gaussian-based length-scale sampling distribution

Applying the Fisher's transformation to length-scale leads to the sampling distribution of length-scales. The calculation is given for the Gaussian-based length-scale, knowing that the parabola-based case is similar.

Actually, for the Gb, correlation must be positive. Thus the correlation sampling distribution has to be limited to the positive correlation part. Let $\chi_{(0,1]}$ be the characteristic function defined on $[-1,1]$; it is equal to one on $(0,1]$ and null otherwise. The random variable associated to positive correlation is $C^+ = \chi_{(0,1]}C$. Its sample distribution is $f_{C^+}(c) = \Lambda(\rho,N)^{-1} f_C(c)$, $c > 0$, where $\Lambda(\rho,N) = \int_0^1 f_C(c)dc$ is the normalization term ; it can be approximated by $\Lambda(\rho,N) \approx 1 - \frac{\Gamma(N)}{\Gamma(N+1/2)\sqrt{2\pi}}\frac{(1-\rho^2)^{N/2}}{\rho}$ (Hotteling, 1953). The change of variable $C^+ = \exp\left(-\frac{\delta x^2}{2L_{Gb}^{N\ 2}}\right)$, with L_{Gb}^N the estimator of the Gb length-scale calculated with a N member ensemble, leads to the sampling distribution

$$f_{L_{Gb}^N}(l) = \frac{\Lambda(\rho,N)^{-1}\delta x^2}{2l^3 \sinh(\frac{\delta x^2}{2l^2})\sigma_Z(\rho,N)\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{\left[Z\left(\rho(l)\right)-\mu_Z(\rho,N)\right]^2}{\sigma_Z(\rho,N)^2}\right\}, \tag{10}$$

with $\rho(l) = \exp\left(-\frac{\delta x^2}{l^2}\right)$. The bottom panel of figure 11 illustrates the frequency distribution and approximation of the sampling distribution for both Pb and Gb length-scales. These results are obtained with the correlation value $\rho \approx 0.8$ of the previous section. It appears that the analytical approximations of sampling distribution are in accordance with the experimental frequency distribution, with a sufficient accuracy.

In the case of an infinite ensemble, the expected value is $L_{Gb}^{\infty} = \frac{\delta x}{\sqrt{-2\,ln(\rho))}}$. In the N-member case, the resulting sampling distribution is positively skewed, and L_{Gb}^N is a positively biased estimator: $b_{Gb}^N = \mathbb{E}(L_{Gb}^N) - L_{Gb}^{\infty} > 0$. Again it can be deduced from the analytical approximation of the sampling distribution that, when N is large there is a convergence to zero in $\mathcal{O}(N^{-1})$ for the bias, and in $\mathcal{O}(N^{-1/2})$ for the standard deviation.

This result is also valid for other length-scale formulae under the assumption that background error distribution is Gaussian.

References

Baker W., Bloom S., Woollen J., Nestler M., Brin E., Schlatter T. and Branstator G. 1987. *Experiments with a three-dimensional statistical objective analysis scheme using FGGE data. Mon. Wea. Rev.*, **115**, 272–296.

Belo Pereira M. and Berre L. 2006. *The use of an Ensemble approach to study the Background Error Covariances in a Global NWP model. Mon. Wea. Rev.*, **134**, 2466–2489.

Bouttier F. 1993. *The dynamics of error covariances in a barotropic model. Tellus*, **45A**, 408–423.

Buehner M. 2005. *Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting. Q.J.R. Meteorol. Soc.*, **131**, 1013–1043.

Courtier P. and Geleyn J.F. 1988. *A global numerical weather prediction model with variable resolution : Application to the shallow-water equations. Q.J.R. Meteorol. Soc.*, **114**, 1321–1346.

Courtier P., Andersson E., Heckley W., Pailleux J., Vasiljević D., Hamrud M., Hollingsworth A., Rabier F. and Fisher M. 1998. *The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. Quart. J. Roy. Meteor. Soc.*, **124**, 1783–1807.

Daley R. 1991. *Atmospheric Data Analysis*. Cambridge University Press. p471.

Deckmyn A. and Berre L. 2005. *A wavelet approach to representing background error covariances in a LAM. Mon. Wea. Rev.*, **133**, 1279-1294.

Fisher R.A. 1953. *On the 'probable error' of a coefficient of correlation deduced from a small sample*. Metron, **1**, 1–32.

Fisher M. and Courtier P. 1995. *Estimating the covariance matrices of analysis and forecast error in variational data assimilation.* ECMWF Technical Memorandum, **220**, 29pp.

Fisher M. 2003. *Background error covariance modelling*. Processing of the ECMWF Seminar on "Recent developments in data assimilation for atmosphere and ocean", Reading, 8–12 September 2003, 45–63.

Gaspari G. and Cohn S. 1999. *Construction of correlation functions in two and three dimensions. Q.J.R. Meteorol. Soc.*, **125**, 723–757.

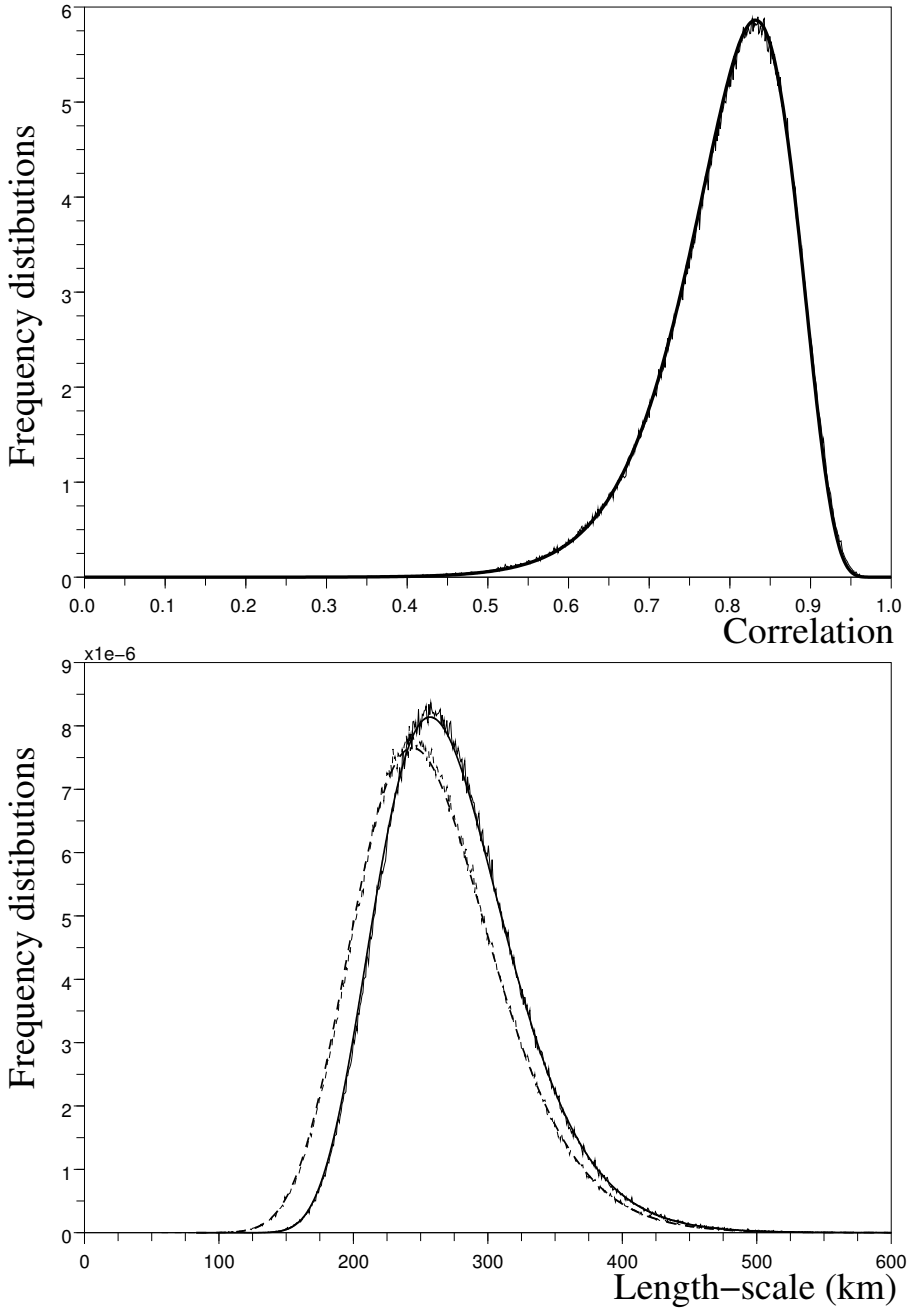


Figure 11. Comparison of approximated and estimated sampling distributions resulting from a 25 member ensemble. Top panel : sampling distribution for correlation $\rho \approx 0.8$, either theoretically approximated (bold solid line), or estimated experimentally (thin solid line). Bottom panel : sampling distribution of both parabola-based and Gaussian-based length-scales for $\delta x = 166\,km$. The theoretical approximation of the parabola-based length-scale (resp. Gaussian-based) is in bold solid line (resp. bold dashed line), while its sampled estimation is in thin solid line (resp. thin dashed line).

Hollingsworth A. 1987. *Short- and medium-range numerical weather prediction*. Collection of papers presented at the WMO/IUGG symposium, Tokyo, 4–8 August 1986.

Houtekamer P.L., Lefaivre L., Derome J., Ritchie H. and Mitchell H.L. 1996. *A system simulation approach to ensemble prediction. Mon. Wea. Rev.*, **124**, 1225–1242.

Houtekamer P.L. and Mitchell H.L. 2001. *A sequential ensemble Kalman filter for Atmospheric Data Assimilation. Mon. Wea. Rev.*, **129**, 123–137.

Hoteling H. 1953. *New light on the correlation coefficient and its transforms. Journal of the Royal Statistical Society*. Series B (Methodological), **15**, 193–232.

Ingleby B. 2001. *The statistical structure of forecast errors and its representation in The Met. Office Global Model. Q.J.R. Meteorol. Soc.*, **124**, 1783–1807.

Kalnay E. 2002. *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, p364.

Kendall M., Stuart A. and Ord J.K. 1998. *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*. A Hodder Arnold Publication.

Pannekoucke O., Berre L. and Desroziers G. 2007. *Filtering properties of wavelets for local background-error correlations. Q.J.R. Meteorol. Soc.*, **133**, 363–379.

Rabier F., McNally A., Andersson E., Courtier P., Undén P., Eyre J., Hollingsworth A. and Bouttier F. 1998. *The ECMWF implementation of three-dimensional variational assimilation (3D-Var). II: Structure*

functions. *Quart. J. Roy. Meteor. Soc.*, **124**, 1809–1829.

Rabier F., Jarvinen H., Klinker E., Mahfouf J.F. and Simmons A. 2000. *The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics*. *Q.J.R. Meteorol. Soc.*, **126**, 1148–1170.

Veersé F. and Thépaut J-N. 1998. *Multiple-truncation incremental approach for four-dimensional variational data assimilation*. *Q.J.R. Meteorol. Soc.*, **124**, 1889–1908.

Annexe C

A variational assimilation ensemble and the spatial filtering of its error covariances : increase of sample size by local spatial averaging.

Proceedings of ECMWF Workshop on Flow-dependent Aspects of Data Assimilation (Berre L., Pannekoucke O., Desroziers G., Stefanescu S., Chapnik B. and Raynaud L. 2007. 151–168)

A variational assimilation ensemble and the spatial filtering of its error covariances: increase of sample size by local spatial averaging

Loïk Berre[*], Olivier Pannekoucke[*], Gérald Desroziers[*], Simona Ecaterina Ştefănescu[+], Bernard Chapnik[*] and Laure Raynaud[*]

[*] *Météo-France, CNRM/GAME-GMAP, Toulouse, France*
[+] *National Meteorological Administration, LMN, Bucharest, Romania*

Abstract:

An ensemble of perturbed assimilations is a powerful technique to simulate the space and time dynamics of errors in an operational assimilation system. Using an ensemble of variational assimilations is relatively straightforward, as its basic elements can be directly derived from the operational variational scheme. Moreover, with respect to the analysis update of the perturbations, the potential benefit is both to adequately simulate the effect of the operational variational gain matrix, and to take advantage of its realism (e.g. non linear mass/wind balances), of its appropriate filtering, and of its full rank.

This includes available spectral and wavelet covariance tools also, which are believed to ease both an optimal estimation of filtered (possibly hybrid) ensemble covariances, and their use in the construction of the analysis members (including a careful filtering of sampling noise). It is also noticed that with e.g. an ensemble of 3D-Fgat, the analysis part of the assimilation ensemble cost is relatively small, compared to the forecast part.

In order to reduce the amplitude of sampling noise, local spatial averaging can be applied, as it allows the ensemble size to be multiplied by a 2D spatial sample size. A methodology, based on the comparison of statistics of two independent ensembles which have the same size, is also presented to estimate signal and sampling noise statistics. It is shown that the spatial structure of sampling noise is relatively small scale. This justifies the application of spatial filtering.

Following the usual linear estimation theory, it is shown that signal-to-noise ratios can be used also as objective and optimal filtering coefficients. The results indicate that a small ensemble (with typically 3 to 10 members) can provide relevant and robust information about flow-dependent error standard deviations, with e.g. larger values near troughs than near ridges. Comparison with innovation-based estimates and impact experiments tend to support these results. The (expected and effective) localized and positive nature of the flow-dependent modifications and impacts is also shown.

The use of wavelets for correlation modelling can also be seen as a spatial filtering tool applied to raw ensemble correlations, in order to improve their accuracy. These attractive filtering properties are illustrated too.

1 The variational assimilation ensemble at Météo-France

1.1 Simulating the error evolution of a reference assimilation cycle

Using an ensemble of perturbed assimilations is now relatively usual, in order to estimate e.g. climatological error covariances, which can be specified in variational assimilation schemes for instance (e.g. Houtekamer et al 1996, Fisher 2003, Buehner 2005, Berre et al 2006). These climatological estimates may include not only globally averaged covariances, but also local standard deviations, whose impact has been shown to be positive in Belo Pereira and Berre (2006).

The idea is to simulate the error evolution of the reference assimilation system (e.g. the operational 4D-Var assimilation cycle), by (explicitly or implicitly) perturbing observations and the background, in order to estimate the corresponding error covariances. Compared e.g. to the NMC method (Parrish and Derber 1992, Rabier et al 1998), one of the strengths of this approach is the ability to simulate the analysis effect in the error evolution (Berre et al 2006), and in particular the data density and the effect of the (sub-optimal) gain matrix of the reference system.

This can be illustrated by noting that the same basic equation and sub-optimal gain matrix $\mathbf{K}$ are involved in the equations for the analyzed state $\mathbf{x_a}$, for the exact state $\mathbf{x}_\star$, and for the analysis error $\mathbf{e_a} = \mathbf{x_a} - \mathbf{x}_\star$:

$$\mathbf{x_a} = (\mathbf{I} - \mathbf{KH})\mathbf{x_b} + \mathbf{Ky}$$

$$\mathbf{x}_\star = (\mathbf{I} - \mathbf{KH})\mathbf{x}_\star + \mathbf{Ky}_\star$$

$$\mathbf{e_a} = (\mathbf{I} - \mathbf{KH})\mathbf{e_b} + \mathbf{Ke_o}$$

where $\mathbf{y}_\star = \mathbf{Hx}_\star$, and where the third equation is simply deduced from the difference between the first two equations. This analysis error equation indicates that, when simulating the error evolution of the reference system, one should use the reference (sub-optimal) gain matrix $\mathbf{K}$ to transform observation and background perturbations into analysis perturbations. This is exactly what the ensemble of assimilations allows to be done, by implicitly handling differences $\varepsilon_\mathbf{a}$ between analyses which use this reference gain matrix $\mathbf{K}$:

$$\varepsilon_\mathbf{a} = (\mathbf{I} - \mathbf{KH})\varepsilon_\mathbf{b} + \mathbf{K}\varepsilon_\mathbf{o}$$

Moreover, it remains possible to incorporate e.g. flow-dependent ensemble information in this reference $\mathbf{K}$ matrix, for both deterministic and perturbed assimilations (see section 1.3 also). Thus, in the remainder of this paper, it will be suggested that, with respect to the construction of the analysis members, combining an ensemble of variational assimilations and spatial filtering tools may be an efficient approach, in order both to adequately simulate the analysis error step and to optimally filter sampling noise.

1.2 Design of the variational ensemble

Following the aforementioned studies on climatological covariance estimates, a real time variational assimilation ensemble is under development at Météo-France, in order to estimate flow-dependent covariances (as a next logical step). One of the main issues is naturally to decide about strategies with respect to the ensemble size and cost. The design of the ensemble has thus been guided by the following elements.

Firstly, it has been noticed that a small number of members (e.g. 3 to 10) already provides a lot of robust and interesting information. This will be shown e.g. in section 3, and it has been illustrated by Kucukkaraca and Fisher (2006) also.

Moreover, this robustness may be further enhanced by using spatial ergodic properties, in a similar way as in the domain of turbulence (with temporal ergodic properties, see e.g. Monin and Yaglom 1971). The idea is to increase the sample size, by calculating a local spatial average.

Another natural idea is that the full (reference) assimilation system may be approximated for the purpose of the error simulation. This is often achieved by lowering the model resolution, but one may also consider to approximate the assimilation scheme (e.g. by approximating 4D-Var with 3D-Fgat).

Based on these ideas, six global assimilation members are running in nearly real time at Météo-France. They are based on the Arpège model (Courtier and Geleyn 1988), with truncation T359, a stretching factor equal to $c = 1$ (uniform resolution), 46 vertical levels and 3D-Fgat.

Diagnostic studies have indicated that e.g. standard deviation maps were relatively similar between the 3D-Fgat and 4D-Var ensembles, as in particular they reflect the weather situation (e.g. positions of troughs and ridges) and data density effects (plus effects of common parts of the variational gain matrix).

With such a configuration, the cost of the assimilation ensemble is similar to the cost of the operational 4D-Var cycle at Météo-France (with T359 c2.4 L46 for high resolution trajectories). This rather small cost of the assimilation ensemble reflects also the fact that in the 3D-Fgat ensemble, the analysis cost is relatively small, compared to the forecast cost. In other words, running an assimilation ensemble is not much more costly than simply running a forecast ensemble.

Moreover, 3D-Fgat is easy to build from the operational 4D-Var configuration, as the basic elements of 3D-Fgat can be directly derived from 4D-Var. In addition, this configuration will make it easy in the future to consider possible intermediate configurations between 3D-Fgat and the full operational 4D-Var (e.g. an ensemble of 4D-Var with a reduced number of minimizations).

It may be also mentioned that this global assimilation ensemble has provided lateral boundary conditions, during two experimental periods of two weeks, to a regional ensemble with the Aladin limited area model (at 10 km resolution). This Aladin ensemble has provided initial and lateral boundary conditions to a high resolution ensemble with the Arome model at 2.5 km resolution. The results (Desroziers et al 2007) indicate the ability of these ensembles to reflect expected seasonal contrasts in the covariances.

1.3 Comparison with an ETKF-Var hybrid scheme

It may be of interest to compare such a variational assimilation ensemble with a hybrid ETKF-Var scheme, with respect to the gain matrix to be used for the analysis update of the perturbations. In such a hybrid ETKF-Var scheme, the ETKF (Bishop et al 2001) is used to determine how background perturbations are to be updated into analysis perturbations. The corresponding **B** matrix is in fact solely derived from the available ensemble of backgrounds, without any representation of the effect of the actual gain matrix used in the reference 4D-Var run (except for observation operators, which can be taken from 4D-Var).

In contrast, in a variational assimilation ensemble, the gain matrix for the analysis perturbation update adequately represents the actual operators used in the **B** matrix of 4D-Var (e.g. non linear balances, standard deviations and correlations, either climatological, or flow-dependent, or hybrid). In fact, the 3D-Fgat variational gain matrix can still incorporate flow-dependent ensemble information, either in the form of extra control variables (Lorenc 2003, Buehner 2005), or in the standard deviations and correlations (with possibly optimized spatial filtering, as will be shown later on). The potential benefit of the variational assimilation ensemble corresponds thus also to the realism of e.g. non-linear mass/wind balances (Fisher 2003), to the full rank of the

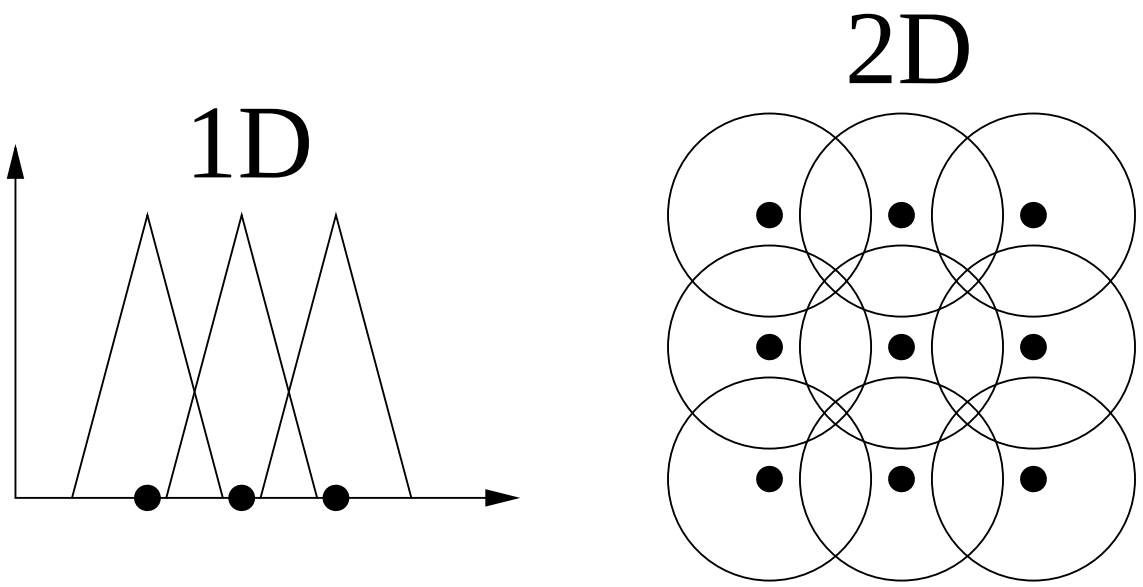


Figure 1: Visualization of the number N_g of nearest gridpoints in simple 1D ($N_g = 3$) and 2D ($N_g = 9$) cases. In 1D, the curves correspond to (simplified) correlation functions. In 2D, the isolines correspond to the distance at which the background error correlation is close to zero.

covariances, and to their optimization facilities.

A common feature is that in both cases, the cost of the assimilation ensemble is more or less dominated by the cost of the forecast (while the analysis part is relatively costless). As mentioned previously, the variational framework offers the possibility also to consider intermediate configurations between 3D-Fgat and 4D-Var in the assimilation ensemble.

2 The concept of local spatial averaging

2.1 Increase of sample size

Said shortly, the basic idea is to MULTIPLY(!) the ensemble size N_e by a number N_g of gridpoint samples, to define a total sample size $N_t = N_e \times N_g$ (over which covariances are calculated).

The simplest case of a local gridpoint average is illustrated in Figure 1 for the 1D and 2D cases, by considering an average over the nearest gridpoints. In the 1D case, $N_g = 3$, while in the 2D case, $N_g = 9$. This means for instance that if $N_e = 6$, then the 2D total sample size will be equal to $N_t = 9 \times 6 = 54$. This corresponds to an increase by almost one order of magnitude.

Depending on the parameter and configuration, it may be also considered to calculate a spatial average over a larger number of gridpoints. Thanks to the 2D geometry on one hand, and because the increase is obtained through a multiplication of N_e by N_g (instead of a simple addition for instance), the increase of sample size can be quite large.

Another attractive feature is that this kind of spatial averaging can be performed efficiently in spectral space, through a costless multiplication by a low-pass filter. This corresponds to the classical equivalence between a multiplication in spectral space and a convolution in gridpoint space (see e.g. Courtier et al 1998, in a different but related context over the sphere).

2.2 Conditions at play

An ideal case for the relevance and efficiency of this local averaging arises from two basic conditions. A first condition is that the statistic of interest (e.g. either the standard deviation, or the correlation, or both) should be nearly homogeneous locally, or at least (more or less) slowly varying. This will ensure that the local spatial average remains representative of the gridpoint statistic of interest.

A second condition is that the background error correlation length-scales are relatively short. This will ensure that the neighbouring gridpoint error realizations are nearly independent, so that there is an effective increase of the number of independent samples (when locally averaging).

If one considers e.g. the estimation of local standard deviations, a synthesis of these two conditions may be expressed as follows. The spatial average is particularly efficient when the typical scales (of geographical variations) of the standard deviation field are larger than the background error correlation length-scales.

Another way to justify spatial filtering is to notice that, experimentally, the sampling noise tends to be relatively small scale, compared to the signal of interest. This is shown e.g. in Figure 6 of Fisher and Courtier (1995) in a simple 1D academic context. This small scale structure of sampling noise will be diagnosed and illustrated here in a real NWP context.

It may be also anticipated that the second condition amounts to considering that the sampling noise correlation length-scale is relatively small. This corresponds to the idea that the sampling noise values (on e.g. the ensemble gridpoint standard deviations) for two neighbouring gridpoints will tend to be uncorrelated, if the associated background error realizations are also uncorrelated (when considering sampling noise as a random process, essentially driven by the random values of the background perturbations).

This can be seen as a connection with the design of an objective and optimal filter, which will be evoked later on: an ideal condition is when the scales (of geographical variations) of the standard deviation field are larger than the typical scales of sampling noise. As will be illustrated below, experimental results suggest that these conditions are nearly met in practice.

2.3 Illustration in a simulated framework

One way to illustrate and understand the properties of sampling noise and of local spatial averaging is to consider a simulated framework, in which a true state can be defined, handled and compared to.

This has been achieved by considering the randomization (Fisher and Courtier 1995, Andersson and Fisher 1998) of the operational version of the Arpège background error covariance matrix. This operational version includes background error standard deviations for vorticity, which vary geographically (according to a climatological average from an ensemble of assimilations (Belo Pereira and Berre 2006)).

These geographical variations are illustrated near the surface in the top left panel of Figure 2. As expected, the error standard deviations are larger over oceanic storm track areas (in the Northern Atlantic, in the Northern Pacific, and in the Southern Hemisphere), and they are smaller over data dense areas such as Northern America and Europe.

The right panels correspond to raw results of the randomization, calculated respectively from 6 and 220 random realizations. One first striking feature is that even with a small 6 member ensemble, it is possible to recognize the main large scale relevant features. On the other hand, for both 6- and 220-member ensembles, residual errors are also noticeable, and the second striking feature is that these (sampling) errors tend to be relatively

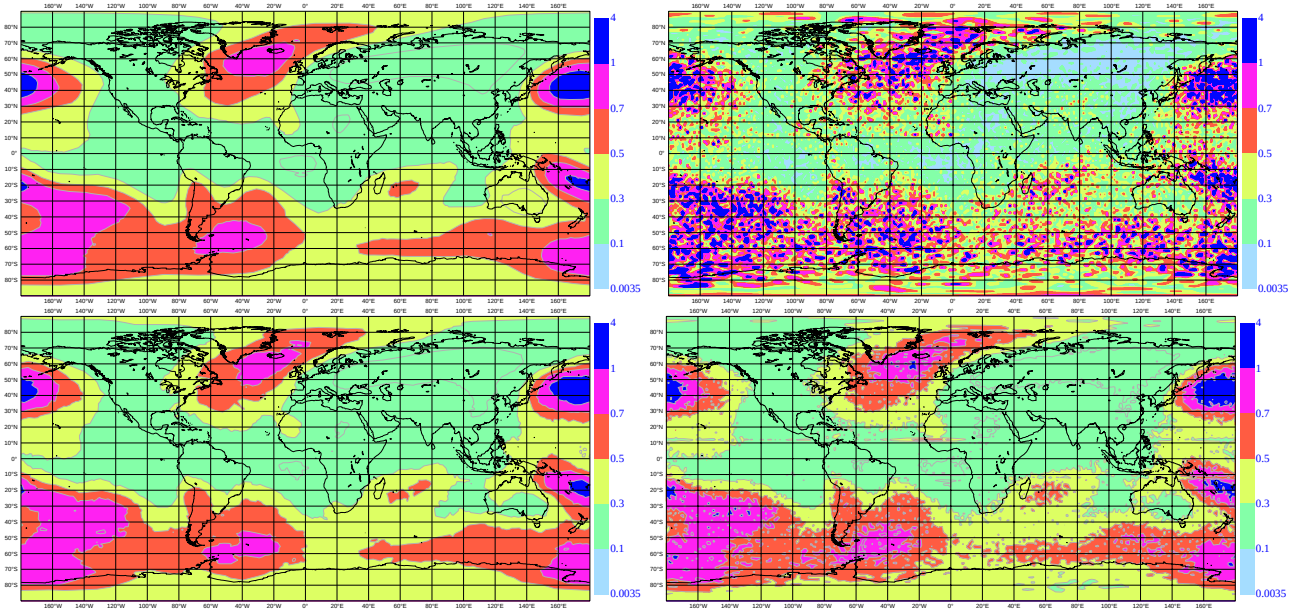


Figure 2: Reference map of background error standard deviations (top left panel) and its randomized estimates: raw estimate with 6 members (top right panel), filtered estimate with 6 members (bottom left panel), and raw estimate with 220 members (bottom right panel). Unit: $10^{-5}s^{-1}$.

small scale (with a larger amplitude when the ensemble is smaller, as expected). In other words, these two features illustrate the fact that the sampling noise is relatively small scale, compared to the signal of interest.

The bottom left panel corresponds to a spatially filtered version of the 6-member standard deviation map (which has been optimized "manually" in this case). As expected, the filtering removes small scale noise, while it preserves the large scale signal of interest. It may be also noticed that this filtered 6-member estimate tends to be even more accurate than the raw 220-member estimate. This increased accuracy reflects the fact that the total sample size N_t of the filtered 6-member estimate is larger than the ensemble size of the raw 220-member estimate.

These results illustrate the potential efficiency of the local spatial averaging. On the other hand, one may wonder if it is possible to objectively diagnose and optimally filter out the amount of sampling noise in the context of a real ensemble of assimilations. This is the object of the next section.

3 Spatial filtering of local standard deviations

3.1 Illustration of signal and noise in a real ensemble of assimilations

A specific methodology has been applied, in order to diagnose the respective amounts of signal and of sampling noise, in a real ensemble of assimilations. The main idea is to compare statistics of two independent ensembles of assimilations which have the same size. Typically, from a first qualitative point of view, their common features can be seen as corresponding to the signal, while their differences correspond to effects of sampling noise.

The results for standard deviation maps are illustrated in Figure 3, for the case of two independent 3-member ensembles. It is striking that similar large scale structures are visible in the two maps, despite the small ensemble size. For instance, small values are visible in the North Atlantic ridge, while values larger by a factor of 2 or 3 can be seen in the neighbouring trough.

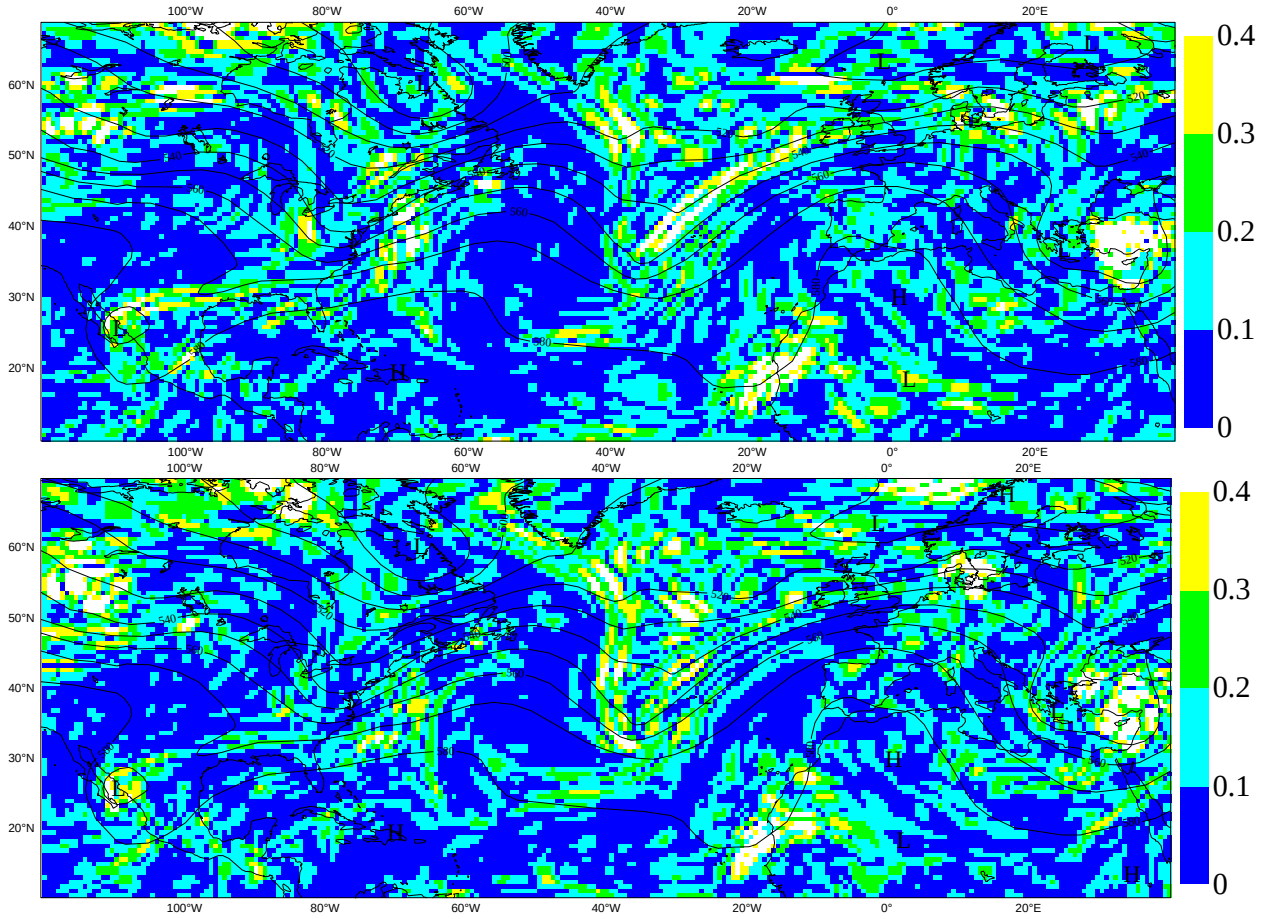


Figure 3: Respective maps of raw standard deviations of vorticity near 500 hPa, derived from two independent three-member ensembles, valid for 11 February 2002 at 12 UTC. Unit: $10^{-4}s^{-1}$.

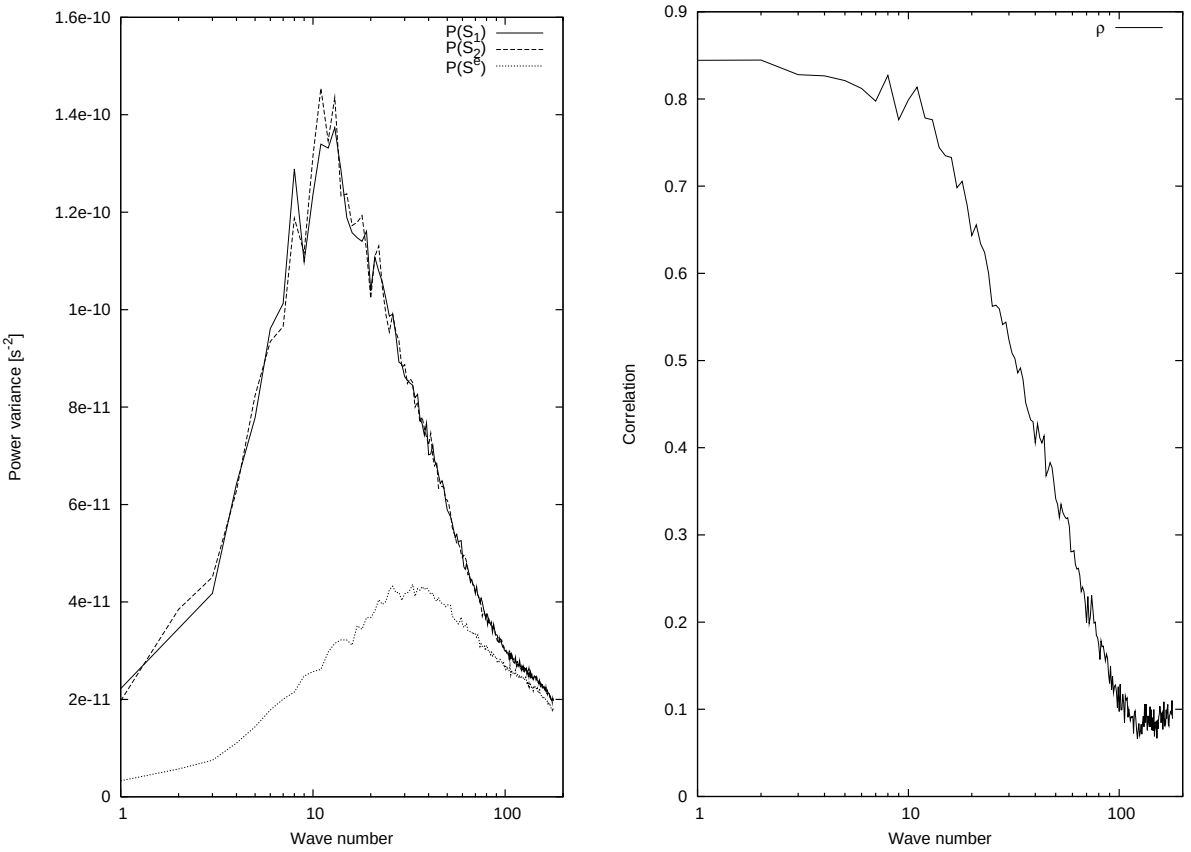


Figure 4: Left panel: power spectra of daily variations of the raw standard deviation field of the first ensemble, namely $\mathbf{P}(\mathbf{S}_1)$ *(full line), of the second ensemble, namely* $\mathbf{P}(\mathbf{S}_2)$ *(dashed line), and of sampling noise* $\mathbf{P}(\mathbf{S}^e)$ *(dotted line). The difference between the first two curves and* $\mathbf{P}(\mathbf{S}^e)$ *corresponds to the signal power* $\mathbf{P}(\mathbf{S}_\star)$*. Right panel: correlation* $\boldsymbol{\rho}$ *between the two ensemble-estimated maps of standard deviation, as a function of total wave number n.*

Conversely, one can also identify small scale details which do not coincide between the two maps. This suggests that the sampling noise is relatively small scale, compared to the signal of interest.

3.2 Formalism and diagnosis of signal and noise

To confirm this, one can try to estimate the amplitude (i.e. variance) of signal and noise, e.g. over a given time period (a 49-day period here, although a single day period could be enough as well). Formally, the standard deviation field $\mathbf{S}_i$ of the ith ensemble (i =1 or 2) can be decomposed as:

$$\mathbf{S}_i = \mathbf{S}_\star + \mathbf{S}_i^e,$$

where $\mathbf{S}_i^e$ is the sampling error, and $\mathbf{S}_\star$ is the noise-free signal. Under the assumption that the sampling noise is a random process (arising from the randomness of observation perturbation values basically), it is uncorrelated with the signal. The variance of $\mathbf{S}_i$ may thus be decomposed as follows:

$$\mathbf{V}(\mathbf{S}_i) = \mathbf{V}(\mathbf{S}_\star) + \mathbf{V}(\mathbf{S}^e) \tag{1}$$

after noticing that $\mathbf{V}(\mathbf{S}_1^e) = \mathbf{V}(\mathbf{S}_2^e)$ (as the two independent ensembles have the same size), which may thus be noted $\mathbf{V}(\mathbf{S}^e)$. This noise variance may be estimated by expanding the difference between the ensemble estimates:

$$\mathbf{S}_1 - \mathbf{S}_2 = (\mathbf{S}_\star + \mathbf{S}_1^e) - (\mathbf{S}_\star + \mathbf{S}_2^e) = \mathbf{S}_1^e - \mathbf{S}_2^e,$$

and thus, under the assumption that the two noises $\mathbf{S}_1^e$ and $\mathbf{S}_2^e$ are two random uncorrelated processes: $\mathbf{V}(\mathbf{S}_1 - \mathbf{S}_2) = \mathbf{V}(\mathbf{S}_1^e) + \mathbf{V}(\mathbf{S}_2^e) - 2\ \mathbf{cov}(\mathbf{S}_1^e, \mathbf{S}_2^e) = 2\ \mathbf{V}(\mathbf{S}^e)$. This means that the noise variance can be estimated as half the variance of $\mathbf{S}_1 - \mathbf{S}_2$:

$$\mathbf{V}(\mathbf{S}^e) = \frac{1}{2}\mathbf{V}(\mathbf{S}_1 - \mathbf{S}_2)$$

While the above formalism is general, it is convenient to consider it in terms of spectral coefficients of the standard deviation maps, in order to diagnose the scale dependence of the signal and noise contributions. The individual modal variances $V(S_i(n,m))$ can be cumulated as a function of total wave number n, to provide the power spectrum $\mathbf{P}(\mathbf{S}_i)$: $P(S_i)[n] = \sum_{m=-n}^{+n} V(S_i(n,m)) = (2n+1)V(S_i)[n]$ (where $V(S_i)[n]$ is the average modal variance).

This power spectrum describes the total contribution of each wave number n to the time variations of the estimated standard deviation field. The left panel of Figure 3 shows the power spectra associated to the two ensembles, $\mathbf{P}(\mathbf{S}_1)$ and $\mathbf{P}(\mathbf{S}_2)$. These two spectra are nearly identical, in accordance with equation (1).

The corresponding power spectrum of sampling noise $\mathbf{P}(\mathbf{S}^e)$ is also shown in this panel (dotted curve). Note also that, according to equation (1), the difference between the first two curves and the third one ($\mathbf{P}(\mathbf{S}^e)$) corresponds to the power of signal $\mathbf{P}(\mathbf{S}_\star)$. It appears that the amount of noise is relatively small in the large scales (where the largest time variations of the standard deviation fields occur), and that it is relatively large in the small scales.

Another way to visualize this scale dependence is to calculate the correlation between the two standard deviation maps, which is a simple function of the noise-to-signal ratio:

$$\boldsymbol{\rho} = \frac{\mathbf{cov}(\mathbf{S}_1, \mathbf{S}_2)}{\sqrt{\mathbf{P}(\mathbf{S}_1)\mathbf{P}(\mathbf{S}_2)}} = \frac{\mathbf{P}(\mathbf{S}_\star)}{\mathbf{P}(\mathbf{S}_\star) + \mathbf{P}(\mathbf{S}^e)} = \frac{\mathbf{1}}{\mathbf{1} + \frac{\mathbf{P}(\mathbf{S}^e)}{\mathbf{P}(\mathbf{S}_\star)}}. \tag{2}$$

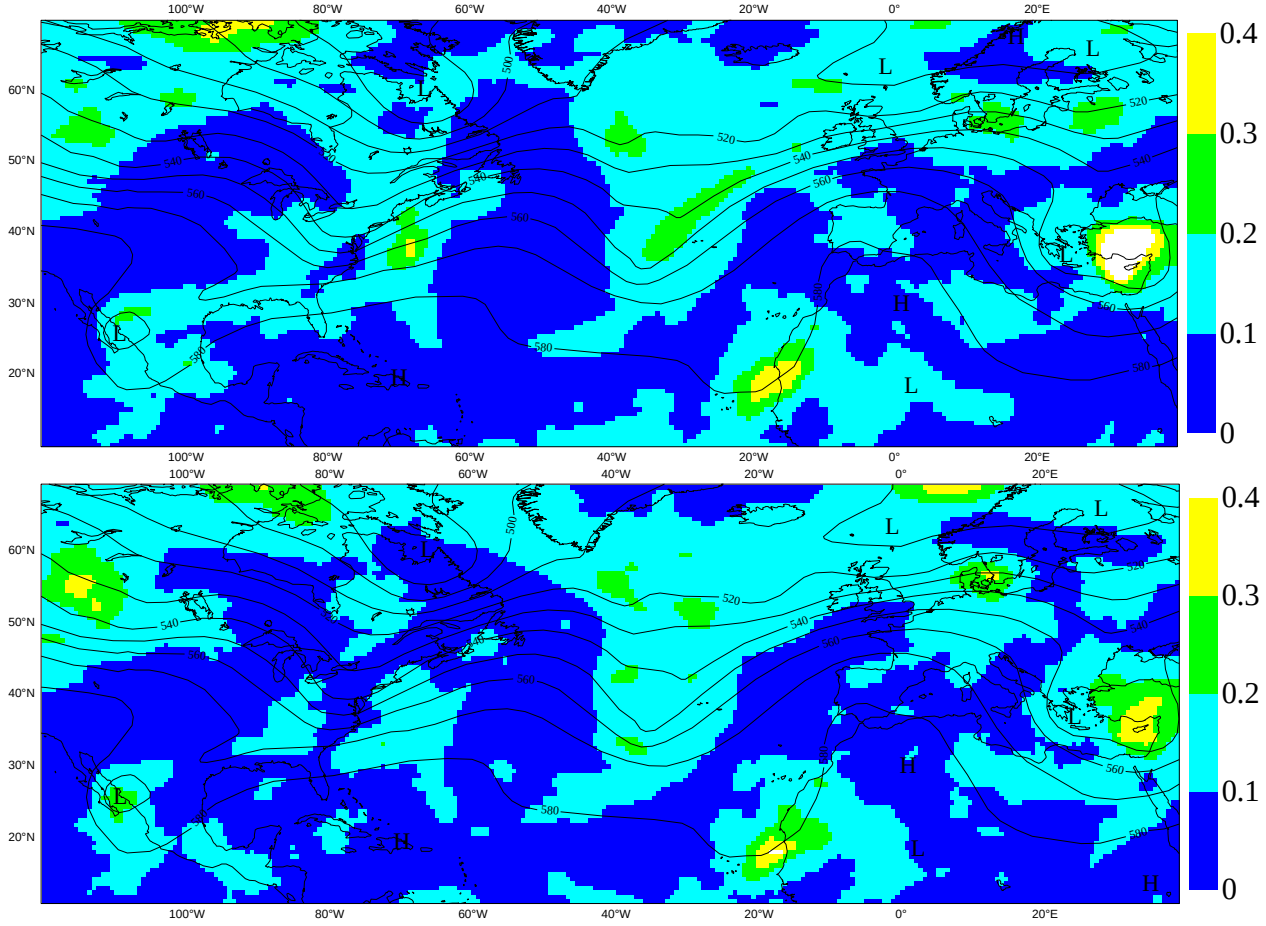


Figure 5: Same as figure 3, after applying the objective and optimal filter. Unit: $10^{-4}s^{-1}$.

The scale dependence of this correlation is shown in the right panel of Figure 4. It is larger than 80% in the large scales, and then it decreases progressively towards small values in the small scales.

These results in Figure 4 are consistent with expectations evoked in the previous section about Figure 3. This is also coherent with studies at ECMWF (Lars Isaksen, personal communication and this volume), showing that a 10-member standard deviation map estimate is similar to a 50-member estimate, except for some small scale sampling noise that could be filtered. All this supports the idea to apply a spatial filter, in order to extract the relevant large scale signal and to remove the small scale sampling noise.

3.3 Design and application of an objective and optimal filter

In order to design an objective and optimal filter, a simple idea is to apply the usual linear estimation theory (which is e.g. at the basis of most data assimilation techniques). A given ensemble estimate $\mathbf{S}_1$ can be seen as a predictor of the signal $\mathbf{S}_\star$, and formally, the best linear estimate of $\mathbf{S}_\star$ is provided by a simple regression equation (with the signal as the predictand):

$$\mathbf{S}_\star \sim \frac{\mathbf{cov}(\mathbf{S}_\star, \mathbf{S}_1)}{\mathbf{V}(\mathbf{S}_1)} \mathbf{S}_1.$$

and then it can be shown easily that the regression coefficent is simply equal to the correlation ρ between the two ensemble estimates $\mathbf{S}_1$ and $\mathbf{S}_2$:

$$\frac{\mathbf{cov}(\mathbf{S}_\star, \mathbf{S}_1)}{\mathbf{V}(\mathbf{S}_1)} = \frac{\mathbf{V}(\mathbf{S}_\star)}{\mathbf{V}(\mathbf{S}_\star) + \mathbf{V}(\mathbf{S}^e)} = \frac{\mathbf{1}}{\mathbf{1} + \frac{\mathbf{V}(\mathbf{S}^e)}{\mathbf{V}(\mathbf{S}_\star)}} = \boldsymbol{\rho}$$

according to equation (2). This means that the correlation $\boldsymbol{\rho}$ is not only a useful diagnostic function of the signal-to-noise ratio. It can be also used as an objective and optimal filter of the raw ensemble standard deviation fields.

The results of this optimal filtering are illustrated in Figure 5, which can be compared to Figure 3. As expected, large scale coherent structures are extracted, while small scale noisy details tend to be removed.

It may be also mentioned that the amplitude of the regression residual error can be estimated, according to: $\mathbf{V}(\boldsymbol{\rho}\mathbf{S}_1 - \mathbf{S}_\star) = \boldsymbol{\rho}\ \mathbf{V}(\mathbf{S}^e)$. This equation and the corresponding results (not shown) indicate that the error reduction (achieved by the regression) is particularly large in the medium and small scales (in accordance with the shape of $\boldsymbol{\rho}$).

3.4 Comparison with innovation-based diagnostics

One way to validate ensemble standard deviation estimates is to compare them with independent estimates, derived from innovation data. In particular, it has been shown by Desroziers et al (2005) that the covariance between the analysis increment $\mathbf{dx}$ (in observation space) and the innovation $\delta\mathbf{y}$ is an estimate of the background error covariance (in observation space):

$$\mathbf{cov}(\mathbf{H}\ \mathbf{dx}, \delta\mathbf{y}) \sim \mathbf{HBH^T}.$$

An example of such a comparison is shown in Figure 6, for a specific date. Note that a simple spatial average has been applied to both ensemble and innovation-based estimates, with a 500 km averaging radius around each observation location. Moreover, as the ensemble standard deviations tend to be larger by a factor 1.4 (on the average) than the innovation-based estimates, the ensemble estimates have been divided by this factor, in order to focus on the comparison of geographical variations.

It is striking that similar spatial structures are visible in the two estimates, with e.g. large values in the Western Pacific, and smaller values in the Southern Atlantic. This relative coherence between these two independent estimates looks rather encouraging. This kind of comparison is believed to be essential in order to validate the ensemble estimates, and/or in order to improve their realism. In particular, as suggested by Daley (1992), such comparisons could provide objective information about model error covariances for instance.

3.5 Expected impacts and effective results

The current operational version of the Arpège 4D-Var experiments uses climatological ensemble standard deviations (as described in Belo Pereira and Berre 2006). In order to anticipate expected impacts of flow-dependent ensemble standard deviations (”of the day”), it is useful to compare examples of the corresponding maps (Figure 7).

It can be noticed firstly that there are similar large scale contrasts in the two maps, with e.g. larger values over the oceanic storm track areas, and smaller values e.g. in the tropics and over the data dense USA. It appears in fact that the strongest modifications of standard deviations are relatively localized spatially. They correspond to increases in local areas such as the Western part of the Northern Pacific, near the Southern coast of Australia, and Western Europe (which coincides with an intense storm over Northern France).

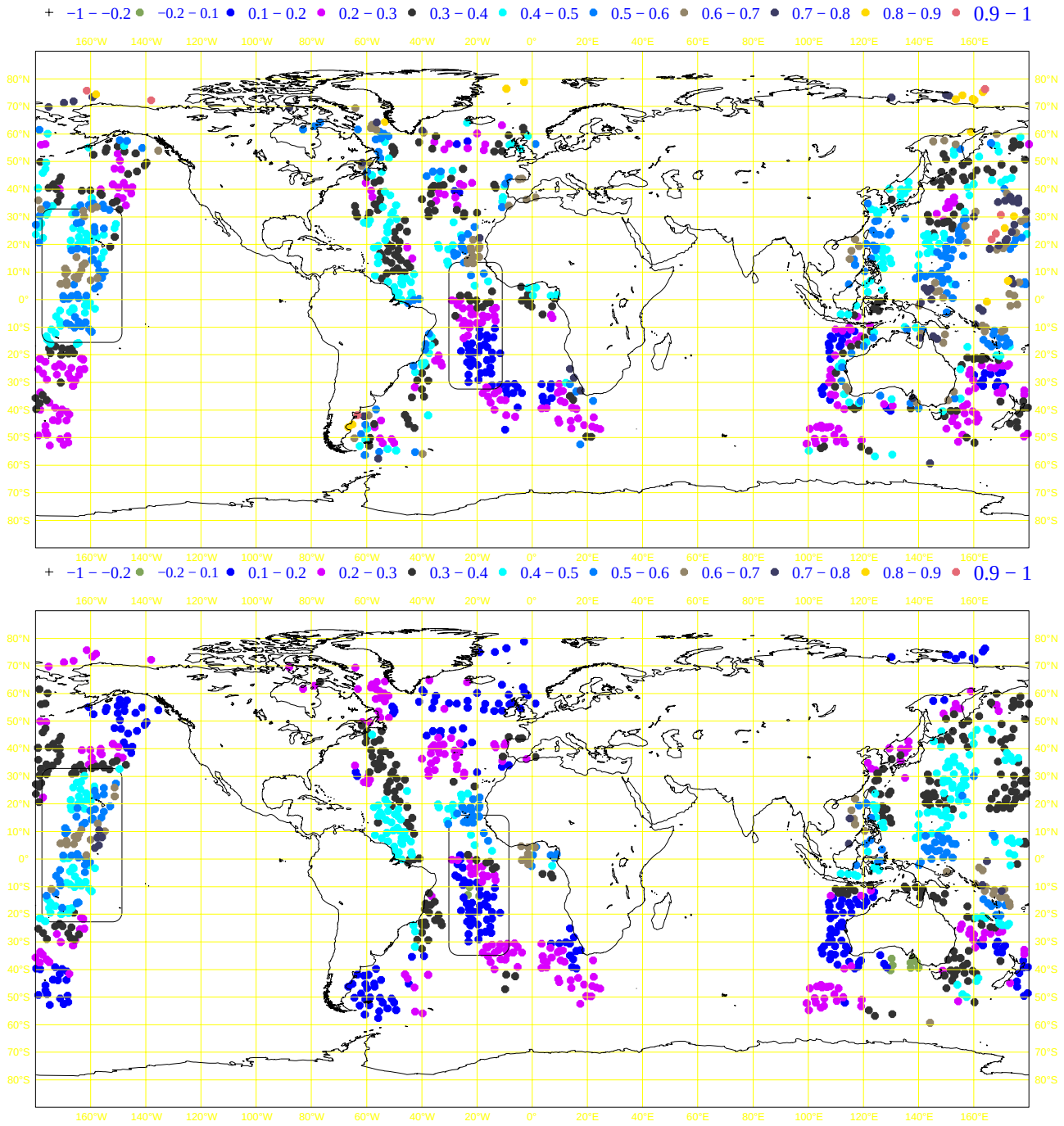


Figure 6: Filtered background error standard deviations for channel HIRS 7, at available observation locations, for one specific date (28/08/2006 at 00h). Top panel: from a six member assimilation ensemble. Bottom panel: from the method based on covariances of residuals. Unit: K.

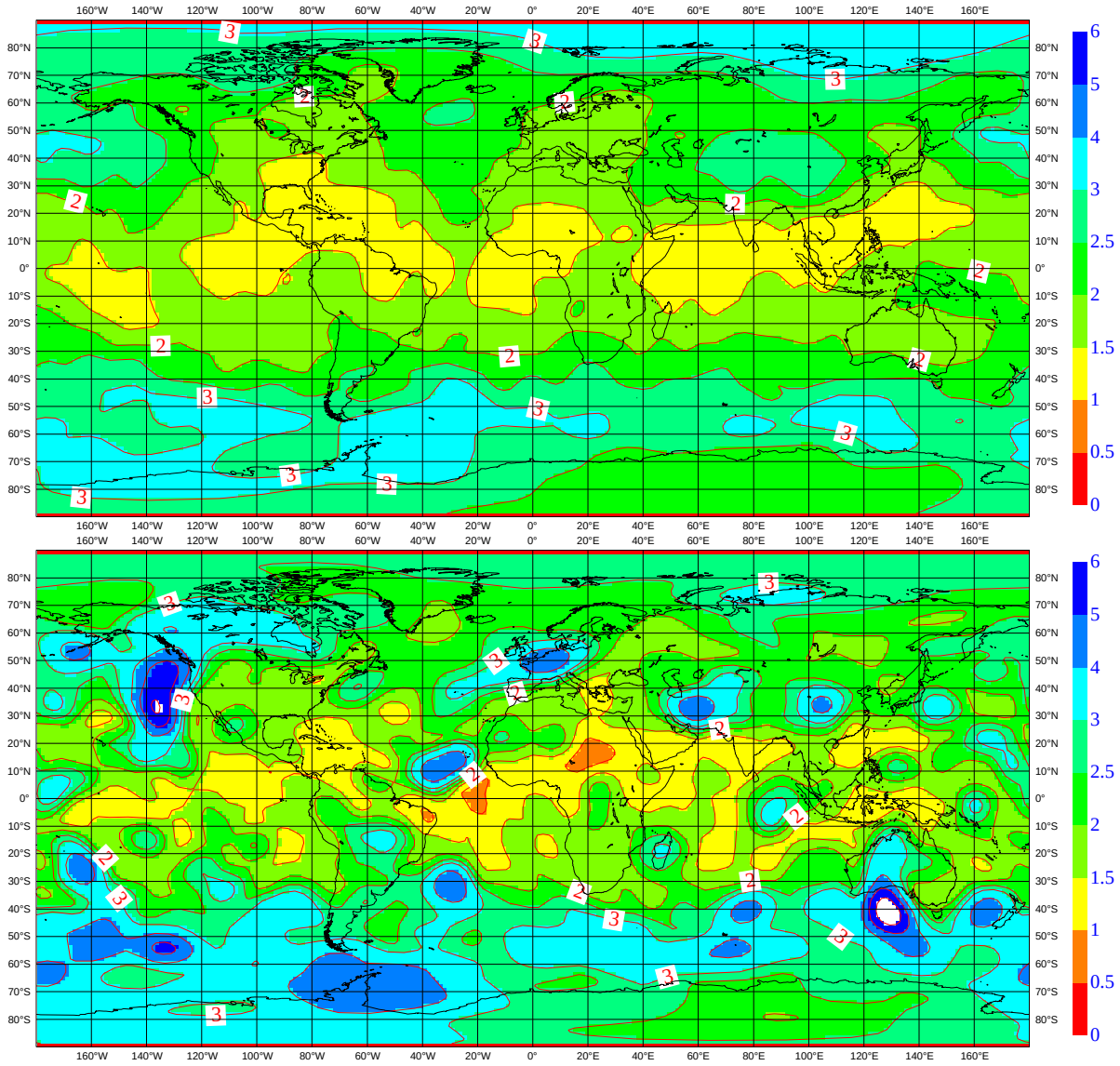


Figure 7: Maps of standard deviations of vorticity near 500 hPa. Top panel: in operations (from the climatological average of an ensemble of assimilations). Bottom panel: from the ensemble of backgrounds valid for 10 December 2006 at 3 UTC. Unit: $10^{-5}s^{-1}$.

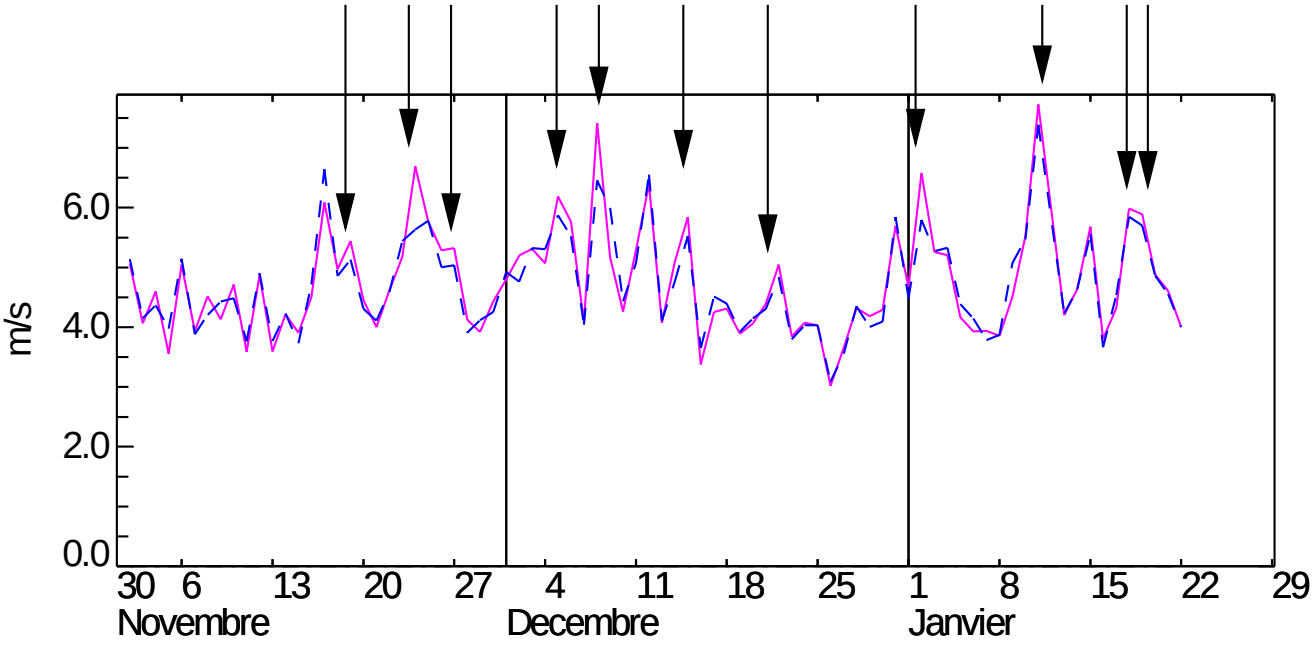


Figure 8: RMS of 24h forecasts of wind at 500 hPa over Europe, obtained from two different Arpège 4D-Var experiments. Pink full curve: with climatological ensemble sigmab's. Blue dashed curve: with flow-dependent ensemble sigmab's of the day. The vertical black arrows materialize the number of reductions of local RMS peaks.

The localized nature of these strong modifications implies that it is not expected to obtain a systematic improvement of forecast scores at any place and at any time, but rather a tendency to reduce local errors connected to intense weather systems.

A three-month experiment has thus been carried out, in order to examine the impact of flow-dependent ensemble standard deviations ("of the day"), against the climatological ensemble standard deviations which are currently used in the operational Arpège 4D-Var.

An example of impact is shown in Figure 8 for RMS of 24h forecasts of wind over Europe. While "good" forecasts are not much affected, there is a tendency to reduce the local maxima of RMS. This positive and localized impact in time is consistent with aforementioned expectations from these flow-dependent local modifications of the standard deviations.

A similar tendency has been observed for local areas such as North-America and Australia-New Zealand, while being less visible when scores are averaged over very large domains such as the extratropical Northern Hemisphere. This illustrates that the impact is localized in space (as expected).

These positive local impacts (in space and time) were also found to be larger for wind than for geopotential, in accordance with possible connections to intense weather systems with large wind speeds. This is also consistent with case studies in Kucukkaraca and Fisher (2006) of intense events such as the second French storm.

4 Spatial filtering of local correlations

4.1 Wavelets as a relevant compromise between two extreme approaches

It may be noticed that there are typically two extreme approaches in correlation modelling.

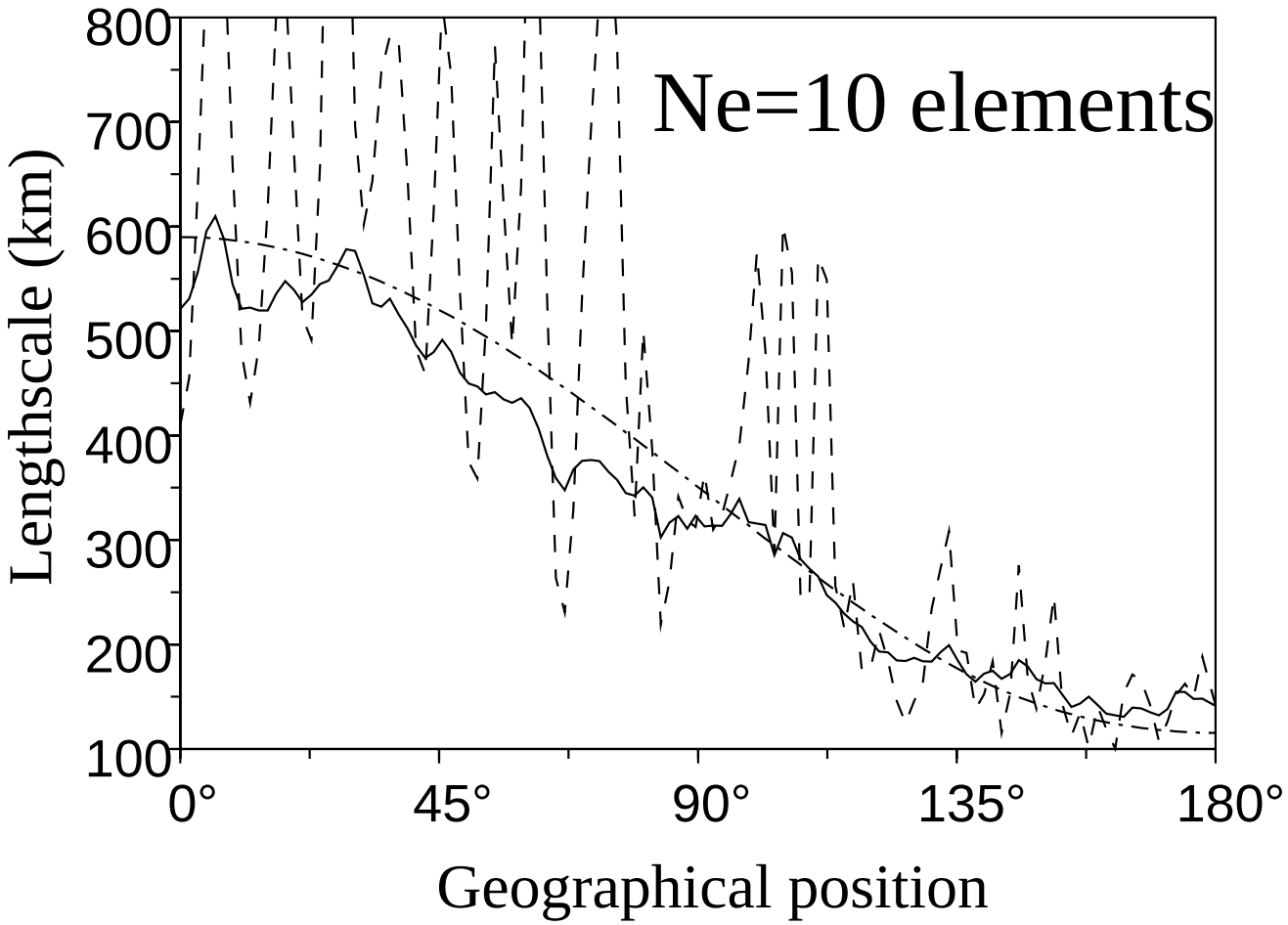


Figure 9: Geographical variations of the correlation length-scales in a 1D academic case: exact (dash-dotted line), raw 10-member estimate (dashed line), and wavelet 10-member estimate (full line).

Ensemble Kalman filters are based on local correlation functions, which are calculated independently for each gridpoint. This looks good with respect to the idea to represent a lot of geographical variations, but it also means that a rather large ensemble is needed, in order to have a large statistical sample.

Variational assimilation schemes often use a spectral diagonal approach, which amounts to calculating a global spatial average of the correlation functions. This is rather good with respect to the idea to use a large spatial sample, but it also means homogeneity, and then there are no geographical variations at all.

In this context, an interesting compromise (to combine heterogeneity and an increased sample size) is to use wavelets, in order to calculate a local average of correlations. This potential has been illustrated by Pannekoucke et al (2007), in a simulated 1D academic case, and also in the context of a real ensemble of assimilations.

4.2 The spatial structure of sampling noise and its filtering in a simulated framework

Figure 9 corresponds to the 1D academic case. The sinusoide is an example of length-scale map, and the noisy dashed curve is the raw ensemble estimation with 10 members. While the main geographical contrast is captured, there is also a lot of noise, which varies much from one gridpoint to the next one. In particular, the spatial structure of this sampling noise is thus relatively small scale (with higher frequencies than for the signal).

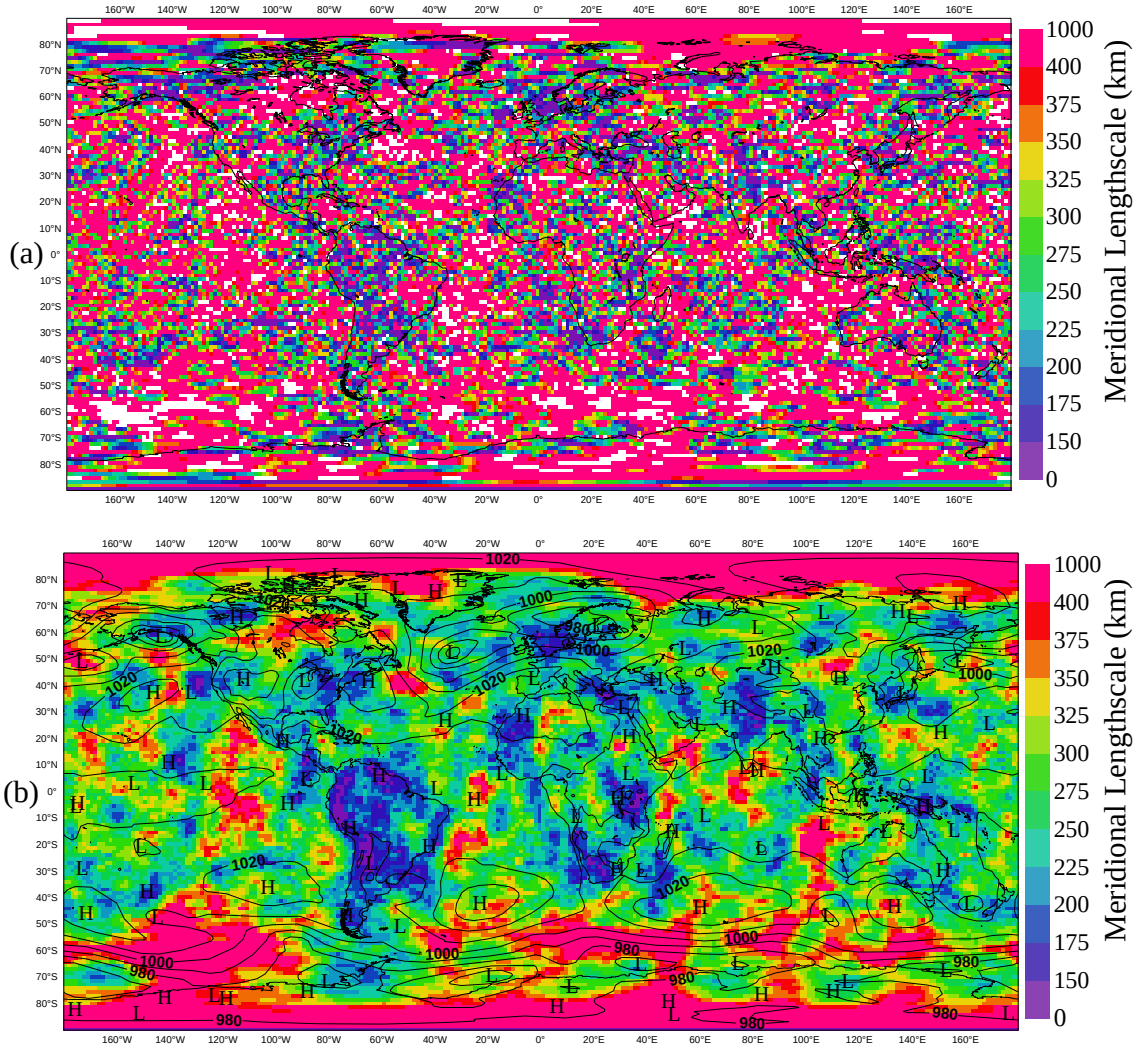


Figure 10: Meridional length-scales (km) for surface pressure on 10 February 2002 at 12 UTC. Top panel: raw length-scales. Bottom panel: wavelet-implied length-scales, superimposed on the background field of surface pressure.

The full line is from the wavelet ensemble estimation, still with 10 members. The sampling noise is much smaller, and the main variation is still represented. This illustrates the potential efficiency of the wavelet local averaging.

4.3 The spatial structure of sampling noise and its filtering in a real NWP context

This is further illustrated in Figure 10 by a map of the length-scales ”of the day”, from a real 6-member assimilation ensemble. The top panel corresponds to the raw ensemble estimation, which is very noisy, even if a few patterns can be identified.

The bottom panel corresponds to the wavelet ensemble estimation, which is much more structured, for instance with smaller length-scales near the Andes mountains, and small length-scales near the Scandinavian low (and larger length-scales in the Southern hemisphere).

Case studies over several days have been carried out (not shown). They tend to indicate that it is important to consider a real time assimilation ensemble, instead of e.g. a simplified background state dependence, in order

to represent effects of both meteorological processes and data density.

5 Conclusions and perspectives: towards optimized flow-dependent covariances

A step-by-step approach has been applied, in order to extract realistic and useful information from an ensemble of variational assimilations. Initially, climatological averages of global covariances and of local standard deviations, from an off-line assimilation ensemble, have been calculated and implemented operationally.

Current work is now devoted to the extraction of flow-dependent information on local standard deviations and correlations, from a real time variational ensemble. Combining such an assimilation ensemble and spatial filtering techniques seems to be a promising way to do this. Standard deviation maps can be easily filtered in spectral space, and a wavelet approach allows the correlation functions to be efficiently filtered.

The spatial filtering is justified by the relatively small scale structure of sampling noise, and it can be optimized objectively with simple statistical techniques. The local spatial averaging allows the sample size to be much increased, because the ensemble size is multiplied by a 2D spatial sample size. Moreover, the spatial filtering is relatively costless. Thus, it may help to reduce the constraint on the required number of ensemble members (to reach a given accuracy).

First impact experiments and comparisons with innovation-based diagnostics are encouraging. A real time ensemble is thus considered to become operational at Météo-France in 2008, to provide flow-dependent background error standard deviations (as a first step). Applications for correlations (using wavelets), assimilation diagnostics and ensemble prediction will be considered too.

6 References

Andersson, E. and M. Fisher, 1998: Background errors for observed quantities and their propagation in time. *Proceedings of the ECMWF workshop on Diagnosing Data Assimilation Systems*, 81-89.

Belo Pereira, M. and L. Berre, 2006: The use of an Ensemble approach to study the Background Error Covariances in a Global NWP Model. *Mon.Wea.Rev.*, **134**, 2466-2489.

Berre, L., S.E., Ştefănescu, and M. Belo Pereira, 2005: The representation of the analysis effect in three error simulation techniques. *Tellus*, **58A**, 196-209.

Bishop, C.H., B. J. Etherton and S. J. Majumdar, 2001: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Mon. Wea. Rev.*, **129**, 420-436.

Buehner, M., 2005: Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting. *Q. J. Roy. Meteor. Soc.*, **131**, 1013 - 1043.

Courtier, P. and J.F. Geleyn, 1988: A global numerical weather prediction model with variable resolution: Application to the shallow-water equations. *Quart. J. Roy. Meteor. Soc.*, **114**, 1321-1346.

Courtier, P., Andersson, E., Heckley, W., Pailleux, J., Vasiljevic, D., Hamrud, M., Hollingsworth, A., Rabier, F. and Fisher, M., 1998: The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part I: Formulation. *Quart. J. Roy. Meteorol. Soc.*, **124**, 1783-1808.

Daley, R., 1992 : Estimating model-error covariances for application to atmospheric data assimilation. *Mon. Wea. Rev.*, **120**, 1735-1746.

Desroziers, G., L. Berre, B. Chapnik, and P. Poli, 2005: Diagnosis of observation, background and analysis error statistics in observation space. *Quart. J. Roy. Meteorol. Soc.*, **131**, 3385-3397.

Desroziers, G., L. Berre, O. Pannekoucke, S.E. Ştefănescu, P. Brousseau, L. Auger and B. Chapnik, 2007: Flow-dependent error covariances from variational assimilation ensembles on global and regional domains. Proceedings of the SRNWP workshop on high resolution data assimilation. To appear in a Hirlam technical report.

Fisher, M. and P. Courtier, 1995: Estimating the covariance matrices of analysis and forecast error in variational data assimilation. *ECMWF technical memorandum* **220**, 28 pages.

Fisher, M., 2003: Background error covariance modelling. *Proceedings of the ECMWF Seminar on Recent developments in data assimilation for atmosphere and ocean, 8-12 September 2003*, 45-63.

Houtekamer, P.L., Lefaivre, L., Derome, J., Ritchie, H. and Mitchell, H.L., 1996: A System Simulation Approach to Ensemble Prediction. *Mon. Wea. Rev.*, **124**, 1225-1242.

Kucukkaraca, E. and M. Fisher, 2006: Use of analysis ensembles in estimating flow-dependent background error variances. ECMWF Technical Memorandum, **492**. European Centre for Medium-range Weather Forecasts (ECMWF), Reading, UK.

Lorenc, A.C., 2003: The potential of the ensemble Kalman filter for NWP - a comparison with 4D-Var. *Quart. J. Roy. Meteor. Soc.*, **129**, 3183 - 3203.

Monin, A.S., and Yaglom, A.M., 1971: Statistical fluid mechanics: mechanics of turbulence. Vol. 1. The MIT Press. Cambridge (USA).

Pannekoucke, O., L. Berre and G. Desroziers, 2007: Filtering properties of wavelets for the local background error correlations. *Q. J. Roy. Meteor. Soc.*, **133**, 363 - 379.

Parrish, D.F. and J.C. Derber, 1992: The National Meteorological Center's spectral statistical interpolation analysis system. *Mon. Wea. Rev.*, **120**, 1747-1763.

Rabier, F., Mc Nally, A., Andersson, E., Courtier, P., Undén, P., Eyre, J., Hollingsworth, A., Bouttier, F., 1998: The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part II: Structure functions. *Quart. J. Roy. Meteor. Soc.*, **124**, 1809-1829.

Annexe D

Ondelettes et modélisation pour la matrice B

RÉSUMÉ

Ce chapitre annexe correspond à un complément sur la modélisation ondelette des covariances d'erreur d'ébauche. Il présente de manière plus détaillée les ondelettes, ainsi que d'autres outils, et leur utilisation pour la modélisation.

Dans le chapitre 3, la modélisation à l'aide de l'hypothèse diagonale dans l'espace spectral a été présentée. Une avancée significative apportée par cette formulation est une modélisation de portée non nulle.

Cependant, cette modélisation ne permet pas la représentation des variations géographiques des portées. Pour ce faire, il est nécessaire d'utiliser d'autres représentations du signal, telles que l'analyse en ondelette.

Ce chapitre présente tout d'abord l'analyse en ondelette continue. Ce type d'analyse est, avant tout, réservé à l'étude d'un signal, plutôt qu'à sa manipulation numérique (filtrage, compression,...). Les ondelettes orthogonales sont ensuite présentées, ainsi que leur utilisation pour la modélisation des covariances. Le choix de l'ondelette apparaît alors important. Une amélioration est apportée par la décomposition en paquet d'ondelettes. Ce type de décomposition généralise la décomposition ondelette orthogonale et offre un dictionnaire de bases. Ce dictionnaire permet alors la recherche de la meilleure base représentant un signal, mais aussi la meilleure base approximant la base de Karhunen-Loève (base qui diagonalise la matrice de covariance). Enfin, les ondelettes sphériques utilisées sont présentées sommairement (voir le chapitre 4 pour plus de détails).

Ainsi, il est montré que les ondelettes permettent bien la modélisation de covariances hétérogènes. La question de la mauvaise modélisation de la variance est abordée : au lieu de modéliser une matrice de corrélation, la diagonale de la matrice résultante n'est pas homogène à 1. Si l'analyse par paquets d'ondelettes permet effectivement une meilleure représentation du signal, la difficulté à construire une matrice de covariance 3D non séparable est discutée.

MOT CLÉS: Covariance, ondelettes continues, ondelettes orthogonales, paquets d'ondelettes, ondelette sphérique.

D.1 Retour sur la modélisation par l'hypothèse diagonale spectrale

L'hypothèse diagonale spectrale est un modèle classique pour les corrélations d'erreur de prévision (Courtier *et al.* , 1998). Cette modélisation est construite à partir de la transformée spectrale. Il s'agit de la transformée de Fourier sur la droite ou le cercle, et de la transformée de Laplace sur la sphère (associée aux harmoniques sphériques).

Pour en comprendre plus clairement les limitations observées dans le chapitre 3, au paragraphe 3.3.2, il est intéressant de revenir sur l'outil principal de cette modélisation : la transformée spectrale.

Dans la suite, les fonctions considérées sont des fonctions définient sur la droite et à valeur réelle ou complexe. Ces fonctions sont supposées être de carré sommable. Il est donc possible

de définir un produit scalaire $< f|g >= \int f^* g$. La norme Euclidienne associée est alors définie par $||f||^2 =< f|f >$.

D.1.1 Transformée de Fourier

Définition

La transformée de Fourier, sur la droite réelle, d'un signal f est définie par

$$\mathcal{F}(f)(\nu) = \int_{-\infty}^{+\infty} f(t) e^{-i2\pi\,\nu t} dt. \tag{D.1}$$

Il s'agit d'une application de $\mathbb{R}$ dans $\mathbb{C}$ de carré sommable. Pour simplifier cette notation, $\mathcal{F}(f)(\nu)$ est également notée $\hat{f}(\nu)$. De plus, le signal initial f peut être retrouvé à l'aide de la transformation inverse

$$\mathcal{F}^{-1}(g)(t) = \int_{-\infty}^{+\infty} g(\nu) e^{i2\pi\,\nu t} d\nu, \tag{D.2}$$

qui est telle que $\mathcal{F}^{-1} \circ \mathcal{F} = 1$. Ainsi, $\mathcal{F}^{-1} \circ \mathcal{F}(f) = f$.

Dans le cas d'un signal périodique discrétisé à pas régulier, la transformée de Fourier correspond à un changement de base orthonormale : passage de la base canonique (dirac) à l'espace des coefficients de Fourier. En particulier, l'opérateur linéaire $\boldsymbol{F}$ représentant cette transformation est alors orthogonal : $\boldsymbol{F}^{-1} = \boldsymbol{F}^T$.

Quelques propriétés

La transformée de Fourier possède différentes propriétés, par exemple la conservation de l'énergie (égalité de Plancherel) :

$$\int_{-\infty}^{+\infty} |f(t)|^2 dt = \int_{-\infty}^{+\infty} |\hat{f}(\nu)|^2 d\nu. \tag{D.3}$$

Une autre propriété est que la transformée de Fourier se comporte particulièrement bien avec la convolution. En effet,

$$\mathcal{F}(h * f) = \mathcal{F}(h)\mathcal{F}(f) \tag{D.4}$$

où le produit de convolution est défini par

$$h * f(x) = \int_{-\infty}^{+\infty} h(x-t) f(t) dt = \int_{-\infty}^{+\infty} h(t) f(x-t) dt. \tag{D.5}$$

Limitation intrinsèque : l'inégalité de Heisenberg

On peut considérer le signal f normalisé suivant $||f||_2 = 1$, avec $||f||_2^2 = \int_{\mathbb{R}} |f|^2 d\lambda$. Alors $|f|^2$ est une densité de probabilité, de moyenne $\mu_t = \int_{-\infty}^{+\infty} t|f(t)|^2 dt$ et d'écart type (caractérisant la dispersion) $\sigma_t = \left(\int_{-\infty}^{+\infty} (t-\mu_t)^2 |f(t)|^2 dt\right)^{1/2}$.

Il est possible de définir de même une dispersion spectrale pour le signal. Par la conservation de l'énergie, il vient que $||\mathcal{F}(f)||_2 = 1$, ce qui justifie la définition de sa moyenne $\mu_\nu = \int_{-\infty}^{+\infty} \nu|\hat{f}(\nu)|^2 d\nu$ et de son écart type $\sigma_\nu = \left(\int_{-\infty}^{+\infty} (\nu-\mu_\nu)^2 |\hat{f}(\nu)|^2 d\nu\right)^{1/2}$.

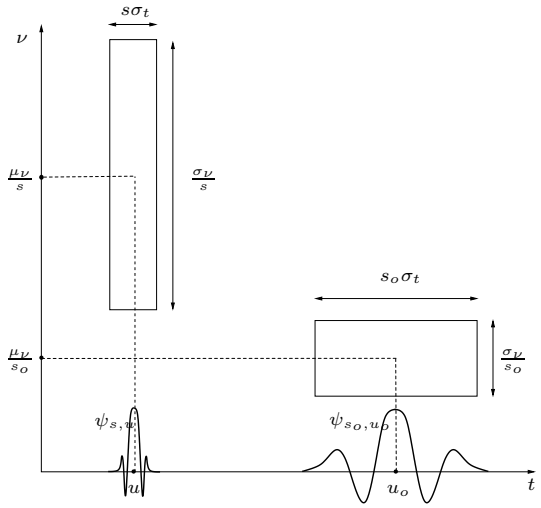


FIG. D.1 – Représentation dans le plan temps-fréquence de la localisation temps-fréquence. Le principe d'incertitude de Heisenberg implique que si l'on augmente la localisation spatiale, on diminue la localisation fréquentielle et réciproquement. Dans le cas présent, nous avons conservation de l'aire $\sigma_\nu\sigma_t$ des rectangles. Les rectangles sont les boîtes de Heisenberg (d'après Mallat (2001)).

La relation de Heisenberg stipule qu'il n'est pas possible d'obtenir une précision arbitrairement grande à la fois en temps et en fréquence. Cette contrainte s'exprime par l'**inégalité de Heisenberg** :

$$\sigma_t\sigma_\nu \geq \frac{1}{4\pi}. \tag{D.6}$$

Cette inégalité ne peut être raffinée puisque l'égalité est atteinte dans le cas où f est une gaussienne. De manière pratique, cela s'interprète comme suit : si on connaît une localisation à une précision σ_t, alors on ne pourra obtenir une précision fréquentielle σ_ν inférieure à $\frac{1}{4\pi\,\sigma_t}$. Cette contrainte est illustrée sur la figure D.1 où chaque "boîte de Heisenberg" représente les zones spatiale et spectrale caractérisant un signal.

Ceci correspond à l'idée simple que pour connaître avec précision la fréquence locale d'un signal il est nécessaire d'en considérer plusieurs périodes, ce qui a pour effet de diminuer la localisation spatiale. A l'inverse, si l'on veut caractériser le signal en un endroit précis, on n'aura pas assez de périodes pour déterminer simultanément sa fréquence.

D.1.2 Conséquence au niveau de l'interprétation d'un signal

D'après le chapitre 3, la modélisation des corrélations, par l'hypothèse diagonale spectrale, interdit la représentation des variations géographiques de la corrélation. Ceci est dû au fait que les fonctions sinus/cosinus ont un support sur l'ensemble du domaine et qu'elles ont la même modulation en tout point. L'analyse spectrale est adaptée à l'étude des signaux périodiques tels que celui représenté sur la figure D.2-(a). Il s'agit d'un *signal stationnaire*. Cette dénomination stationnaire correspond également à une classe de signaux aléatoires dont les propriétés statistiques sont invariantes par translation (dans l'espace ou dans le temps). Ces signaux sont caractérisés par leur énergie spectrale correspondant à la transformée de Fourier de leur autocorrélation en points de grille (théorème de Wiener-Khintchine). Le spectre de ce signal représenté sur la figure D.2-(b) est très simple : le spectre est discret, avec seulement quelques raies représentant les composantes spectrales du signal. A l'inverse, la figure D.2-(c) représente un signal plus complexe dont l'information est localisée à certains endroits. Le spectre d'énergie, Fig. D.2-(d), est complexe. D'une part il est plein, *i.e.* il n'y a plus de raies comme dans l'exemple précédent, et toutes les composantes spectrales sont excitées. Par contre, il est difficile de déduire du spectre la localisation spatiale à l'origine de certaines contributions énergétiques.

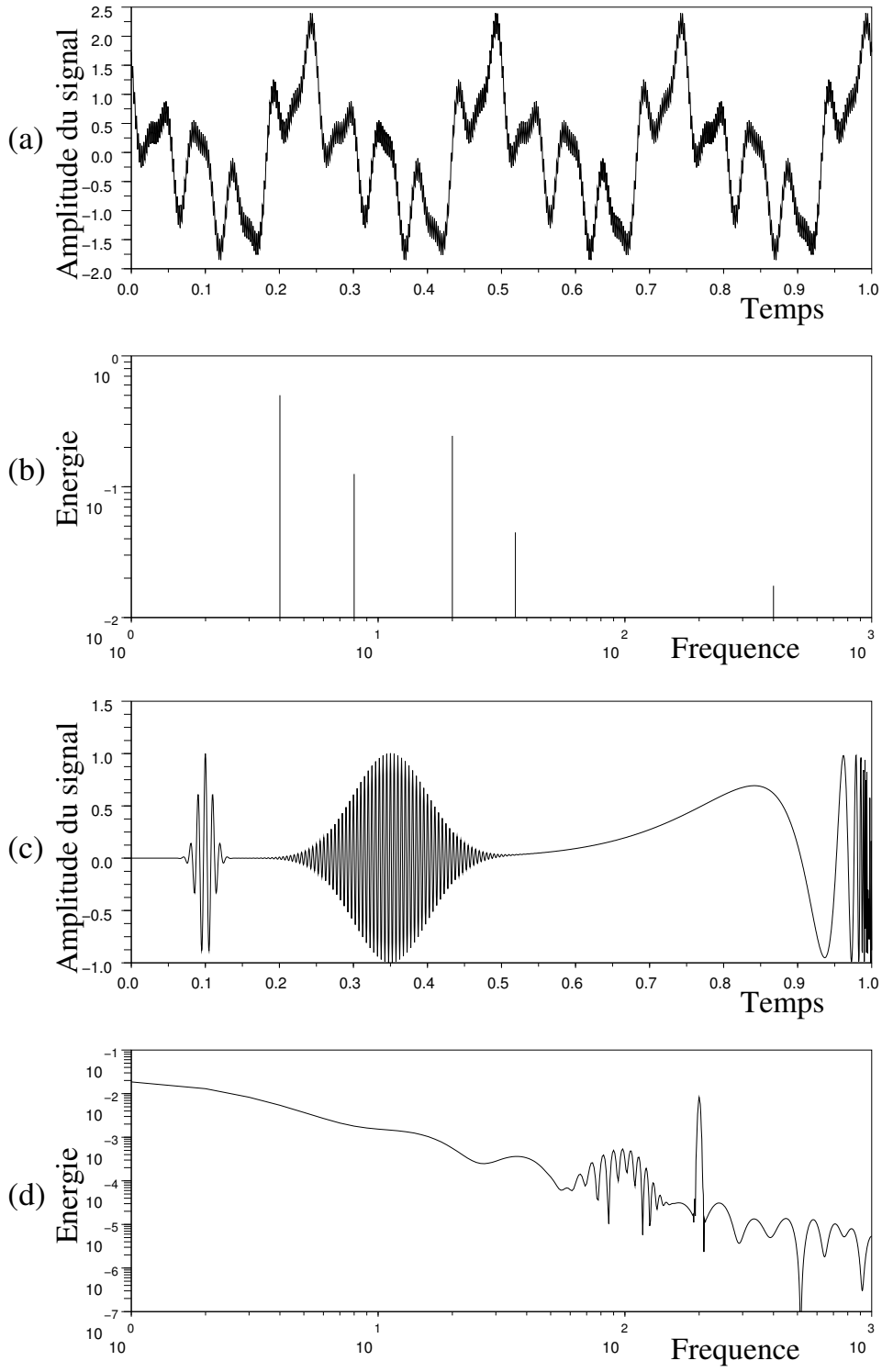


FIG. D.2 – Exemple de signal stationnaire (a) [resp. instationnaire (c)] et son spectre de Fourier (b) [resp. (d)].

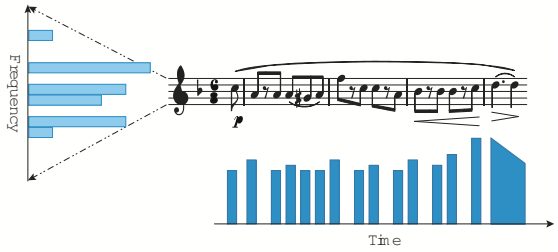


FIG. D.3 – Représentation spatiale et spectrale d'une partition de musique. Le graphique de l'évolution temporelle indique le moment où une note est jouée ainsi que son intensité, mais ne donne pas la fréquence. Le spectre de fréquence indique quelles sont les fréquences contenues dans le signal, mais ce spectre n'indique pas à quel moment une fréquence est jouée (d'après Fisher, newsletter ECMWF Winter 2005-2006).

Ainsi, toute information locale est noyée dans la moyenne et il n'est plus possible, alors, de la représenter. L'idéal serait de pouvoir spécifier localement l'information d'échelle. Naturellement une telle représentation doit être compatible avec l'inégalité de Heisenberg (intrinsèque).

Une partition de musique est une manière de définir un signal à la fois en temps et en fréquence. En effet, la figure D.3, illustre une telle représentation, et ses descriptions associées : spatiale et spectrale. La description spatiale permet de localiser l'information temporelle et l'intensité, mais ne donne pas accès à l'information fréquentielle. A l'inverse, le spectre de fréquence fournit cette information fréquentielle, mais ne décrit pas le moment où une fréquence donnée est jouée.

Un outil mathématique "récent" permet d'effectuer une analyse locale en espace et en fréquence, dans les conditions de localisation spécifiée par la contrainte de l'inégalité de Heisenberg. Il s'agit des ondelettes (pour l'histoire de la découverte des ondelettes voir Burke-Hubbard, 2000, p62-68).

D.2 L'analyse position-fréquence : application à la modélisation des corrélations

Le terme *analyse position-fréquence* est employé pour spécifier que le signal est étudié à la fois en espace et en fréquence. Ce type d'analyse regroupe toutes les méthodes d'analyse du signal en dehors de l'analyse harmonique de Fourier. En particulier, cela correspond *e.g.* à l'analyse en ondelettes continues, en ondelettes orthogonales ou en paquets d'ondelettes (et pseudo-cosinus locaux). Il existe d'autres méthodes d'analyse position-fréquence, par exemple : la transformée de Fourier fenêtrée, la décomposition en cosinus-locaux (particulièrement adaptée au signaux stationnaires par morceaux, ou localement-stationnaires). Toutes ces transformations cherchent à atteindre un même objectif : représenter au mieux un signal non stationnaire, en vue de son analyse ou de son traitement.

Contrairement à la transformation de Fourier, l'analyse position-fréquence développe un signal 1D en un signal 2D. Cette expansion est similaire à l'écriture temps-fréquence d'une partition musicale qui représente des notes (donc des structures fréquentielles), que l'on doit jouer à certain moment (donc localisées dans le temps) (cf Fig.D.3).

Si l'utilisation des ondelettes est ici fortement orientée par l'application visée, il existe d'autres domaines d'application en physique. Par exemple, elles sont utilisées dans l'extraction de structures cohérentes en turbulence (McWilliam, 1984 ; Farge, 1992). Cette application à la turbulence profite des propriétés de filtrage dans les bases ondelettes (Donoho et Johnstone,

1994). Ces thèmes ne sont pas abordés dans cette thèse.

Le formalisme des ondelettes est maintenant introduit sous une forme allégée. Le choix d'une telle présentation est d'exposer, avant tout, les idées, et de les illustrer sur des exemples utiles pour la question traitée : la modélisation des covariances. Le lecteur souhaitant approfondir les notions introduites ici peut se reporter à l'ouvrage très complet de Mallat (Mallat, 2001). Le formalisme des ondelettes sur la droite, introduit ici, ne correspond pas au formalisme usuel, généralement utilisé pour une telle introduction. Cependant, il est proche de la manière dont sont définies les ondelettes utilisées par Fisher (2003) sur la sphère, ce qui explique ce choix de mise en forme.

D.2.1 Ondelettes continues

Définition

Une transformation ondelette correspond à l'analyse d'un signal par des variantes d'une fonction ψ oscillante, de moyenne nulle, d'énergie finie et vérifiant la *condition d'admissibilité*

$$C_\psi = \int_0^{+\infty} \frac{|\hat{\psi}(\nu)|^2}{\nu} d\nu < +\infty, \tag{D.7}$$

avec $\hat{\psi}$ la transformée de Fourier de la fonction ψ. L'analyse d'un signal f correspond à des produits scalaires de ce signal, avec des versions dilatées et translatées de la fonction ψ. La dilatation permet d'analyser le signal en spectre pour une localisation spatiale donnée. La translation permet d'étudier les variations géographiques du signal.

Ainsi, la fonction translatée-dilatée de coefficient de translation u et de coefficient de dilatation s est donnée par $\psi_{s,u}(x) = \frac{1}{\sqrt{s}}\psi\left(\frac{x-u}{s}\right)$. s désigne l'information d'échelle et u celle de position. Cette fonction s'écrit également $\psi_{s,u}(x) = \psi_s(x-u)$ avec $\psi_s(x) = \frac{1}{\sqrt{s}}\psi\left(\frac{x}{s}\right)$, la fonction dilatée de la fonction ψ. L'analyse d'un signal par une telle fonction n'est autre que l'analyse dans le plan temps-fréquence telle que représentée sur la figure D.1.

Dans cette définition des fonctions $\psi_{s,u}$, le groupe des homothéties-translations a été introduit implicitement. Il existe en effet un lien intime entre les groupes de transformations et les ondelettes.

Le coefficient ondelette, pour le signal f, associé à cette fonction est donné par $f_{s,u} =< \psi_{s,u}|f >$. Il s'agit d'une "corrélation" entre le signal analysant $\psi_{s,u}$ et le signal analysé f. Une autre interprétation est celle d'une moyenne locale de poids $\psi_{s,u}$. Ainsi, la transformée ondelette d'un signal f 1D est un signal 2D. Les coefficients ondelettes caractérisent deux informations : l'une locale et l'autre spectrale. La transformée ondelette développe l'information 1D du signal en une information 2D.

Une autre manière d'écrire cette décomposition est d'introduire la fonction symétrique $\bar{\psi}_{s,u}$ de $\psi_{s,u}$, telle que $\bar{\psi}_{s,u}(x) = \bar{\psi}_{s,u}(-x)$. On a alors $f(s,u) = (\bar{\psi}_s * f)(u)$, ce qui définit la fonction

$$f_s = \bar{\psi}_s * f. \tag{D.8}$$

A ce stade, plusieurs angles d'approche de cette transformation ondelette se complètent.

Le premier est l'aspect de moyenne spatiale locale. En effet, La fonction f_s correspond à une version moyennée du signal initial f par le noyau $\bar{\psi}_s$ de la convolution, dont le support spatial est caractérisé par l'extension spatiale de la boîte de Heisenberg associée à $\bar{\psi}_s$. Il s'agit des détails du signal f à l'échelle s.

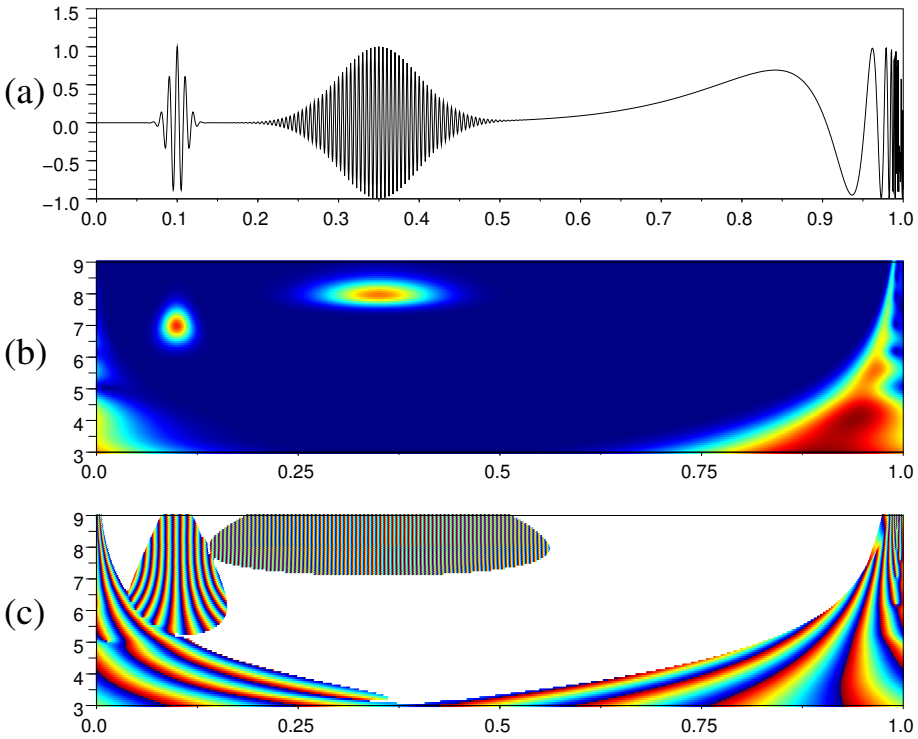


FIG. D.4 – Analyse ondelette d'un signal structuré avec une ondelette complexe de Morlet. (a) signal d'origine, (b) module de la transformée ondelette et (c) phase de la transformée ondelette : les zones en blanc correspondent aux zones dont le module est nul ou inférieur à un seuil très petit. L'axe des ordonnées caractérise l'échelle ; les grandes échelles sont en bas et les petites échelles en haut. Ainsi cet axe est associé à la représentation de l'information de fréquence avec les basses fréquences en bas et les hautes fréquences en haut.

Le deuxième correspond aux aspects de filtrage. En effet, ce formalisme basé sur la convolution permet de faire le lien entre l'analyse ondelette et le filtrage linéaire associé au formalisme des produits de convolution. En particulier, $\bar{\psi}_s$ s'interprète comme un filtre passe-bande : il permet d'extraire les informations spectrales contenues dans la bande spectrale spécifiée par la boîte de Heisenberg associée à $\bar{\psi}_s$.

De même que la transformation de Fourier, la transformation ondelettes est inversible : à partir de l'ensemble des versions f_s de f vues à des échelles différentes s, il est possible de reconstituer le signal initial f en le synthétisant à partir des informations aux différentes échelles tel que

$$f(u) = \frac{1}{C_\psi} \int_0^{+\infty} (\psi_s * f_s)(u) \frac{ds}{s^2}. \tag{D.9}$$

Illustration

La figure D.4 représente l'analyse ondelette du signal structuré représenté sur la figure D.2-(c). L'ondelette utilisée ici correspond à l'ondelette complexe de Morlet

$$\psi(x) = exp(-\frac{x^2}{2} + 5\imath x),$$

avec $\imath^2 = -1$.

Ainsi, le signal 1D initial est transformé en un signal 2D. Sur cette représentation du signal, le module (Fig. D.4-(b)) permet de localiser les structures en donnant leur localisation spatiale (axe des abscisses) et leur localisation spectrale (axe des ordonnées, représentant l'échelle ou la fréquence locale). Le carré de l'amplitude du module est lié à l'énergie de la structure locale. L'information de phase (Fig. D.4-(c)) permet de déterminer la fréquence locale du signal.

Les structures localisées au voisinage de 0.1 et de 0.35 apparaissent clairement sur la représentation du module. Le première structure est de plus petite échelle spatiale mais ses oscillations sont de plus basses fréquences par rapport à la seconde structure. La structure finale dont l'oscillation augmente au voisinage de l'abscisse 1 est également très remarquable sur la représentation du module : les échelles excitées sont de plus en plus petites. Cependant, étant près du bord, cette portion du signal provient des conditions aux bords périodiques. C'est également ce qui se passe au voisinage de 0 : il apparaît que le module présente de l'énergie alors qu'il n'y a pas de structure apparente. Il s'agit d'un artefact associé à la structure localisée en 1 (cône d'influence).

L'utilisation d'ondelettes complexes permet de séparer l'information d'amplitude et de phase. C'est ce type d'ondelette qui est habituellement utilisé pour l'analyse de signaux. En revanche pour le traitement numérique, ce sont plutôt les ondelettes réelles qui sont utilisées et en particulier les ondelettes orthogonales.

Utilisation des ondelettes continues dans la modélisation des corrélations

Les ondelettes continues (peut maniables) ne sont pas adaptées à la modélisation des corrélations par l'hypothèse diagonale. Cependant, dans le cas où l'ondelette est à spectre fini (nul au delà d'un certain nombre d'onde), alors une version discrète des ondelettes continues peut être construite. Cette approche est celle utilisée sur le cercle et présentée dans le chapitre 4 (Pannekoucke *et al.* , 2007). Dans ce contexte, les ondelettes sont dites à *bande limitée*. Cette approche est similaire à la décomposition ondelette sur la sphère.

D.2.2 Ondelettes orthogonales

Définition

Comme il a été mentionné dans les paragraphes précédents, la transformée ondelette développe un signal 1D en un signal 2D. Le signal 2D ne contient pas plus d'information que le signal 1D : il y a donc des couplages (ou dépendance) entre les coefficients. Cette dépendance provient de l'enchevêtrement des boîtes associés à deux fonction ψ_{s_1,u_1} et ψ_{s_2,u_2} avec s_1 proche de s_2 et u_1 proche de u_2.

Ainsi, l'analyse dans le plan temps-fréquence n'est pas optimale, puisque la représentation 2D est redondante. La théorie des ondelettes orthogonales, et des multirésolutions associées, permet de montrer qu'il est possible de réduire le nombre de coefficients à un jeu fini, tel que les coefficients soient indépendants les uns des autres. Cette représentation n'est valable que pour certaines ondelettes. Dans ce cas, l'analyse n'est nécessaire qui suivant un sous-ensemble $\psi_{j,i}$ d'ondelettes, avec i et j des indices discrets associés respectivement à la position et l'échelle, tels que les fonctions $\psi_{j,i}$ soient deux à deux orthogonales, la norme de $\psi_{j,i}$ étant 1. Ceci s'écrit $< \psi_{j,i} | \psi_{k,l} >= \delta_{j,k}\delta_{i,l}$, avec $\delta_{i,l}$ le symbole de Kroenecker.

En particulier, cet ensemble dénombrable de fonctions permet de former un pavage régulier optimal (au sens où il n'y a pas de redondance) pour le plan temps-fréquence. Pour

un signal discret de taille n, et de manière similaire au cas d'un signal périodique discrétisé dans le cas de Fourier, la transformation ondelette correspond à un changement de base orthonormale. Les ondelettes orthogonales sont associées à deux filtres : un passe-haut "g" et un passe-bas "h". Ces deux filtres permettent de séparer l'information spectrale : l'information résultant du filtrage passe-haut est orthogonale à celle résultant du passe-bas (Mallat, 1989). Formellement, si un signal f composé de $N = 2^J$ coefficients en points de grille appartient à un espace $\boldsymbol{V}_0$, alors $\boldsymbol{V}_0$ est décomposé sous la forme d'une somme directe orthogonale $\boldsymbol{V}_0 = \boldsymbol{V}_1 \oplus^{\perp} \boldsymbol{W}_1$, avec $h(f) \in \boldsymbol{V}_1$ et $g(f) \in \boldsymbol{W}_1$. $h(f)$ correspond à une information basse résolution et $g(f)$ correspond aux détails. Une décomposition ondelette orthogonale correspond à un filtrage en cascade du signal. Une manière possible pour calculer, en pratique, les coefficients ondelette est de périodiser l'ondelette. En particulier, un sous-échantillonnage est alors utilisé pour se prémunir d'un repliement du spectre (aliasing). Au final, l'espace initial $\boldsymbol{V}_0$ s'écrit alors $\boldsymbol{V}_0 = \boldsymbol{V}_J \oplus^{\perp} \boldsymbol{W}_J \oplus^{\perp} \boldsymbol{W}_{J-1} \oplus^{\perp} \cdots \oplus^{\perp} \boldsymbol{W}_1$. Dans cette écriture, l'espace $\boldsymbol{W}_j$ correspond au niveau de détail au niveau j.

La figure D.5 représente quelques fonctions ondelettes orthogonales ainsi que leur spectre de puissance. Deux types d'ondelette peuvent être distingués selon que l'ondelette est régulière ou non. Par exemple, l'ondelette de Haar (a) est discontinue et donc peu régulière. Le spectre de puissance de ce filtre présente bien une structure de passe-bande, mais avec plus de lobes. À l'opposé, l'ondelette Coiflet-5 (d) est plus régulière. Son spectre correspond effectivement à celui d'un passe-bande.

Algorithme d'analyse en ondelettes orthogonales

La figure D.6 représente les étapes de la décomposition orthogonale. Cette décomposition correspond à l'utilisation en cascade des filtres miroirs conjugués : passe-haut "g" et passe-bas "h". Les arbres binaires représentant les décompositions successives sont représentés sur la colonne de gauche. Les plans temps-fréquence (ou position-fréquence) et les boîtes de Heisenberg associés sont représentés sur la colonne de droite. Le signal initial en points de grille (a) est filtré successivement au cours des étapes (b - e). Pour un point de grille donné en (a), *e.g.* le point de grille 5, l'amplitude du signal en ce point correspond à l'amplitude de chaque composante spectrale relative à la transformation de Fourier du signal nul en tout point sauf en 5, où sa valeur correspond à celle du signal initial. Ainsi, cette uniformité de l'amplitude sur le spectre est caractérisée par une boîte de Heisenberg localisée au point 5 (rectangle grisé).

Les coefficients ondelettes correspondent à l'amplitude de l'information caractérisée par les boîtes de Heisenberg représentées en (e). Chaque étape, pour passer de (a) à (e), correspond à une décomposition haute fréquence/basse fréquence de l'information basse fréquence issue de l'étape précédente. Ces étapes sont représentées sur l'arbre binaire à gauche où chaque branche montante (resp. descendante) correspond au filtrage passe haut associé à "g" (resp. passe-bas associé à "h"). Les feuilles correspondent (points noirs) aux coefficients dans la décomposition. Des 16 informations en points de grille initiales, 16 coefficients ondelettes sont obtenus.

Illustration sur un exemple

La figure D.7 représente une erreur d'ébauche générée (a), compatible avec la matrice de covariance d'erreur d'ébauche introduite au chapitre 2, et représentée sur les figures 2.1 et 2.2. En particulier, cette erreur de prévision présente des variations d'échelle bien marquées : les oscillations sont de plus haute fréquence au voisinage de 180 °qu'au voisinage de 0 °. Ceci est à relier à la zone Z2 (résp. Z1), représentée sur la figure 2.2, correspondant à une région où les

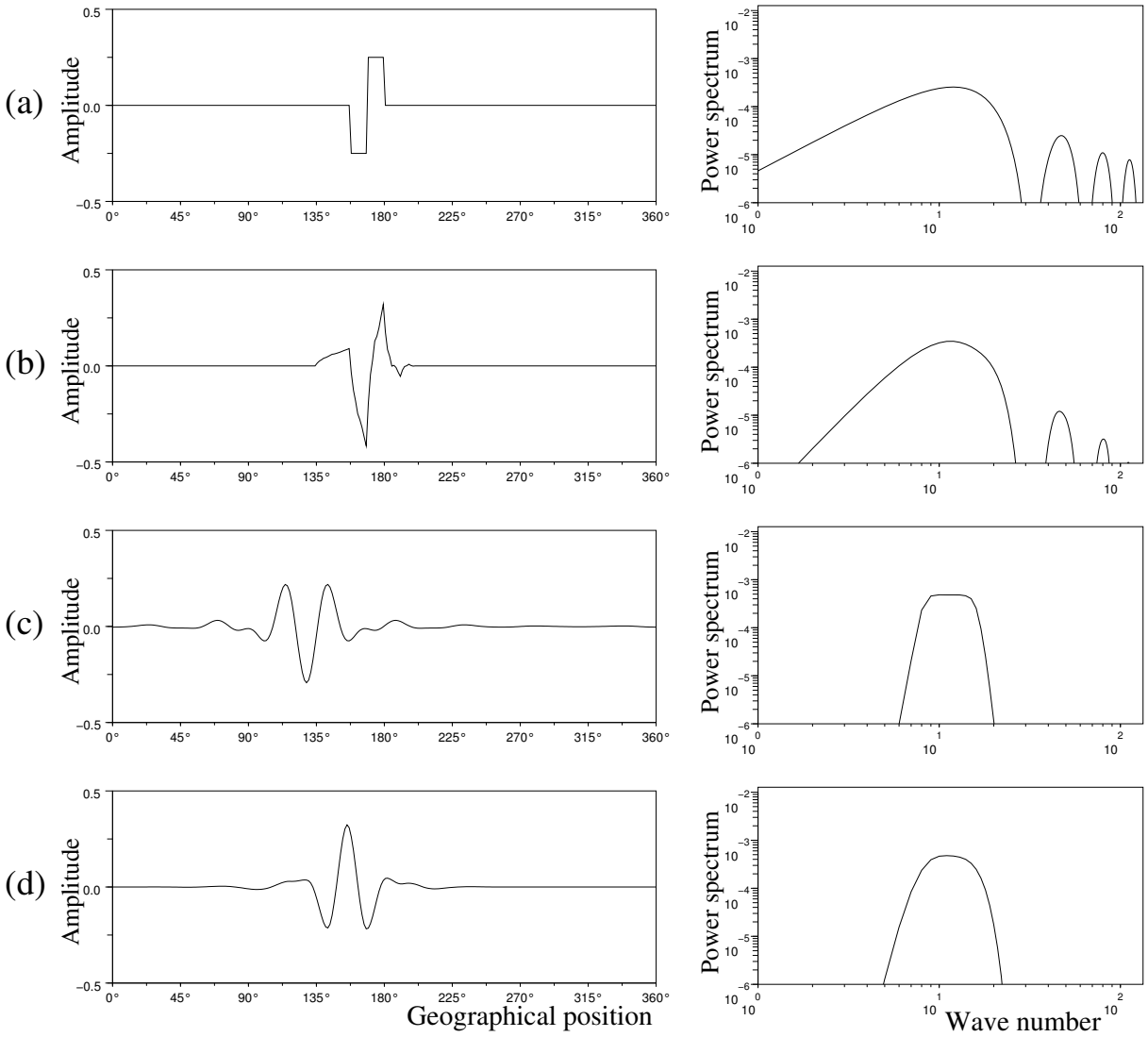


FIG. D.5 – Représentation de quelques ondelettes orthogonales (colonne de gauche) et de leur spectre de puissance (colonne de droite) : ondelette de Haar (a), ondelette de Daubechie (à 2 moments nuls) (b), ondelette de Battle-Lemarie (c) et Coiflet-5 (d). Une ondelette est une fonction oscillante localisée à la fois en espace et en fréquence.

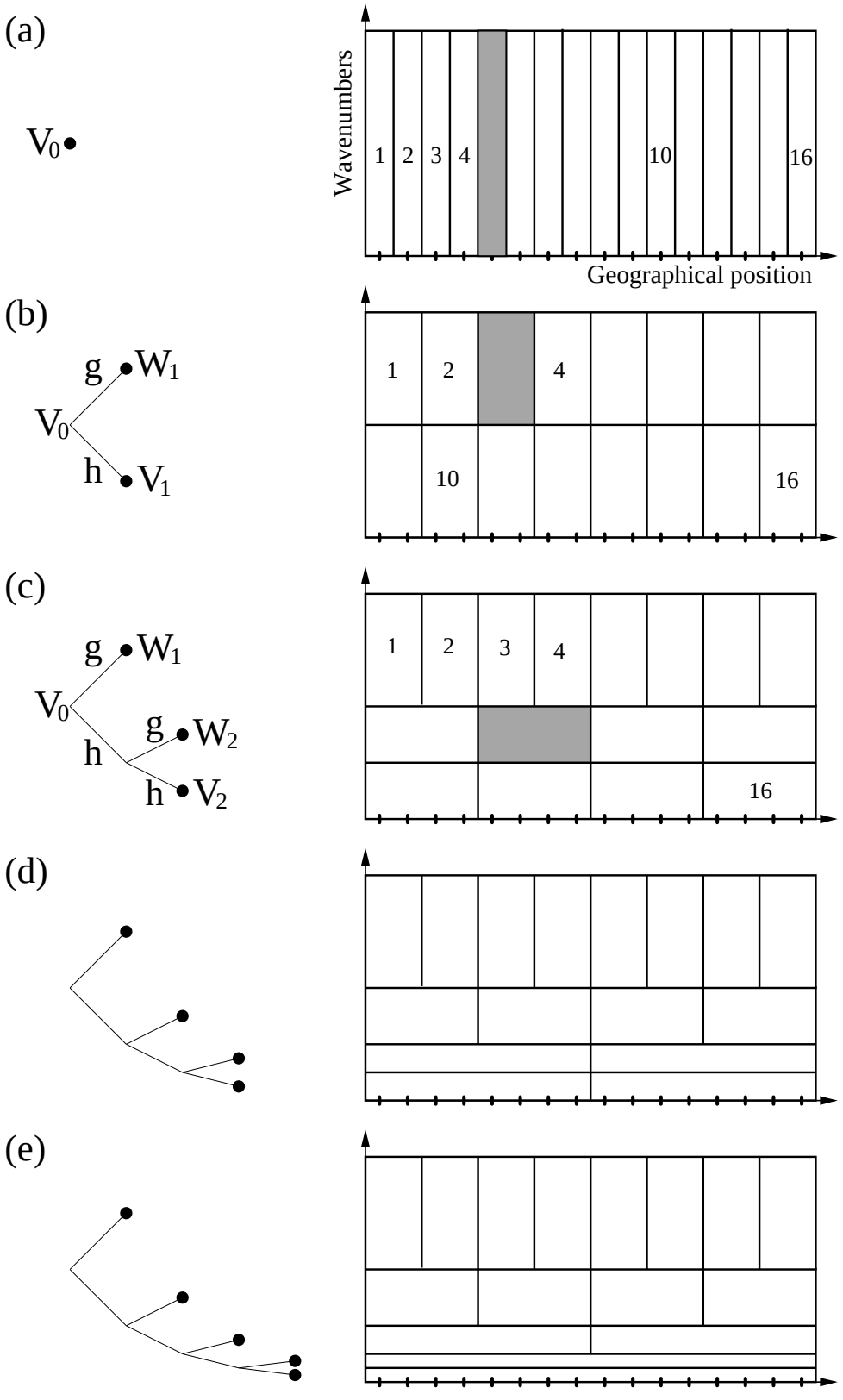


FIG. D.6 – Étapes de l'algorithme de décomposition en ondelettes orthogonales, à l'aide des filtres miroirs conjugués : passe-haut "g" et passe-bas "h". Le signal point de grille (a) est filtré suivant les arbres binaires représentés à gauche, pour obtenir au final les coefficients ondelettes (e) (voir le texte pour les détails).

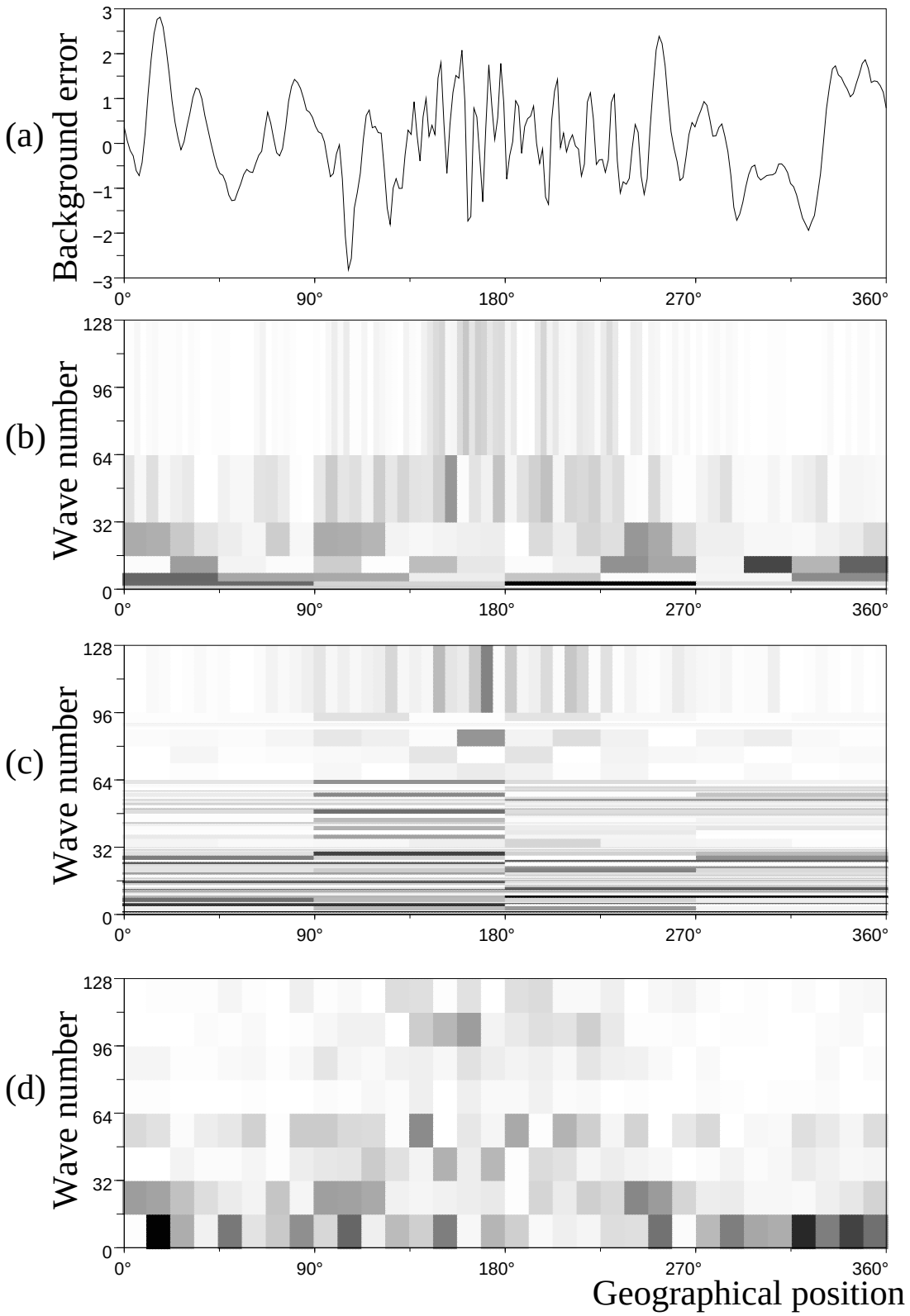


FIG. D.7 – Exemple de différentes décompositions admissibles (représentées dans le plan temps-fréquence à l'aide des boîtes de Heisenberg) pour un signal (a) correspondant à une erreur d'ébauche compatible avec la matrice de corrélation associé aux Fig. 2.1 et Fig. 2.2 : coefficients ondelettes (b), coefficients de la décomposition en paquets d'ondelettes la plus adaptée au signal (c), et coefficients de la décomposition en pseudo-cosinus locaux (d). Les niveaux de gris correspondent à l'amplitude du coefficient (en valeur absolue, blanc pour le minimum et noir pour la maximum) (résultats utilisant l'ondelette de Daubechies-2).

portées sont courtes (resp. larges). Le graphique (b) représente la décomposition ondelette du signal (a) à l'aide de l'ondelette orthogonale de Daubechies-2. L'amplitude des coefficients suit une échelle de gris : blanc pour le minimum et noir pour le maximum.

L'analogie entre la décomposition ondelette et une partition de musique, comme la présente Fisher (2006) (Fig. D.3) prend ici tout son sens. Les coefficients, représentés par l'amplitude et leur zone caractéristique dans le plan tremps-fréquence, apparaissent comme des notes jouées à certains instants. Ainsi, il est aisé de lire les coefficients ondelettes (b). Au voisinage de 0 °, le signal comporte avant tout des basses fréquences, caractérisées par une excitation des coefficients ondelettes associés à des fréquences basses. Puis le signal comporte des fréquences plus élevées au voisinage de 180 °, avec des coefficients ondelettes plus élevés dans les petites échelles. Pour finir, le signal redevient de plus grande échelle.

Utilisation des ondelettes orthogonales dans la modélisation des corrélations

La base orthonormale fournie par la décomposition en ondelettes orthogonales peut être utilisée pour la modélisation des corrélations sous l'hypothèse diagonale. Dans ce cas, le principal avantage est que l'information étant décrite à la fois en espace et en fréquence, il est alors possible de représenter des variations géographiques (espace) caractérisées par des variations de portées (fréquence).

La figure D.8 présente le résultat de la modélisation d'une matrice de corrélation de référence (a) (seules quelques fonctions de corrélation ont été représentées) avec l'hypothèse diagonale dans l'espace des coefficients ondelettes pour les ondelettes présentées sur la figure D.5.

Une première remarque primordiale est que les fonctions de corrélation ne sont pas les mêmes et qu'elles varient géographiquement. De plus ces variations correspondent bien à celles de la référence (a).

Ensuite, il apparaît que les corrélations modélisées sont d'autant plus proches de la référence que l'ondelette utilisée est lisse. Ainsi l'ondelette de Haar fournit une modélisation plutôt grossière (b) comparée à la Coiflet-5 (e). Ceci s'explique simplement par le fait que le signal de référence est relativement lisse.

D'autre part, dans toutes les modélisations, les corrélations ont du mal à être nulles comme l'est la référence, avec par exemple des oscillations autour de zéro sur (d).

Une autre propriété, commune à toutes les expériences, est que la matrice résultante de l'hypothèse diagonale n'est pas une matrice de corrélation. Ainsi, il faut alors normaliser par les variances résultantes de la modélisation pour obtenir une matrice de corrélation valide (non montré ici). En général, les variances résultantes de la modélisation directe sont sous-estimées. Cette observation a été faite dans le cas de l'utilisation d'autres ondelettes (Fisher, 2004 ; Pannekoucke *et al.* , 2007). Cette sous-estimation est à relier à la sous-estimation du spectre. Ce point est abordé plus en détail dans la section suivante.

Dans ces exemples, il apparaît donc que le choix de l'ondelette conditionne la qualité de la modélisation résultante. L'analyse par ondelette orthogonale n'est autre que la représentation du signal dans une autre base (tout comme la transformée de Fourier). Or, une propriété très intéressante pour la modélisation et qu'il est possible de construire un grand nombre de bases à l'aide d'une ondelette orthogonale. Il s'agit de la décomposition par paquets d'ondelettes. Reste alors à déterminer la base la plus appropriée : celle qui se rapproche le plus de la base de Karhunen-Loève (base qui diagonalise la matrice de covariance).

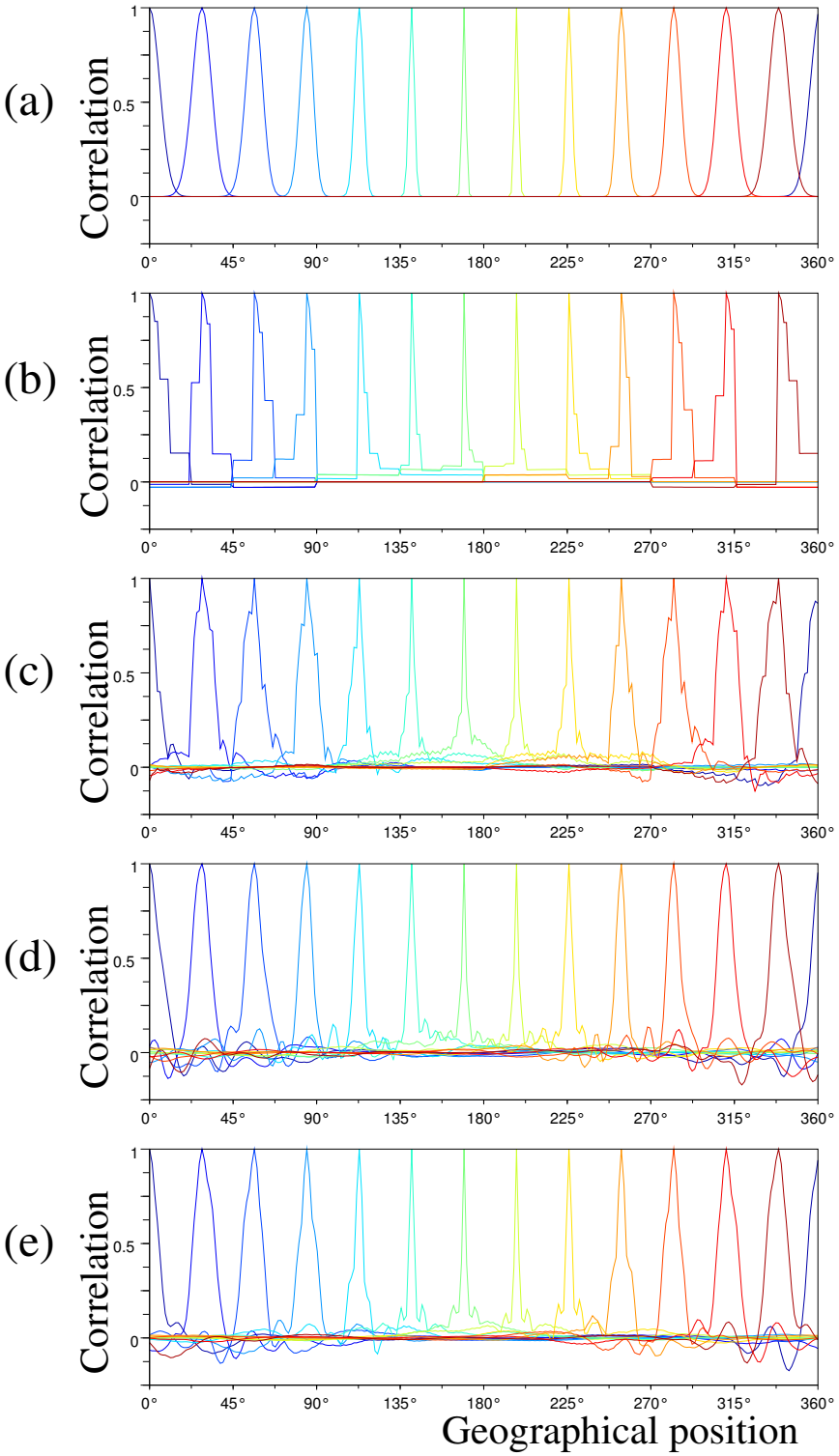


FIG. D.8 – Exemple de modélisation d'une matrice de corrélation de référence (a) à l'aide de l'hypothèse diagonale dans l'espace des coefficients ondelette pour les ondelettes orthogonales de Haar (b), Daubechies-2 (c), Battle-Lemarie (d) et Coiflet-5 (e). Ici, seules quelques fonctions de corrélation sont représentées (une couleur par fonction).

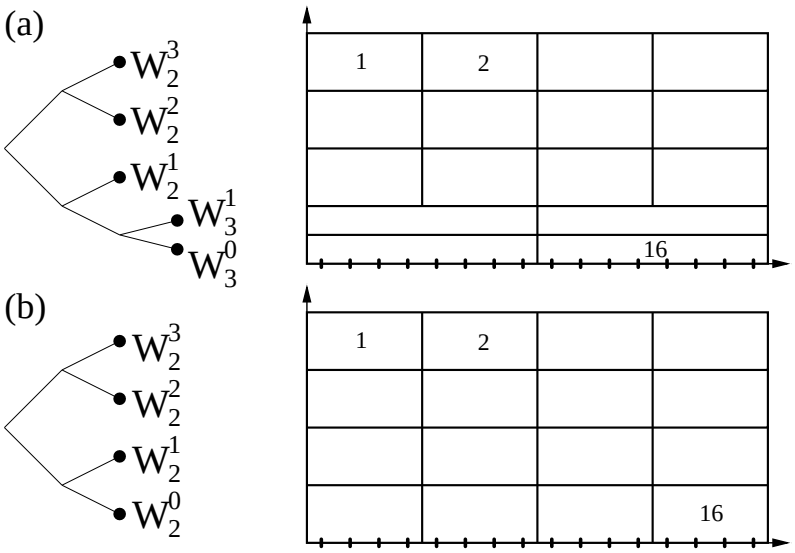


FIG. D.9 – Arbre binaire et pavage du plan temps-fréquence correspondant à une décomposition en paquets d'ondelettes orthogonales (a) et au cas particulier d'une décomposition en quasi-cosinus local (b). Le formalisme des espaces dans l'arbre binaire correspond au formalisme généralisé $\boldsymbol{W}_j^p$.

D.2.3 Paquet d'ondelettes et pseudo-cosinus locaux

Définition

L'idée apportée par les ondelettes est, au fond, la suivante : trouver une base orthogonale adaptée pour représenter le signal. Le terme "adaptée" signifie que la base permet de décrire le signal avec peu de coefficients significativement non nuls. Ainsi, le changement de base permet de passer d'une base orthogonale à une autre dans laquelle le signal se projette sur un nombre plus limité de coefficients significatifs.

Cependant, une base d'ondelette orthogonale n'est pas nécessairement la meilleure base. Une propriété intéressante est que, à partir de l'ondelette orthogonale, il est possible de construire d'autres bases qui sont plus adaptées. En effet, rappelons que dans le cas d'une décomposition en ondelettes orthogonales, chaque étape de la construction des coefficients ondelettes correspond à un filtrage passe-haut/passe-bas de l'information basse fréquence issue de l'étape précédente (cf Fig. D.6). Or il est également possible de filtrer l'information haute fréquence. Dans ce cas, l'information de position est dépréciée au profit de l'information fréquentielle. Pour ce contexte, un nouveau formalisme est introduit.

Les espaces correspondants aux feuilles de l'arbre sont maintenant notés $\boldsymbol{W}_j^p$, avec j le niveau de détail et p le numéro de la feuille au niveau j.

Chaque étape de la décomposition d'une feuille en une information basse fréquence/haute fréquence est alors donnée par la somme directe orthogonale $\boldsymbol{W}_j^p = \boldsymbol{W}_{j+1}^{2p} \oplus^{\perp} \boldsymbol{W}_{j+1}^{2p+1}$. Ici, $\boldsymbol{W}_{j+1}^{2p}$ correspond à l'espace de projection de l'information basse-fréquence et $\boldsymbol{W}_{j+1}^{2p+1}$ correspond à l'information haute-fréquence (détails au niveau j pour la feuille p).

La figure D.9-(a) représente l'arbre binaire et le pavage du plan temps-fréquence associé dans le cas où l'information haute-fréquence à la profondeur 1 est séparée en une composante haute-fréquence/basse-fréquence. Ce type de décomposition correspond à ce qui est appelé *une décomposition en paquets d'ondelettes*. Le cas particulier d'un arbre binaire dont les feuilles sont toutes situées à une même profondeur, correspond à *une décomposition en pseudo cosinus locaux*. Une telle décomposition est schématisée sur la figure D.9-(b).

Il est à noter que chaque feuille décrit un espace vectoriel. De plus l'espace vectoriel, associé à la discrétisation, correspond à la somme directe orthogonale des espaces vectoriels associés aux feuilles.

Dictionnaire de bases et meilleure base

Ainsi, la décomposition par paquets d'ondelettes permet de construire des nouvelles bases orthonormales. À chaque base orthonormale est associée un arbre binaire. Ainsi, le nombre de bases orthonormales correspond au nombre d'arbres binaires différents. Pour un arbre de profondeur J, il est alors possible de construire plus de $2^{2^{J-1}}$ bases orthonormales différentes (Mallat, 2001). D'autre part, le nombre de vecteurs différents est $P = Nlog_2(N)$ (ou $P = JN$).

Pour un signal donné, il est intéressant de rechercher la base dans laquelle le signal est "le mieux représenté". Naturellement, au niveau informationnel, toutes les bases représentent correctement le signal. Mais il peut exister des bases pour lesquelles l'information contenue dans un signal ne se projette significativement que sur un nombre limité de vecteurs de base. Ainsi, cette recherche est effectuée *e.g.* en minimisant une fonction coût.

Le coût de la représentation d'un signal discrétisé f dans une base $\mathcal{B}^\alpha = \{g_n^\alpha\}_{n\in[1,N]}$, pour une fonction concave Φ donnée est défini par

$$C_\Phi(f, \mathcal{B}^\alpha) = \sum_{n=1}^{N} \Phi\left(\frac{|< g_n^\alpha | f >|^2}{||f||^2}\right). \tag{D.10}$$

Les fonctions Φ couramment utilisées sont la fonction d'entropie $\Phi(x) = -x\,log(x)$, ou encore les coûts l^p $(p < 2)$ $\Phi(x) = x^{p/2}$. Avec $\Phi(0) = 0$, le coût est borné tel que

$$\Phi(1) \leq C_\Phi(f, B^\alpha) \leq N\Phi\left(\frac{1}{N}\right).$$

Ainsi, la meilleure base pour la fonction Φ est $\mathcal{B}^\alpha = Arg\,Min_{\mathcal{B}^\lambda\in\mathcal{D}}\left\{C_\Phi(f, \mathcal{B}^\lambda)\right\}$, *i.e.* celle qui minimise le coût, dans le dictionnaire de bases $\mathcal{D}$ (voir l'annexe pour la manière de déterminer la meilleure base dans un dictionnaire construit à partir d'un arbre binaire).

Illustration sur un exemple

La figure D.7-(c) représente la décomposition par paquet d'ondelettes (pour l'ondelette de Daubechies-2) du signal représenté sur le graphique (a). La base orthonormale construite, et représentée ici, correspond à la base minimisant le coût pour la fonction entropie. L'algorithme utilisé ici est basé sur une approche récursive profitant de l'orthogonalité des sous-espaces associés au feuille des arbres binaires (Coifman et Wickerhauser, 1992). Cet algorithme profite d'une propriété particulière de la fonction coût (voir annexe).

La figure D.7-(d) représente la décomposition en pseudo-cosinus locaux, pour une profondeur d'arbre de 4. La meilleure base est représentée par son arbre sur la figure D.10.

Utilisation des paquets d'ondelettes orthogonales dans la modélisation des corrélations

Pour déterminer la meilleure base approximant la base de Karhunen-Loève et appartenant à un dictionnaire de bases donné, il est possible d'utiliser l'algorithme décrit par Mallat *et al.* (1998). Comme dans le cas de la recherche de la meilleure base pour un signal donné, cet algorithme consiste à minimiser une fonction coût (Mallat *et al.* , 1998 ; Donoho *et al.* , 2003). Les notations utilisées dans la suite sont celles de Mallat (2001).

Une fonction coût adaptée à ce problème est donnée par

$$C(\boldsymbol{B}, \mathcal{B}) = -\sum_k |p(k)|^2, \tag{D.11}$$

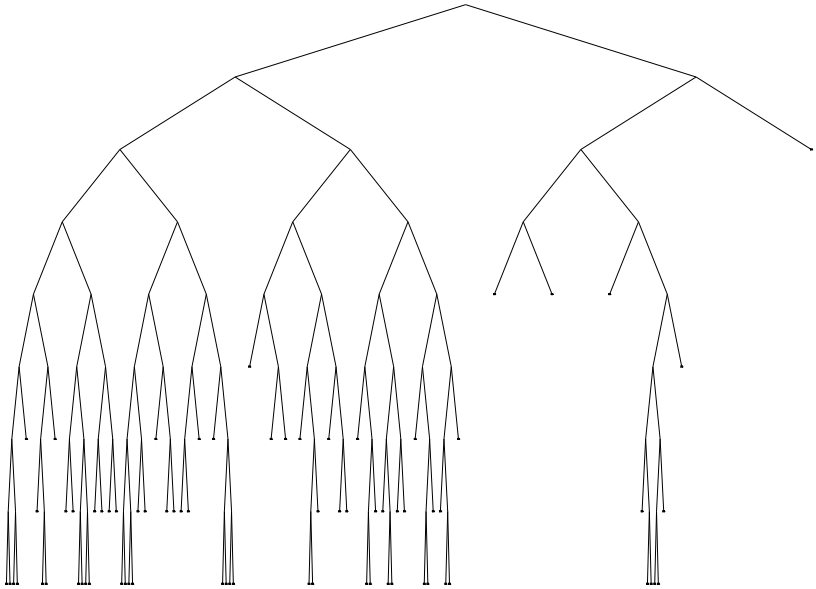


FIG. D.10 – Arbre binaire pour la meilleure base de l'erreur de prévision représentée sur la figure D.7-(a). L'information basse fréquence correspond aux feuilles de gauche. La racine de l'arbre correspond à l'information point de grille. Les feuilles de l'arbre correspondent aux détails pour un niveau donné.

où les quantités $p(k) =< \boldsymbol{g}_k | \boldsymbol{B}\boldsymbol{g}_k >$ correspondent à la diagonale de $\boldsymbol{B}$ dans la base $\mathcal{B} = \{\boldsymbol{g}_k\}$ (voir en annexe l'origine de cette fonction coût et la consistance du problème de minimisation).

La figure D.11 représente les matrices de corrélation modélisées à l'aide de l'hypothèse diagonale dans l'espace de la meilleure base de paquets d'ondelettes : pour les ondelettes de Haar, Daubechies-2, Battle-Lemarie et Coiflet-5. La comparaison avec la figure D.8 permet d'observer les améliorations obtenues. Les fonctions de corrélation semblent mieux localisées pour l'ondelette de Haar. Cependant, les autres modèles présentent toujours des oscillations erronées.

Une manière de quantifier les résultats est de calculer l'erreur relative (en pourcentage) de l'approximation $\boldsymbol{B}_{model}$ de $\boldsymbol{B}$: $e_{model} = 100 \frac{||\boldsymbol{B}_{model}||^2 - ||\boldsymbol{B}||^2}{||\boldsymbol{B}||^2}$. La norme utilisée ici étant celle de Hilbert (voir annexe), elle correspond, pour une matrice de covariance $\boldsymbol{\Gamma}$ donnée, à l'opposé de $C(\boldsymbol{\Gamma}, \mathcal{B}^{KL})$ où $\mathcal{B}^{KL}$ est une base de Karhunen-Loève de $\boldsymbol{\Gamma}$. La figure D.12 représente cette erreur relative, pour les différentes ondelettes, et pour trois modèles : hypothèse diagonale ondelette sans renormalisation (courbe avec carrés), hypothèse diagonale dans la meilleure base sans normalisation (courbe avec croix) et avec normalisation (courbe avec triangles noirs). La normalisation consiste à multiplier la matrice modélisée par l'inverse de sa diagonale, afin d'assurer des corrélations égales à 1 sur la diagonale. Le choix de l'ondelette apparaît fondamental : plus de 20% d'erreur pour l'ondelette de Haar et moins de 9% d'erreur pour celle de Battle-Lemarié. La représentation dans la meilleure base apporte une amélioration notable pluisque pour l'ondelette de Battle-Lemarié, le pourcentage d'erreur passe de 8.7% à 2.3%. La normalisation en points de grille permet d'améliorer ce résultat : seulement 1.3% d'erreur. En particulier, la normalisation permet de combler une part de l'énergie perdue par l'approximation diagonale dans une base autre que celle de Karhunen-Loève.

D.3 Commentaires sur la modélisation 3D

Dans ce qui précède, seul un champ horizontal a été considéré. Pour l'atmosphère 3D, il est nécessaire de faire correspondre les niveaux entre eux, et de construire une formulation non-séparable.

L'hypothèse diagonale dans l'espace des coefficients ondelettes permet de représenter les

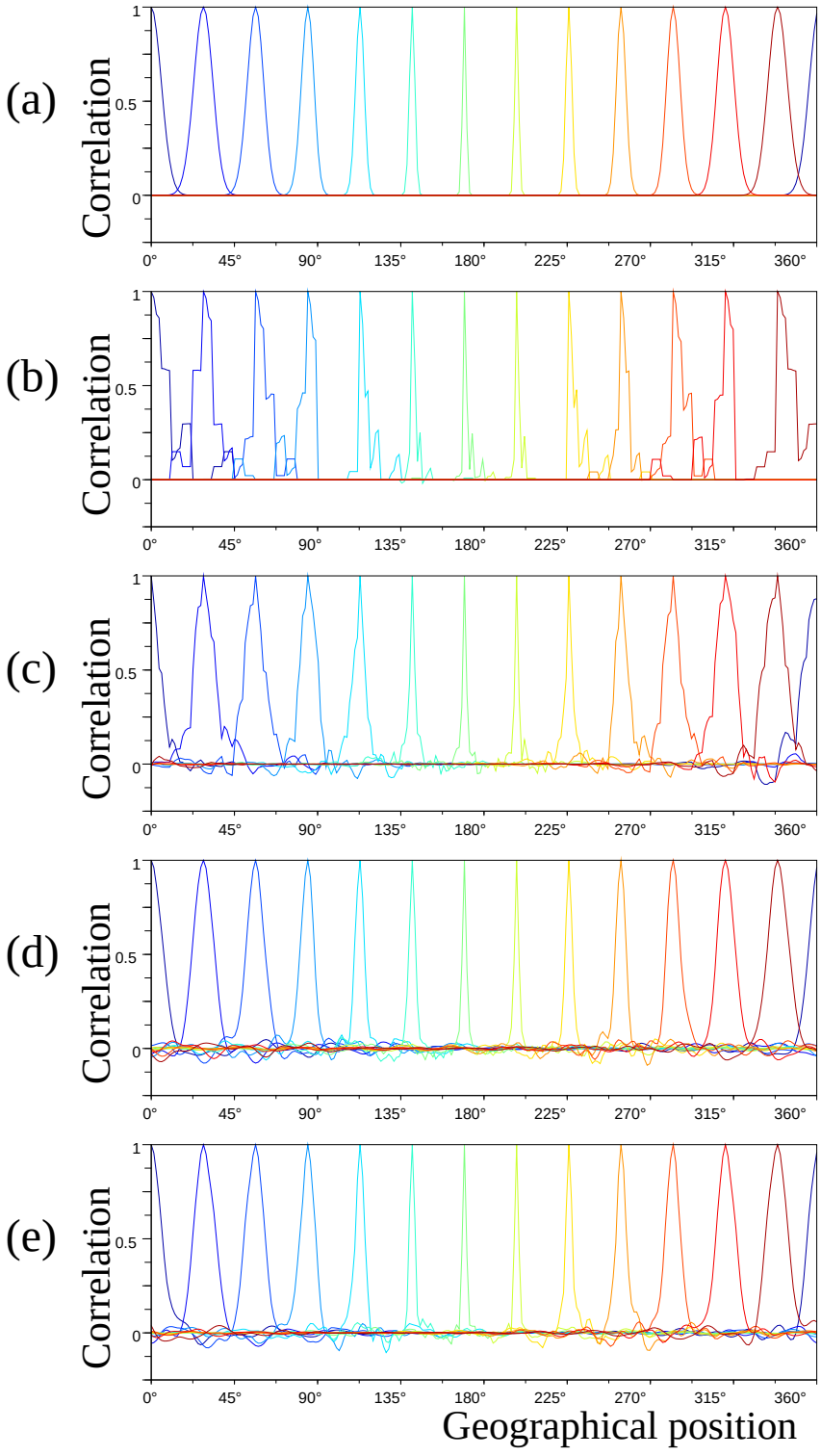


FIG. D.11 – Exemple de modélisation d'une matrice de corrélation de référence (a) à l'aide de l'hypothèse diagonale dans l'espace des coefficients pour la meilleure base de paquets d'ondelettes, pour les ondelettes orthogonales de Haar (b), Daubechies-2 (c), Battle-Lemarie (d) et Coiflet-5 (e). Ici, seules quelques fonctions de corrélation sont représentées (une couleur par fonction).

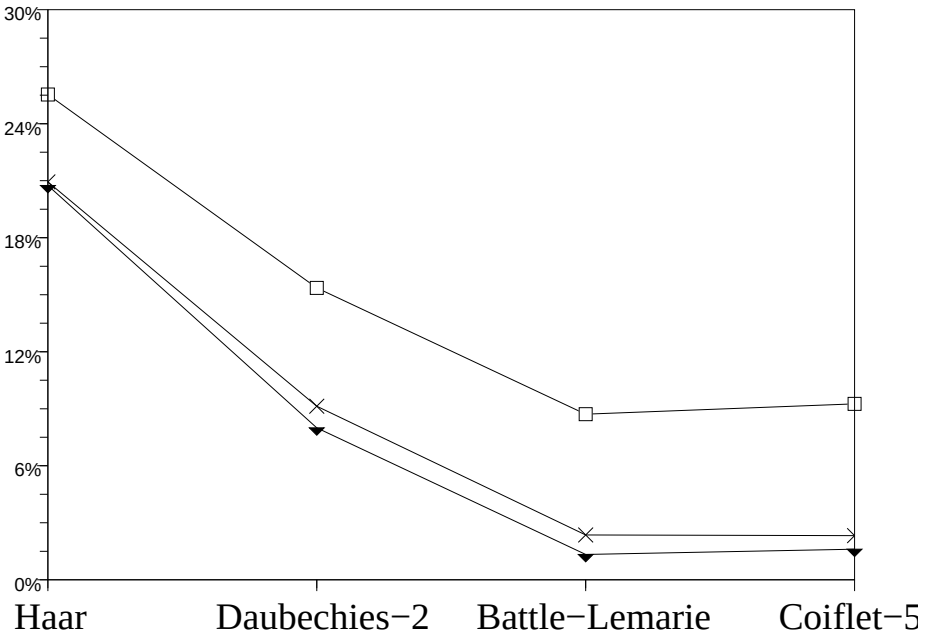


FIG. D.12 – Erreur relative de l'approximation de la matrice $\boldsymbol{B}$ de référence pour les ondelettes de Haar, Daubechises-2, Battle-Lemarié et Coiflet-5 : approximation dans la base d'ondelette orthogonale (courbe avec carrés), approximation dans la meilleure base de paquets d'ondelettes sans normalisation (courbe avec croix) et avec normalisation (courbe avec triangles).

variations géographique des fonctions de covariance. Il est relativement simple de construire une matrice de corrélation non-séparable à partir d'une modélisation ondelette sur l'horizontale (Fisher, 2001). De manière similaire au cas spectral (Courtier *et al.* , 1998 ; Derber et Bouttier, 1999) il suffit de spécifier les corrélations tridimentionnelles $\boldsymbol{C}_\alpha$, pour un champ α, sous la forme d'une matrice diagonale par bloc, où chaque bloc diagonal est une matrice de corrélation s'écrivant cette fois-ci $\mathbf{H}^{1/2}_{\alpha,i,j}\mathbf{V}_{\alpha,i,j}\mathbf{H}^{1/2}_{\alpha,i,j}$ et ne dépendant que des position i et de l'échelle j (dans le cas spectral, il n'y a qu'une dépendance en fonction du nombre d'onde total).

Si l'utilisation d'une meilleure base de paquets d'ondelettes améliore la représentation de la matrice de covariance 1D (ou 2D dans le cas d'un modèle à aire limitée), il reste difficile de concevoir une matrice 3D non-séparable. En effet, dans ce cas, il est nécessaire de faire correspondre des directions principales différentes d'un niveau de l'atmosphère à un autre. Ainsi, la modélisation non-séparable comme indiquée dans le cas ondelette ne semble pas applicable directement.

La géométrie sphérique, très particulière, rend mal adaptés les outils de type ondelettes orthogonales et paquets d'ondelettes. Ces outils sont en effet adaptés au tore (bi-périodique ou bi-Fourier), mais ne savent pas bien prendre en compte la re-connexion des maillages aux pôles. Si des utilisations des ondelettes 2D bi-périodiques ont déjà été réalisées pour la sphère (Auger et Tangborn, 2004), il est possible d'utiliser des ondelettes spécialement construites pour la sphère (Fisher, 2003).

D.4 "Ondelette" sur la sphère

Les ondelettes sur la sphère S^2 diffèrent notablement de celles sur la droite. En particulier, si la notion de translation admet un équivalent sur la sphère, à savoir le groupe des rotations

$SO(3)$, la dilatation n'est, quant à elle, pas possible au sens introduit sur la droite. En effet, toute dilatation autour d'un point implique une contraction pour le voisinage des points aux antipodes (Holschneider, 1990).

Cependant, il est tout de même possible de définir des familles de fonctions à la fois localisées en espace et en fréquence. Pour ce faire, la définition des ondelettes est donnée dans l'espace spectral associé à la sphère, plus particulièrement l'espace de Legendre (Freeden et Windheuser, 1996 ; Freeden et Schreiner, 1998). Dans ce cas, par analogie au cas 1D issu du formalisme Eq. (D.8), les ondelettes sont définies par leur filtre associé. Le cas particulier d'une fonction sur la sphère définie uniquement par son spectre de Legendre correspond au cas d'une fonction radiale (ou zonale), *i.e.* une fonction possédant un axe de révolution.

Les coefficients ondelettes sont alors définis comme la convolution du signal à analyser avec la famille de fonctions ondelettes. Sur la sphère, la convolution d'un signal avec un autre, est définie sur le groupe des rotations $SO(3)$. Par exemple, la convolution des fonctions f et g définient sur la sphère est donnée par $(g * f)(r) = \int_{S^2}(g \circ r^{-1})(\zeta) f(\zeta) d\omega(\zeta)$ avec $r \in SO(3)$ et $d\omega$ la métrique sur la sphère ; $g * f$ est bien une fonction définie sur $SO(3)$. Dans le cas particulier de la convolution où la fonction analysante est radiale, la convolution se simplifie en une fonction définie sur $SO(3)/SO(2) \equiv S^2$, *i.e.* la sphère (Boer, 1994 ; Courtier *et al.* , 1998). Dans ce cas particulier, si Ψ_j désigne une ondelette radiale, et f désigne la fonction à analyser, il vient la convolution $f_j = \Psi_j * f$ telle que $f_{j_n^m} = \Psi_n f_n^m$, avec Ψ_n le spectre de Legendre de Ψ, f_n^m le spectre en harmonique sphérique de f.

Une relation similaire à l'équation (D.9) permet de reconstruire le signal à partir de la décomposition ondelette.

L'idée d'utiliser les ondelettes pour modéliser la matrice de covariance d'erreur de prévision par l'hypothèse diagonale a été explorée par Fisher (2003) sur la sphère (Fisher, 2004) et Deckmyn et Berre (2005) sur un bi-Fourier (pour Aladin). Les ondelettes sphériques utilisées par Fisher ne sont pas orthogonales (à noter qu'il n'existe pas d'ondelettes orthogonales sur la sphère, au sens de celle des paragraphes précédents). Il s'agit des frames (Fisher, 2004 ; Daubechies, 1992) : version discrète d'ondelette continue, dont les coefficients sont dépendants les uns des autres. Ces ondelettes et leur utilisation pour la modélisation sont étudiées en détail au chapitre 4.

La modélisation relative à l'utilisation de ces ondelettes sphériques prend difficilement en compte l'anisotropie. Une raison principale provient du caractère radial des fonctions ondelettes. Il serait possible d'améliorer cette modélisation par l'utilisation d'ondelettes non radiales (Torrésani, 1995). Cependant, la convolution doit alors être prise sur le groupe des rotations en dimension 3 $SO(3)$.

D.5 Conclusions

Au cours de ce chapitre, les idées fondatrices des ondelettes ont été présentées. En particulier, leur utilisation dans la modélisation des covariances a été introduite. Le choix de l'ondelette s'est avérée importante, pour approximer au mieux la base de Karhunen-Loève : la base qui diagonalise la matrice de covariance. L'utilisation de la décomposition en paquets d'ondelettes et le dictionnaire de bases associé ont montré un gain dans l'approximation de la base de Karhunen-Loève.

Même si ces outils ne sont pas disponibles sur la sphère (peut être que ces versions équivalentes pourront être construites dans le futur), leur utilisation a révélé des comportements observés avec d'autres formulations ondelette, en particulier celle présentée au chapitre 4.

Ainsi, ce chapitre a permis de faire un bilan non exhaustif de la manière d'utiliser d'autres représentations du signal que la transformée de Fourier (ou de Laplace sur la sphère).

Enfin, il est à noter qu'il existe d'autres utilisations des ondelettes pour la modélisation des fonctions de covariance. En effet, il est possible d'utiliser les propriétés des ondelettes orthogonales pour la compression du signal. Dans ce cas, seule l'information de "grande échelle" est conservée, en supprimant les détails au-delà d'une certaine échelle. Ainsi, la matrice étant de rang réduit, il est moins coûteux de la propager temporellement à l'aide des équations du filtre de Kalman (Auger et Tangborn, 2004 ; Tangborn, 2004).

D.6 Annexe : Approximation de la base de Karhunen-Loève

On rappelle que la base de Karhunen-Loève (KL) est la base qui diagonalise une matrice de covariance. En pratique une telle base n'est pas connue et le nombre d'échantillons pour la calculer précisément est trop petit. En particulier, si aucune information supplémentaire n'est fournie, il est possible de montrer qu'un bon estimateur pour la matrice recherchée est la matrice nulle. Naturellement, cette matrice nulle n'a aucun sens dans les applications.

Ainsi, il est donc préférable de fixer une base adaptée, prise comme approximation de la base KL. Dans ce cas, il suffit d'estimer les puissances spectrales de la matrice de covariance dans cette base.

Les paragraphes suivants s'intéressent à la manière de construire une bonne approximation de la base KL dans un dictionnaire de bases, dans le cas où la matrice est connue (cas test pour expérience numérique analytique) et dans le cas où la matrice n'est pas connue et doit être estimée à l'aide d'échantillons.

Cas où la matrice de covariance est connue

Dans l'espace des matrices carrées de taille n, la norme de Hilbert (ou de Froebenius) d'une matrice $\boldsymbol{B}$ est définie par $||\boldsymbol{B}|| = \sqrt{Trace\,(\boldsymbol{B}^T\boldsymbol{B})}$. Cette norme vérifie $||\boldsymbol{B}||^2 = \sum_{i,j} |b_{ij}|^2$. Cette norme est invariante pour tout changement de base orthonormale (pour laquelle la matrice de passage $\boldsymbol{O}$ est orthogonale, *i.e.* $\boldsymbol{O}^T = \boldsymbol{O}^{-1}$).

Soit $\mathcal{B}$ une base orthonormale, les termes diagonaux de $\boldsymbol{B}$ dans la base $\mathcal{B} = \{\boldsymbol{g}_k\}$ sont notés

$$p(k) =< \boldsymbol{g}_k|\boldsymbol{B}\boldsymbol{g}_k > .$$

La matrice $\boldsymbol{B}$ est approximée par la matrice $\boldsymbol{B}^d$, diagonale dans la base $\mathcal{B}$, et égale à la diagonale de $\boldsymbol{B}$ représentée dans $\mathcal{B}$. La matrice $\boldsymbol{B}^d$ correspond à la modélisation de la matrice $\boldsymbol{B}$ par l'hypothèse diagonale dans la base $\mathcal{B}$.

Ainsi, la norme de Hilbert de la différence des deux matrices peut s'écrire

$$||\boldsymbol{B}^d - \boldsymbol{B}||^2 = ||\boldsymbol{B}||^2 + C(\boldsymbol{B}, \mathcal{B}), \tag{D.12}$$

avec $C(\boldsymbol{B}, \mathcal{B}) = -||\boldsymbol{B}^d||^2 = -\sum_k |p(k)|^2$. Cette dernière quantité est minorée suivant

$$C(\boldsymbol{B}, \mathcal{B}) \geq -||\boldsymbol{B}||^2.$$

Il y a égalité dans le cas où la base correspond à la base de Karhunen-Loève $\mathcal{B}^{KL}$. En effet, dans une base $\mathcal{B}$ autre que $\mathcal{B}^{KL}$, les contributions hors diagonale intervenant dans le calcul de $||\boldsymbol{B}||^2$ sont mises à zéro dans le calcul de $||\boldsymbol{B}^d||^2$, conduisant à l'inégalité $||\boldsymbol{B}^d||^2 \leq ||\boldsymbol{B}||^2$.

Ainsi le problème de la minimisation du coût $C(\boldsymbol{B}, \mathcal{B})$ pour $\mathcal{B}$ une base d'un dictionnaire $\mathcal{D}$ a bien un sens et admet une solution. Cette quantité apparaît naturellement comme étant adaptée au problème de la recherche de la meilleure base approximant la base de Karhunen-Loève dans un dictionnaire donné.

Cette fonction coût peut s'étendre à n'importe quelle projection de $\boldsymbol{B}$. Par exemple, en considérant un sous-espace $\boldsymbol{W}_j^p$ muni d'une base orthogonale $\mathcal{B}_j^p = \{\boldsymbol{g}_m^{p,j}\}$ où $m \in I_j^p$ avec I_j^p un ensemble d'indices. Alors la projection de $\boldsymbol{B}$ sur cet espace et pour cette base orthonormale s'écrit

$$P_{\mathcal{B}_j^p}(\boldsymbol{B}) = \sum_{m \in I_j^p} < \boldsymbol{g}_m^{p,j} | \boldsymbol{B} \boldsymbol{g}_m^{p,j} > \ | \boldsymbol{g}_m^{p,j} >< \boldsymbol{g}_m^{p,j} |,$$

où $|\boldsymbol{g}_m^{p,j} >< \boldsymbol{g}_m^{p,j}|$ désigne le projecteur orthogonal associé au vecteur $\boldsymbol{g}_m^{p,j}$. Ainsi, le coût de cette approximation est alors $C(\boldsymbol{B}, \mathcal{B}_j^p) = \sum_{m \in I_j^p} | < \boldsymbol{g}_m^{p,j} | \boldsymbol{B} \boldsymbol{g}_m^{p,j} > |^2$.

Ainsi, il apparaît que la fonction coût est *additive*, au sens où, si une base $\mathcal{B}$ est la réunion de deux bases orthonormale $\mathcal{B}_1$ et $\mathcal{B}_2$, *i.e.* $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$, alors

$$C(\boldsymbol{B}, \mathcal{B}_1 \cup \mathcal{B}_2) = C(\boldsymbol{B}, \mathcal{B}_1) + C(\boldsymbol{B}, \mathcal{B}_2). \tag{D.13}$$

Dans le cas où le dictionnaire de bases considéré peut être contruit à partir d'un arbre binaire de sous-espaces orthogonaux (comme c'est le cas pour les paquets d'ondelette), cette propriété est fondamentale. En effet, elle permet alors une recherche déterministe de la meilleure base approximant la base KL d'après l'algorithme de Coifman et Wickerhauser (1992) (Donoho *et al.* , 2003).

La base de $\boldsymbol{W}_j^p$ construite au travers de l'arbre binaire est notée $\mathcal{B}_j^p$. Soit $\mathcal{O}_j^p$ la meilleure base de l'espace $\boldsymbol{W}_j^p$. On rappelle que l'arbre fournit la décomposition $\boldsymbol{W}_j^p = \boldsymbol{W}_{j+1}^{2p} \oplus^{\perp} \boldsymbol{W}_{j+1}^{2p+1}$. Ainsi, la meilleure base est construite récursivement d'après l'algorithme suivant :

$$\mathcal{O}_j^p = \begin{cases} \mathcal{B}_j^p & \text{si une feuille de l'arbre est atteinte} \\ \mathcal{O}_{j+1}^{2p} \cup \mathcal{O}_{j+1}^{2p+1} & \text{si } C(\boldsymbol{B}, \mathcal{O}_{j+1}^{2p}) + C(\boldsymbol{B}, \mathcal{O}_{j+1}^{2p+1}) < C(\boldsymbol{B}, \mathcal{B}_j^p) \\ \mathcal{B}_j^p & \text{si } C(\boldsymbol{B}, \mathcal{O}_{j+1}^{2p}) + C(\boldsymbol{B}, \mathcal{O}_{j+1}^{2p+1}) \geq C(\boldsymbol{B}, \mathcal{B}_j^p) \end{cases} . \tag{D.14}$$

Remarque : l'algorithme présenté ici correspond également à celui de la recherche de la meilleure base. Dans ce cas, la fonction coût est adaptée à cette recherche particulière : le coût est alors $C_{\Phi}(f, \mathcal{B})$ (voir section D.2.3).

Cas où la matrice de covariance est estimée à partir d'un ensemble

Cet algorithme peut être adapté au cas où la matrice $\boldsymbol{B}$ n'est pas connue de manière exacte mais uniquement via un ensemble d'erreurs de prévision ${\boldsymbol{\varepsilon}^b}_n$, $n \in [1, N_e]$. La matrice estimant $\boldsymbol{B}$ est alors $\overline{\boldsymbol{B}} = \frac{1}{N_e} \sum_n {\boldsymbol{\varepsilon}^b}_n {\boldsymbol{\varepsilon}^b}_n^T$.

Dans ce cas, il s'agit de minimiser une fonction coût similaire à celle du paragraphe ci-dessus, avec $\overline{p}(k) = \frac{1}{N_e} \sum_n | < \boldsymbol{g}_k | {\boldsymbol{\varepsilon}^b}_n > |^2$ (Donoho *et al.* , 2003). D'autre part, $\overline{p}(k)$ correspond à l'estimation de la diagonale de $\boldsymbol{B}$ dans la base $\mathcal{B}$. Cette quantité correspond également à la variance suivant la direction $\boldsymbol{g}_k$. Pour la base de Fourier, il s'agit de la puissance spectrale, ou variance spectrale du signal. Pour la base d'ondelettes, il s'agit de la variance ondelette.

Annexe E

Complément sur les frames et leur utilisation

Complément sur les frames et point de vue matriciel

Cette partie n'est pas fondamentale pour la compréhension de ce qui a été développé dans le chapitre 4 mais elle permet une description plus détaillée des algorithmes et de leur mise en oeuvre numérique.

Dans ce paragraphe, les fonctions sont des éléments d'un espace de Hilbert $(\mathcal{H}, < \mid >)$. L'opérateur linéaire $\boldsymbol{T}$, défini par $\forall m \in \mathcal{M}, (\boldsymbol{T}f)_m =< \tilde{\psi}_m | f >$, est à valeur dans

$$l^2(\mathcal{M}) = \left\{ (c_m)_{m\in\mathcal{M}} \text{ tel que } \sum_{m\in\mathcal{M}} |c_m|^2 < +\infty \right\}.$$

Daubechie montre que $\boldsymbol{T}^*\boldsymbol{T}$ est inversible et que

$$\psi_m = (\boldsymbol{T}^*\boldsymbol{T})^{-1} \tilde{\psi}_m. \tag{E.1}$$

Mallat (2000) montre encore que si $\left\{\tilde{\psi}_m, m \in \mathcal{M}\right\}$ est une frame dont les vecteurs sont linéairement dépendants, alors $\boldsymbol{ImT}$ est strictement inclus dans $l^2(\mathcal{M})$ et $\boldsymbol{T}$ possède une infinité d'inverses à gauche $\boldsymbol{T}^\dagger$ tels que $\forall f \in \mathcal{H}, \boldsymbol{T}^\dagger\boldsymbol{T}f = f$. Dans ce cas, il est possible de définir le pseudo-inverse $\boldsymbol{T}^{-1}$ comme étant l'inverse à gauche qui est nul sur $\boldsymbol{ImT}^\perp$ *i.e.* tel que $\forall c \in \boldsymbol{ImT}^\perp \subset l^2(\mathcal{M}), \boldsymbol{T}^{-1}c = 0$. De plus l'expression du pseudo-inverse est donnée par

$$\boldsymbol{T}^{-1} = (\boldsymbol{T}^*\boldsymbol{T})^{-1}\boldsymbol{T}^*, \tag{E.2}$$

(matrice inverse de Moore-Penrose) ; c'est également l'inverse à gauche de norme $|||\cdot|||$ minimum, avec $|||\cdot|||$ la norme subordonnée aux norme de Hilbert pour l'espace de départ et d'arrivée. Ainsi, pour tout inverse à gauche $\boldsymbol{T}^\dagger$ on a $|||\boldsymbol{T}^{-1}||| \leq |||\boldsymbol{T}^\dagger|||$. Il est à noter que $\boldsymbol{ImT} \subsetneq l^2(\mathcal{M})$ implique que toute liste $c \in l^2(\mathcal{M})$ n'est pas la transformée par $\boldsymbol{T}$ d'un signal $f \in \mathcal{H}$. Pour obtenir une liste $\tilde{c}$ qui soit la transformée par $\boldsymbol{T}$ d'un signal $f \in \mathcal{H}$, il est possible de construire $\tilde{c}$ comme étant la projection orthogonale de c sur $\boldsymbol{ImT}$. Cette projection est calculée à l'aide du projecteur orthogonal sur $\boldsymbol{ImT}$, défini par $\boldsymbol{P_T} = \boldsymbol{T}\boldsymbol{T}^{-1}$, tel que $\tilde{c} = \boldsymbol{P_T}c$. Or $\boldsymbol{ImT}^\perp \subset \boldsymbol{KerT}^*$ donc en écrivant $c = \tilde{c} + \tilde{c}^\perp$ avec $\tilde{c}^\perp \in \boldsymbol{ImT}^\perp$ il vient $\boldsymbol{T}^*\tilde{c}^\perp = 0$ soit $\boldsymbol{T}^{-1}c = \boldsymbol{T}^{-1}\tilde{c}$. L'égalité de Pythagore $||c||^2 = ||\tilde{c}||^2 + ||\tilde{c}^\perp||^2$ implique finalement que $||\tilde{c}|| \leq ||c||$. Soit pour $\tilde{c} = \boldsymbol{T}f$ on a $\tilde{c} = Arg\ min\{||c|| \text{ tel que } c \in l^2(\mathcal{M}) \text{ et } \boldsymbol{T}^{-1}c = f\}$. Ceci s'interprète comme suit : de toutes les manières d'écrire f dans $l^2(\mathcal{M})$, $\tilde{c} = \boldsymbol{T}f$ est la plus économique (celle qui a le moins d'énergie).

A l'aide de ce formalisme matriciel, il est possible de reformuler le formalisme de l'hypothèse diagonale. Finalement, d'un point de vue matriciel, l'hypothèse diagonale s'écrit simplement

$$\boldsymbol{B}_d(x',x) = \boldsymbol{B} : \boldsymbol{\Phi}_{(x,x')} = \sum_{z,z'} \boldsymbol{B}(z',z)\boldsymbol{\Phi}_{(x',x)}(z',z), \tag{E.3}$$

où le produit doublement contracté de deux matrices $\boldsymbol{A}$ et $\boldsymbol{B}$, est défini par $\boldsymbol{A} : \boldsymbol{B} = Tr\,(\boldsymbol{AB})$. En particulier, cette expression peut être modifiée pour ne dépendre cette fois que de $\boldsymbol{\Phi}^T_{x',x}$ (ici l'exposant ^{T} désigne la transposition). En effet, dans le cas purement réel, $\boldsymbol{A} : \boldsymbol{B} = Tr\,(\boldsymbol{AB}) = \boldsymbol{A}^T : \boldsymbol{B}^T$, d'où $\boldsymbol{B} : \boldsymbol{\Phi}_{x,x'} = \boldsymbol{B}^T : \boldsymbol{\Phi}^T_{(x,x')}$. De la symétrie de la matrice $\boldsymbol{B}$, il vient alors

$$\boldsymbol{B}_d(x',x) = \boldsymbol{B} : \boldsymbol{\Phi}^T_{(x,x')}. \tag{E.4}$$

Donc finalement,

$$\boldsymbol{B}_d(x',x) = \boldsymbol{B} : \frac{1}{2}\left(\boldsymbol{\Phi} + \boldsymbol{\Phi}^T\right)_{(x,x')}. \tag{E.5}$$

De la définition de l'opérateur $\boldsymbol{T}$ il s'en déduit que $\tilde{\psi}_m = \boldsymbol{T^*}\boldsymbol{\delta}_m$ et $\tilde{\psi}_m = \boldsymbol{T^{-1}}\boldsymbol{\delta}_m$. D'où finalement

$$\begin{aligned}
\boldsymbol{\Phi}_{(x',x)}(z',z) &= \sum_m \tilde{\psi}^*_m(z)\tilde{\psi}_m(z')\psi_m(x')\psi^*_m(x), \\
&= \sum_m \left(\boldsymbol{T^*}\boldsymbol{\delta}_m\right)^*(z)\ \left(\boldsymbol{T^*}\boldsymbol{\delta}_m\right)(z')\ \left(\boldsymbol{T^{-1}}\boldsymbol{\delta}_m\right)(x')\ \left(\boldsymbol{T^{-1}}\boldsymbol{\delta}_m\right)^*(x), \\
&= \sum_m \left(\boldsymbol{T}\boldsymbol{\delta}_z\right)_m\ \left(\boldsymbol{T}\boldsymbol{\delta}_{z'}\right)^*_m\ \left(\boldsymbol{T^{-*}}\boldsymbol{\delta}_{x'}\right)^*_m\ \left(\boldsymbol{T^{-*}}\boldsymbol{\delta}_x\right)_m.
\end{aligned}$$

Cette dernière expression est particulièrement adaptée pour la calcul numérique des matrices de poids.

Illustration de l'hypothèse diagonale dans le plan

Pour finir cette sous-section, il est intéressant d'illustrer l'hypothèse diagonale dans le cas d'une frame très simple : celle introduite par Daubechie dans le cas du plan. Soit

$$\left\{\psi_m = e^{\iota m\frac{2\pi}{3}}, m \in [0,2]\right\},$$

cette famille constitue une frame du plan complexe (Daubechie 1998). La frame duale étant dans ce cas $\tilde{\psi}_m = \frac{2}{3}\psi_m$. L'opérateur de transformation est alors

$$\boldsymbol{T} = \sqrt{\frac{2}{3}}\begin{pmatrix} 1 & 0 \\ -1/2 & \sqrt{3}/2 \\ -1/2 & -\sqrt{3}/2 \end{pmatrix}, \tag{E.6}$$

Cette fois, le quasi-inverse correspond à la transposée de $\boldsymbol{T}$ soit $\boldsymbol{T}^{-1} = \boldsymbol{T}^T$ tel que $\boldsymbol{T}^{-1}\boldsymbol{T} = \boldsymbol{I}$. En revanche, $\boldsymbol{T}\boldsymbol{T}^{-1} \neq \boldsymbol{I}$. Considérons la matrice de corrélation $\boldsymbol{C}$ définie par

$$\boldsymbol{C} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}. \tag{E.7}$$

Cette matrice est représentée dans l'espace des coefficients de la frame par $\boldsymbol{C}_f = \boldsymbol{T}\boldsymbol{C}\boldsymbol{T}^T$ telle que

$$\boldsymbol{C}_f = \begin{pmatrix} 0.6666667 & 0.0130768 & -0.6797435 \\ 0.0130768 & 0.3202565 & -0.3333333 \\ -0.6797435 & -0.3333333 & 1.0130768 \end{pmatrix}. \tag{E.8}$$

Appliquée à cette matrice, l'hypothèse diagonale correspond à ne conserver que la diagonale $\boldsymbol{D}_f$ de $\boldsymbol{C}_f$. La matrice ainsi modélisée est alors $\boldsymbol{C}_d = \boldsymbol{T}^{-1}\boldsymbol{D}_f\boldsymbol{T}^{-T}$ dont les coefficients sont

$$\boldsymbol{C}_d = \begin{pmatrix} 0.6666667 & 0.2 \\ 0.2 & 0.6666667 \end{pmatrix}, \tag{E.9}$$

qui n'est pas une matrice de corrélation, puisque la diagonale n'est pas l'identité. Cette matrice est également obtenue à l'aide des produits doublement contractés $\boldsymbol{C}_{di,j} = \boldsymbol{C} : \boldsymbol{\Phi}_{(i,j)}$ où les matrices $\left\{\boldsymbol{\Phi}_{(i,j)}\right\}_{(i,j)\in[1,4]^2}$ désignent les fonctions poids. En retournant dans l'espace des coefficients de frame, la matrice correspondante est $\boldsymbol{C}_{df} = \boldsymbol{T}\boldsymbol{C}_d\boldsymbol{T}^T$ dont les coefficients sont

$$\boldsymbol{C}_{df} = \begin{pmatrix} 0.4444444 & -0.1067522 & -0.3376923 \\ -0.1067522 & 0.3289744 & -0.2222222 \\ -0.3376923 & -0.2222222 & 0.5599145 \end{pmatrix}. \tag{E.10}$$

Il apparaît que cette matrice n'est pas diagonale. Ceci peut sembler surprenant puisqu'elle est issue de l'hypothèse diagonale dans l'espace des coefficients de la frame, et à la décomposition $\boldsymbol{C}_{df} = \boldsymbol{T}\boldsymbol{C}_d\boldsymbol{T}^T = \boldsymbol{T}\boldsymbol{T}^{-1}\boldsymbol{D}_f\left(\boldsymbol{T}\boldsymbol{T}^{-\mathbf{1}}\right)^T$. Cependant, dans cette décomposition, $\boldsymbol{T}\boldsymbol{T}^{-1}$ n'est pas l'identité, au contraire, c'est ici une matrice pleine

$$\boldsymbol{T}\boldsymbol{T}^{-1} = \begin{pmatrix} 0.6666667 & -0.3333333 & -0.3333333 \\ -0.3333333 & 0.6666667 & -0.3333333 \\ -0.3333333 & -0.3333333 & 0.6666667 \end{pmatrix}. \tag{E.11}$$

En résumé, l'*hypothèse diagonale dans la frame conduit à une matrice pleine dans le plan mais également dans l'espace de la frame*. La diagonale de cette matrice ne ressemble pas à la diagonale D_f initiale. Cet exemple montre qu'il est délicat d'agir directement sur les coefficients dans la frame. En effet, contrairement au cas d'une base orthogonale où les informations sont découplées, ici les couplages entre les coefficients induisent des liens impossibles à isoler. De plus, il apparaît que la matrice résultante $\boldsymbol{C}_d$ n'est pas une matrice de corrélation (la diagonale n'est pas homogène égale à 1). Il apparaît donc que la modélisation des corrélations par l'hypothèse diagonale dans une frame est complexe.

Modélisation des structures locales de covariance des erreurs de prévision à l'aide des ondelettes

Thèse de l'Université Toulouse III - Paul Sabatier
Discipline : Météorologie
Auteur : **Olivier PANNEKOUCKE**
Directeur de thèse : **Gérald DESROZIERS**, *Co-directeur de thèse :* **Loïk BERRE**

CNRM-GAME
42 avenue Coriolis, 31057 Toulouse cedex, France

Résumé

La représentation des variations spatio-temporelles des fonctions de covariance d'erreur d'ébauche reste un problème majeur dans les algorithmes d'assimilation. Dans cette thèse le diagnostic des variations géographiques des corrélations locales est introduit via le diagnostic de la portée locale. L'estimation de cette portée ainsi que les propriétés de l'estimation sont étudiés en détail. Ce travail utilise des ondelettes sphériques, suivant la formulation introduite par Mike Fisher (ECMWF), pour modéliser les fonctions de corrélation locale "du jour". Il est montré que cette formulation moyenne spatialement les corrélations locales, permettant de réduire le bruit d'échantillonnage. D'autre part, cette formulation ondelette fournit une estimation robuste même pour un petit ensemble. Elle est aussi capable de capturer la dynamique spatio-temporelle des corrélations, ceci est illustré à l'aide de la dynamique des portées locales du jour.

Mots clés : Assimilation de données, ensemble d'analyses perturbées, estimation des covariances d'erreur d'ébauche, dépendance à l'écoulement, longueur de portée, ondelette sphérique, modélisation ondelette des covariances d'erreur d'ébauche.

Local structure modeling of forecast error covariances using wavelets

Abstract

The spatio-temporal representation of background error covariances is one of the major problems in data assimilation algorithms. In this thesis, the diagnosis of geographical variations of the local correlation is introduced through the local length-scale diagnosis. The length-scale estimation and the properties of this estimation are studied in details. In this work spherical wavelets are used, according to the formulation introduced by Mike Fisher (ECMWF), in oder to model the local correlation functions "of the day". It is shown that this formulation offers a spatial average of the local correlation that reduces the sampling noise. Moreover, this wavelet formulation provides a robust estimation even for a small ensemble. This formulation is also able to catch the spatio-temporal dynamic of correlation, it is illustrated with the length-scale.

Keywords : data assimilation, ensemble of perturbed analysis, estimation of background error covariances, flow dependence, length-scale, spherical wavelets, wavelet model of background error covariances.