

Utilisation de statistiques d'erreur de prévision  
hétérogènes pour l'analyse du brouillard dans le  
modèle de prévision numérique du temps à fine  
échelle AROME

Benjamin Ménétrier

Stage encadré par Thibaut Montmerle (CNRM-GAME/GMAP)  
et Jean-François Mahfouf (CNRM-GAME/GMME/TURBAU)

Mars - Juin 2010

Stage effectué au Centre National de Recherches Météorologiques  
dans le Groupe de Modélisation pour l'Assimilation et la Prévision,  
pour le second semestre du M2 "Océan, Atmosphère, Climat et  
Observations Spatiales" (ENS-UPMC)



# Résumé

La prévision météorologique opérationnelle repose de plus en plus sur l'utilisation de modèles numériques. En raison de la nature chaotique du fluide atmosphérique, l'initialisation des modèles est d'une importance cruciale pour la qualité de la prévision. L'assimilation de données consiste à utiliser de manière optimale toutes les informations disponibles (prévisions antérieures, observations, relations d'équilibre, modèle numérique, ...) afin de produire l'analyse qui servira d'état initial pour les prévisions. Pour combiner ces différentes informations, on utilise une méthode variationnelle - décrite dans l'introduction - pour laquelle il est important de connaître les incertitudes des sources d'information. Or, à cause de la taille considérable des systèmes actuels, l'incertitude sur les prévisions à court terme utilisées comme ébauches pour l'assimilation ne peut pas être connue explicitement chaque fois que cela est nécessaire. La seule solution est donc de modéliser cette incertitude, désignée sous le nom de "statistiques d'erreur de prévision" ou plus simplement de "matrice  $\mathbf{B}$ ". Cette modélisation est abordée dans le chapitre 2, en particulier pour le modèle à fine échelle AROME de Météo-France. Une propriété fondamentale des statistiques d'erreur de prévision est leur hétérogénéité sur le domaine représenté. Si les modélisations opérationnelles récentes de la matrice  $\mathbf{B}$  possèdent cette propriété pour certains modèles globaux (ex : IFS-ARPEGE), ce n'est pas encore le cas pour le modèle à fine échelle AROME de Météo-France. L'objectif de cette étude est donc d'introduire dans les statistiques d'erreur de prévision une hétérogénéité rudimentaire, en testant une méthode déjà connue mais jamais implémentée, basée sur l'utilisation de masques géographiques. Dans chaque zone sélectionnée par une masque, une matrice  $\mathbf{B}$  homogène est appliquée, ce qui rend la matrice  $\mathbf{B}$  totale hétérogène (fin du chapitre 2). Comme le montre le chapitre 3 où sont estimées et analysées des statistiques d'erreur de prévision pour les cas de brouillard et de non-brouillard, l'hétérogénéité de la matrice  $\mathbf{B}$  devrait être très marquée dans certaines situations météorologiques. Pour estimer la validité de la méthode des masques géographiques, plusieurs expériences de réalisme croissant sont décrites et commentées dans le chapitre 4. Le problème du raccordement des statistiques aux frontières entre masques y est également soulevé. L'impact de l'introduction d'une hétérogénéité dans la matrice  $\mathbf{B}$  est visible sur les analyses, mais reste encore faible sur la qualité des prévisions pour les situations de brouillard. Ces conclusions ainsi que les nombreuses perspectives dans le domaine des statistiques d'erreur de prévision à fine échelle sont évoquées dans la dernière partie.

# Table des matières

<b>1</b>	<b>Introduction : l'assimilation de données par méthode variationnelle</b>	<b>4</b>
1.1	Notations usuelles . . . . .	4
1.2	Minimisation de la fonction coût . . . . .	5
1.3	Interpolation optimale . . . . .	5
<b>2</b>	<b>Modélisation des statistiques d'erreur de prévision pour l'assimilation de données</b>	<b>7</b>
2.1	Propriétés et modélisation des statistiques d'erreur de prévision . . .	7
2.1.1	Influence de la matrice $\mathbf{B}$ sur l'analyse . . . . .	7
2.1.2	Caractéristiques de la matrice $\mathbf{B}$ idéale . . . . .	8
2.1.3	Modélisation de la matrice $\mathbf{B}$ . . . . .	9
2.2	Assimilation de données dans le modèle AROME . . . . .	12
2.2.1	Caractéristiques principales du modèle . . . . .	12
2.2.2	Cycles d'assimilation . . . . .	12
2.2.3	Modèle de matrice $\mathbf{B}$ dans AROME . . . . .	14
2.3	Statistiques d'erreur de prévision hétérogènes . . . . .	16
2.3.1	Limites de l'hypothèse diagonale spectrale . . . . .	16
2.3.2	Utilisation spatialisée de statistiques d'erreur de prévision . .	16
2.3.3	Résultats escomptés . . . . .	17
<b>3</b>	<b>Estimation de statistiques d'erreur de prévision hétérogènes pour les situations de brouillard</b>	<b>18</b>
3.1	Calcul de la matrice $\mathbf{B}$ à partir d'un ensemble d'analyses perturbées .	18
3.1.1	Calibration de la matrice $\mathbf{B}$ . . . . .	18
3.1.2	Méthode ensembliste . . . . .	19
3.1.3	Génération d'un ensemble AROME . . . . .	19
3.2	Estimation des zones de brouillard . . . . .	20
3.2.1	Définition des critères de détection de brouillard . . . . .	20
3.2.2	Evaluation des critères avec le produit CARIBOU . . . . .	22
3.2.3	Résultats et choix du critère . . . . .	24
3.3	Statistiques d'erreur de prévision . . . . .	26
3.3.1	Division du domaine . . . . .	26
3.3.2	Ecart-types . . . . .	27
3.3.3	Variance expliquées . . . . .	28
3.3.4	Covariances croisées . . . . .	29
3.3.5	Corrélations verticales . . . . .	31

3.3.6	Conclusions sur les statistiques d'erreur de prévision . . . . .	32
<b>4</b>	<b>Utilisation de statistiques d'erreur de prévision hétérogènes pour l'analyse du brouillard</b>	<b>33</b>
4.1	Expériences avec une seule observation . . . . .	33
4.1.1	Description des expériences . . . . .	33
4.1.2	Incréments de température . . . . .	33
4.1.3	Incréments d'humidité spécifique . . . . .	35
4.2	Raccordement des statistiques aux frontières entre masques . . . . .	36
4.2.1	Problématique . . . . .	36
4.2.2	Impact du lissage, cas idéalisé . . . . .	37
4.2.3	Impact du lissage, cas réaliste . . . . .	38
4.3	Expériences réelles . . . . .	39
4.3.1	Cas non-cyclés . . . . .	40
4.3.2	Cas cyclés . . . . .	43
<b>5</b>	<b>Annexes</b>	<b>51</b>
5.1	Algorithmes variationnels . . . . .	51
5.1.1	Interpolation optimale . . . . .	51
5.1.2	Dimension temporelle . . . . .	51
5.1.3	Lissage bayésien . . . . .	54
5.2	Démonstrations de propriétés de la matrice <b>B</b> . . . . .	56
5.2.1	Existence de racines carrées . . . . .	56
5.2.2	Homogénéité et isotropie des fonctions de corrélation dans l'hypothèse diagonale spectrale . . . . .	56
5.3	Figures complémentaires . . . . .	59
5.3.1	Ecarts-types de la vorticité et de la divergence . . . . .	59
5.3.2	Longueurs de corrélation horizontale . . . . .	60
5.3.3	Incréments à 800 hPa . . . . .	61

# Chapitre 1

## Introduction : l'assimilation de données par méthode variationnelle

Cette partie récapitule les bases théoriques de l'assimilation de données par méthode variationnelle, couramment utilisée depuis les années 1990. La dérivation complète des résultats principaux et le détail exhaustif des propriétés de ces méthodes sont disponibles dans bon nombre d'ouvrages comme [Bouttier et Courtier, 1999], [Kalnay, 2003] ou [Bocquet, 2009], mais ne sont pas l'objet de la présente étude.

### 1.1 Notations usuelles

Dans les modèles numériques de prévision du temps, les champs continus du modèle physique d'atmosphère sont discrétisés pour obtenir  $n = m \times N_g$  variables discrètes, où  $m$  représente le nombre de champs pronostiques dans le modèle et  $N_g$  la taille de la base utilisée pour discrétiser les champs. La discrétisation peut avoir lieu dans l'espace physique (valeurs aux points de grilles) ou dans l'espace spectral (coefficients spectraux). Certains centre météorologiques (Met Office, NCAR, CMC) ont choisi la première solution, d'autres (Météo-France, ECMWF, JMA, NCEP) la seconde. Les  $n$  variables discrète sont classées dans une matrice colonne appelée vecteur d'état et notée

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{pmatrix}, \quad (1.1)$$

où chaque  $\mathbf{x}_i$  est un vecteur colonne de  $N_g$  composantes représentant la discrétisation du champ d'indice  $i$ . Bien que la réalité physique, continue, ne puisse être représentée parfaitement par ce vecteur d'état discret, on suppose qu'il existe un vecteur d'état dit état vrai et noté  $\mathbf{x}^t$  qui représente au mieux la réalité. Le but de l'assimilation de données est de déterminer la meilleure approximation de l'état vrai appelée analyse et notée  $\mathbf{x}^a$ . Pour cela, on dispose généralement de deux sources d'information :

1. une estimation de l'état vrai appelée ébauche et notée  $\mathbf{x}^b$ , généralement fournie par prévision effectuée à une date antérieure,

2. des observations de l'état du système regroupées dans une matrice colonne de taille  $p$  notée  $\mathbf{y}^o$ .

L'ébauche n'est qu'une première approximation de l'état vrai, erronée, et on peut donc définir une erreur d'ébauche, aussi appelée erreur de prévision, donnée par  $\varepsilon^b = \mathbf{x}^b - \mathbf{x}^t$ . Cette erreur peut être décrite par une fonction de densité de probabilité (PDF) qui est inconnue. En raison de la complexité de cette PDF, on se contente de travailler avec ses deux premiers moments : l'espérance, que l'on suppose nulle (erreur sans biais), et les variances-covariances regroupées sous la forme d'une matrice de taille  $n \times n$  :  $\mathbf{B} = E(\varepsilon^b \varepsilon^{bT})$ .

De même, les observations sont entachées d'erreurs provenant du bruit instrumental ou des imperfections de l'opérateur d'observation  $H$ , qui permet de passer de l'espace des variables du modèle à celui des observations. On peut définir une erreur d'observation donnée par  $\varepsilon^o = \mathbf{y}^o - H(\mathbf{x}^t)$ . On suppose que la PDF de cette erreur possède aussi une espérance nulle, et on définit la matrice des ses variances-covariances de taille  $p \times p$  :  $\mathbf{R} = E(\varepsilon^o \varepsilon^{oT})$ .

Les matrices  $\mathbf{B}$  et  $\mathbf{R}$  sont cruciales pour déterminer l'analyse puisqu'elles caractérisent au premier ordre l'incertitude des données disponibles, ce qui permet de combiner ces informations de manière optimale.

## 1.2 Minimisation de la fonction coût

Les méthodes variationnelles en assimilation de données reposent sur la théorie de l'optimisation. Il s'agit de minimiser une fonctionnelle appelée fonction coût qui mesure l'écart séparant l'état du système des contraintes que l'on souhaite imposer. Ces contraintes peuvent être de natures diverses : proximité par rapport à une ébauche ou à des observations, respect de certains équilibres, filtrage d'ondes de gravité, prise en compte d'erreurs de modèle ou de traitement des observations, etc... Exceptée la méthode de Newton-Raphson (coûteuse en raison de la taille des systèmes étudiés en géophysique), la seule méthode pour trouver le minimum de la fonction coût est de descendre pas à pas vers celui-ci à partir d'un point initial, comme indiqué sur la figure 1.1. Plusieurs méthodes existent pour déterminer la direction et le pas de descente, de sophistication (vitesse de convergence) et de coût de calcul variables : plus forte pente, gradient conjugué, quasi-Newton (BFGS)... Ces méthodes ne seront donc pas détaillées davantage mais peuvent être approfondies dans [\[Bocquet, 2009\]](#).

## 1.3 Interpolation optimale

En première approche, on cherche seulement une interpolation optimale entre une ébauche et des observations. La fonctionnelle peut donc se scinder en deux termes :

$$J(\mathbf{x}) = J^b(\mathbf{x}) + J^o(\mathbf{x}) \quad (1.2)$$

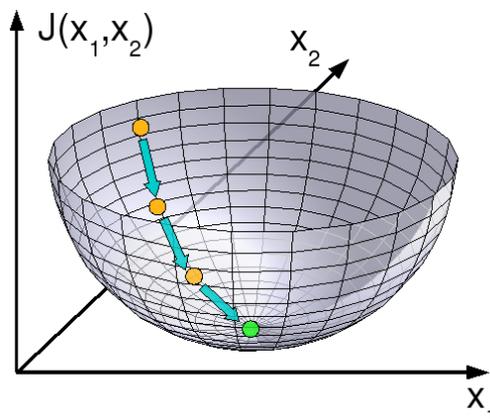


FIGURE 1.1 – Minimisation pas à pas d'une fonction coût dans un cas bidimensionnel quadratique

avec

$$J^b(\mathbf{x}) = \frac{1}{2} \|\mathbf{x} - \mathbf{x}^b\|_{\mathbf{B}^{-1}}^2 \quad (1.3)$$

$$J^o(\mathbf{x}) = \frac{1}{2} \|\mathbf{y}^o - H(\mathbf{x})\|_{\mathbf{R}^{-1}}^2 \quad (1.4)$$

Pour faciliter la minimisation de  $J$  qui nécessite le calcul de son gradient, on choisit pour les normes  $\|\cdot\|_{\mathbf{B}^{-1}}$  et  $\|\cdot\|_{\mathbf{R}^{-1}}$  des formes quadratiques issues de produits scalaires pondérés par les matrices  $\mathbf{B}^{-1}$  et  $\mathbf{R}^{-1}$ , qui sont symétriques définies positives par construction et représentent la "confiance" dans les données (inverse de l'erreur),

$$\|\mathbf{x} - \mathbf{x}^b\|_{\mathbf{B}^{-1}}^2 = (\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}^b) \quad (1.5)$$

$$\|\mathbf{y}^o - H(\mathbf{x})\|_{\mathbf{R}^{-1}}^2 = [\mathbf{y}^o - H(\mathbf{x})]^T \mathbf{R}^{-1} [\mathbf{y}^o - H(\mathbf{x})] \quad (1.6)$$

Le gradient de  $J(\mathbf{x})$  est alors donné par

$$\nabla J(\mathbf{x}) = \mathbf{B}^{-1} \mathbf{x} - \mathbf{H}^T \mathbf{R}^{-1} [\mathbf{y}^o - H(\mathbf{x})] \quad (1.7)$$

ce qui nécessite donc le codage de l'adjoint  $\mathbf{H}^T$  de l'opérateur d'observation linéarisé autour de  $\mathbf{x}$ . Cette approche heuristique peut être justifiée sous certaines hypothèses par une démarche probabiliste, donnée dans l'annexe 5.1.1. D'autre part, on a supposé dans ce paragraphe que les observations comme l'ébauche étaient disponibles à l'instant exacte de l'analyse, ce qui n'est pas le cas dans la réalité. La dimension temporelle peut être intégrée précisément dans la fonction coût (4D-VAR), ce qui génère des difficultés et des coûts de calculs importants. Plusieurs approximations de précisions moindres mais de coûts de calcul plus abordables (4D-Inc, 3D-FGAT, 3D-VAR et 3D-Inc) sont déduites dans l'annexe 5.1.2.

Alors que la matrice  $\mathbf{R}$  qui pondère le terme d'écart aux observations dans la fonction coût a une structure généralement simple (diagonale), la matrice  $\mathbf{B}$  qui pondère le terme d'écart à l'ébauche a une structure beaucoup plus riche. La suite de cette étude lui est entièrement consacrée.

# Chapitre 2

## Modélisation des statistiques d'erreur de prévision pour l'assimilation de données

### 2.1 Propriétés et modélisation des statistiques d'erreur de prévision

Les articles [\[Bannister, 2008a\]](#) et [\[Bannister, 2008b\]](#) sont une référence claire et complète concernant les propriétés de la matrice  $\mathbf{B}$  et les méthodes de mesure (première partie), ainsi que la modélisation de cette matrice (seconde partie). Les premiers chapitres de [\[Pannecoucke, 2008\]](#) apportent aussi une présentation simple des méthodes de modélisation de la matrice  $\mathbf{B}$ . Les éléments majeurs de ces ouvrages sont repris dans cette section.

#### 2.1.1 Influence de la matrice $\mathbf{B}$ sur l'analyse

Dans le cas où l'opérateur d'observation est linéaire, on peut annuler le gradient de la fonction coût pour obtenir une expression explicite de l'analyse. Celle-ci est équivalente au BLUE (Best Linear Unbiased Estimator), obtenu en minimisant la variance totale de l'analyse (trace de la matrice de covariance de l'analyse) sous les hypothèses de linéarité, d'absence de biais et de décorrélation des erreurs d'ébauche et d'observation. L'équation du BLUE est

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{y}^o - H(\mathbf{x})) \quad (2.1)$$

La correction apportée à l'ébauche  $\mathbf{y}$  est une fonction linéaire de l'innovation, donnée par le gain de Kalman :

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}, \quad (2.2)$$

L'information apportée par le vecteur d'innovation est donc diffusée horizontalement et verticalement, et transmise à tous les degrés de liberté du système par l'intermédiaire de la matrice de gain, où la matrice  $\mathbf{B}$  joue un rôle crucial. Comme elle est le dernier opérateur à agir dans le gain, elle définit l'espace vectoriel dans lequel peut

évoluer l'incrément d'analyse. En cela, elle joue un rôle de filtrage spatial et doit garantir que l'incrément d'analyse respecte les équilibres dominants de l'écoulement (géostrophisme, hydrostatisme). La matrice  $\mathbf{B}$  est donc d'une grande importance pour assurer la précision et la cohérence de l'analyse, et doit être déterminée avec beaucoup de soin. Malheureusement, plusieurs obstacles d'importance compliquent cette tâche.

1. Le problème principal est que dans la définition  $\mathbf{B} = E[(\mathbf{x}^b - \mathbf{x}^t)(\mathbf{x}^b - \mathbf{x}^t)^T]$ , l'état vrai est inconnu. Il est donc impossible de calculer  $\mathbf{B}$ , on ne peut qu'en estimer une approximation.
2. On peut tenter de calculer  $\mathbf{B}$  à partir d'estimations de l'erreur de prévision  $(\mathbf{x}^b - \mathbf{x}^t)$ , mais cela nécessite un grand nombre d'échantillons indépendants pour limiter le bruit et conserver le signal. De plus, au moins  $n$  estimations sont nécessaires pour assurer que  $\mathbf{B}$  soit de plein rang. C'est un point très important, car si  $\mathbf{B}$  n'est pas de plein rang, l'incrément d'analyse évolue dans un sous-espace de dimension réduite, ce qui limite la précision de la correction apportée à l'ébauche.
3. Dans les modèles numériques de prévision du temps actuels, la taille  $n$  du vecteur d'état est de l'ordre de  $10^7$ . La matrice  $\mathbf{B}$  possède donc  $n^2 \sim 10^{15}$  éléments, ce qui est considérable !

En pratique, il est donc impossible de calculer et de stocker tous les coefficients de la matrice  $\mathbf{B}$ , et il apparaît nécessaire de modéliser cette matrice. Afin de pouvoir juger de la pertinence de la modélisation développée, les caractéristiques principales de la matrice  $\mathbf{B}$  sont exposées dans la section suivante.

## 2.1.2 Caractéristiques de la matrice $\mathbf{B}$ idéale

### Aspect multivarié

Si le vecteur d'état  $\mathbf{x}$  contient  $m$  champs pronostiques  $\mathbf{x}_i$ , alors la matrice  $\mathbf{B}$  peut être divisée en  $m^2$  sous-matrices :

$$\mathbf{B} = \begin{pmatrix} AC[\mathbf{x}_1] & CC[\mathbf{x}_1, \mathbf{x}_2] & \dots & CC[\mathbf{x}_1, \mathbf{x}_m] \\ CC[\mathbf{x}_2, \mathbf{x}_1] & AC[\mathbf{x}_2] & \dots & CC[\mathbf{x}_2, \mathbf{x}_m] \\ \vdots & \vdots & \ddots & \vdots \\ CC[\mathbf{x}_m, \mathbf{x}_1] & CC[\mathbf{x}_m, \mathbf{x}_2] & \dots & AC[\mathbf{x}_m] \end{pmatrix} \quad (2.3)$$

où les sous-matrices  $AC$  sont des fonctions d'auto-covariance pour un champ donné (fonctions univariées) et les sous-matrices  $CC$  des fonctions de covariance croisée entre deux champs (fonctions multivariées). Ces dernières traduisent les interactions entre les champs physiques, et en particulier les équilibres dominants de l'écoulement atmosphérique. La modélisation de ces covariances croisées sera abordée dans la section 2.1.3, à l'aide de l'opérateur de balance. On s'intéresse donc désormais uniquement aux fonctions d'auto-covariance.

## Variance et fonction de corrélation

Par souci de simplicité, on supposera pour la suite de ce paragraphe que le modèle ne compte qu'une seule variable  $\mathbf{x}$ , et que la matrice  $\mathbf{B}$  est ainsi univariée. Pour la  $i^{\text{ème}}$  composante de  $\mathbf{x}$ , notée  $x_i$ , la dispersion statistique de l'erreur d'ébauche est quantifiée par la variance  $(\sigma_i^b)^2 = E[(x_i^b - x_i^t)^2]$ . Celle-ci correspond au  $i^{\text{ème}}$  coefficient diagonal de la matrice  $\mathbf{B}$ . En notant  $\mathbf{\Sigma}$  la matrice diagonale des écarts-types  $\sigma_i^b$ , on peut normaliser  $\mathbf{B}$  pour obtenir une matrice de corrélation donnée par  $\mathbf{C} = \mathbf{\Sigma}^{-1}\mathbf{B}\mathbf{\Sigma}^{-1}$ . La  $i^{\text{ème}}$  colonne de  $\mathbf{C}$  représente alors une fonction de corrélation variant entre 0 et 1 qui donne la structure spatiale associée au degré de liberté  $x_i$ . Plusieurs propriétés caractérisent les variances et les fonctions de corrélation :

- Les variances et les fonctions de corrélation sont hétérogènes : elles varient selon leur localisation dans le domaine.
- Les fonctions de corrélation sont anisotropes : elles ne possèdent pas de centre de symétrie.
- Les fonctions de corrélation sont non-séparables verticalement : leur structure verticale dépend de la localisation et de l'échelle horizontale.

## Dépendance à l'écoulement et aux observations

Pour être exacte, la matrice  $\mathbf{B}$  devrait varier en fonction de l'écoulement atmosphérique et des observations assimilées (nature, densité et précision). L'algorithme du filtre de Kalman (KF) fournit une expression explicite de cette actualisation (cf. [Bocquet, 2009]), donnant pour un modèle parfait :

$$\mathbf{B}_{k+1} = \mathbf{M}_k(\mathbf{I} - \mathbf{B}_k\mathbf{H}_k^T(\mathbf{H}_k\mathbf{B}_k\mathbf{H}_k^T + \mathbf{R}_k)^{-1}\mathbf{H}_k)\mathbf{B}_k\mathbf{M}_k^T \quad (2.4)$$

où l'indice  $k$  représente le numéro du cycle d'assimilation. On constate que la matrice  $\mathbf{B}$  n'est pas statique et qu'elle dépend des valeurs prises au cycle précédent par :

- elle-même ( $\mathbf{B}_k$ ),
- le modèle linéarisé ( $\mathbf{M}_k$ ),
- l'opérateur d'observation ( $\mathbf{H}_k$ ),
- l'incertitude sur les observations ( $\mathbf{R}_k$ ).

Etant donné la taille des systèmes opérationnels, le coût de calcul de l'équation (2.4) est prohibitif, et il est donc impossible de faire évoluer la matrice  $\mathbf{B}$  toute entière à chaque cycle d'assimilation. Cependant, la prise en compte des "variances du jour", c'est-à-dire l'utilisation d'une carte des variances  $\mathbf{\Sigma}$  mise à jour à chaque cycle est un premier pas - déjà opérationnel pour le modèle global ARPEGE - vers une dépendance complète à l'écoulement et aux observations [Raynaud *et al.*, 2009].

### 2.1.3 Modélisation de la matrice $\mathbf{B}$

#### Transformation de la variable de contrôle

On peut montrer qu'il existe une infinité de racines carrées notées  $\mathbf{B}^{1/2}$  telles que  $\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{T/2}$  (cf. annexe 5.2.1). Le principe de la transformation de la variable de contrôle (CVT) est de pré-conditionner la minimisation en choisissant l'une de ces

racines carrées pour renormaliser la variable de contrôle dans les formulations incrémentales. Ainsi, la matrice  $\mathbf{B}$  est remplacée par l'identité  $\mathbf{I}$  dans le terme d'ébauche de la fonction coût. Pour simplifier les notations, on utilise une fonction coût incrémentale générique sans multi-incrémentation ( $\mathbf{x}^g = \mathbf{x}^b$ ,  $\delta\mathbf{x}^b = 0$ ), qui peut être adaptée directement aux cas particuliers du 4D-Inc, du 3D-FGAT ou du 3D-Inc :

$$J(\delta\mathbf{x}) = \frac{1}{2}\|\delta\mathbf{x}\|_{\mathbf{B}^{-1}}^2 + \frac{1}{2}\|\mathbf{d} - \mathbf{H}\delta\mathbf{x}\|_{\mathbf{R}^{-1}}^2 \quad (2.5)$$

On pose alors  $\delta\mathbf{x} = \mathbf{B}^{1/2}\chi$ , la variable  $\chi$  étant implicitement incrémentale, même si elle n'est pas précédée d'un  $\delta$ . Cette transformation donne

$$J(\chi) = \frac{1}{2}\|\chi\|_{\mathbf{I}}^2 + \frac{1}{2}\|\mathbf{d} - \mathbf{H}\mathbf{B}^{1/2}\chi\|_{\mathbf{R}^{-1}}^2 \quad (2.6)$$

La matrice  $\mathbf{B}$  n'apparaît plus dans la fonction coût, mais sa racine carrée intervient maintenant dans le terme de rappel aux observations.

### Principe de la modélisation

La matrice  $\mathbf{B}^{1/2}$  ne pouvant être calculée explicitement, elle est approchée par un modèle noté  $\mathbf{B}_0^{1/2}$ . Cette matrice devra être telle que  $\delta\mathbf{x} = \mathbf{B}_0^{1/2}\chi$ , où  $\chi$  est la variable de contrôle dont les composantes sont statistiquement indépendantes (covariances nulles) et normées (variances unitaires). Sous ces hypothèses, il sera alors possible de comparer les caractéristiques connues de  $\mathbf{B}$  avec celles de la matrice induite par la modélisation :

$$\mathbf{B}^{\text{ic}} = E(\delta\mathbf{x}\delta\mathbf{x}^T) = \mathbf{B}_0^{1/2}E(\chi\chi^T)\mathbf{B}_0^{T/2} = \mathbf{B}_0^{1/2}\mathbf{B}_0^{T/2} \quad (2.7)$$

Dans le modèle  $\mathbf{B}_0^{1/2}$  retenu, la variable de contrôle  $\chi$  peut se décomposer en  $m'$  paramètres  $\chi_i$  de taille  $N_g$  :

$$\chi = \begin{pmatrix} \chi_1 \\ \vdots \\ \chi_{m'} \end{pmatrix}, \quad (2.8)$$

Il est à noter que le nombre  $m'$  de paramètres dans la variable de contrôle  $\chi$  peut être différent du nombre  $m$  de champs dans l'incrément  $\delta\mathbf{x}$ , car la matrice  $\mathbf{B}^{1/2}$  de taille  $(mN_g) \times (m'N_g)$  n'est pas nécessairement carrée. Plusieurs étapes sont généralement nécessaires pour modéliser  $\mathbf{B}_0^{1/2}$ , ce qui équivaut à une composition d'opérateurs. Typiquement, on a

$$\begin{pmatrix} \delta\mathbf{x}_1 \\ \vdots \\ \delta\mathbf{x}_m \end{pmatrix} = \mathbf{K}_p\mathbf{B}_S^{1/2} \begin{pmatrix} \chi_1 \\ \vdots \\ \chi_{m'} \end{pmatrix} \quad (2.9)$$

Cependant, le rôle des opérateurs est plus explicite pour la matrice  $\mathbf{B}_0^{-1/2}$  :

$$\begin{pmatrix} \chi_1 \\ \vdots \\ \chi_{m'} \end{pmatrix} = \mathbf{B}_S^{-1/2} \mathbf{K}_p^{-1} \begin{pmatrix} \delta \mathbf{x}_1 \\ \vdots \\ \delta \mathbf{x}_m \end{pmatrix} \quad (2.10)$$

La composition des deux opérateurs inversés doit ainsi décorrélérer et normaliser l'incrément  $\delta \mathbf{x}$  pour donner  $\chi$ .

### Opérateur de balance

L'opérateur de balance  $\mathbf{K}_p$  traduit le caractère multivarié (liant plusieurs champs) de la matrice  $\mathbf{B}_0^{-1/2}$ . Son inverse  $\mathbf{K}_p^{-1}$  transforme l'incrément  $\delta \mathbf{x}$  en un vecteur  $\tilde{\chi}$  de  $m'$  paramètres  $\tilde{\chi}_i$  discrétisés sur la même base que les  $\delta \mathbf{x}_i$ . Les paramètres  $\tilde{\chi}$  sont supposés décorrélés entre eux, ce qui signifie que la matrice de covariance de  $\tilde{\chi}$  ne contient plus que des sous-matrices d'auto-corrélation :

$$\begin{pmatrix} \tilde{\chi}_1 \\ \vdots \\ \tilde{\chi}_{m'} \end{pmatrix} = \mathbf{K}_p^{-1} \begin{pmatrix} \delta \mathbf{x}_1 \\ \vdots \\ \delta \mathbf{x}_m \end{pmatrix} \quad (2.11)$$

avec

$$E(\tilde{\chi} \tilde{\chi}^T) = \begin{pmatrix} \mathbf{B}_{S, \tilde{\chi}_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{B}_{S, \tilde{\chi}_{m'}} \end{pmatrix} \quad (2.12)$$

L'opérateur de balance est obtenu par des régressions statistiques sur un ensemble de différences de prévisions ou par des lois analytiques.

### Transformation spatiale

La transformation spatiale  $\mathbf{B}_S^{1/2}$  définit les variances et les structures de corrélation pour chaque paramètre  $\tilde{\chi}_i$ . Son inverse  $\mathbf{B}_S^{-1/2}$  normalise chaque paramètre  $\tilde{\chi}_i$ , puis le projette sur une base de modes spatiaux. Ces modes sont choisis de telle façon que les composantes des projections soient décorrélées au mieux.

$$\begin{pmatrix} \chi_1 \\ \vdots \\ \chi_{m'} \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{S, \tilde{\chi}_1}^{-1/2} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{B}_{S, \tilde{\chi}_{m'}}^{-1/2} \end{pmatrix} \begin{pmatrix} \tilde{\chi}_1 \\ \vdots \\ \tilde{\chi}_{m'} \end{pmatrix} \quad (2.13)$$

avec

$$E(\chi \chi^T) = \mathbf{I} \quad (2.14)$$

## Hypothèse diagonale spectrale

La matrice  $\mathbf{B}$  est formulée dans le même espace (physique ou spectral) que le vecteur d'état. Cependant, quelque soit l'espace choisi, de nombreux centres (Météo-France, ECMWF, MetOffice, ...) basent leur transformation spatiale sur l'hypothèse diagonale spectrale. Celle-ci suppose que les erreurs d'ébauche sont décorréliées dans la base des harmoniques (sphériques pour les modèles globaux ou bi-Fourier pour les modèles à aire limitée) pour la direction horizontale et dans une base de mode verticaux (EOF : fonction propres empiriques calculées préalablement) pour la direction verticale. Dans ce cas, on peut montrer que la matrice de corrélation déduite de la matrice induite  $\mathbf{B}^{\text{ic}}$  est homogène et isotrope (démonstration dans l'annexe 5.2.2 pour un cas unidimensionnel), et que les corrélations horizontales et verticales sont non-séparables.

## 2.2 Assimilation de données dans le modèle AROME

### 2.2.1 Caractéristiques principales du modèle

Le modèle AROME (Application de la Recherche à l'Opérationnel à Méso-Echelle, [Seity *et al.*, 2010]) est un modèle à aire limitée (1470 km x 1250 km) en projection Lambert centré sur la France, possédant une résolution d'environ 2.5 km et 60 niveaux verticaux (NB : les niveaux sont indexés par altitude décroissante). La figure 2.1 montre le domaine géographique couvert et la finesse du relief représenté par le modèle, tandis que la figure 2.2 donne l'altitude et la pression des niveaux-modèle.

Par un schéma numérique semi-implicite semi-lagrangien, il résout les équations non-hydrostatiques pour de nombreux champs pronostiques : les composantes horizontales de la vitesse ( $u, v$ ), la divergence verticale ( $\zeta$ ), la température  $T$ , la pression de surface  $P_s$ , l'écart  $P'$  à la pression hydrostatique donnée par  $(T, P_s)$ , l'humidité spécifique  $q$ , l'énergie cinétique turbulente ( $TK E$ ), ainsi que pour cinq espèces d'hydrométéores du schéma micro-physique : l'eau liquide nuageuse  $q_c$  et pluvieuse  $q_r$ , la glace nuageuse  $q_i$ , le graupel  $q_g$  et la neige  $q_s$ . Les champs sont représentés spectralement (bi-Fourier) avec une troncature elliptique, même si de nombreux calculs sont effectués en points de grille (paramétrisations physiques). Dans la configuration retenue pour cette étude, le modèle ARPEGE fournit les conditions aux limites du domaine AROME :

- conditions aux limites latérales,
- analyse de surface calculée par interpolation spatiale,
- relaxation vers ARPEGE dans les niveaux supérieurs d'AROME.

### 2.2.2 Cycles d'assimilation

Le modèle AROME constitue un saut de résolution important par rapport aux modèles globaux (passage de 10 km à 2.5 km de résolution). Un système d'assimilation de données spécifique a donc été développé à partir du schéma 3D-VAR du modèle ALADIN-FRANCE [Fischer *et al.*, 2005]. Pour être précises, les prévisions AROME doivent être initialisées avec des données récentes. En raison de la complexité et du

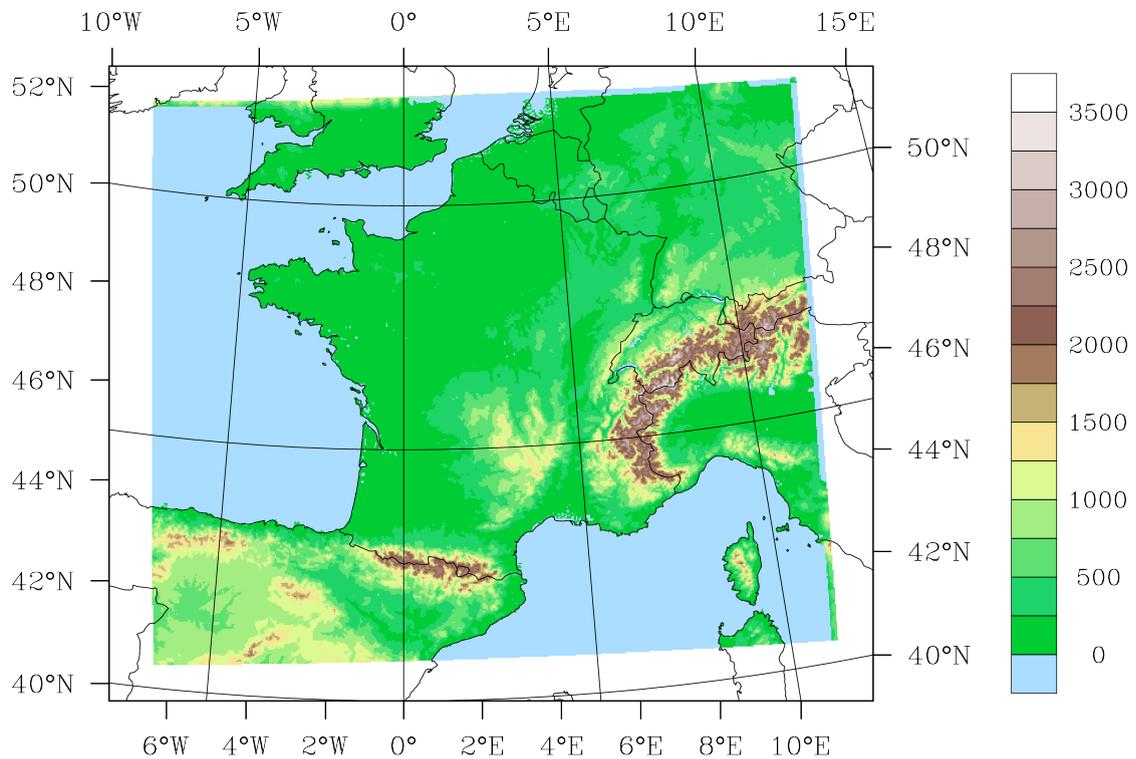


FIGURE 2.1 – Relief (altitude en mètres) du modèle AROME dans le domaine France XL. Les aberrations sur les bords correspondent à des zones "non-physiques", dédiées aux rappels vers les champs coupleurs.

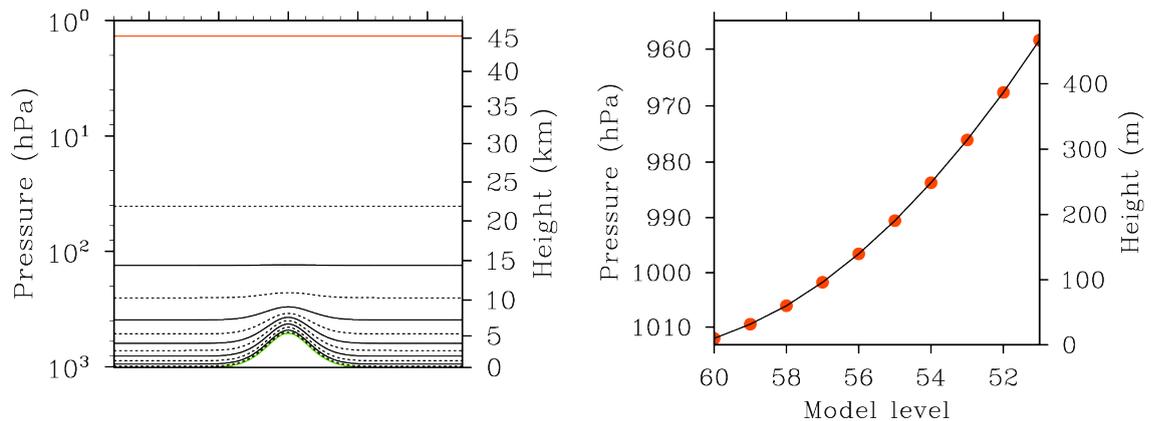


FIGURE 2.2 – Droite : niveaux pleins de la coordonnée verticale hybride au dessus d'un relief : surface (vert), niveaux multiples de 10 (trait plein) et de 5 (tireté), dernier niveau (rouge). Gauche : pression et altitude des 10 premiers niveaux pour une atmosphère standard.

coût d'un schéma 4D-VAR aux échelles résolues par AROME, le schéma 3D-VAR a été retenu, en espérant que son absence de dimension temporelle soit compensée par la fréquence des cycles d'assimilation (RUC : Rapid Update Cycle, [Brousseau](#)

*et al.*, 2008]). En plus des observations assimilées habituellement par les autres modèles, l'analyse d'altitude AROME traite des observations disponibles avec une haute résolution spatio-temporelle et considère en plus les réflectivités et vent Doppler issus du réseau de RADAR météorologiques national [Montmerle et Faccani, 2009]. L'analyse des champs de surface AROME est obtenue en interpolant l'analyse de surface ARPEGE. Les champs tridimensionnels analysés à la résolution du modèle sont :

- la vorticit    $\zeta$ ,
- la divergence  $\eta$ ,
- la masse repr  sent  e par le couple temp  rature/pression de surface  $(T, P_s)$ ,
- l'humidit   sp  cifique  $q$ .

Les autres champs tridimensionnels (TKE, champs non-hydrostatiques et micro-physiques) sont cycl  s    partir de l'  bauche. Dans la fonction co  t sous forme incr  mentale, la variable de contr  le est donc

$$\delta \mathbf{x} = \begin{pmatrix} \delta \zeta \\ \delta \eta \\ (\delta T, \delta P_s) \\ \delta q \end{pmatrix} \quad (2.15)$$

### 2.2.3 Mod  le de matrice B dans AROME

#### Op  rateur de balance

L'op  rateur de balance employ      M  t  o-France est le m  me que celui du ECMWF, d  crit dans [Derber et Bouttier, 1999], avec toutefois une sophistication suppl  mentaire dans le traitement de l'humidit   sp  cifique [Berre, 2000] :

$$\begin{pmatrix} \delta \zeta \\ \delta \eta \\ (\delta T, \delta P_s) \\ \delta q \end{pmatrix} = \begin{pmatrix} \mathbf{I} & 0 & 0 & 0 \\ \mathcal{MH} & \mathbf{I} & 0 & 0 \\ \mathcal{NH} & \mathcal{P} & \mathbf{I} & 0 \\ \mathcal{QH} & \mathcal{R} & \mathcal{S} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \delta \tilde{\zeta} \\ \delta \tilde{\eta}_u \\ (\delta \tilde{T}, \delta \tilde{P}_s)_u \\ \delta \tilde{q}_u \end{pmatrix} \quad (2.16)$$

On suppose que les erreurs associ  es aux champs   quilibr  s (modes de Rossby) sont en grande partie d  corr  l  es des erreurs associ  es aux champs non-  quilibr  s (modes d'inertie-gravit  , d  not  s avec un tilde). Plusieurs hypoth  ses sous-tendent cette formulation :

- Par ajustement g  ostrophique, la vorticit    $\delta \zeta$  est en grande partie   quilibr  e,
- L'op  rateur  $\mathcal{H}$  provient d'une   quation d'  quilibre g  ostrophique lin  aire (LBE) qui donne le champ de "masse"  $P_b$  en   quilibre avec la vorticit    $\delta \zeta$ .
- La divergence   quilibr  e est reli  e    ce champ de pression par l'op  rateur de regression  $\mathcal{M}$ , le r  sidu   tant la divergence non-  quilibr  e  $\delta \tilde{\eta}_u$ .
- De m  me, le champ de masse  $(\delta T, \delta P_s)$  a une partie   quilibr  e reli  e    la masse   quilibr  e par l'op  rateur de regression  $\mathcal{N}$ , et une partie non-  quilibr  e partiellement obtenue par  $\delta \tilde{\eta}_u$  via l'op  rateur de regression  $\mathcal{P}$ , donnant un r  sidu  $(\delta \tilde{T}, \delta \tilde{P}_s)_u$ .
- Pour les mod  les globaux (ARPEGE-IFS), l'incr  ment d'humidit   sp  cifique  $\delta \tilde{q}$  est d  coupl   des variables dynamiques ( $\mathcal{Q} = \mathcal{R} = \mathcal{S} = 0$ ). Cependant, l'op  rateur de balance a   t     tendu    M  t  o-France pour les mod  les    aire limit  e ALADIN et

AROME, en reliant  $\delta\tilde{q}$  à la pression équilibrée et aux composantes non-équilibrées  $\delta\tilde{\eta}_u$  et  $(\delta\tilde{T}, \delta\tilde{P}_s)_u$ , donnant un résidu  $\delta\tilde{q}_u$ .

### Transformation spatiale

La transformation spatiale des modèles de Météo-France et du ECMWF suit l'hypothèse diagonale spectrale. Ainsi, la sous-matrice relative à chaque paramètre  $\tilde{\chi}_i$  a pour expression :

$$\mathbf{B}_{s,\tilde{\chi}_i}^{1/2} = (\mathbf{S}^{-\text{T}}\mathbf{V}^{1/2}\mathbf{S}^{\text{T}})\mathbf{W}^{1/2}(\mathbf{E}\mathbf{D}^{1/2}) \quad (2.17)$$

Comme dans la section 2.1.3, l'effet de la transformation spatiale est plus clair si l'on considère son inverse, qui décorrèle chaque paramètre  $\tilde{\chi}_i$  en le projetant sur une base de modes spectraux et verticaux, tout en le normalisant :

$$\chi_i = (\mathbf{D}^{-1/2}\mathbf{E}^{\text{T}})\mathbf{W}^{-1/2}(\mathbf{S}^{-\text{T}}\mathbf{V}^{-1/2}\mathbf{S}^{\text{T}})\tilde{\chi}_i \quad (2.18)$$

Le paramètre  $\tilde{\chi}_i$  est représenté à chaque niveau-modèle sous forme spectrale, il subit donc une composition de six opérateurs :

1. L'opérateur  $\mathbf{S}^{\text{T}}$  est une transformation de Fourier inverse, niveau par niveau, qui passe le paramètre  $\tilde{\chi}_i$  dans l'espace physique.
2. La matrice diagonale  $\mathbf{V}^{-1/2}$  est une carte des inverses des écarts-types, qui normalise les fonctions de covariance horizontales dans l'espace physique.
3. L'opérateur  $\mathbf{S}^{-\text{T}}$  est une transformation de Fourier direct, niveau par niveau, qui repasse le paramètre dans l'espace spectral.
4. L'opérateur  $\mathbf{W}^{-\text{T}/2}$  décorrèle horizontalement le paramètre, niveau par niveau.
5. L'opérateur  $\mathbf{E}^{\text{T}}$  projette le paramètre sur une base de modes propres verticaux (EOF), différente pour chaque mode spectral horizontal.
6. Une fois projetés, les modes verticaux sont normalisés par  $\mathbf{D}^{-1/2}$ , afin qu'ils aient une variance unitaire.

La matrice induite par la modélisation est :

$$\mathbf{B}^{\text{ic}} = \mathbf{B}_0^{1/2}\mathbf{B}_0^{\text{T}/2} = \mathbf{K}_p\mathbf{B}_s^{\text{ic}}\mathbf{K}_p^{\text{T}} \quad (2.19)$$

où la partie spatiale pour un paramètre donnée se décompose en

$$\mathbf{B}_s^{\text{ic}} = \mathbf{S}^{-\text{T}}\{\mathbf{V}^{1/2}[\mathbf{S}^{\text{T}}\mathbf{W}^{1/2}\mathbf{E}\mathbf{D}\mathbf{E}^{\text{T}}\mathbf{W}^{-1/2}\mathbf{S}]\mathbf{V}^{1/2}\}\mathbf{S}^{-1} \quad (2.20)$$

Dans cette formulation, les matrices  $\mathbf{S}^{-\text{T}}$  et  $\mathbf{S}^{-1}$  situées en dehors des accolades servent à exprimer la matrice  $\mathbf{B}_s^{\text{ic}}$  dans l'espace spectral aux niveaux-modèle. L'intérieur des accolades représente donc une matrice de covariance univariée dans l'espace physique. On en reconnaît la forme, décrite dans la section 2.1.2 :  $\mathbf{B}_s^{\text{ic}} = \mathbf{S}^{-\text{T}}\{\mathbf{\Sigma}\mathbf{C}\mathbf{\Sigma}\}\mathbf{S}^{-1}$ , avec la matrice diagonale des écart-types  $\mathbf{\Sigma} = \mathbf{V}^{1/2}$  et la matrice de corrélation  $\mathbf{C}_s^{\text{ic}} = \mathbf{S}^{\text{T}}\mathbf{E}\mathbf{D}\mathbf{E}^{\text{T}}\mathbf{S}$ . L'hypothèse diagonale spectrale apparaît explicitement dans cette expression. Tous comme les relations utilisées dans l'opérateur de balance, ces matrices sont issues de statistiques effectuées sur des ensemble de différences de prévision. Les propriétés de cette matrice de corrélation sont celles décrites pour un cas unidimensionnel dans l'annexe 5.2.2 : homogénéité et isotropie. Ces propriétés ne correspondent pas à la réalité, et doivent être corrigées par une modélisation plus sophistiquée.

## 2.3 Statistiques d'erreur de prévision hétérogènes

### 2.3.1 Limites de l'hypothèse diagonale spectrale

La stationnarité, l'homogénéité et l'isotropie de la matrice  $\mathbf{B}^{\text{ic}}$  modélisée par la méthode sus-décrite sont des inconvénients majeurs. Plusieurs stratégies ont été envisagées pour pallier à ces défauts, et sont brièvement évoquées ici. La dépendance temporelle de la matrice  $\mathbf{B}$  peut ainsi être prise en compte par un opérateur de balance utilisant des relations analytique d'ordre deux linéarisées autour de l'ébauche [Fisher, 2003], ou par une transformation spatiale sur des grilles irrégulières [Desroziers, 1997]. D'autre part, l'hétérogénéité et l'anisotropie peuvent aussi être introduite par ces grilles irrégulières, ou plus généralement en abandonnant l'hypothèse diagonale spectrale pour d'autres techniques de transformation spatiale : formulation en ondelettes [Fisher, 2003], filtres récursifs [Purser *et al.*, 2003], opérateurs de diffusions [Weaver et Courtier, 2001],... L'utilisation totale ou partielle d'une estimation des statistiques d'erreur de prévision issue d'un ensemble de prévisions/analyses quotidien permettrait un meilleur traitement des structure spatiales et temporelles de la matrice  $\mathbf{B}$  ("erreurs du jour", hétérogénéité, anisotropie), mais le coût de ce type de méthode est très élevé, notamment pour les modèles d'échelle convective comme AROME. A défaut d'avoir un ensemble suffisant, il est alors nécessaire de procéder à un filtrage optimisé pour éliminer le bruit d'échantillonnage inhérent à un ensemble de petite taille. Une méthode plus simple est employée dans cette étude, elle a l'avantage d'un coût raisonnable et d'une application directe aux algorithmes opérationnels.

### 2.3.2 Utilisation spatialisée de statistiques d'erreur de prévision

Suivant une idée proposée par [Courtier *et al.*, 1998] mais qui n'a jamais été implémentée dans les systèmes opérationnels, on peut exprimer la matrice  $\mathbf{B}$  comme une combinaison linéaire de  $N$  termes :

$$\mathbf{B} = \sum_{i=1}^N \mathbf{F}_i^{1/2} \mathbf{B}_i \mathbf{F}_i^{T/2} \quad (2.21)$$

Chaque  $\mathbf{B}_i$  représente les statistiques d'erreur d'ébauche pour une situation météorologique donnée, et chaque  $\mathbf{F}_i$  définit le sous-domaine où la matrice  $\mathbf{B}_i$  est appliquée. En pratique,  $N$  masques géographiques bidimensionnels sont calculés avant l'analyse. Ces cartes binaires de 0 et de 1 sont ensuite transformées en matrices diagonales  $\mathbf{D}_i$ , où chaque élément est défini en fonction de la position horizontale correspondante dans la variable de contrôle. On doit s'assurer que la somme des  $\mathbf{D}_i$  donne bien la matrice identité. Comme les matrices  $\mathbf{B}_i$  sont définies dans l'espace spectral, il est nécessaire d'opérer des transformées de Fourier sur les  $\mathbf{D}_i$ , ce qui donne

$$\mathbf{F}_i^{1/2} = \mathbf{S} \mathbf{D}_i^{1/2} \mathbf{S}^{-1} \quad (2.22)$$

Une racine carrée de cette matrice  $\mathbf{B}$  est facilement obtenue car

$$\mathbf{B} = \mathbf{B}^{1/2} \mathbf{B}^{T/2} = \begin{pmatrix} \mathbf{F}_1^{1/2} \mathbf{B}_1^{1/2} & \dots & \mathbf{F}_N^{1/2} \mathbf{B}_N^{1/2} \end{pmatrix} \begin{pmatrix} \mathbf{B}_1^{T/2} \mathbf{F}_1^{T/2} \\ \vdots \\ \mathbf{B}_N^{T/2} \mathbf{F}_N^{T/2} \end{pmatrix} \quad (2.23)$$

Il faut noter que cette matrice  $\mathbf{B}^{1/2}$  est de taille  $n \times Nn$ . Cela signifie que la variable  $\chi$  dans la formulation CVT et le gradient de la fonction coût sont maintenant de taille  $Nn$  :

$$\delta \mathbf{x} = (\mathbf{F}_1^{1/2} \mathbf{B}_1^{1/2} \dots \mathbf{F}_N^{1/2} \mathbf{B}_N^{1/2}) \chi \quad (2.24)$$

$$J(\chi) = \frac{1}{2} \|\chi\|_{\mathbf{I}}^2 + \frac{1}{2} \|\mathbf{d} - \mathbf{H}(\mathbf{F}_1^{1/2} \mathbf{B}_1^{1/2} \dots \mathbf{F}_N^{1/2} \mathbf{B}_N^{1/2}) \chi\|_{\mathbf{R}^{-1}}^2 \quad (2.25)$$

$$\nabla J(\chi) = \chi - \begin{pmatrix} \mathbf{B}_1^{T/2} \mathbf{F}_1^{T/2} \\ \vdots \\ \mathbf{B}_N^{T/2} \mathbf{F}_N^{T/2} \end{pmatrix} \mathbf{H}^T \mathbf{R}^{-1} (\mathbf{d} - \mathbf{H}(\mathbf{F}_1^{1/2} \mathbf{B}_1^{1/2} \dots \mathbf{F}_N^{1/2} \mathbf{B}_N^{1/2}) \chi) \quad (2.26)$$

### 2.3.3 Résultats escomptés

L'avantage de cette méthode est de permettre une hétérogénéité des variances et des fonctions de corrélation dans les  $N$  sous-domaines géographiques, avec une dépendance de la géométrie de ces sous-domaines à la situation météorologique, puisque les masques  $\mathbf{F}_i$  sont déterminés à chaque minimisation, à partir d'observations, de l'ébauche, ou d'autres critères. Le prix à payer pour cette amélioration est l'allongement de la variable de contrôle et du gradient de la fonction coût, proportionnel au nombre de sous-domaines considérés. Cette méthode a déjà été employée avec succès dans [\[Montmerle et Berre, 2010\]](#) en distinguant les cas de convection précipitante et d'air clair dans le modèle à fine échelle AROME. L'objectif de cette étude est de conduire la même méthodologie en distinguant cette fois les zones de brouillard hivernal des zones de bonne visibilité.

# Chapitre 3

## Estimation de statistiques d'erreur de prévision hétérogènes pour les situations de brouillard

### 3.1 Calcul de la matrice $\mathbf{B}$ à partir d'un ensemble d'analyses perturbées

#### 3.1.1 Calibration de la matrice $\mathbf{B}$

Comme cela a été montré dans la section précédente, il est nécessaire de modéliser la matrice  $\mathbf{B}$  pour rendre possible son utilisation dans les systèmes variationnels d'assimilation de données. Cependant, une étape préalable dite de calibration doit être effectuée pour déterminer les éléments de cette modélisation. Dans la CVT retenue, il faut ainsi calculer :

- les opérateurs de régression  $\mathcal{M}$ ,  $\mathcal{N}$ ,  $\mathcal{P}$ ,  $\mathcal{Q}$ ,  $\mathcal{R}$  et  $\mathcal{S}$  de l'opérateur de balance,
- la carte des variances dans l'espace physique  $\mathbf{V}$ ,
- les matrices  $\mathbf{D}$  et  $\mathbf{E}$ , qui sont respectivement les valeurs propres et les vecteurs propres de la matrice de covariance des erreurs verticales.

L'étape de calibration suppose que l'on connaisse une estimation de la matrice  $\mathbf{B}$ , afin d'en déduire les éléments de modélisation. Plusieurs approches ont été explorées pour calculer les variances et les covariances d'erreur de prévision, que [Fisher, 2003] regroupe en deux catégories qui consistent à

1. extraire l'information sur les statistiques d'erreur de prévision à partir de l'information disponible (les vecteurs d'innovation),
2. construire des quantités dont les statistiques devraient être similaire à celles des erreurs de prévision.

Historiquement, deux méthodes se sont imposées : la méthode de [Hollingworth et Lonnerberg, 1986] dans la première catégorie, et la méthode NMC introduite par [Parrish et Derber, 1992] dans la seconde. N'étant pas utilisées dans cette étude, ces méthodes ne seront pas décrites ici. En effet, on leur a préféré une troisième méthode basée sur un ensemble d'analyses perturbées.

### 3.1.2 Méthode ensembliste

Le principe de cette méthode est résumé dans [Berre *et al.*, 2006] : l'utilisation d'une prévision perturbée comme ébauche pour l'assimilation d'observations perturbées fournit une analyse perturbée elle aussi, et cette analyse engendre une prévision perturbée lorsqu'elle est employée comme condition initiale par le modèle de prévision. Ainsi, en raison de la croissance des erreurs dans les cycles prévision/analyse successifs, l'erreur de prévision obtenue après un certain nombre de cycle n'est quasiment plus corrélée à la perturbation initiale qui a été introduite, mais elle est représentative de l'erreur de prévision "climatologique". Tandis que la perturbation des observations est explicite, celle de l'ébauche est implicite via le cyclage. La seconde idée importante est que l'on peut estimer la matrice  $\mathbf{B}$  à partir d'un ensemble de  $N$  prévisions simultanées  $(\mathbf{x}_i^b)_{1 \leq i \leq N}$ . Cet ensemble peut être obtenu en effectuant des cycles analyse/prévision simultanément pour  $N$  membres issus de champs perturbés indépendamment. Deux méthodes de calculs sont alors possibles :

1. On suppose que la moyenne de l'ensemble  $\langle \mathbf{x}_i^b \rangle$ , bien qu'elle ne représente pas un état réel, est représentative de l'état vrai. Les covariances d'erreur sont alors estimées en prenant les différences de chaque membre avec cette moyenne :

$$\mathbf{B} \simeq \langle (\mathbf{x}_i^b - \langle \mathbf{x}_i^b \rangle)(\mathbf{x}_i^b - \langle \mathbf{x}_i^b \rangle)^T \rangle \quad (3.1)$$

Par définition, l'ensemble des différences  $(\mathbf{x}_i^b - \langle \mathbf{x}_i^b \rangle)$  ne possède pas de biais.

2. On suppose que les erreurs de prévision des membres de l'ensemble sont décorréliées entre elles. Les covariances d'erreur sont alors estimées en prenant les différences entre deux membres de l'ensemble :

$$\begin{aligned} \langle (\mathbf{x}_i^b - \mathbf{x}_j^b)(\mathbf{x}_i^b - \mathbf{x}_j^b)^T \rangle &= \langle (\varepsilon_i^b - \varepsilon_j^b)(\varepsilon_i^b - \varepsilon_j^b)^T \rangle \\ &= \langle \varepsilon_i^b(\varepsilon_i^b)^T \rangle + \langle \varepsilon_j^b(\varepsilon_j^b)^T \rangle + 2\langle \varepsilon_i^b(\varepsilon_j^b)^T \rangle \\ &\simeq 2\mathbf{B} \end{aligned}$$

d'où

$$\mathbf{B} \simeq \frac{1}{2} \langle (\mathbf{x}_i^b - \mathbf{x}_j^b)(\mathbf{x}_i^b - \mathbf{x}_j^b)^T \rangle \quad (3.2)$$

Même s'il n'existe que  $N - 1$  paires  $(i, j)$  indépendantes, il est préférable d'effectuer les statistiques sur  $N$  paires circulaires  $(i, i + 1[N])$ , ce qui assure l'absence de biais dans l'ensemble des différences.

La seconde méthode sera retenue dans cette étude pour calibrer la matrice  $\mathbf{B}$ .

### 3.1.3 Génération d'un ensemble AROME

Le modèle AROME est initialisé à partir du modèle ARPEGE par adaptation dynamique (interpolation), et reste couplé à celui-ci pour obtenir les conditions aux limites du domaine. Grace au système de prévision d'ensemble à courte échéance basée sur ARPEGE (AEARP, [Desroziers *et al.*, 2008]), on peut obtenir un ensemble de conditions initiales et de champs coupleurs perturbés qui seront également utilisés dans les cycles prévision/analyse suivant. Dans l'expérience présente, le système AEARP possède six membres à un résolution moyenne (troncature T359, résolution

( $\sim 55$  km) et il assimile les observations perturbées par un schéma 3D-VAR FGAT. Parmi les réseaux disponibles pour AROME (RUC tous les 3h), celui de 6h semble le plus intéressant pour étudier le brouillard, car c'est généralement l'heure d'occurrence maximale [Remy, 2009]. Pour calibrer la matrice  $\mathbf{B}$  dans les cas de brouillard, on doit donc disposer de prévisions à 3h d'échéance issues du réseau de 3h. Cependant, pour obtenir des prévisions précises avec une variabilité de petite échelle suffisante au sein de l'ensemble, il est préférable d'initialiser AROME en début de nuit (18h) et d'effectuer ensuite trois cycles de prévision/analyse jusqu'à 3h avant de faire la prévision finale qui mène à 6h. Cette opération est répétée à plusieurs dates favorables sélectionnées pendant l'hiver 2009-2010, constituant un ensemble de 126 membres (21 dates  $\times$  6 membres).

## 3.2 Estimation des zones de brouillard

### 3.2.1 Définition des critères de détection de brouillard

Pour calibrer les matrices  $\mathbf{B}$  dans des situations de brouillard ou de non-brouillard à partir de l'ensemble de prévisions, il faut pouvoir distinguer les zones de brouillard à partir des sorties du modèle. La définition internationale du brouillard stipule que la visibilité horizontale doit être inférieure à 1000 m [NOAA, 2005]. Cependant, la visibilité n'est pas un champ du modèle, et il n'existe pas de définition consensuelle du brouillard en fonction de paramètres du modèle, celle-ci dépendant notamment des échelles résolues. Dans le modèle AROME, le schéma microphysique comporte des paramétrisations de la fraction nuageuse et du contenu en eau nuageuse dans la maille à partir :

- de la température potentielle,
- de la quantité d'eau non-précipitante,
- des variances de ces paramètres fournies par le schéma général de turbulence.

Plusieurs paramètres pour la détection du brouillard ont donc été envisagés.

#### Nébulosité basse

La nébulosité basse fournie par le modèle correspond au complémentaire de la fraction libre de nuages pour une colonne de hauteur donnée, en supposant une répartition aléatoire des nuages au sein de la maille. Si la fraction nuageuse au niveau  $k$  est notée  $CF_k$  et que le niveau maximum considéré dans le modèle a pour indice  $N_{mod}$ , la nébulosité basse (indice L pour "low") est donnée par

$$NEB_L = 1 - \prod_{k=60}^{N_{mod}} (1 - CF_k). \quad (3.3)$$

#### Nébulosité proche de la surface

Il est rapidement apparu que le calcul de cette nébulosité basse dans le modèle avait une extension verticale trop importante et englobait les stratus bas en plus des brouillards. La nébulosité proche de la surface (indice NS pour "near surface") a

donc été défini en utilisant la même méthode avec un nombre réduit de niveau  $N_{top}$  défini par l'utilisateur, donnant

$$\text{NEB}_{\text{NS}}(N_{top}) = 1 - \prod_{k=60}^{N_{top}} (1 - \text{CF}_k). \quad (3.4)$$

Il faut noter que cette nébulosité est une fonction croissante du nombre de niveaux considéré.

### Visibilité intégrée

La définition optique de la visibilité [Kunkel, 1984] est donnée par

$$\text{VIS} = \frac{\ln(\varepsilon)}{\beta}, \quad (3.5)$$

où  $\varepsilon$  est un seuil de contraste généralement fixé à 0.02 et  $\beta$  est le coefficient d'extinction. Comme le rapporte [Gultepe *et al.*, 2007], ce coefficient a souvent été modélisé en fonction du contenu en eau liquide  $LWC$  par des lois de puissance :

$$\beta = a(LWC)^b \quad (3.6)$$

avec des variations significatives pour les coefficients  $a$  et  $b$  selon les auteurs et les situations météorologiques concernées. Pour cette étude, on adopte les valeurs  $(a, b) = (144.7, 0.88)$  indiquées dans [Kunkel, 1984] et reprises dans le modèle unidimensionnel de brouillard COBEL. [Gultepe *et al.*, 2007] note qu'il serait plus réaliste de tenir compte de la concentration en gouttelettes  $N_D$ , mais ce champ doit aussi être paramétrisé dans AROME, ce qui limite la pertinence de cette approche. L'intégration verticale est obtenue en pondérant la visibilité  $\text{VIS}_k$  de chaque niveau par l'épaisseur en pression  $\Delta P_k$  de la couche, d'où la définition intégrée de la visibilité dans la couche de surface :

$$\text{VIS}_{\text{NS}}(N_{top}) = \left( \sum_{k=60}^{N_{top}} \Delta P_k \right)^{-1} \sum_{k=60}^{N_{top}} \text{VIS}_k \Delta P_k \quad (3.7)$$

### Humidité relative intégrée

Un dernier critère est basé sur l'humidité relative, qui n'est pas paramétrisée mais diagnostiquée à partir de  $(q, T, P)$  dans le modèle. La méthode d'intégration verticale est semblable à celle utilisée pour la visibilité, et en notant  $\text{RH}_k$  l'humidité relative au niveau  $k$ , l'humidité relative intégrée s'obtient par

$$\text{RH}_{\text{NS}}(N_{top}) = \left( \sum_{k=60}^{N_{top}} \Delta P_k \right)^{-1} \sum_{k=60}^{N_{top}} \text{RH}_k \Delta P_k \quad (3.8)$$

### 3.2.2 Evaluation des critères avec le produit CARIBOU

Pour définir le masque de brouillard, il faut choisir :

- un des paramètres physiques définis dans la section précédente,
- l'altitude d'intégration si nécessaire,
- un seuil de détection.

L'évaluation des différents cas possibles a été conduite avec pour référence le produit CARIBOU (Cartographie de l'Analyse du Risque de Brume et brOUillard, [Guidard et Tzanos, 2007]). Il s'agit d'une fusion de données issues des analyses de surface VARPACK (humidité et vent) et des observations satellitaires géostationnaires (classification nuageuse produite par le Centre de Météorologie Spatiale), RADAR (lame d'eau) et *in situ* (visibilité). Il faut garder à l'esprit que les analyses CARIBOU sont loin d'être une référence parfaite. Leur défaut principal, mis en évidence lors de contrôles de qualité, est l'extension trop importante des zones de risques moyen ou fort, biais dont il faudra tenir compte dans l'évaluation des critères. De plus, l'analyse CARIBOU ne distingue pas les brumes (visibilité inférieure à 5000 m) des brouillard (visibilité inférieure à 1000 m) et peut être faussée par la présence de nuages d'altitude. Un produit plus précis sur ce point est en cours de développement à Météo-France, mais n'était pas encore disponible au moment de cette étude. On considère donc que les cartes CARIBOU fournissent la meilleure information disponible à l'heure actuelle sur la localisation des zones de brouillard, même si sa fiabilité est parfois mise en défaut. L'algorithme de qualification du risque de CARIBOU est représenté de manière simplifiée sur la figure 3.1, accompagné d'un exemple de carte de risque. La méthodologie employée pour diagnostiquer du brouillard prévu par AROME est la suivante :

1. Pour un critère de détection donné, le masque de brouillard est calculé pour chacun des six membres de l'ensemble.
2. Si un point géographique appartient au masque pour au moins trois des six membres, alors il est retenu dans le masque global.
3. Le masque global est comparé aux zones de risque fort ou moyen des données CARIBOU.

On dénombre les points dans chacune des quatre configurations possibles pour établir la table de contingence 3.1.

	Brouillard selon AROME	Non-brouillard selon AROME
Brouillard selon CARIBOU	h (hits)	m (misses)
Non-brouillard selon CARIBOU	fa (false alarms)	cr (correct rejections)

TABLE 3.1 – Table de contingence calculée pour les différents critères

La corrélation entre les masques obtenus par chaque critère et les données CARIBOU est alors évaluée par quatre scores :

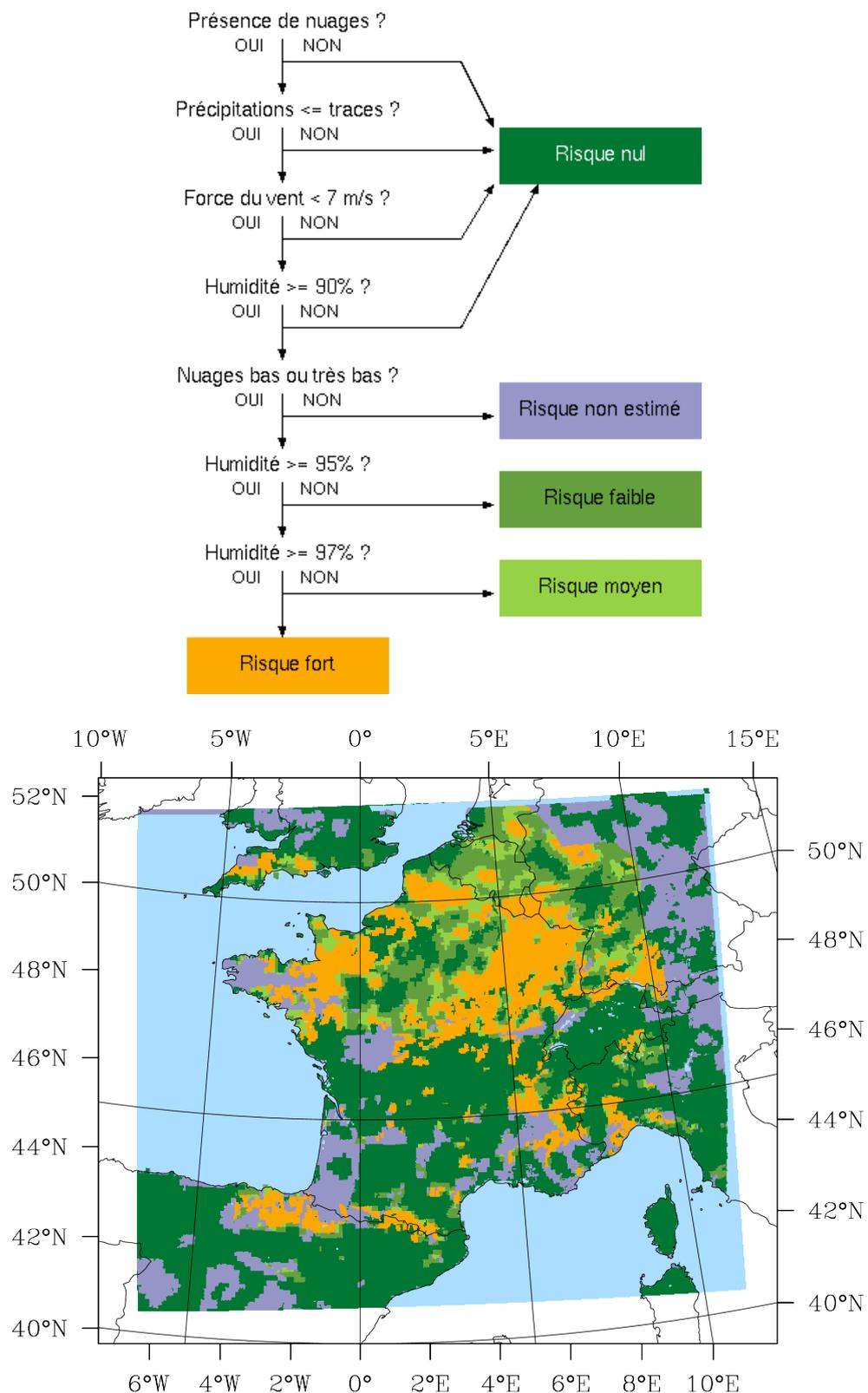


FIGURE 3.1 – Haut : algorithme de décision CARIBOU (version simplifiée); Bas : carte de risque de brume et brouillard brouillard CARIBOU pour le 01/10/2009 à 6h.

– La probabilité de détection :

$$\text{POD} = \frac{h}{h + m}, \quad (3.9)$$

à valeurs dans  $[0, 1]$ , et qui vaut 1 dans le meilleur des cas.

– Le taux de fausse alerte :

$$\text{FAR} = \frac{fa}{h + fa}, \quad (3.10)$$

à valeurs dans  $[0, 1]$ , et qui vaut 0 dans le meilleur des cas.

– L'indice de biais en fréquence :

$$\text{FBI} = \frac{h + fa}{h + m}, \quad (3.11)$$

à valeurs dans  $\mathbb{R}$ , et qui vaut 1 dans le meilleur des cas.

– La proportion de prévisions correctes :

$$\text{PC} = \frac{h + cr}{h + fa + m + cr}, \quad (3.12)$$

à valeurs dans  $[0, 1]$ , et qui vaut 1 dans le meilleur des cas.

### 3.2.3 Résultats et choix du critère

Tous les scores ont été analysés pour les différents critères, mais ne seront pas tous représentés ici. Seules les conclusions principales seront détaillées :

1. La nébulosité basse  $\text{NEB}_L$  est rejetée car elle conduit à un  $\text{FBI} \sim 2$ , ce qui est considérable sachant que les alertes CARIBOU sont déjà trop étendues. Bien que le  $\text{POD}$  et le  $\text{FAR}$  soient bons, le  $\text{PC}$  est très inférieur à ceux obtenus pour les autres critères.
2. La nébulosité proche de la surface  $\text{NEB}_{NS}$  présentent de très bons scores pour une intégration sur trois ou quatre niveaux. Ces scores sont très peu sensibles au seuil de détection choisi.
3. Les masques obtenus par le critère de visibilité  $\text{VIS}_{NS}$  sont trop peu étendus quelque soit l'épaisseur d'intégration, ce qui conduit à un  $\text{POD}$  trop faible et à un  $\text{FBI} \leq 1$ .
4. L'humidité relative  $\text{RH}_{NS}$  donne de bons résultats lorsque le seuil de détection est élevé ( $\sim 99\%$ ), mais présente une grande sensibilité au choix de ce seuil, ce qui n'est pas souhaitable.

Une illustration de ces conclusions est donnée par la figure 3.2 pour une date donnée.

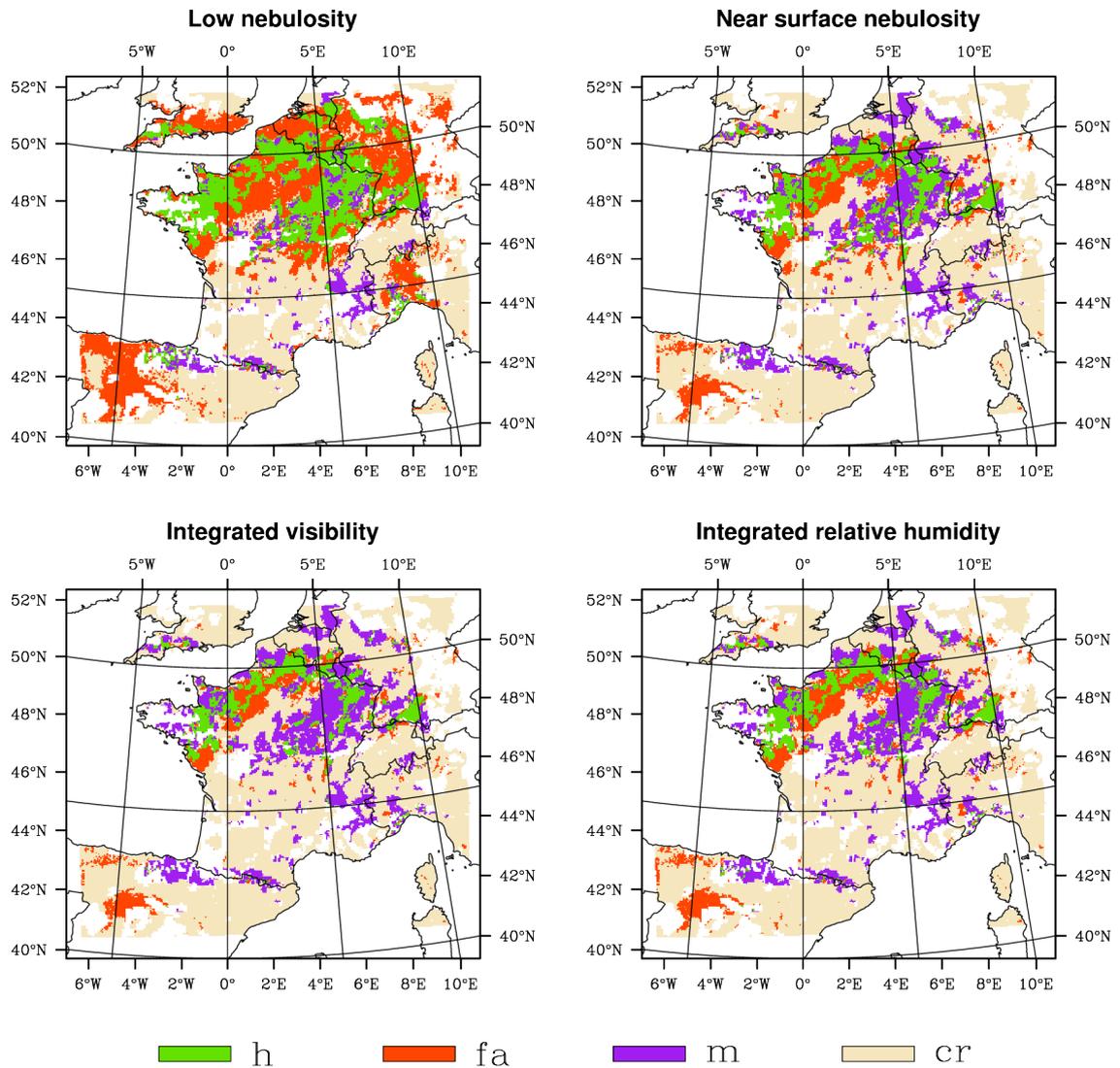


FIGURE 3.2 – Localisation des valeurs de la table de contingence pour différents critères : nébulosité basse (seuil = 0.9), nébulosité proche de la surface (seuil = 0.9), visibilité intégrée (seuil = 1000m) et humidité intégrée (seuil = 99%).

La représentation des scores sur un diagramme POD vs. FAR est aussi très instructive puisqu'elle permet en même temps d'évaluer le  $FBI = \frac{POD}{1-FAR}$ . Ainsi,  $POD = 1 - FAR$  pour  $FBI = 1$ , ce qui signifie que les points situés près de la diagonale descendante ont un biais en fréquence faible. La figure 3.3 donne les résultats à différents niveaux pour des seuils de détection optimaux. Compte tenu des arguments précédents et de cette figure, le critère d'appartenance au masque de brouillard retenu est le suivant :

$$NEB_{NS}(N_{top} = 58) \geq 0.9 \quad (3.13)$$

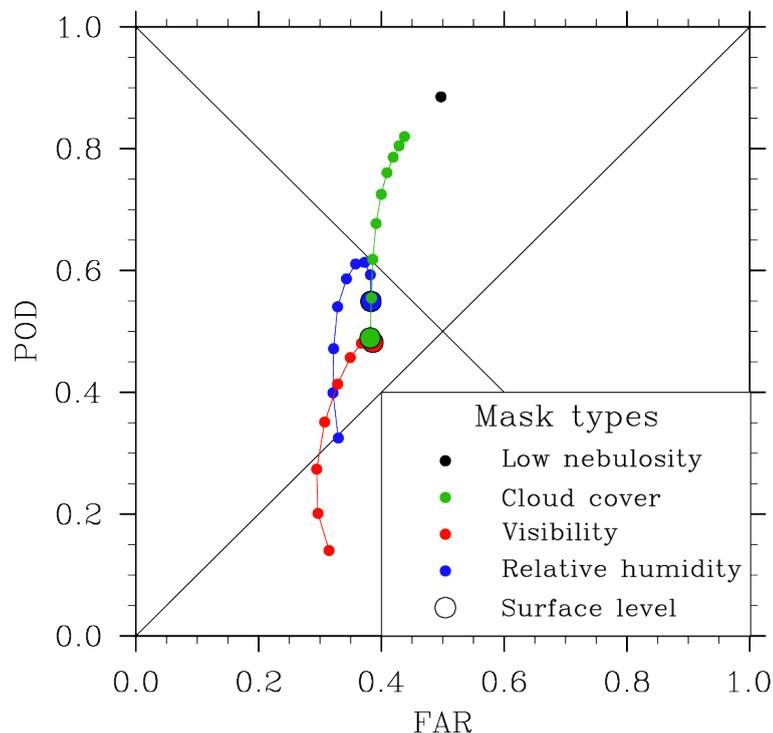


FIGURE 3.3 – Représentation des scores pour différents critères avec les mêmes seuils que pour la figure 3.2 : nébulosité basse (noir), nébulosité proche de la surface (vert), visibilité intégrée (bleu) et humidité intégrée (rouge). Les grand disques donnent les scores en surface, et les petits disques à des niveaux supérieurs.

### 3.3 Statistiques d'erreur de prévision

#### 3.3.1 Division du domaine

Pour chaque membre de l'ensemble, on définit un masque de brouillard et son complémentaire (non-brouillard), ce qui fait apparaître trois zones distinctes quand on compare deux membres  $i$  et  $j$  :

- zone de brouillard dans les deux membres, représentée par l'opérateur  $\delta_{ij}^{f,f}$
- zone de non-brouillard dans les deux membres, représentée par l'opérateur  $\delta_{ij}^{nf,nf}$
- zone de brouillard dans l'un des membres, non-brouillard dans l'autre, représentée par l'opérateur  $\delta_{ij}^{f,nf}$

Les opérateurs  $\delta_{ij}$  sont appliqués à la différence  $\varepsilon_{ij} = \mathbf{x}_i^b - \mathbf{x}_j^b$ , qu'il multiplie par 1 dans la zone qu'ils représentent et par 0 dans son complémentaire. Ainsi, le champ de différence peut être décomposé en trois termes :

$$\varepsilon_{ij} = \delta_{ij}^{f,f} \varepsilon_{ij} + \delta_{ij}^{nf,nf} \varepsilon_{ij} + \delta_{ij}^{f,nf} \varepsilon_{ij} \quad (3.14)$$

Un problème se pose au bord des masques, où les transitions entre 0 et 1 des opérateurs  $\delta_{ij}$  entraînent un raccourcissement des longueurs de corrélation. Pour limiter cet effet, on lisse les structures spatiales des différents termes en appliquant un flou gaussien  $\mathbf{G}$  aux opérateurs  $\delta_{ij}$ , effectué par convolution avec un noyau gaussien

bidimensionnel :

$$\varepsilon_{ij} \simeq [\mathbf{G} * \delta_{ij}^{f,f}] \varepsilon_{ij} + [\mathbf{G} * \delta_{ij}^{nf,nf}] \varepsilon_{ij} + [\mathbf{G} * \delta_{ij}^{f,nf}] \varepsilon_{ij} \quad (3.15)$$

Deux matrices  $\mathbf{B}$  sont alors calculées, une pour les situations de brouillard et une pour les situations de non-brouillard. Dans le calcul des statistiques pour chaque situation, on doit tenir compte de l'aire du masque considéré pour pallier aux effets de sous-échantillonnage. Les variances obtenues pour chaque situation ont donc subi une inflation par  $\mathcal{A}_i^{-1}$  où  $\mathcal{A}_i$  représente l'aire moyenne du masque d'indice  $i$ . Les sections suivantes présentent les points clés de la nouvelle calibration : carte des écarts-type, opérateur de balance à travers la variance expliquée et les covariances croisées, et enfin structures spatiales. Ceux-ci sont comparés à la version opérationnelle de la matrice  $\mathbf{B}$ , qui a été calibrée par la même méthode sur l'ensemble du domaine et sur un plus grand nombre de situations météorologiques.

### 3.3.2 Ecart-types

La figure 3.4 présente les moyennes spectrales des écarts-type d'erreur de prévision pour l'humidité spécifique et la température. Il apparaît clairement que dans les

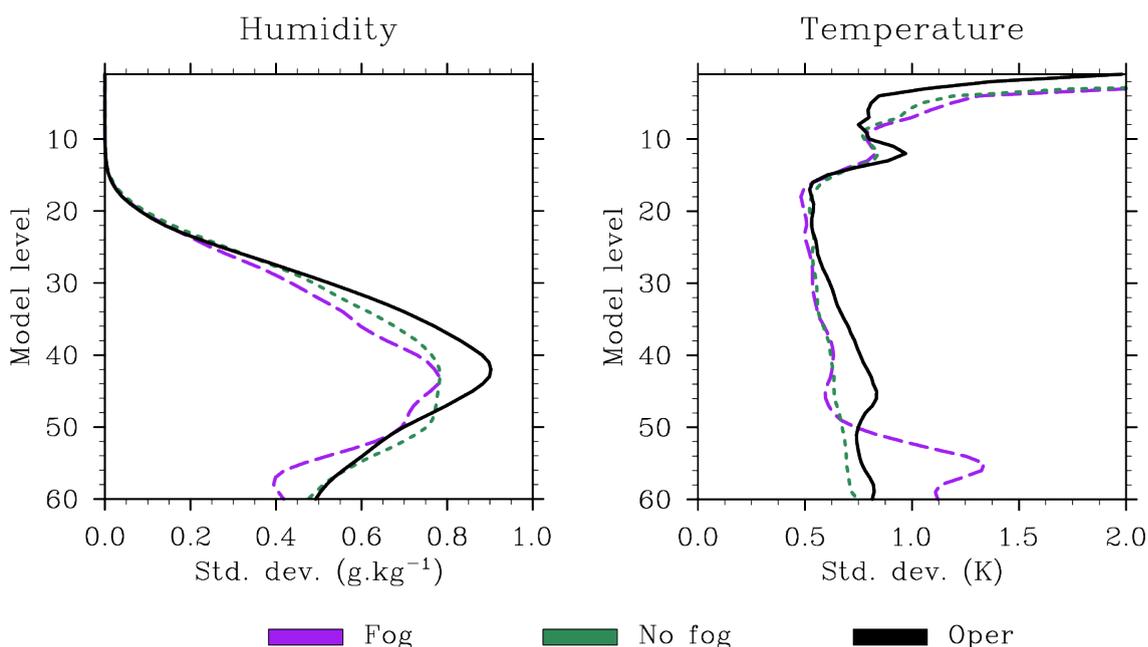


FIGURE 3.4 – Profils moyens des écarts-type d'erreur de prévision pour l'humidité spécifique (à gauche) et la température (à droite) pour les trois cas étudiés : situations de brouillard, de non-brouillard et version opérationnelle.

zones de brouillard, l'erreur est plus faible pour l'humidité mais plus importante pour la température. On distingue nettement les erreurs associées à la forte restratification au sommet de la couche de brouillard, qui a lieu en moyenne aux alentours du niveau 57 ( $z \sim 100$  m). Les profils de divergence et de vorticité (cf. annexe 5.3.1) semblent

indiquer que les zones de brouillard sont dynamiquement plus calmes (vents faibles), générant des erreurs plus faibles, mais des artefacts liés au sous-échantillonnage, déjà pris en compte par une inflation, ne sont pas à exclure pour ces variables.

### 3.3.3 Variance expliquées

Dans l'opérateur de balance utilisé dans le modèle AROME (cf. 2.2.3), l'incrément d'humidité spécifique est décomposé en trois termes équilibrés et un résidu non-équilibré :

$$\delta q = \mathcal{QH}\delta\tilde{\zeta} + \mathcal{R}\delta\tilde{\eta}_u + \mathcal{S}(\delta\tilde{T}, \delta\tilde{P}_s)_u + \delta\tilde{q}_u \quad (3.16)$$

Sachant que la matrice de covariance de chaque paramètre  $\tilde{\chi}_i$  est notée  $\mathbf{B}_{s,\tilde{\chi}_i}^{\text{ic}}$ , on peut exprimer la matrice d'auto-covariance pour  $q$  par :

$$\mathbf{B}_q^{\text{ic}} = \mathcal{QH}\mathbf{B}_{s,\tilde{\zeta}}^{\text{ic}}(\mathcal{QH})^T + \mathcal{R}\mathbf{B}_{s,\tilde{\eta}_u}^{\text{ic}}\mathcal{R}^T + \mathcal{S}\mathbf{B}_{s,(\tilde{T},\tilde{P}_s)_u}^{\text{ic}}\mathcal{S}^T + \mathbf{B}_{s,\tilde{q}_u}^{\text{ic}} \quad (3.17)$$

Ces quatre termes se retrouvent respectivement dans les variances  $(\sigma_q^b)^2$  de  $\mathbf{B}_q^{\text{ic}}$  :

$$(\sigma_q^b)^2 = (\sigma_{q,\tilde{\zeta}}^b)^2 + (\sigma_{q,\tilde{\eta}_u}^b)^2 + (\sigma_{q,(\tilde{T},\tilde{P}_s)_u}^b)^2 + (\sigma_{q,\tilde{q}_u}^b)^2 \quad (3.18)$$

La variance expliquée par chaque terme est alors simplement le rapport de ce terme au total. La figure 3.5 donne les moyennes spectrales des profils de variance expliquée pour l'humidité spécifique dans les trois cas étudiés. Lorsque le brouillard est présent,

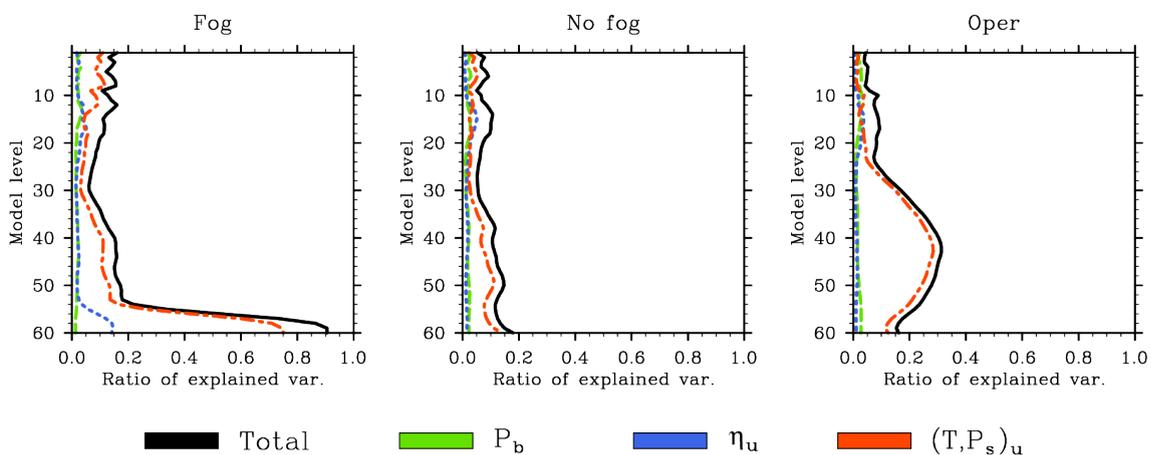


FIGURE 3.5 – Profils de rapport de variance expliquée pour l'humidité spécifique, dans les situations de brouillard (à gauche), de non-brouillard (au centre) et dans la version opérationnelle (à droite).

la variance de l'humidité spécifique est expliquée à 75% par le couple  $(\tilde{T}, \tilde{P}_s)_u$  près de la surface, ce qui traduit une très forte dépendance de l'erreur de prévision d'humidité à l'erreur de prévision de température. Cette dépendance s'explique naturellement par la saturation en vapeur d'eau du milieu qui corrèle fortement l'humidité à la température. La divergence non-équilibrée  $\tilde{\eta}_u$  semble aussi jouer un rôle dans la variance

de l'humidité pour les zones de brouillard, probablement par les mouvements verticaux qu'elle induit dans la couche saturée (raisonnement présenté dans la section suivante). La vorticit   n'a elle aucune influence sur l'humidit  , quelque soit la zone consid  r  e. Dans le cas op  rationnel, qui d  crit une plus grande vari  t   de situations m  t  orologiques, il semble qu'un lien important apparaissent entre l'humidit   sp  cifique et le couple  $(\tilde{T}, \tilde{P}_s)_u$  des niveaux 55    25. Cela pourrait correspondre aux cas d'instabilit   dynamique (passage de d  pressions), qui n'ont pas   t   repr  sent  s dans l'ensemble utilis   pour calibrer les statistiques de brouillard et de non-brouillard, o   seul des cas stable sont pr  sents.

Si l'on admet que la forte variance de  $q$  expliqu  e par  $(\tilde{T}, \tilde{P}_s)_u$  traduit la pr  sence de brouillard, il est int  ressant de tracer cette variance en fonction de l'altitude et de l'  chelle horizontale. La figure 3.6 indique ainsi que les structures de brouillard sont pr  sentes    des longueurs d'onde allant de  $\lambda \sim 8$  km     $\lambda \sim 500$  km, et que les brouillard de grande   chelle sont aussi les plus   pais. Leur   paisseur  $\Delta h$  varie en effet de  $\Delta h \sim 50$  m pour  $\lambda \sim 10$  m jusqu'    $\Delta h \sim 150$  m     $\lambda \sim 400$  km.

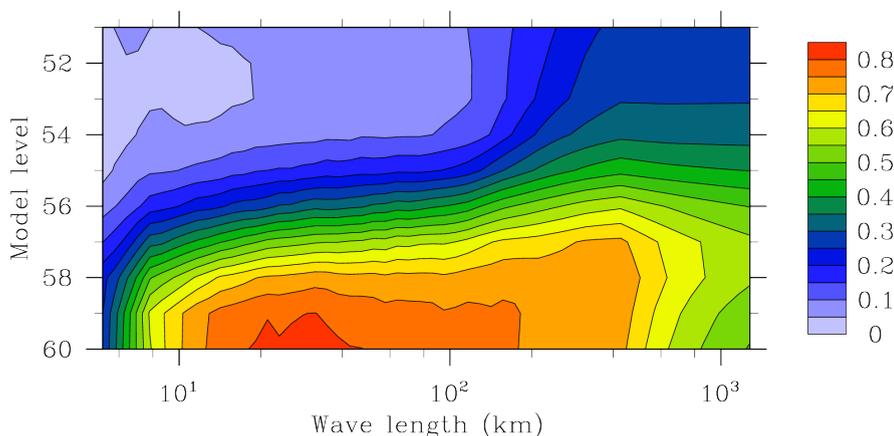


FIGURE 3.6 – Rapport de variance expliqu  e pour l'humidit   sp  cifique par le couple  $(\tilde{T}, \tilde{P}_s)_u$ , en fonction de l'altitude (10 premiers niveaux) et de la longueur d'onde, dans les situations de brouillard.

### 3.3.4 Covariances crois  es

On a vu que la variance des erreurs de pr  vision d'humidit   sp  cifique pouvait   tre expliqu  e en grande partie par les erreurs de temp  rature et de divergence dans les zones de brouillard. Cette d  pendance s'exprime par les covariances crois  es entre ces champs, donn  es dans les figures 3.7 et 3.8.

Sur la premi  re, une structure sym  trique positive appara  t pour les zones de brouillard sur les quatre premiers niveaux, ce qui signifie qu'un incr  ment positif de temp  rature dans cette couche y entraine un incr  ment positif d'humidit   sp  cifique (et vice versa). Ceci r  sulte directement de la saturation en vapeur d'eau qui les lie. La partie asym  trique du graphe indique elle que la temp  rature dans la couche de brouillard est corr  l  e statistiquement    l'humidit   sp  cifique dans toute la colonne,

alors que la température dans l'air au-dessus du brouillard n'influence pas l'humidité, car il n'y a ni saturation au-dessus du brouillard ni forçage dynamique possible à travers l'inversion. Pour les zones de non-brouillard et pour l'opérationnel, le lien entre température est beaucoup moins intense. Ces résultats sont globalement en accord avec ceux présentés dans [Remy, 2009], où des covariances verticales et des corrélations croisées sont estimées pour initialiser un filtre de Kalman-Bucy. La méthode NMC de [Parrish et Derber, 1992] est employée pour cela, en générant un ensemble dit "du pauvre" de prévisions d'origines différentes mais valables à une même échéance. Cependant, la comparaison des résultats est rendue difficile par la différence de résolution verticale :

- modèle 1D de couche limite COBEL : 30 niveaux entre 50 cm et 1360 m, dont 20 dans les 200 premiers mètres,
- premiers niveaux du modèle AROME : 18 niveaux entre 17 m et 1400 m, dont 6 dans les 200 premiers mètres

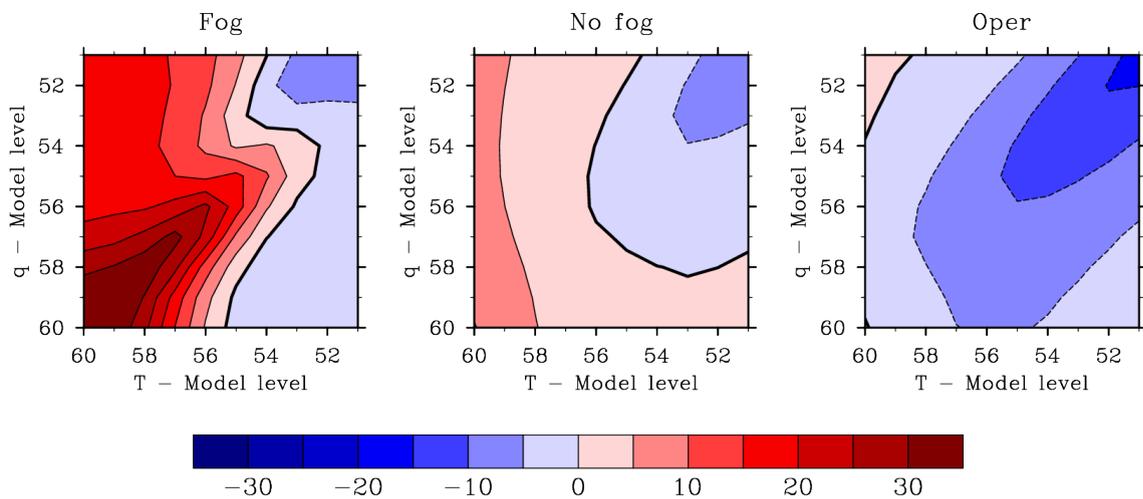


FIGURE 3.7 – Covariance croisée entre l'humidité spécifique (ordonnée) et le couple température/pression (abscisse) de surface sur les dix premiers niveaux, dans les zones de brouillard (à gauche) et de non-brouillard (à droite). Unité :  $10^5 \text{ kg.kg}^{-1}.\text{K}$ .

La figure 3.8 est aussi très explicite sur le lien entre divergence et humidité dans les zones de brouillard. Une convergence dans la couche de brouillard (altitude inférieure au niveau 57) entraîne une ascendance de l'air saturé, ce qui humidifie tout la colonne. Pour l'air situé au-dessus du brouillard, une convergence à un niveau donné n'a pas d'effet sur l'humidité à ce niveau, mais elle assèche les couches situées en-dessous et humidifie les couches situées au-dessus. Ce processus est compatible avec l'hypothèse d'une convergence équilibrée par des mouvements verticaux qui advectent la vapeur d'eau au sein d'une couche stratifiée en humidité (plus humide au sol, plus sèche en altitude). Pour les zones de non-brouillard, la convergence à n'importe quel niveau entraîne une humidification moindre de toute la colonne, sans qu'une raison claire ne se distingue.

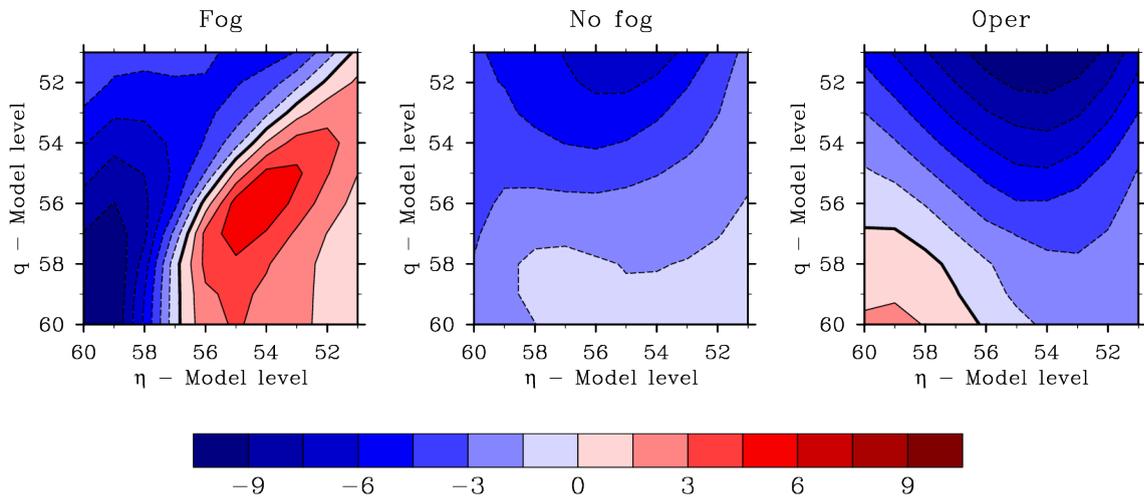


FIGURE 3.8 – Covariance croisée entre l’humidité spécifique (ordonnée) et la divergence (abscisse) sur les dix premiers niveaux, dans les zones de brouillard (à gauche) et de non-brouillard (à droite). Unité :  $10^9 \text{ kg.kg}^{-1}.\text{s}^{-1}$ .

### 3.3.5 Corrélations verticales

Les longueurs de corrélation horizontale sont quasiment semblables quelque soit le sous-domaine considéré, particulièrement dans les basses couches (cf. annexe 5.3.2). On s’intéresse donc plus particulièrement aux corrélations verticales dans les basses couches, qui doivent être différentes d’un domaine à l’autre. Comme le montre la figure 3.9, le raccourcissement attendu des longueurs de corrélation verticale est particulièrement net dans les zones de brouillard. Les erreurs de températures à un

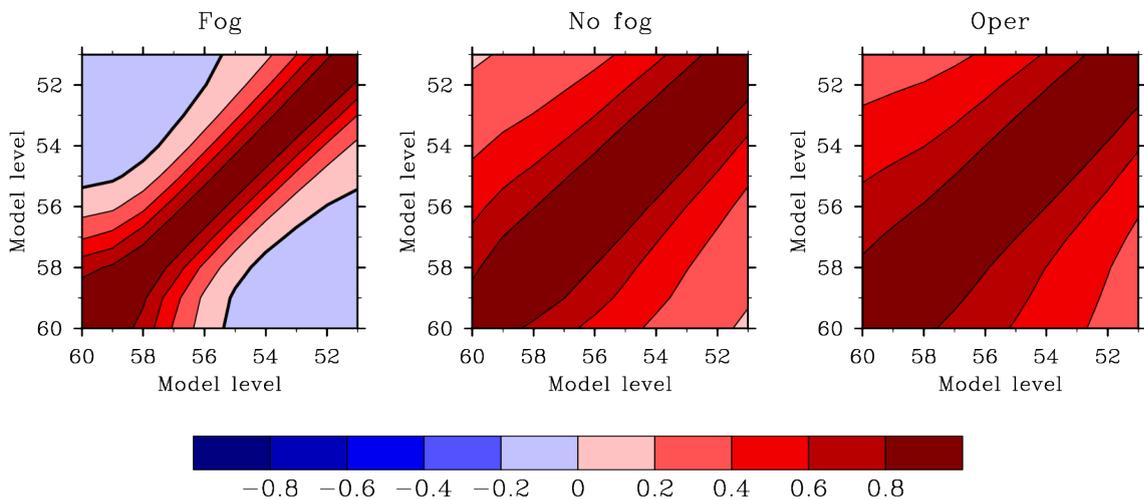


FIGURE 3.9 – Auto-corrélation de température sur les dix premiers niveaux, dans les zones de brouillard (à gauche) et de non-brouillard (à droite).

niveau  $k$  sont ainsi complètement décorréées de celles situées aux niveaux  $k \pm 4$  (différence d'altitude d'environ 150m). Ce resserement des corrélations verticales est aussi visible pour les autres champs (non-représentées ici), mais moins marqué. Ceci traduit une barrière dynamique forte au sommet de la couche de brouillard : une zone de forte stabilité statique sépare l'air situé près du sol qui est brassé par la turbulence (température potentielle  $\theta$  constante) et l'air situé au-dessus dont le profil est stable. Cette structure est bien visible dans les profils de température potentielle de [Remy, 2009], mais pas dans ses cartes de covariances verticales où l'échelle verticale logarithmique masque le raccourcissement des longueurs de corrélation verticale.

### 3.3.6 Conclusions sur les statistiques d'erreur de prévision

Les covariances croisées et les corrélations verticales ont été décrites dans les deux sections précédentes pour les basses couches seulement (dix premiers niveaux). En effet, c'est dans cette zone que se concentrent les différences entre les trois situations étudiées (brouillard, non-brouillard et opérationnel). Pour les niveaux supérieurs, les covariances croisées et les corrélations verticales sont significativement semblables quelque soit la situation.

Si on se concentre sur les niveaux inférieurs, il apparait clairement que les statistiques du cas opérationnel, appliquées à toutes les situations météorologiques sans distinction dans un formulation homogène de la matrice  $\mathbf{B}$ , sont erronées pour les cas de brouillard. L'assimilation d'observations acquises dans les basses couches en présence de brouillard conduit donc à une analyse sous-optimale, et cela explique pourquoi la température et l'humidité à 2 m ne sont pas assimilées en opérationnel dans le modèle ALADIN, et ont un impact variable sur l'analyse dans le modèle AROME. La formulation hétérogène de la matrice  $\mathbf{B}$  devrait permettre d'améliorer cet impact pour les cas de brouillard.

# Chapitre 4

## Utilisation de statistiques d'erreur de prévision hétérogènes pour l'analyse du brouillard

### 4.1 Expériences avec une seule observation

#### 4.1.1 Description des expériences

Pour évaluer l'impact sur l'analyse des différentes matrices  $\mathbf{B}$  (situations de brouillard, de non-brouillard et version opérationnelle), un premier test consiste à comparer les incréments d'analyse engendrés par une même et unique observation. Comme les modifications apportées à la matrice  $\mathbf{B}$  dans les situations de brouillard concernent les variances de la température, les covariances croisées humidité/température et humidité/divergence, ainsi que les longueurs de corrélation verticale dans les basses couches, on choisit d'assimiler des observations conventionnelles de température dans les expériences. Pour obtenir un signal clair, on introduit donc une innovation (différence de ébauche à l'observation) de 2 K à trois altitudes : 2 m, 950 hPa et 800 hPa, et on compare les incréments d'analyse obtenus.

Dans la formulation de la matrice  $\mathbf{B}$  suivant l'hypothèse diagonale spectrale, les fonctions de corrélation sont homogènes et isotropes en suivant les niveaux-modèle. Cependant les niveaux-modèle suivent le relief (cf. figure 2.2), et la figure 4.1 montre que l'incrément est fortement déformé en utilisant l'altitude comme coordonnée verticale. Pour faciliter l'analyse des résultats, on préférera donc examiner les incréments d'analyse en coupes verticales avec les niveaux-modèle comme coordonnée verticale.

#### 4.1.2 Incréments de température

Pour les innovations situées à 800 hPa (cf. annexe 5.3.3), seules des différences d'intensité et d'extension mineures apparaissent dans les incréments d'analyses. Des disparités bien plus significatives apparaissent entre les incréments d'analyse lorsque les innovations affectent les basses couches. Ainsi, comme le montre la figure 4.2, les longueurs de corrélation verticale sont plus courtes dans les zones de brouillard que

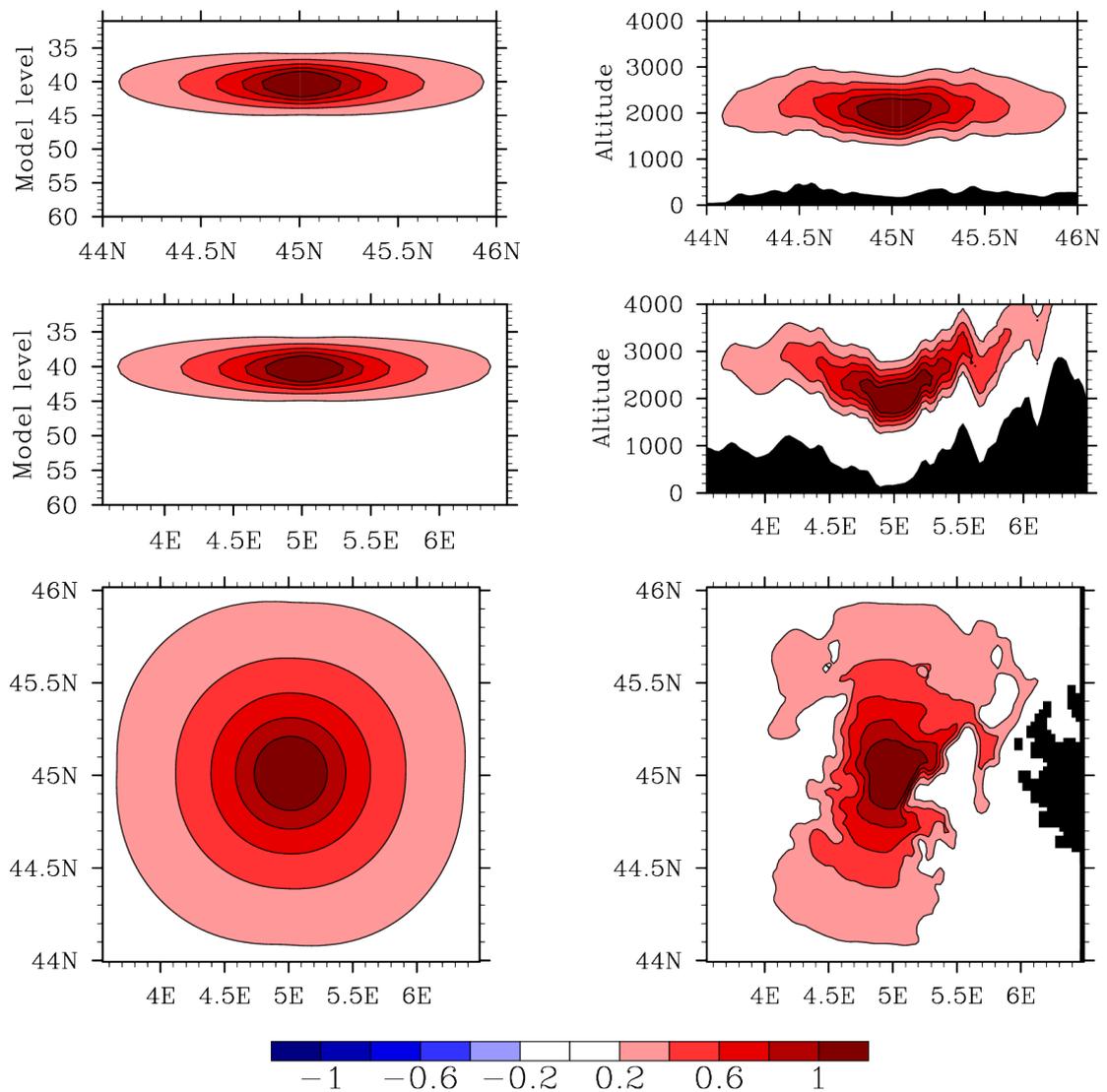


FIGURE 4.1 – Incrément d’analyse généré par une innovation de 2K à 800 hPa situés dans le sillon rhodanien. En haut : coupe latitude-niveau modèle, coupe longitude-niveau modèle et carte au niveau 40. En bas : coupe latitude-altitude, coupe longitude-altitude et carte à 2000 m.

dans les zones de non-brouillard ou dans la version opérationnelle. L’utilisation d’une formulation hétérogène de la matrice  $\mathbf{B}$  permet donc de restreindre la propagation de l’information apportée par l’innovation à travers le sommet de la couche de brouillard, où la forte restabilisation joue le rôle de barrière dynamique. L’incertitude de l’ébauche sur la température en surface est plus importante dans les situations de brouillard que de non-brouillard (cf. figure 3.4), ce qui explique l’impact plus fort de l’observation dans le premier cas que dans le second. Comme l’avaient indiqué les diagnostics effectués sur les nouvelles matrices  $\mathbf{B}$ , les longueurs typiques de corrélation horizontale varient peu d’une version à l’autre.

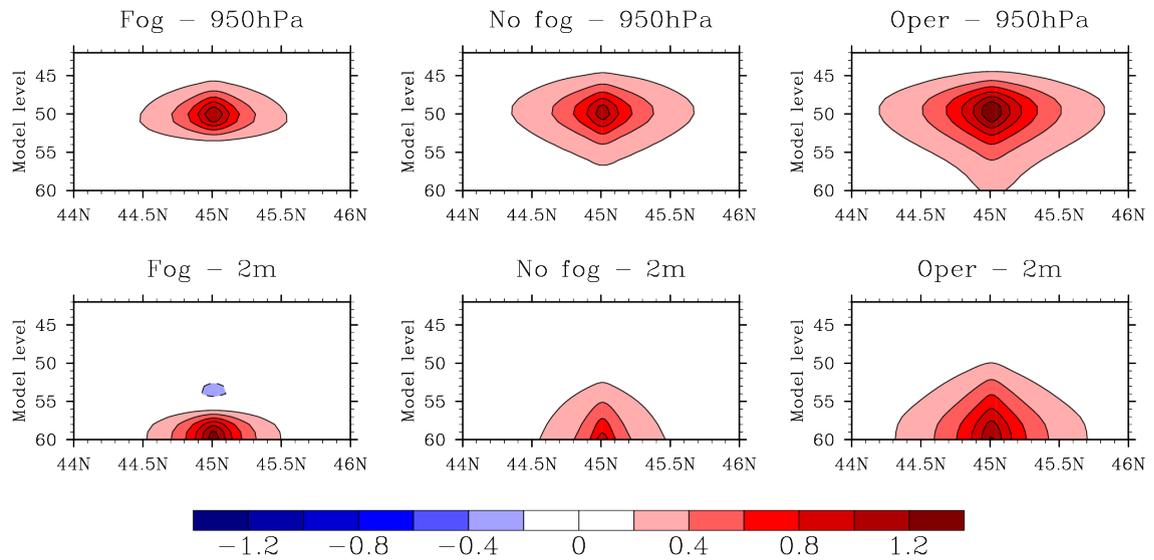


FIGURE 4.2 – Incrément d’analyse en température généré par une innovation de 2K à 950 hPa (en haut) ou à 2 m (en bas). De gauche à droite : situation de brouillard, de non-brouillard, et cas opérationnel

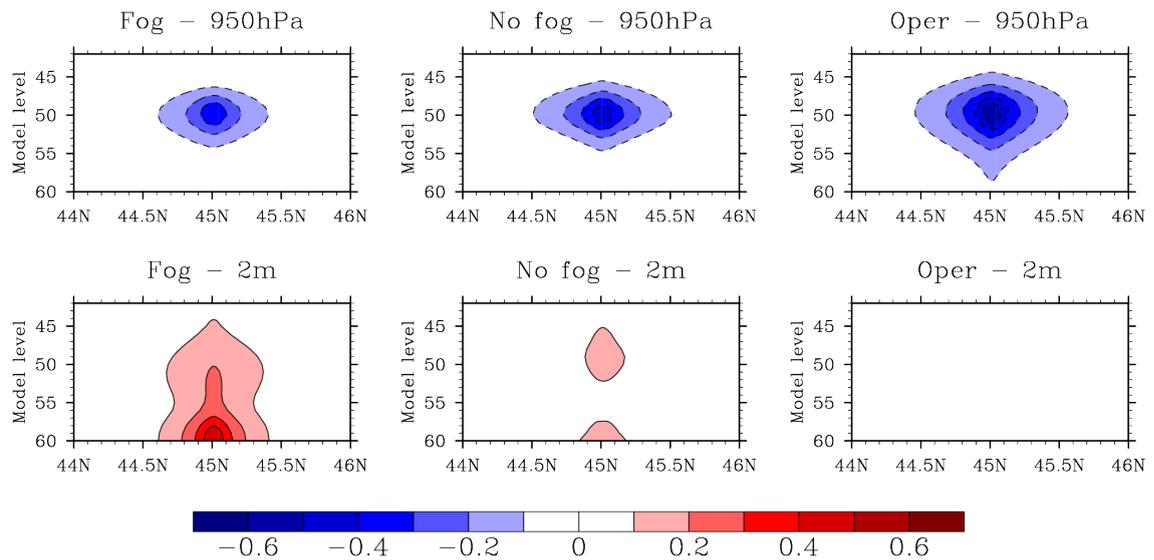


FIGURE 4.3 – Incrément d’analyse en humidité spécifique généré par une innovation de 2K à 950 hPa (en haut) ou à 2 m (en bas). De gauche à droite : situation de brouillard, de non-brouillard, et cas opérationnel

### 4.1.3 Incréments d’humidité spécifique

L’intensité des incréments d’humidité spécifique générés par les innovations de température, qui sont donnés par la figure 4.3, dépend du couplage entre les deux champs. Celui-ci est fourni par la variance expliquée donnée à la figure 3.5. La

structure verticale des incréments peut quant-à elle être directement interprétée en terme de covariance croisée entre l'humidité spécifique et la température, donnée par la figure 3.7. Les structures spatiales des incréments d'humidité spécifique et de température sont semblables pour les innovations de température à 950 hPa mais très différentes pour les innovations de température à 2 m. Alors que la barrière dynamique du sommet de la couche de brouillard est visible dans le premier cas, elle est complètement absente dans le second. En effet, une innovation de température à 2 m ne génère quasiment aucun incrément d'humidité spécifique dans les cas opérationnel et de non-brouillard (covariance nulle ou faible), mais provoque un incrément d'humidité spécifique important sur les dix premier niveaux dans les situations de brouillard (covariances importantes sur toutes la base de la colonne).

## 4.2 Raccordement des statistiques aux frontières entre masques

### 4.2.1 Problématique

Avec la formulation retenue en 4.3.2 pour appliquer les statistiques d'erreur hétérogènes, les covariances sont discontinues à la frontière entre deux masques. Afin d'éviter ces discontinuités et d'assurer la propagation de l'information issue d'une innovation entre les différents masques, il est nécessaire de lisser les contours des masques pour supprimer les sauts entre 0 et 1. La méthode retenue pour cette opération est la convolution par un noyau gaussien  $G$  dont la longueur caractéristique  $L$  doit être déterminée empiriquement. Pour un masque  $\delta(x, y)$  continu défini sur un domaine infini ou périodique, on aurait ainsi

$$[G * \delta](x, y) = \iint \delta(x, y) G(x - x', y - y') dx'dy' \quad (4.1)$$

avec

$$G(x, y) = \frac{1}{2\pi L^2} \exp\left(-\frac{x^2 + y^2}{2L^2}\right) \quad (4.2)$$

Deux propriété intéressantes de cette convolution sont la conservation de l'intégrale du masque (si celle-ci existe) :

$$\iint [G * \delta](x, y) dx dy = \iint \delta(x, y) dx dy \quad (4.3)$$

et l'encadrement des valeurs du masque :

$$\inf(\delta) \leq \inf(G * \delta) \leq \sup(G * \delta) \leq \sup(\delta) \quad (4.4)$$

Un problème se pose lorsque le domaine considéré est borné, ce qui est le cas du domaine AROME, puisqu'il faut choisir :

- la taille effective finie du noyau gaussien,
- la méthode d'application du noyau aux limites du domaine.

Dans cette étude, le noyau fini est défini sur un intervalle  $[-3L, 3L]$ , pour s'assurer que 99% de l'intégrale du noyau est conservée, et une version tronquée renormalisée de ce noyau est appliquée dans la zone d'extension  $3L$  aux limites du domaine. Bien que cela ne permette pas de conserver l'intégrale du masque (une déformation du noyau serait nécessaire), l'encadrement des valeurs du masque entre 0 et 1 est assurée et le lissage du masque binaire est efficace, comme le montre la figure 4.4.

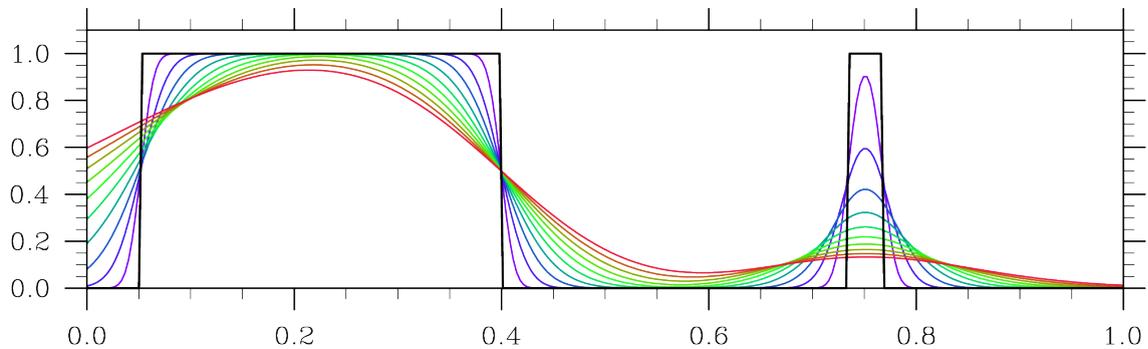


FIGURE 4.4 – Masque arbitraire (noir) et résultats de sa convolution par un noyau gaussien de longueur caractéristique variant linéairement de 0.01 à 0.1 (du violet au rose).

Dans le cas de figure considéré, on constate que la méthode utilisée aux limites du domaine tend à augmenter l'intégrale du masque (à gauche de la figure 4.4), notamment pour les noyaux gaussiens de grande longueur caractéristique. On remarque aussi que les signaux isolés comme le pic situé à droite sur la figure sont fortement amortis par ces mêmes noyaux.

#### 4.2.2 Impact du lissage, cas idéalisé

Pour étudier l'impact du lissage sur les variances et les écarts-types, une expérience idéalisée est mise en oeuvre. Un domaine unidimensionnel est divisé en deux sous-domaines où le masque vaut 0 ou 1. Les statistiques d'erreur sont homogènes et gaussiennes dans chaque sous-domaine, mais se distinguent d'un domaine à l'autre par la variation d'un seul paramètre : la variance ou la longueur de corrélation. La figure 4.5 montre les covariances obtenues à intervalles réguliers dans le domaine, alors que le masque est lissé par des noyaux gaussiens de longueur caractéristique variable.

Le lissage du masque semble peu influencer les variances, mais régularise les fonctions de corrélation à la frontière entre deux masques. Les valeurs du masque comprises entre 0 et 1 entraînent une combinaison des variances et des fonctions de corrélation des deux matrices  $\mathbf{B}$  concernées, ce qui ajoute l'anisotropie à l'hétérogénéité. Cette propriété est souhaitable (cf. 2.1.2), mais elle n'apparaît dans notre formulation qu'aux frontières de masques.

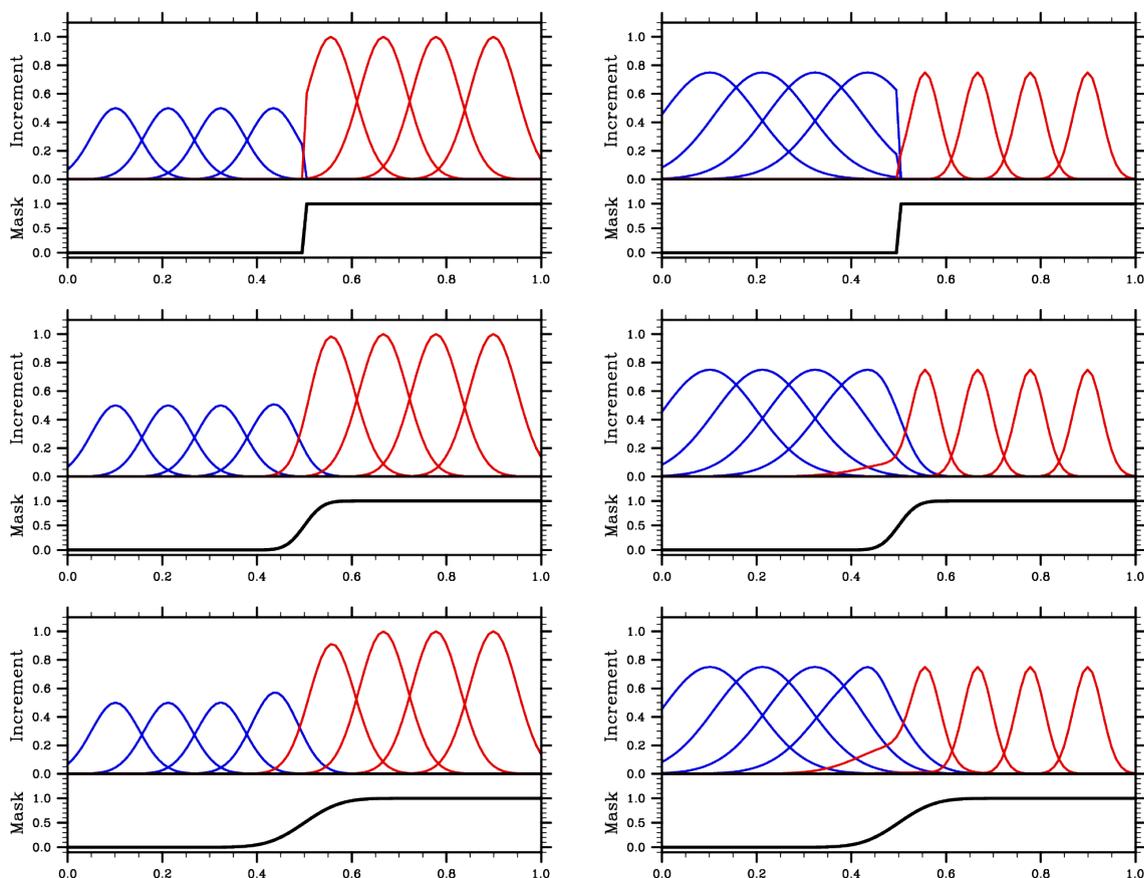


FIGURE 4.5 – Covariances d'une matrice  $\mathbf{B}$  idéalisée (voir texte pour les détails) et masque appliqué. Gauche : variances variables et longueur de corrélation identiques ; droite : variances identiques et longueurs de corrélation variables. Haut : pas de lissage du masque ; centre : lissage avec  $L = 0.03$  ; bas : lissage avec  $L = 0.06$ .

### 4.2.3 Impact du lissage, cas réaliste

La longueur caractéristique  $L$  du noyau gaussien doit être choisie empiriquement, de telle sorte que l'information se propage suffisamment entre les masques, sans pour autant perdre l'hétérogénéité de la matrice  $\mathbf{B}$ . D'après l'expérience idéalisée précédente, on peut supposer que ces propriétés sont réunies quand la longueur caractéristique du noyau gaussien est de l'ordre de la longueur de corrélation moyenne des matrices  $\mathbf{B}$  "en contact". On retient une longueur caractéristique de noyau  $L$  de quatre points de grille (environ 10 km) ce qui est légèrement supérieur aux longueurs de corrélation typiques pour la température et l'humidité spécifique, semblables pour les cas de brouillard et de non-brouillard (cf. figure 5.3, dans l'annexe 5.3.2). La figure 4.6 montre que cette démarche est efficace. Si l'innovation est située loin des frontières du masque, l'incrément généré est régulier, alors qu'il est tronqué si l'innovation se trouve proche d'une frontière abrupte de masque. Le lissage de cette frontière permet de régulariser l'incrément, tout en conservant les caractéristiques propres à chaque sous-domaine, par exemple la longueur de corrélation verticale plus courte dans les zone de brouillard que de non-brouillard.

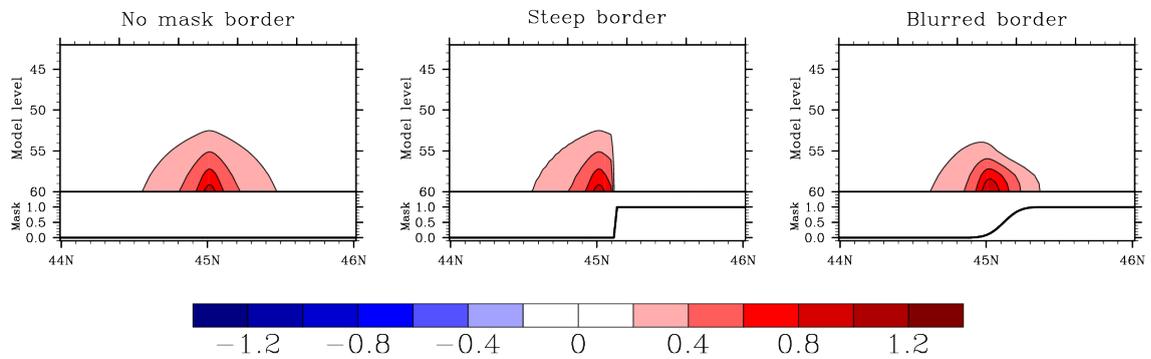


FIGURE 4.6 – Incréments de température générés par une innovation de 2 K à 45°N. De gauche à droite : masque uniforme, frontière de masque abrupte à 45,1°N, frontière de masque lissée par un noyau gaussien de longueur caractéristique 10 km. Le masque vaut 0 pour les zones de non-brouillard et 1 pour les zones de brouillard.

Comme l'illustre la figure 4.4, la convolution par le noyau gaussien lisse les contours du masque et diminue l'impact des zones de faible extension, sans perdre la structure des zones brouillard significatives.

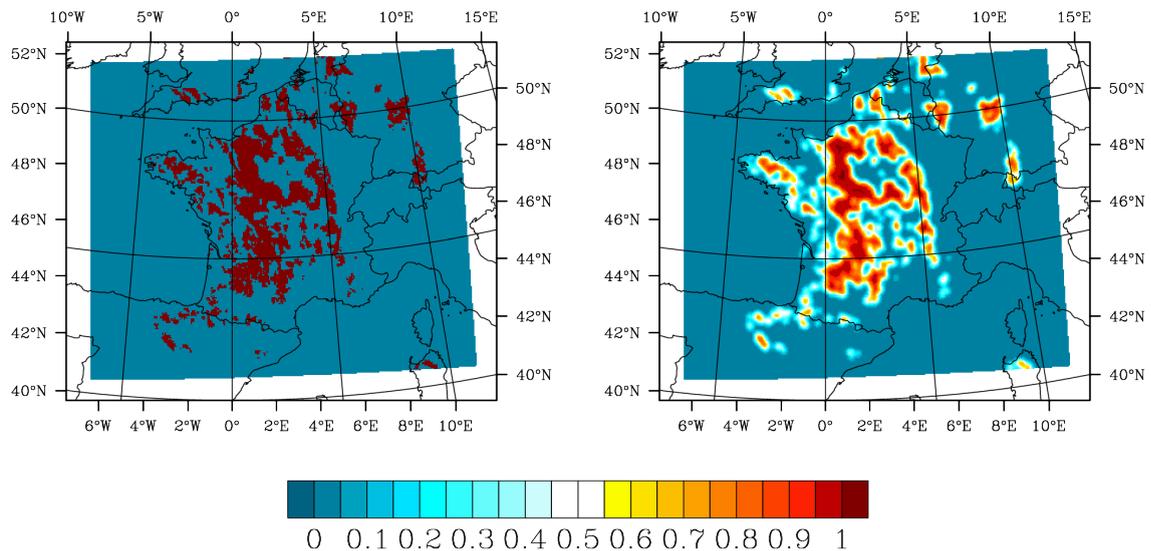


FIGURE 4.7 – Gauche : masque obtenu avec le critère (3.13) le 23/10/2010 à 6h ; droite : même masque après convolution par un noyau gaussien de longueur caractéristique 10 km.

### 4.3 Expériences réelles

Pour tester l'efficacité de la formulation hétérogène de la matrice  $\mathbf{B}$ , les nouvelles statistiques d'erreur de prévision calculées dans la section 3 sont employées avec des

masques lissés pour des cas réels, la chaîne opérationnelle servant de référence. La semaine du 23 au 30 octobre 2009, qui présente plusieurs cas de brouillards étendus sur la France, est retenue pour les tests.

### 4.3.1 Cas non-cyclés

Pour comparer les incréments d'analyse obtenus avec différentes formulations de la matrice  $\mathbf{B}$ , on doit partir de la même ébauche avant d'assimiler les observations. Il est donc nécessaire de ne pas cyclo l'expérience. Ainsi, une analyse est effectuée à 6h pour tous les jours sélectionnés, avec la matrice  $\mathbf{B}$  hétérogène et avec la matrice  $\mathbf{B}$  opérationnelle.

#### Etude de cas particuliers

Sur la coupe de la figure 4.8, des innovations de température très négative à 2 m provoquent un fort incrément négatif de température. Alors que cet incrément se propage jusqu'au niveau 52 environ dans le cas opérationnel, la formulation hétérogène de la matrice  $\mathbf{B}$  modifie cette propagation : elle est légèrement plus importante dans la zone de non-brouillard et beaucoup plus réduite dans la zone de brouillard, au nord de 46,2°N, où un dipôle d'incrément de signes opposés renforce la stabilisation des basses couches. On remarque aussi que l'innovation de température positive isolée à 47,2° a un impact plus fort dans le cas hétérogène que dans le cas opérationnel, car la variances des erreurs de prévision de température y est plus élevée, en particulier dans les basses couches.

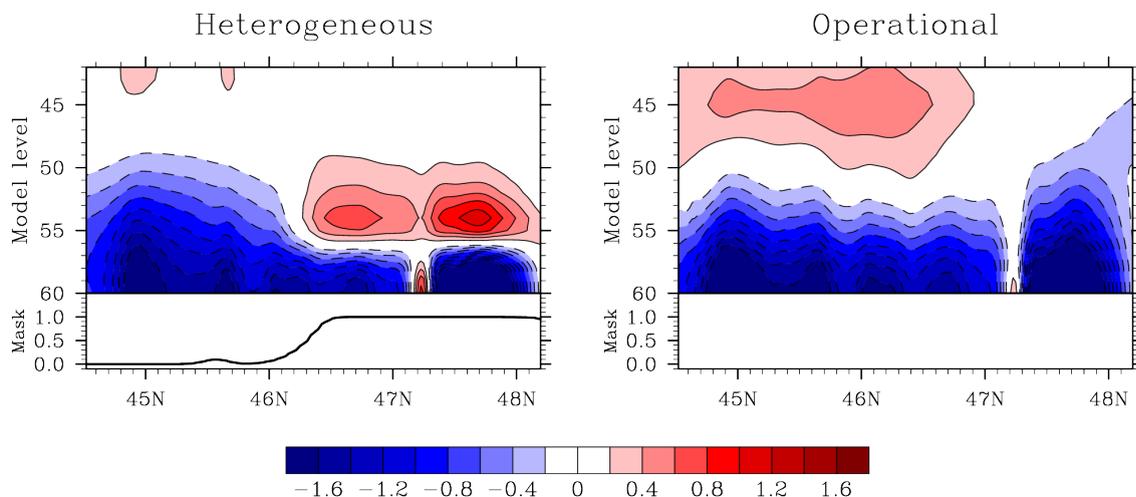


FIGURE 4.8 – Coupe méridienne à 2°E de l'incrément de température obtenu le 27/10/2010 à 6h. Gauche : formulation hétérogène de la matrice  $\mathbf{B}$  ; droite : version opérationnelle.

Cette spatialisation de la propagation de l'information est également très nette sur les cartes d'incrément à un niveau-modèle fixé. La figure 4.9 montre les incréments de températures à différents niveaux pour la version hétérogène (avec le masque de la

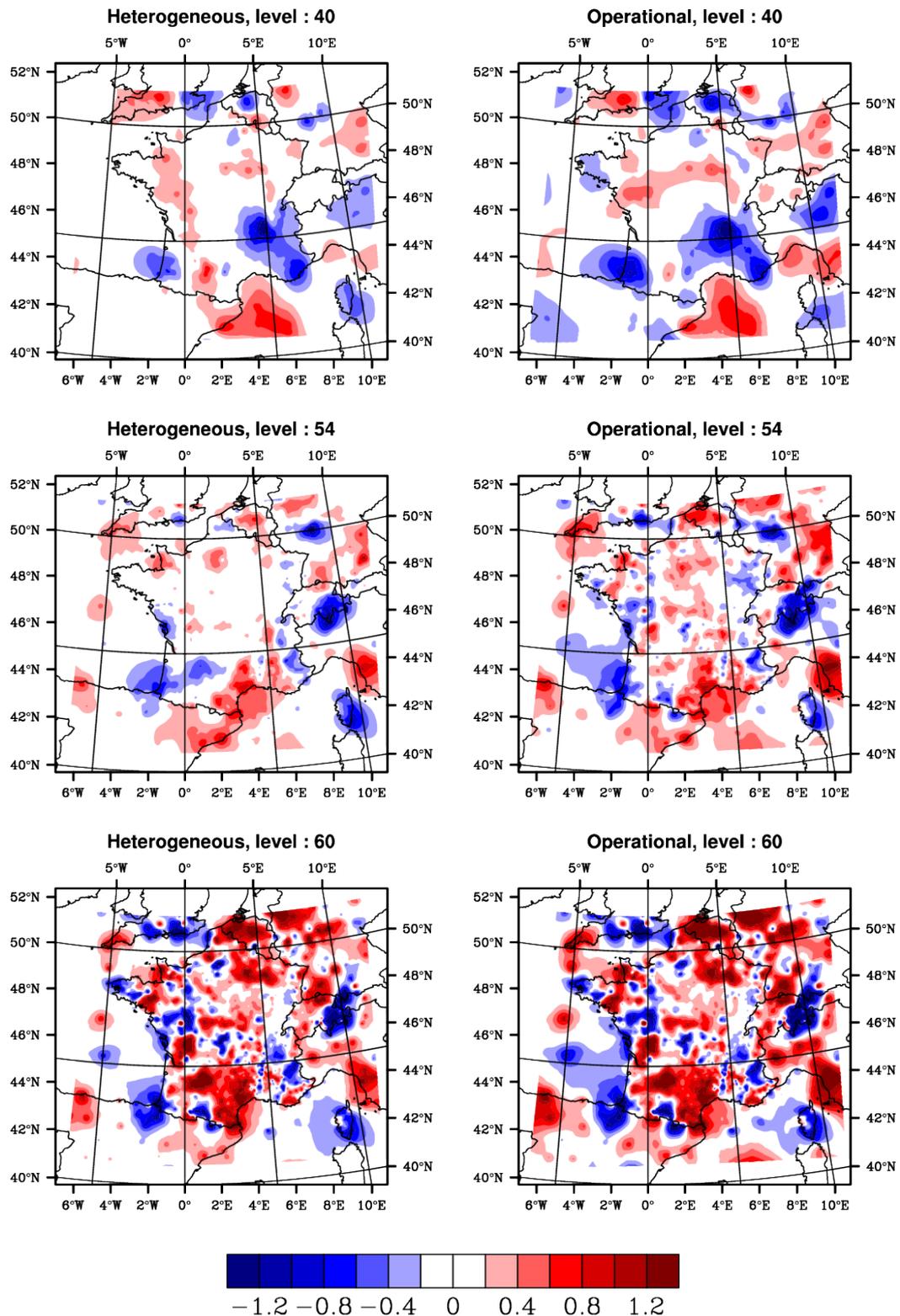


FIGURE 4.9 – Cartes de l'incrément de température obtenu le 23/10/2010 à 6h. Gauche : formulation hétérogène de la matrice  $\mathbf{B}$  ; droite : version opérationnelle. De haut en bas : cartes aux niveaux-modèles 40 53 et 60.

figure 4.4) et pour la version opérationnelle. Alors que les incréments de température sont relativement proches aux niveaux 60 (au sol) et 40 (à 800hPa) pour les deux formulations, l'effet des statistiques d'erreur de prévision hétérogène est très net au niveau 53 (juste au-dessus de la couche de brouillard). Dans les zones couvertes par le masque de la figure 4.4, l'incrément est de faible amplitude. On peut retenir les cas du Sud-Cotentin (hors du masque) et du Midi Toulousain (dans le masque), qui présentent tous les deux un incrément positif important en surface. Conformément aux statistiques détaillées en fin de section 3, les deux incréments sont de signes opposés dans la version hétérogène alors qu'ils sont de même signe dans la version opérationnelle, homogène.

### Moyenne et amplitude des incréments

Les effets de la formulation hétérogène sur les incréments issus des innovations au sol sont visibles non seulement sur des cartes et des coupes de cas particuliers, mais aussi par les statistiques effectuées sur tout le domaine pendant la semaine-test. Ainsi, les figures 4.10 et 4.11 donnent les profils de la moyenne et de l'amplitude RMS des incréments de température et d'humidité spécifique respectivement. Dans les deux cas, il est intéressant de noter que les incréments dans les zones de brouillard ont une forte amplitude au sol et que celle-ci diminue très rapidement jusqu'au niveau 55 environ puis reste assez faible au-delà. Ceci tend à confirmer l'existence de la barrière dynamique du sommet de la couche de brouillard dans les cas réels. On remarque également que les incréments de température ont tendance à réchauffer les basses couches dans les zones de brouillard, alors qu'ils les refroidissent pour les zones de non-brouillard et pour l'opérationnel. Il semble aussi que le modèle assèche trop les basses couches, puisque les incréments d'humidité spécifique sont toujours positifs.

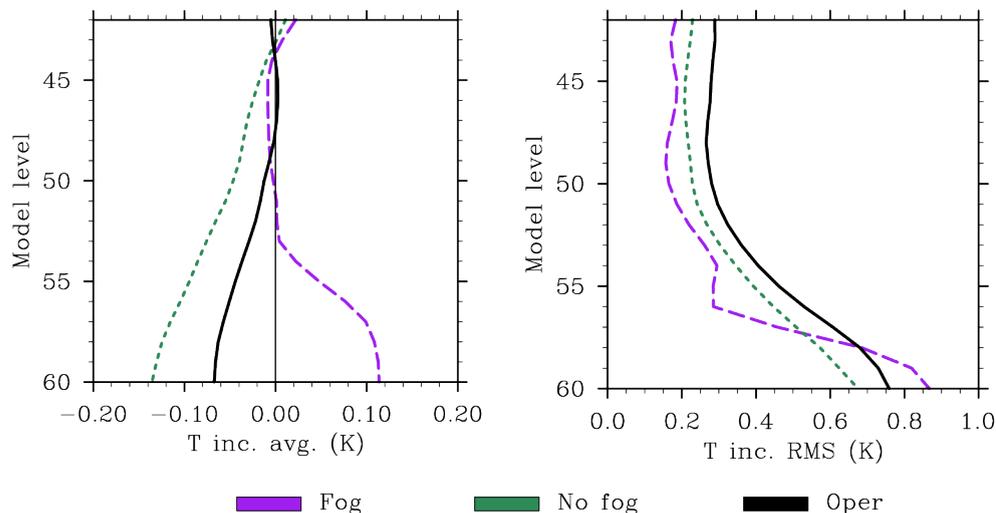


FIGURE 4.10 – Gauche : profil de la moyenne des incréments de température ; droite : profil de l'amplitude RMS des incréments de température.

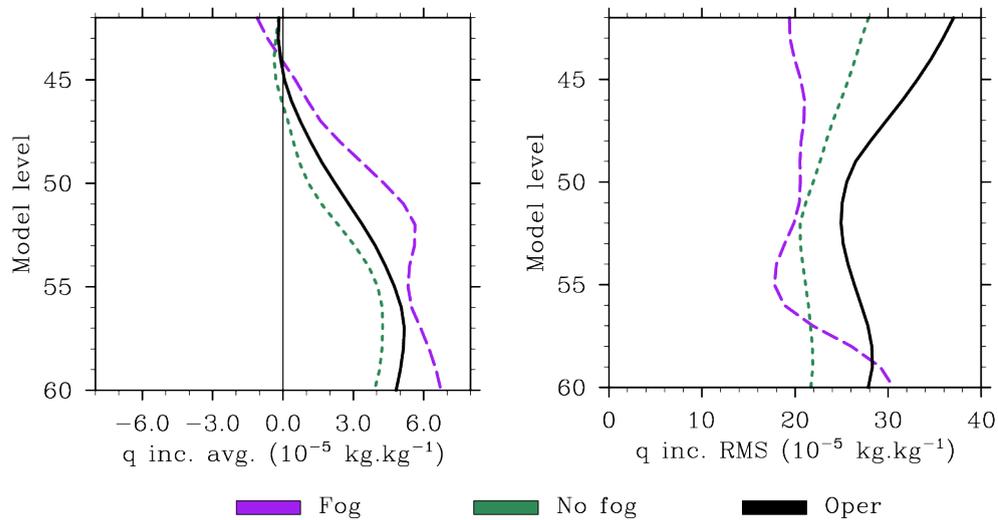


FIGURE 4.11 – Gauche : profil de la moyenne des incréments d’humidité spécifique ; droite : profil de l’amplitude RMS des incréments d’humidité spécifique.

### 4.3.2 Cas cyclés

Comme il est difficile d’estimer directement la qualité d’une analyse, puisque c’est la meilleure approximation que l’on connaisse de l’état de l’atmosphère, il faut procéder à des cycles de prévision/assimilation pour pouvoir tester l’efficacité de la formulation hétérogène par rapport à la version opérationnelle. Deux versions de la matrice  $\mathbf{B}$  hétérogène ont été testées. Dans la première, c’est la matrice  $\mathbf{B}$  homogène des zones de non-brouillard calibrée en 3 qui est utilisée dans les zones de non-brouillard, alors que dans la seconde, c’est la matrice  $\mathbf{B}$  homogène opérationnelle qui y est appliquée. Cette deuxième version devrait être plus performante, surtout dans les cas où l’extension du brouillard est faible, car la matrice  $\mathbf{B}$  opérationnelle a été calibrée sur un plus grand nombre de situations météorologiques. Dans les deux versions, c’est bien sûr la matrice  $\mathbf{B}$  spécifique aux zones de brouillard qui est appliquée dans les zones de brouillard.

En raison du peu de temps disponible pour effectuer les simulations cyclées, qui sont longues et coûteuses, seule la période du 22/10-18h au 26/10-9h, a pu être traitée. Bien que les cas de brouillards étendus se situent principalement sur ces quelques jours, cet échantillon de 30 cycles de prévision/assimilation reste beaucoup trop court pour tirer des conclusions définitives. En effet, l’efficacité d’une méthode d’assimilation de données ne peut être jugée qu’à partir d’un nombre important de cas, puisque l’optimalité recherchée pour l’interpolation entre ébauche et observations est une notion statistique.

### Scores COMPAS

Les scores traditionnels, développés à Météo-France par la division COMPAS, ont été calculés pour les 4 jours de test. Il ressort pour l'instant que l'impact sur les prévisions à 3h est neutre ou légèrement positif. La figure 4.12 montre par exemple des profils de différences entre les ébauches (prévisions à 3h) et les observations effectuées par des radiosondages. Si l'impact est neutre pour la température, il est légèrement positif pour l'humidité spécifique, avec une amélioration significative du biais dans les basses couches pour la seconde version de la matrice **B** hétérogène.

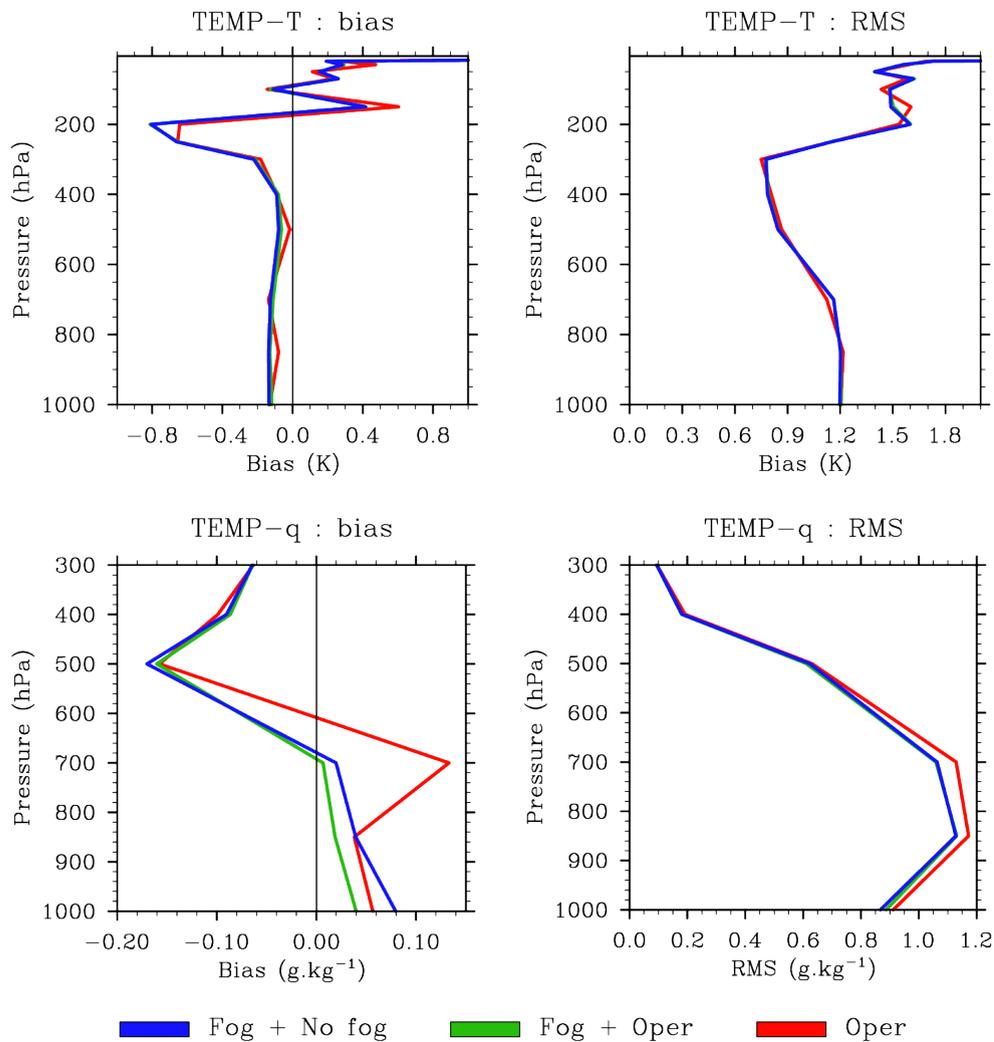


FIGURE 4.12 – Différences entre l'ébauche et les observations de radiosondages (TEMP), pour la température (en haut) et l'humidité spécifique (en bas). Gauche : biais ; droite : amplitude RMS.

### Scores CARIBOU

Les scores CARIBOU décrits à la section 3.2.2 permettent d'estimer la qualité de l'extension spatiale des prévisions de brouillard. Le critère 3.13 a été repris pour ce calcul. Sur la figure 4.13, on remarque un fort cycle diurne des scores POD et FAR, avec de meilleures performances quand le brouillard est le plus étendu (entre 3 h et 9 h). Les versions hétérogènes de la matrice  $\mathbf{B}$  sont moins performantes que la version opérationnelle au maximum de l'extension du brouillard (6 h), mais plus efficaces quand le brouillard est peu étendu (12 h, 18 h). On constate que le modèle sous-estime systématiquement l'étendue du brouillard par rapport à l'analyse CARIBOU, en particulier l'opérationnel entre 12 h et 18 h. Toutefois, il convient de se souvenir des limitations de l'utilisation de CARIBOU comme référence pour les phénomènes de brouillard, exposées lors de la définition des scores.

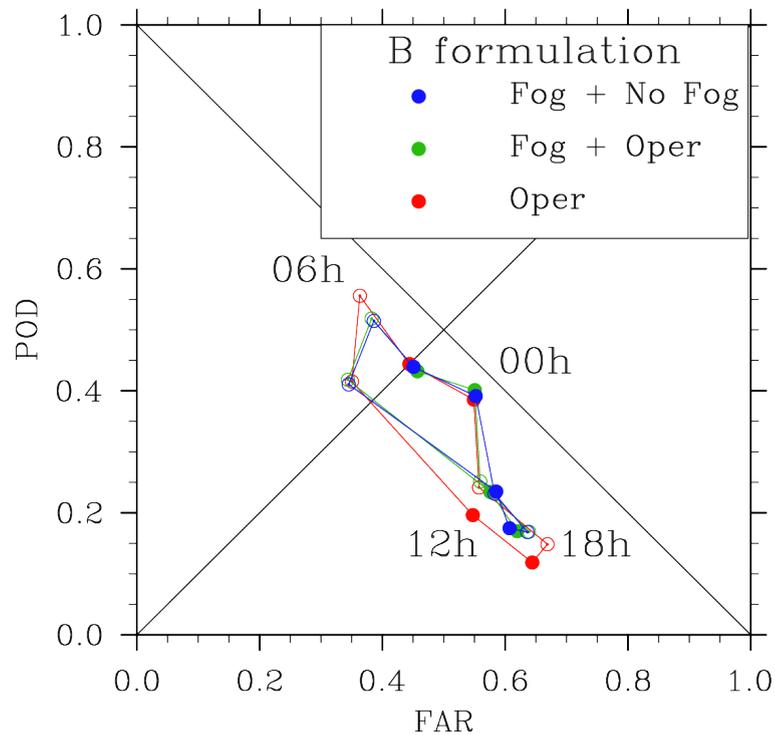


FIGURE 4.13 – Evolution diurne des scores de prévision de brouillard à 3h, par rapport à CARIBOU. Les réseaux synoptiques (00h, 06h, 12h, 18h) sont indiqués par des disques pleins.

# Conclusions et perspectives

## Conclusions

Les statistiques d'erreur de prévision, contenues dans la "matrice  $\mathbf{B}$ ", sont d'une importance cruciale pour l'assimilation de données par méthode variationnelle. Cependant, pour des raisons évidentes de coût de calcul et de stockage, elles doivent être modélisées. Dans les versions opérationnelles de ces modèles de matrice  $\mathbf{B}$ , les fonctions de covariance sont généralement homogènes et isotropes, ce qui limite leur réalisme. Nous avons donc essayé d'introduire une hétérogénéité rudimentaire dans les fonctions de covariance en utilisant des matrices homogènes distinctes selon les situations météorologiques, à l'aide de masques géographiques. En se focalisant sur les cas de brouillard terrestre, on a pu constater que de nombreux paramètres des statistiques d'erreur (variances, corrélations verticales, couplages entre champs) variaient fortement en fonction de la présence de brouillard ou non. Ceci rend donc pertinente l'utilisation de matrices hétérogènes pour ces situations météorologiques. La méthode des masques géographiques a montré son efficacité dans des cas idéalisés où une seule observation est assimilée. Après une légère modification nécessaire pour raccorder les statistiques d'erreur aux frontières de masques, on a pu constater un impact très net de l'hétérogénéité sur l'incrément d'analyse dans les cas réels. La méthode répond donc aux attentes en terme de spatialisation des incréments d'analyse. Son impact sur la qualité des prévision n'a pas encore été complètement estimé, faute de temps, mais les premiers résultats de simulations cyclées semblent encourageant.

## Perspectives

L'analyse fine d'un nombre important de simulations cyclées est requise pour déterminer avec exactitude l'impact sur la prévision de la méthode des masques géographiques. Quelque soit cet impact, plusieurs points de la méthode des masques géographiques peuvent encore être améliorés. La production des masques géographiques est un point particulièrement sensible et non-systématique de la méthode, puisque c'est à l'utilisateur de définir explicitement ces masques. Seuls les masques issus de l'analyse CARIBOU ont été testé jusqu'à présent, mais on peut imaginer d'autres sources, comme l'ébauche, voire une combinaison de ces sources. Le filtrage des masques obtenus peut aussi être perfectionné, notamment par des techniques issues du traitement d'image non-linéaire (opérateurs d'ouverture/fermeture). L'estimation des statistiques d'erreur de prévision pour chaque situation météorologique peut

aussi profiter des progrès effectués dans le domaine de la modélisation de l'opérateur de balance et de la transformation spatiale de la matrice  $\mathbf{B}$ . Plus généralement, de nombreuses méthodes pour rendre la matrice  $\mathbf{B}$  hétérogène, anisotrope et dépendante de l'écoulement sont en cours de développement dans les centres de recherche météorologiques (cf. ). Cependant, ces recherches ont surtout porté jusqu'à présent sur les modèles globaux, et de nombreuses voies restent à explorer pour les modèles à méso-échelle.

## Remerciements

Je tiens à remercier vivement Thibaut Montmerle pour la confiance qu'il m'a accordé, pour son encadrement constructif et sa patience envers toutes mes questions. Je remercie aussi Jean-François Mahfouf et Thierry Bergot pour leurs explications sur les phénomènes de brouillard. Je remercie enfin toute l'équipe du GMAP pour son accueil chaleureux et stimulant, dont j'espère encore profiter dans les années à venir.

# Bibliographie

- [Bannister, 2008a] BANNISTER, R. N. (2008a). A review of forecast error covariance statistics in atmospheric variational data assimilation. i : Characteristics and measurements of forecast error covariances. *Quarterly Journal of the Royal Meteorological Society*, 134:1951–1970.
- [Bannister, 2008b] BANNISTER, R. N. (2008b). A review of forecast error covariance statistics in atmospheric variational data assimilation. ii : Modelling the forecast error covariance statistics. *Quarterly Journal of the Royal Meteorological Society*, 134:1971–1996.
- [Berre, 2000] BERRE, L. (2000). Estimation of synoptic and mesoscale forecast error covariances in a limited area model. *Monthly Weather Review*, 128:644–667.
- [Berre *et al.*, 2006] BERRE, L., STEFANESCU, S. E. et PEREIRA, M. B. (2006). The representation of the analysis effect in three error simulation techniques. *Tellus*, 58A:196–209.
- [Bocquet, 2009] BOCQUET, M. (2009). *Introduction aux principes et méthodes de l'assimilation de données en géophysique*.
- [Bouttier et Courtier, 1999] BOUTTIER, F. et COURTIER, P. (1999). Data assimilation concepts and methods. *In Meteorological Training Course Lecture Series*. ECMWF.
- [Brousseau *et al.*, 2008] BROUSSEAU, P., BOUTTIER, F., HELLO, G., SEITY, Y., FISCHER, C., BERRE, L., MONTMERLE, T., AUGER, L. et MALARDEL, S. (2008). A prototype convective-scale data assimilation system for operation : the Arome-RUC. Rapport technique, HIRLAM.
- [Courtier *et al.*, 1998] COURTIER, P., ANDERSSON, E., HECKLEY, W., VASILJEVIC, D., HAMRUD, M., HOLLINGSWORTH, A., RABIER, F., FISHER, M. et PAILLEUX, J. (1998). The ecmwf implementation of three-dimensional variational assimilation (3d-var). i : Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124:1783–1807.
- [Derber et Bouttier, 1999] DERBER, J. et BOUTTIER, F. (1999). A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus*, 51A:195–221.
- [Desroziers, 1997] DESROZIERS, G. (1997). A coordinate change for data assimilation in spherical geometry of frontal structures. *Monthly Weather Review*, 125:3030–3038.
- [Desroziers *et al.*, 2008] DESROZIERS, G., BERRE, L., PANNEKOUCKE, O., STEFANESCU, S., BROUSSEAU, P., AUGER, L., CHAPNIK, B. et RAYNAUD, L. (2008).

Flow-dependent error covariances from variational assimilation ensembles on global and regional domains. Rapport technique, HIRLAM.

- [Fischer *et al.*, 2005] FISCHER, C., MONTMERLE, T., BERRE, L., AUGER, L. et STEFANESCU, S. (2005). An overview of the variational assimilation in the ALADIN/France numerical weather-prediction system. *Quarterly Journal of the Royal Meteorological Society*, 131:3477–3492.
- [Fisher, 2003] FISHER, M. (2003). Background error covariance modeling. *In Proc. ECMWF Seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean*, pages 45–63, Reading, U.K.
- [Guidard et Tzanos, 2007] GUIDARD, V. et TZANOS, D. (2007). Analysis of fog probability from a combination of satellite and ground observation data. *Pure and Applied Geophysics*, 164:1207–1220.
- [Gultepe *et al.*, 2007] GULTEPE, I., TARDIF, R., MICHAELIDES, S., CERMAK, J., BOTT, A., BENDIX, J., MÜLLER, M., PAGOWSKI, M., HANSEN, B., ELLROD, G., JACOBS, W., TOTH, G. et COBER, S. (2007). Fog research : A review of past achievements and future perspectives. *Pure and Applied Geophysics*, 164:1121–1159.
- [Hollingworth et Lonnerberg, 1986] HOLLINGWORTH, A. et LONNERBERG, P. (1986). The statistical structure of short-range forecast errors as determined from radiosonde data. part i : The wind field. *Tellus*, 38A:111–136.
- [Kalnay, 2003] KALNAY, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press.
- [Kunkel, 1984] KUNKEL, B. (1984). Parameterization of droplet terminal velocity and extinction coefficient in fog models. *Journal of Applied Meteorology*, 23:34–41.
- [Montmerle et Berre, 2010] MONTMERLE, T. et BERRE, L. (2010). Diagnosis and use of heterogeneous background error covariances at mesoscale. *Quarterly Journal of the Royal Meteorological Society*, in press.
- [Montmerle et Faccani, 2009] MONTMERLE, T. et FACCANI, C. (2009). Mesoscale assimilation of radial velocities from doppler radar in a pre-operational framework. *Monthly Weather Review*, 137:1937–1953.
- [NOAA, 2005] NOAA (2005). *Federal Meteorological Handbook No. 1 (FCM-H1-2005)*, 104 pp. National Oceanic and Atmospheric Administration.
- [Pannecoucke, 2008] PANNECOUCKE, O. (2008). *Modélisation des structures locales de covariance des erreurs de prévision à l'aide des ondelettes*. Thèse de doctorat, Université Toulouse III - Paul Sabatier.
- [Parrish et Derber, 1992] PARRISH, F. D. et DERBER, J. C. (1992). The national meteorological center's spectral statistical-interpolation analysis system. *Monthly Weather Review*, 120:1747–1763.
- [Purser *et al.*, 2003] PURSER, R., WU, W., PARRISH, D. et ROBERTS, N. (2003). Numerical aspects of the application of recursive filters to variational statistical analysis. part ii : Spatially inhomogeneous and anisotropic general covariances. *Monthly Weather Review*, 131:1536–1548.

- [Raynaud *et al.*, 2009] RAYNAUD, L., BERRE, L. et DESROZIERS, G. (2009). Objective filtering of ensemble-based background error variances. *Quarterly Journal of the Royal Meteorological Society*, 135:1177–1199.
- [Remy, 2009] REMY, S. (2009). *Prévision locale des faibles visibilités pour l'aéronautique*. Thèse de doctorat, Université Toulouse III - Paul Sabatier.
- [Seity *et al.*, 2010] SEITY, Y., BROUSSEAU, P., MALARDEL, S., HELLO, G., BÉ-NARD, P., BOUTTIER, F., LAC, C. et MASSON, V. (2010). The arome-france convective scale operational model. *Monthly Weather Review*, submitted.
- [Weaver et Courtier, 2001] WEAVER, A. et COURTIER, P. (2001). Correlation modelling on the sphere using a generalized diffusion equation. *Quarterly Journal of the Royal Meteorological Society*, 127:1815–1846.

# Chapitre 5

## Annexes

### 5.1 Algorithmes variationnels

#### 5.1.1 Interpolation optimale

Pour définir la fonction coût en 1.3, on a utilisé des normes quadratiques issues de produits scalaires pondérés par des matrices symétriques définies positives. On justifie ce choix en supposant que les PDF des erreurs d'ébauche et d'observation peuvent être approximées par des fonctions gaussiennes multivariées [Bannister, 2008a]. En effet, si l'ébauche seule est connue, l'état du système est décrit par une PDF *a priori* :

$$p^b(\mathbf{x}) \propto \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b)\right\}, \quad (5.1)$$

et la probabilité conditionnelle d'observer  $\mathbf{y}^o$  étant donné le vecteur d'état  $\mathbf{x}$  est

$$p^o(\mathbf{y}^o|\mathbf{x}) \propto \exp\left\{-\frac{1}{2}[\mathbf{y}^o - H(\mathbf{x})]^T \mathbf{R}^{-1}[\mathbf{y}^o - H(\mathbf{x})]\right\}. \quad (5.2)$$

L'analyse est alors décrite par une PDF *a posteriori* tenant compte de l'information apportée par les observations, calculée par la loi de Bayes :

$$\begin{aligned} p^a(\mathbf{x}|\mathbf{y}^o) &\propto p^b(\mathbf{x})P^o(\mathbf{y}^o|\mathbf{x}) \\ &\propto \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) \right. \\ &\quad \left. - \frac{1}{2}[\mathbf{y}^o - H(\mathbf{x})]^T \mathbf{R}^{-1}[\mathbf{y}^o - H(\mathbf{x})]\right\}. \end{aligned} \quad (5.3)$$

Il est courant de choisir le maximum de vraisemblance comme estimateur de l'analyse, même si ce n'est pas le seul envisageable. Déterminer l'analyse revient donc à maximiser la PDF  $p^a(\mathbf{x}|\mathbf{y}^o)$ , ce qui est équivalent à minimiser la fonction coût  $J$ .

#### 5.1.2 Dimension temporelle

Dans les systèmes opérationnels, des observations  $\mathbf{y}_i^o$  sont acquises à des instants successifs  $t_i$  tout au long d'une période  $[t_0, t_N]$  appelée fenêtre d'assimilation. L'état

du système peut être avancé de tout instant  $t_i$  vers tout autre instant  $t_j > t_i$  en appliquant le modèle noté  $M : \mathbf{x}(t_j) = M_{i \rightarrow j}[\mathbf{x}(t_i)]$ . Ce dernier est supposé parfait (contrainte forte), et la trajectoire de l'analyse au sein de la fenêtre d'assimilation est donc entièrement déterminée par sa valeur en  $t_0$ . Ainsi, la variable de contrôle de la minimisation est réduite à  $\mathbf{x} = \mathbf{x}(t_0)$ . La non-simultanéité des observations peut être complètement intégrée dans la fonction coût (4D-VAR), mais des approximations moins coûteuses sont généralement utilisées.

#### 4D-VAR / 4D-Inc

L'algorithme le plus précis est le 4D-VAR, pour lequel la date d'acquisition des observations est pleinement prise en compte, en composant l'opérateur d'observation avec le modèle :

$$J(\mathbf{x}) = \frac{1}{2} \|\mathbf{x} - \mathbf{x}^b\|_{\mathbf{B}^{-1}}^2 + \frac{1}{2} \sum_{i=0}^N \|\mathbf{y}_i^o - H_i \circ M_{0 \rightarrow i}(\mathbf{x})\|_{\mathbf{R}_i^{-1}}^2 \quad (5.4)$$

La forme de cette fonction coût peut aussi être justifiée sous certaines hypothèses, elle représente en fait un lissage bayésien, démontré dans l'annexe 5.1.3. Le gradient de la fonction coût 4D-VAR est alors donné par

$$\nabla J(\mathbf{x}) = \mathbf{B}^{-1} \mathbf{x} - \sum_{i=0}^N \mathbf{M}_{0 \rightarrow i}^T \mathbf{H}_i^T \mathbf{R}_i^{-1} [\mathbf{y}_i^o - H_i \circ M_{0 \rightarrow i}(\mathbf{x})] \quad (5.5)$$

Des difficultés inhérentes à la forme 4D-VAR de la fonction coût surgissent alors :

- l'apparition de minima locaux dans la fonction coût dus à la non-linéarité de  $H_i \circ M_{0 \rightarrow i}$ , qui compliquent singulièrement la minimisation itérative,
- la nécessité de développer et de maintenir l'adjoint du modèle  $\mathbf{M}_{0 \rightarrow i}^T$ , ce qui est particulièrement délicat pour les modèles de nuage non-hydrostatiques dont certaines paramétrisations (comme la micro-physique) sont fortement non-linéaires et donc difficilement linéarisables,
- le coût de calcul dû aux nombreuses intégrations du modèle et de son adjoint.

Des approximations successives permettent de passer outre ces difficultés, au prix d'une perte de précision dans l'analyse.

Pour éviter les minima locaux, il est préférable de minimiser une approximation quadratique de la fonction coût complète, obtenue en linéarisant  $H_i \circ M_{0 \rightarrow i}$  autour d'un état de référence  $\mathbf{x}^g$  :

$$H_i \circ M_{0 \rightarrow i}[\mathbf{x}^g + \delta \mathbf{x}] \simeq H_i \circ M_{0 \rightarrow i}(\mathbf{x}^g) + \mathbf{H}_i \mathbf{M}_{0 \rightarrow i} \delta \mathbf{x} \quad (5.6)$$

où  $\mathbf{H}_i$  désigne l'opérateur d'observation linéarisé autour de  $M_{0 \rightarrow i}[\mathbf{x}^g]$  et  $\mathbf{M}_{0 \rightarrow i}$  le modèle linéarisé autour de  $\mathbf{x}^g$ . La nouvelle variable de contrôle de la fonction coût est l'incrément  $\delta \mathbf{x} = \mathbf{x} - \mathbf{x}^g$ , d'où l'approximation dite incrémentale de la fonction coût donnée par

$$J(\delta \mathbf{x}, \mathbf{x}^g) = \frac{1}{2} \|\delta \mathbf{x} - \delta \mathbf{x}^b\|_{\mathbf{B}^{-1}}^2 + \frac{1}{2} \sum_{i=0}^N \|\mathbf{d}_i - \mathbf{H}_i \mathbf{M}_{0 \rightarrow i} \delta \mathbf{x}\|_{\mathbf{R}_i^{-1}}^2 \quad (5.7)$$

où  $\mathbf{d}_i = \mathbf{y}_i^o - H_i \circ M_{0 \rightarrow i}(\mathbf{x}^g)$  est le vecteur d'innovation (différence entre référence et observations). Son gradient est donné par

$$\nabla J(\delta \mathbf{x}, \mathbf{x}^g) = \mathbf{B}^{-1} \mathbf{x} - \sum_{i=0}^N \mathbf{M}_{0 \rightarrow i}^T \mathbf{H}^T \mathbf{R}^{-1} [\mathbf{d}_i - \mathbf{H}_i \mathbf{M}_{0 \rightarrow i} \delta \mathbf{x}] \quad (5.8)$$

Il est possible de minimiser des approximations successives de la fonction coût, selon une méthode dite multi-incrémentale. Pour la première minimisation, la référence  $\mathbf{x}_0^g$  est choisie égale à l'ébauche  $\mathbf{x}^b$ . Pour les étapes suivantes, c'est le résultat de la minimisation précédente qui donne la référence :  $\mathbf{x}_{k+1}^g = \mathbf{x}_k^g + \text{Argmin}(J(\delta \mathbf{x}, \mathbf{x}_k^g))$ , comme illustré sur la figure 5.1. En opérationnel, on utilise souvent des variables d'état de résolutions de plus en plus fines dans les minimisations successives [Bouttier et Courtier, 1999].

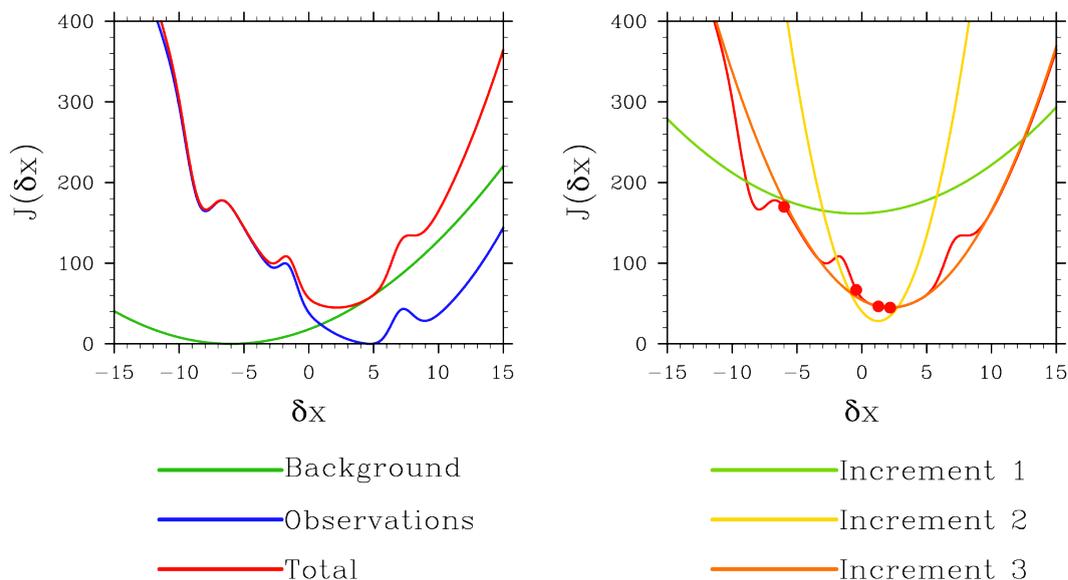


FIGURE 5.1 – Gauche : fonction coût totale avec ses composantes (unités arbitraires), dans le cas d'un opérateur d'observation non-linéaire donnant une fonction coût d'observation légèrement non-quadratique; Droite : fonction coût totale avec ses approximations incrémentales successives. Les points rouges représentent les valeurs de  $\mathbf{x}^g$  qui convergent vers  $\mathbf{x}^a$ , l'analyse.

### 3D-FGAT / 3D-VAR / 3D-Inc

Le calcul du modèle linéaire tangent et de son adjoint sont toujours nécessaires dans l'approximation 4D-Inc. Pour s'en passer, on peut simplement les remplacer par l'identité. Cette approximation donne l'algorithme 3D-FGAT (First Guess at Appropriate Time), dont la fonction coût est

$$J(\delta \mathbf{x}, \mathbf{x}^g) = \frac{1}{2} \|\delta \mathbf{x} - \delta \mathbf{x}^b\|_{\mathbf{B}^{-1}}^2 + \frac{1}{2} \sum_{i=0}^N \|\mathbf{d}_i - \mathbf{H}_i \delta \mathbf{x}\|_{\mathbf{R}_i^{-1}}^2 \quad (5.9)$$

avec  $\mathbf{d}_i = \mathbf{y}_i^o - H_i \circ M_{0 \rightarrow i}(\mathbf{x}^g)$  et  $\mathbf{H}_i$  linéarisé autour de  $M_{0 \rightarrow i}[\mathbf{x}^g]$ . Son gradient est donné par

$$\nabla J(\delta \mathbf{x}, \mathbf{x}^g) = \mathbf{B}^{-1} \mathbf{x} - \sum_{i=0}^N \mathbf{H}^T \mathbf{R}^{-1} [\mathbf{d}_i - \mathbf{H}_i \delta \mathbf{x}] \quad (5.10)$$

La fonction coût du 3D-FGAT demande cependant l'intégration du modèle complet pour calculer  $\mathbf{d}_i$  et  $\mathbf{H}_i$ . Là encore, on peut se passer définitivement du modèle en le remplaçant par l'identité, mais on compare alors les observations  $\mathbf{y}_i^o$  à la variable de contrôle en  $t_0$  et non en  $t_i$ . Pour que cette approximation soit valable, il faut que les  $t_i$  des observations assimilées soient suffisamment proches de  $t_0$  ("cutoff court") et que les observations soient moyennées à l'instant  $t_0$ . On définit ainsi l'algorithme 3D-VAR, dont la fonction coût est

$$J(\mathbf{x}) = \frac{1}{2} \|\mathbf{x} - \mathbf{x}^b\|_{\mathbf{B}^{-1}}^2 + \frac{1}{2} \|\mathbf{y}^o - H[\mathbf{x}]\|_{\mathbf{R}^{-1}}^2 \quad (5.11)$$

dont le gradient est

$$\nabla J(\mathbf{x}) = \mathbf{B}^{-1} \mathbf{x} - \mathbf{H}^T \mathbf{R}^{-1} [\mathbf{y}^o - H(\mathbf{x})] \quad (5.12)$$

Sa version incrémentale, appelée 3D-Inc, est obtenue en linéarisant l'opérateur d'observation autour de  $\mathbf{x}^g$  :  $H_i[\mathbf{x}^g + \delta \mathbf{x}] \simeq H_i[\mathbf{x}^g] + \mathbf{H} \delta \mathbf{x}$ , ce qui donne

$$J(\delta \mathbf{x}, \mathbf{x}^g) = \frac{1}{2} \|\delta \mathbf{x} - \delta \mathbf{x}^b\|_{\mathbf{B}^{-1}}^2 + \frac{1}{2} \|\mathbf{d} - \mathbf{H} \delta \mathbf{x}\|_{\mathbf{R}^{-1}}^2 \quad (5.13)$$

avec  $\mathbf{d}_i = \mathbf{y}_i^o - H_i[\mathbf{x}^g]$ . Le gradient est alors donné par

$$\nabla J(\delta \mathbf{x}, \mathbf{x}^g) = \mathbf{B}^{-1} \mathbf{x} - \mathbf{H}^T \mathbf{R}^{-1} [\mathbf{d} - \mathbf{H} \delta \mathbf{x}] \quad (5.14)$$

La dépendance temporelle (quatrième dimension du 4D-VAR et du 4D-Inc) a été amoindrie dans l'algorithme 3D-FGAT, et a disparu pour le 3D-VAR et le 3D-Inc.

### 5.1.3 Lissage bayésien

Dans le cas où les observations sont distribuées dans une fenêtre d'assimilation  $[t_0, t_n]$ , on peut résumer le problème d'assimilation par

$$\mathbf{x}_j = M_{i \rightarrow j}(\mathbf{x}_i) + \varepsilon_i^m, \quad 0 \leq i < j \leq N \quad (5.15)$$

$$\mathbf{y}_i^o = H_i(\mathbf{x}_i^t) + \varepsilon_i^o, \quad 0 \leq i \leq N \quad (5.16)$$

avec  $\mathbf{x}_i = \mathbf{x}(t_i)$  pour alléger les notations. [Bocquet, 2009] montre que ce problème non-linéaire peut être traité par une formulation variationnelle 4D-VAR à contrainte faible si  $\varepsilon_i^m \neq 0$  ou à contrainte forte en l'absence d'erreur modèle  $\varepsilon_i^m$ . La démonstration est reprise intégralement dans la suite de ce paragraphe.

L'objectif est de déterminer la PDF  $p(\mathbf{X}_i|\mathbf{Y}_i)$  par lissage bayésien, où  $\mathbf{X}_i$  et  $\mathbf{Y}_i$  sont les ensembles d'états du système et d'observations de  $t_0$  à  $t_i$  :

$$\mathbf{X}_i = \{\mathbf{x}_i, \mathbf{x}_{i-1}, \dots, \mathbf{x}_0\} \quad (5.17)$$

$$\mathbf{Y}_i = \{\mathbf{y}_i^o, \mathbf{y}_{i-1}^o, \dots, \mathbf{y}_0^o\} \quad (5.18)$$

$$(5.19)$$

La loi de Bayes donne

$$p(\mathbf{X}_i|\mathbf{Y}_i) \propto p(\mathbf{X}_i)p(\mathbf{Y}_i|\mathbf{X}_i) \quad (5.20)$$

On cherche une relation de récurrence donnant  $p(\mathbf{X}_i|\mathbf{Y}_i)$  en fonction de  $p(\mathbf{X}_{i-1}|\mathbf{Y}_{i-1})$ . Le premier facteur de (5.20) peut être explicité en notant que la suite de variables aléatoires  $\mathbf{X}_i$  est markovienne :

$$\begin{aligned} p(\mathbf{X}_i) &= p(\mathbf{x}_i, \mathbf{X}_{i-1}) \\ &= p(\mathbf{x}_i|\mathbf{X}_{i-1})p(\mathbf{X}_{i-1}) \\ &= p(\mathbf{x}_i|\mathbf{x}_{i-1})p(\mathbf{X}_{i-1}) \end{aligned} \quad (5.21)$$

De même pour le second terme, en notant que  $\mathbf{y}_i$  et  $\mathbf{Y}_{i-1}$  sont indépendantes :

$$\begin{aligned} p(\mathbf{Y}_i|\mathbf{X}_i) &= p(\mathbf{y}_i, \mathbf{Y}_{i-1}|\mathbf{X}_i) \\ &= p(\mathbf{y}_i|\mathbf{X}_i)p(\mathbf{Y}_{i-1}|\mathbf{X}_i) \\ &= p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{X}_{i-1})p(\mathbf{Y}_{i-1}|\mathbf{x}_i, \mathbf{X}_{i-1}) \\ &= p(\mathbf{y}_i|\mathbf{x}_i)p(\mathbf{Y}_{i-1}|\mathbf{X}_{i-1}) \end{aligned} \quad (5.22)$$

Ainsi, on obtient

$$\begin{aligned} p(\mathbf{X}_i|\mathbf{Y}_i) &\propto p(\mathbf{X}_i)p(\mathbf{Y}_i|\mathbf{X}_i) \\ &\propto p(\mathbf{x}_i|\mathbf{x}_{i-1})p(\mathbf{X}_{i-1})p(\mathbf{y}_i|\mathbf{x}_i)p(\mathbf{Y}_{i-1}|\mathbf{X}_{i-1}) \\ &\propto p(\mathbf{x}_i|\mathbf{x}_{i-1})p(\mathbf{y}_i|\mathbf{x}_i)p(\mathbf{X}_{i-1}|\mathbf{Y}_{i-1}) \end{aligned} \quad (5.23)$$

$$(5.24)$$

De cette récurrence, on déduit

$$p(\mathbf{X}_i|\mathbf{Y}_i) \propto p(\mathbf{x}_0) \prod_{j=0}^{i-1} p(\mathbf{x}_{j+1}|\mathbf{x}_j) \prod_{j=0}^i p(\mathbf{y}_j|\mathbf{x}_j) \quad (5.25)$$

où  $p(\mathbf{x}_0)$  est la PDF de l'ébauche à l'instant  $t_0$ .

Aucune hypothèse sur les PDFs n'a encore été utilisée. Cependant, il est nécessaire de les spécifier pour parvenir à calculer l'estimateur retenu pour l'analyse (maximum de vraisemblance dans notre cas). On suppose donc à présent que l'erreur d'ébauche  $\varepsilon^b$ , les erreurs-modèle  $\varepsilon_i^m$  et les erreurs d'observations  $\varepsilon_i^o$  constituent des bruits blancs gaussien d'espérance nulle et de matrices de covariance  $\mathbf{B}$ ,  $\mathbf{Q}_i$  et  $\mathbf{R}_i$  respectivement.

Lorsque toutes les observations sont assimilées, cela donne

$$\begin{aligned}
p(\mathbf{X}_N | \mathbf{Y}_N) &\propto \exp \left\{ -\frac{1}{2} (\mathbf{x}_0 - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}^b) \right. \\
&\quad - \frac{1}{2} \sum_{i=0}^{N-1} [\mathbf{x}_{i+1}^o - M_{i \rightarrow i+1}(\mathbf{x}_i)]^T \mathbf{Q}_i^{-1} [\mathbf{x}_{i+1}^o - M_{i \rightarrow i+1}(\mathbf{x}_i)] \\
&\quad \left. - \frac{1}{2} \sum_{i=0}^N [\mathbf{y}_i^o - H_i(\mathbf{x}_i)]^T \mathbf{R}_i^{-1} [\mathbf{y}_i^o - H_i(\mathbf{x}_i)] \right\}. \tag{5.26}
\end{aligned}$$

Ainsi, en présence d'erreurs-modèle, l'analyse est donnée par la minimisation de la fonction coût 4D-VAR à contrainte faible, dont la variable de contrôle est  $\mathbf{X}_N$  :

$$\begin{aligned}
J(\mathbf{X}_N) &= \frac{1}{2} (\mathbf{x}_0 - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}^b) \\
&\quad + \frac{1}{2} \sum_{i=0}^{N-1} [\mathbf{x}_{i+1}^o - M_{i \rightarrow i+1}(\mathbf{x}_i)]^T \mathbf{Q}_i^{-1} [\mathbf{x}_{i+1}^o - M_{i \rightarrow i+1}(\mathbf{x}_i)] \\
&\quad + \frac{1}{2} \sum_{i=0}^N [\mathbf{y}_i^o - H_i(\mathbf{x}_i)]^T \mathbf{R}_i^{-1} [\mathbf{y}_i^o - H_i(\mathbf{x}_i)] \tag{5.27}
\end{aligned}$$

Pour un modèle parfait, le second terme disparaît et la variable de contrôle est réduite à  $\mathbf{x}_0$ . On retrouve alors la fonction coût 4D-VAR à contrainte forte donnée par une approche heuristique au paragraphe 5.1.2.

## 5.2 Démonstrations de propriétés de la matrice $\mathbf{B}$

### 5.2.1 Existence de racines carrées

Par construction, la matrice  $\mathbf{B}$  est symétrique définie positive, donc diagonalisable dans une base orthonormale. En notant  $\mathbf{F}$  la matrice orthonormale des vecteurs propres de  $\mathbf{B}$  ( $\mathbf{F}^T \mathbf{F} = \mathbf{I}$ ) et  $\Lambda$  la matrice diagonale contenant ses valeurs propres, on a

$$\mathbf{B} = \mathbf{F} \Lambda \mathbf{F}^T = \mathbf{F} \Lambda^{1/2} \mathbf{Q}^T \mathbf{Q} \Lambda^{1/2} \mathbf{F}^T \tag{5.28}$$

où  $\mathbf{Q}^T$  est une matrice orthonormale quelconque vérifiant  $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$ . Il existe donc une infinité de racines carrées de la forme  $\mathbf{B}^{1/2} = \mathbf{F} \Lambda^{1/2} \mathbf{Q}^T$ . Les racines carrées d'une matrice de taille  $n \times n$  peuvent être de taille  $n \times p$ , où  $p$  n'est pas nécessairement égal à  $n$ .

### 5.2.2 Homogénéité et isotropie des fonctions de corrélation dans l'hypothèse diagonale spectrale

Dans les modèles spectraux comme ceux de Météo-France, la matrice de corrélation  $\mathbf{C} = \Sigma^{-1} \mathbf{B} \Sigma^{-1}$  est représentée par l'hypothèse diagonale spectrale, qui suppose que les coefficients spectraux résultant de la transformation des fonctions de corrélation contenues dans  $\mathbf{C}$  sont décorrélés.

**Propriété :** *Les fonctions de corrélations sont homogènes et isotropes si et seulement si la matrice de corrélation s'écrit sous la forme  $\mathbf{C} = \mathbf{F}^{-1}\mathbf{D}\mathbf{F}$ , où  $\mathbf{F}$  et  $\mathbf{F}^{-1}$  sont des transformées spectrales directe et inverse et  $\mathbf{D}$  une matrice diagonale.*

La base spectrale utilisée varie selon le modèle, mais pour comprendre la nature de cette propriété tout en simplifiant les calculs, on se propose de démontrer ce résultat dans un cas unidimensionnel périodique régulier.

### Transformée de Fourier unidimensionnelle

Soit une série  $(A_n)_{0 \leq n \leq N-1}$  représentant une fonction unidimensionnelle  $A$  périodique de période  $N\Delta x$  discrétisée sur l'axe réel aux points  $x_n = n\Delta x$  :  $A_n = A(x_n)$ . On suppose  $N$  pair. Dans ce cas discret, l'analyse et la synthèse de Fourier sont :

$$a_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} A_n \exp(-2i\pi \frac{kn}{N}) \quad , \quad 0 \leq k \leq N-1 \quad (5.29)$$

$$A_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} a_k \exp(2i\pi \frac{kn}{N}) \quad , \quad 0 \leq n \leq N-1 \quad (5.30)$$

On définit alors les matrices de Vandermonde directes et inverses :

$$\mathbf{F} = \frac{1}{\sqrt{N}} \begin{pmatrix} \omega^{0.0} & \omega^{0.1} & \dots & \omega^{0.(N-1)} \\ \omega^{1.0} & \omega^{1.1} & \dots & \omega^{1.(N-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \omega^{(N-1).0} & \omega^{(N-1).1} & \dots & \omega^{(N-1).(N-1)} \end{pmatrix} \quad \text{et} \quad \mathbf{F}^{-1} = \mathbf{F}^* \quad (5.31)$$

où  $\omega = \exp(\frac{-2i\pi}{N})$  est la racine  $N^{\text{ième}}$  de l'unité. En posant

$$\mathbf{A} = \begin{pmatrix} A_0 \\ \vdots \\ A_{N-1} \end{pmatrix} \quad , \quad \mathbf{a} = \begin{pmatrix} a_0 \\ \vdots \\ a_{N-1} \end{pmatrix} \quad (5.32)$$

on peut écrire (5.29) et (5.30) sous la forme

$$\mathbf{a} = \mathbf{F}\mathbf{A} \quad (5.33)$$

$$\mathbf{A} = \mathbf{F}^{-1}\mathbf{a} \quad (5.34)$$

### Homogénéité et isotropie

Supposons qu'une matrice  $\mathbf{C}$  est exprimée sous la forme  $\mathbf{C} = \mathbf{F}^{-1}\mathbf{D}\mathbf{F}$ , où  $\mathbf{D}$  est une matrice diagonale. Les matrices  $\mathbf{F}$  et  $\mathbf{F}^{-1}$  étant symétriques, il est évident par construction que  $\mathbf{C}$  est aussi symétrique. Pour que  $\mathbf{C}$  soit définie positive, il est nécessaire et suffisant que tous les coefficients de  $\mathbf{D}$  soient réels et strictement positifs, car ils représentent le spectre de  $\mathbf{C}$ . La dernière condition pour assurer que  $\mathbf{C}$  soit une matrice de corrélation est que ses coefficients diagonaux soient unitaires. Cette

condition est remplie si et seulement si la moyenne des coefficients de  $\mathbf{D}$  vaut 1. En effet, le coefficient d'indices  $(i, j)$  de la matrice  $\mathbf{C}$  a pour expression

$$C_{ij} = \frac{1}{N} \sum_{k=0}^{N-1} \omega^{-i.k} D_k \omega^{k.j} = \frac{1}{N} \sum_{k=0}^{N-1} \omega^{(j-i).k} D_k \quad (5.35)$$

d'où  $C_{ii} = \frac{1}{N} \sum_{k=0}^{N-1} D_k$ .

On constate alors que le coefficient  $C_{ij}$  ne dépend que de l'écart  $(j - i)$ , donc de la distance  $(x_j - x_i)$ , mais pas explicitement de la position de ces points. Les fonctions de corrélation sont donc homogènes (indépendantes de la position). Enfin, on note que  $C_{ji} = \overline{C_{ij}} = C_{ij}$ , ce qui signifie que les fonctions de corrélation sont isotropes.

### Forme diagonale spectrale

Supposons que  $\mathbf{C}$  soit une matrice de corrélation homogène et isotrope. Soit  $\mathbf{D}$  la matrice donnée par  $\mathbf{D} = \mathbf{F}\mathbf{C}\mathbf{F}^{-1}$ . Le coefficient d'indices  $(i, j)$  de la matrice  $\mathbf{D}$  a pour expression

$$D_{ij} = \frac{1}{N} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} \omega^{i.k} C_{kl} \omega^{-j.l} = \frac{1}{N} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} C_{kl} \omega^{i.k-j.l} \quad (5.36)$$

Or, on a les propriétés suivantes, aux modulus sur les indices près :

- $\mathbf{C}$  est symétrique, donc  $C_{kl} = C_{lk}$ ,
- $\mathbf{C}$  est homogène, donc  $C_{k(k+d)}$  ne dépend que de la distance  $d$ ,
- $\mathbf{C}$  est isotrope, donc  $C_{k(k+d)} = C_{k(k-d)}$ ,

On peut noter  $C_d = C_{k(k+d)} = C_{k(k-d)} = C_{(k+d)k} = C_{(k-d)k}$ , ce qui permet d'obtenir après quelques manipulations :

$$\begin{aligned} D_{ij} &= \frac{1}{N} \sum_{d=0}^{N/2-1} C_d \sum_{k=0}^{N-1} (\omega^{i.k-j.(k+d)} + \omega^{i.(k+d)-j.k}) \\ &= \frac{1}{N} \sum_{d=0}^{N/2-1} C_d \sum_{k=0}^{N-1} \omega^{(i-j).k} (\omega^{i.d} + \omega^{-j.d}) \\ &= \frac{1}{N} \sum_{d=0}^{N/2-1} C_d (\omega^{i.d} + \omega^{-j.d}) \sum_{k=0}^{N-1} \omega^{(i-j).k} \end{aligned} \quad (5.37)$$

Ainsi, les coefficients diagonaux de  $\mathbf{D}$  sont réels et ont pour expression :

$$D_{ii} = \frac{2}{N} \sum_{d=0}^{N/2-1} C_d \cos\left(\frac{2id\pi}{N}\right). \quad (5.38)$$

alors que les coefficients non-diagonaux sont tous nuls, car pour  $i \neq j$  :

$$\sum_{k=0}^{N-1} \omega^{(i-j).k} = 0 \quad (5.39)$$

On a donc  $\mathbf{C} = \mathbf{F}^{-1}\mathbf{D}\mathbf{F}$  avec  $\mathbf{D}$  diagonale réelle. Comme dans le cas précédent, les coefficients de  $\mathbf{D}$  sont positifs pour assurer que  $\mathbf{C}$  soit définie positive et leur moyenne vaut 1 pour que  $\mathbf{C}$  soit une matrice de corrélation.

## 5.3 Figures complémentaires

### 5.3.1 Ecart-types de la vorticité et de la divergence

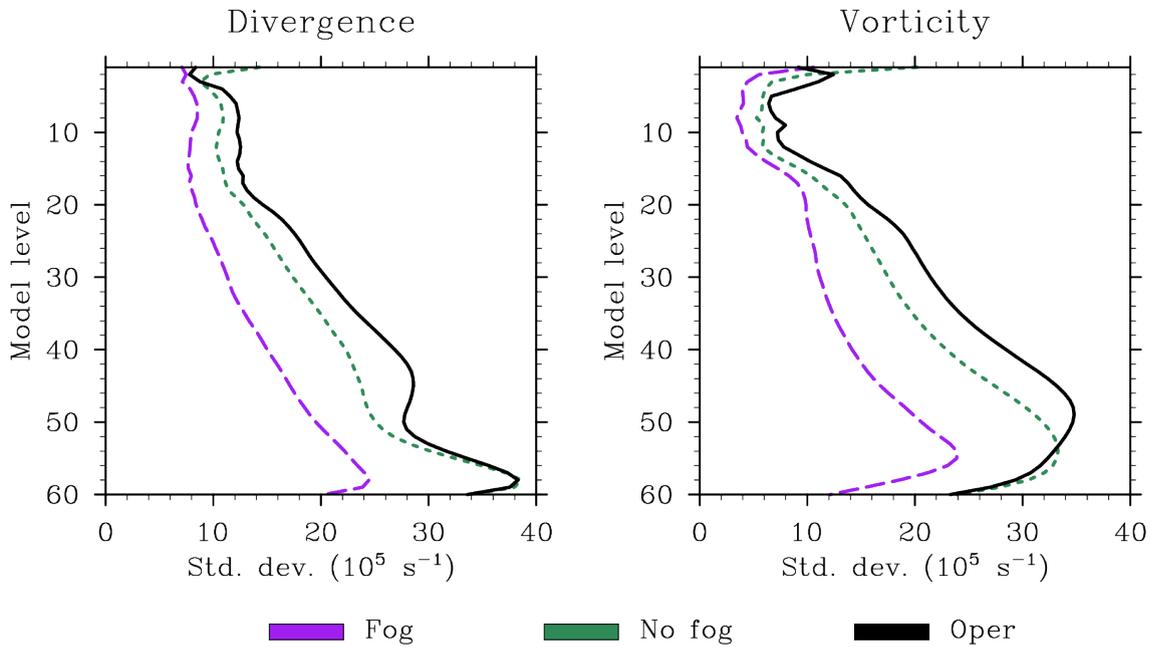


FIGURE 5.2 – Profils moyens des écart-types d’erreur de prévision pour la divergence (à gauche) et la vorticité (à droite) pour les trois cas étudiés : situations de brouillard, de non-brouillard et version opérationnelle.

### 5.3.2 Longueurs de corrélation horizontale

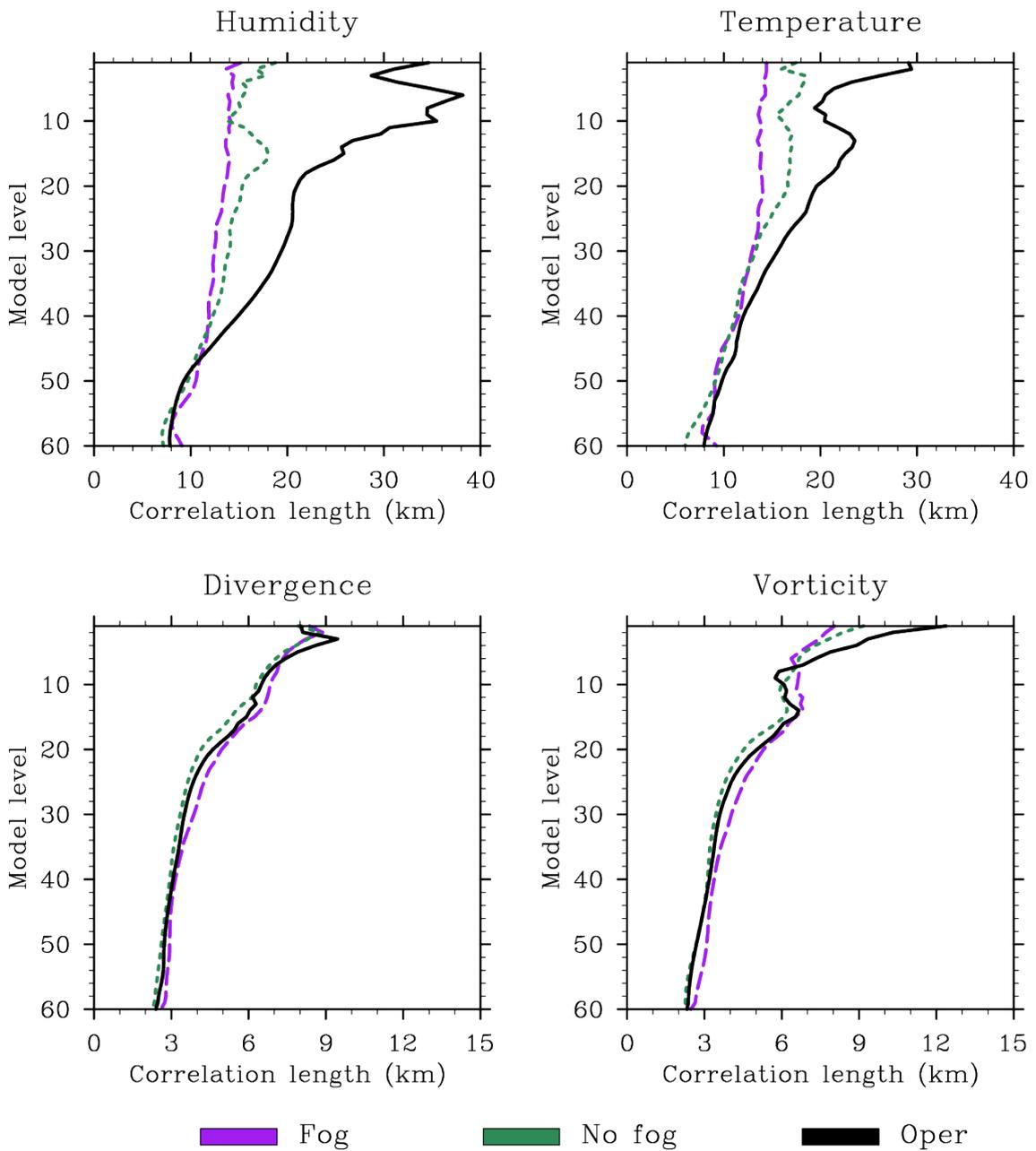


FIGURE 5.3 – Profils moyens des longueurs de corrélation horizontale pour la température (en haut à gauche), l'humidité relative (en haut à droite), la divergence (en bas à gauche) et la vorticité (en bas à droite) pour les trois cas étudiés : situations de brouillard, de non-brouillard et version opérationnelle.

### 5.3.3 Incréments à 800 hPa

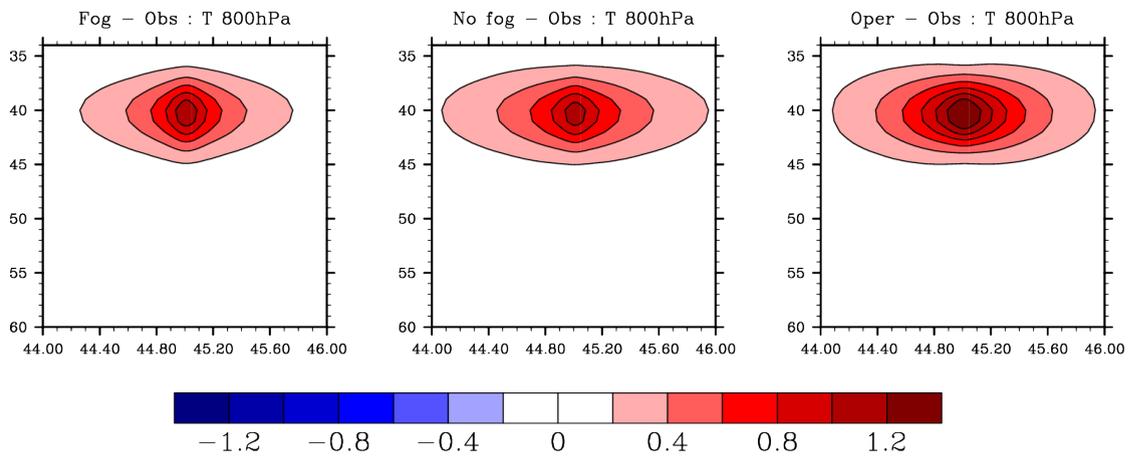


FIGURE 5.4 – Incrément d’analyse en température généré par une innovation de 2K à 800 hPa. De gauche à droite : situation de brouillard, de non-brouillard, et cas opérationnel

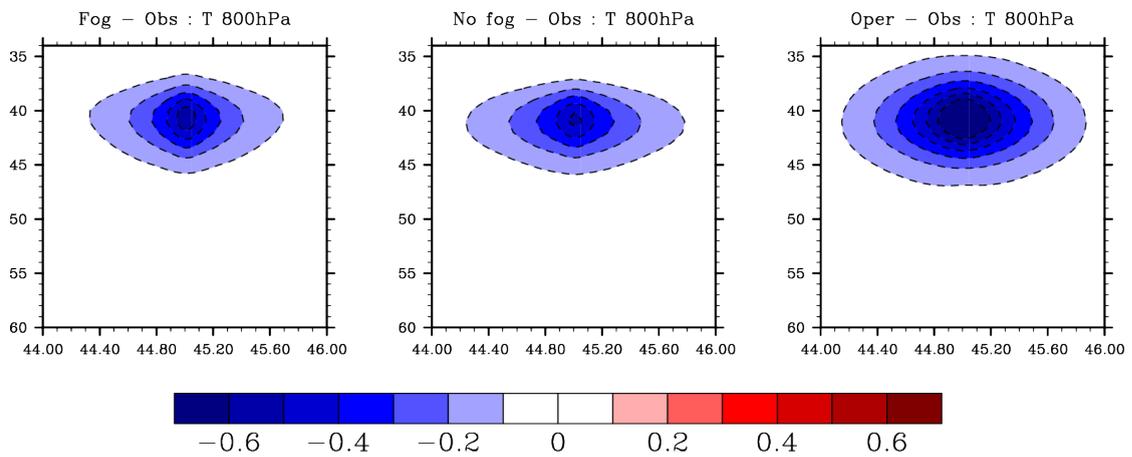


FIGURE 5.5 – Incrément d’analyse en humidité spécifique généré par une innovation de 2K à 800 hPa. De gauche à droite : situation de brouillard, de non-brouillard, et cas opérationnel