Estimation des incertitudes et prévision des risques en qualité de l'air

Damien Garaud

Cerea

Centre d'Enseignement et de Recherche en Environnement Atmosphérique

Thèse de doctorat de l'Université Paris-Est

Présentée et soutenue publiquement le 14 décembre 2011 par

Thèse de doctorat

Damien Garaud

pour l'obtention du diplôme de docteur
de l'Université Paris-Est

Spécialité : Sciences et techniques de l'environnement

Estimation des incertitudes et prévision des risques en qualité de l'air

---

Jury composé de

Pr Matthias Beekmann	CNRS	président
Pr Serge Guillas	University College London	rapporteur
Dr Jean-Luc Ponche	Université de Strasbourg	rapporteur
Dr Frédérik Meleux	INERIS	examinateur
Dr Isabelle Herlin	INRIA	directeur de thèse
Dr Vivien Mallet	INRIA	co-directeur de thèse

Remerciements

Je tiens à remercier tout d'abord EDF R&D (en particulier le département MFEE, Luc Musson-Genon et Damien Bilbault) et l'Association Nationale de la Recherche et de la Technologie d'avoir financé cette thèse.

Merci à Bruno Sportisse et Christian Seigneur de m'avoir accueilli dans leur laboratoire.

Un très grand merci à Vivien Mallet pour sa confiance, sa patience, sa pédagogie et sa disponibilité. Par son encadrement, il m'a permis de progresser et m'a transmis son intérêt pour les mathématiques appliquées, l'informatique et le logiciel libre.

Merci à Isabelle Herlin de m'avoir accueilli au sein de l'équipe-projet CLIME et d'avoir accepté de devenir ma directrice de thèse en cours de route.

Je remercie les rapporteurs, Serge Guillas et Jean-Luc Ponche, ainsi que les autres membres du jury, d'avoir accepté d'évaluer mon travail.

Merci à Hélène Marfaing et Christophe Debert d'Airparif pour les données concernant les incertitudes liées aux instruments de mesure de la qualité de l'air.

Une chaleureuse pensée pour tous mes collègues de l'INRIA et du CEREA. Ceux qui sont déjà passés par là : Irène, Jérôme, Laëtitia, Lin, Elsa, Rachid, Maya, Norah, Marc, Yelva, Karine, Dominique, Jean-Paul, Etienne, Marilyne et Édouard. Ceux qui vont y passer : Victor qui a cédé à la tentation, Karim, Florian, Régis, Ève, Reza, Cédric, Yiguo et Nicolas (première personne que je rencontre IRL qui connaisse SMT et Crunchbang !). Un merci à Nathalie et Véronique pour leur gentillesse, leur aide et leur disponibilité. Merci à Youngseob, co-bureau qui a soutenu le lendemain de ma soutenance, et qui, en plein séjour italien m'a fait découvrir la Corée à sa façon. Une pensée aussi pour Anne, Nicolas, Kévin, Claire, Meryem et Pierre. Merci donc à tous pour leur bonne humeur et les échanges, tant sur le plan humain que scientifique, que j'ai eus avec chacun d'entre eux.

A great thanks to Richard for improving my english, his proofreadings, happy mood, sense of humour and interesting discussions.

Un clin d'œil à mes actuels collègues Logilabiens avec qui j'ai passé mes derniers mois de thèse.

Une pensée à ma famille, à mes amis, aux cafés du troisième, à leur soutien et leurs encouragements. Enfin, une amoureuse pensée à ma femme. Merci à elle pour tout son soutien, sa patience, sa compréhension et le quotidien qu'elle m'offre depuis plusieurs années.

Résumé Ce travail porte sur l'estimation des incertitudes et la prévision de risques en qualité de l'air. Il consiste dans un premier temps à construire un ensemble de simulations de la qualité de l'air qui prend en compte toutes les incertitudes liées à la modélisation. Des ensembles de simulations à l'échelle continentale ou régionale sont générés automatiquement. Ensuite, les ensembles générés sont calibrés par une méthode d'optimisation combinatoire qui sélectionne un sous-ensemble représentatif de l'incertitude ou performant (fiabilité et résolution) pour des prévisions probabilistes. Ainsi, il est possible d'estimer et de prévoir des champs d'incertitude sur les concentrations d'ozone ou de dioxyde d'azote, ou encore d'améliorer la fiabilité des prévisions e dépassement de seuil. Cette approche est ensuite comparée avec la calibration d'un ensemble Monte Carlo. Ce dernier, moins dispersé, est moins représentatif de l'incertitude. Enfin, on a pu estimer la part des erreurs de mesure, de représentativité et de modélisation de la qualité de l'air. Ces travaux ont été appliqués à l'impact de centrales thermiques, afin de quantifier l'incertitude sur les impacts estimés.

Abstract This work is about uncertainty estimation and risk prediction in air quality. Firstly, we build a multimodel ensemble of air quality simulations which can take into account all uncertainty sources related to air quality modeling. Ensembles of photochemical simulations at continental and regional scales are automatically generated. Then, these ensemble are calibrated with a combinatorial optimization method. It selects a sub-ensemble which is representative of uncertainty or shows good resolution and reliability for probabilistic forecasting. This work shows that it is possible to estimate and forecast uncertainty fields related to ozone and nitrogen dioxide concentrations or to improve the reliability of threshold exceedance predictions. The approach is compared with Monte Carlo simulations, calibrated or not. The Monte Carlo approach appears to be less representative of the uncertainties than the multimodel approach. Finally, we quantify the observational error, the representativeness error and the modeling errors. The work is applied to the impact of thermal power plants, in order to quantify the uncertainty on the impact estimates.

Table des matières

Introduction

Modèles de chimie-transport

Les modèles de chimie-transport sont des modèles mathématiques qui prennent en compte des phénomènes atmosphériques complexes, liés à l'interaction de plusieurs espèces chimiques entre elles et avec leur environnement, dans le but d'estimer la concentration d'espèces chimiques dans l'atmosphère.

Ces modèles peuvent se décliner à différentes échelles de temps et d'espace. Les hypothèses et les modèles mathématiques utilisés ne sont pas les mêmes selon les besoins. Le suivi d'une espèce passive autour de sa zone d'émission à une échelle de quelques centaines de mètres est, d'un point de vue numérique, formulé différemment du le transport réactif de plusieurs espèces à une échelle continentale.

Outre les problèmes d'échelle de temps et d'espace, la formulation du modèle va aussi dépendre de la cible observée : traceur passif, photochimie, aérosols, radionucléides, ... La photochimie fait intervenir différentes espèces chimiques présentes dans l'atmosphère, qui vont réagir entre elles et dont certaines (par exemple, le dioxyde d'azote) sont sujettes à la photolyse. L'étude des aérosols ajoute une composante multiphasique puisqu'il faut prendre en compte la formation des particules fines (nucléation), l'interaction entre les particules (coagulation), et les échanges entre phases (condensation et évaporation).

Quels que soient le problème abordé et les cibles observées, ces modèles ont été développés dans le but principal d'étudier l'impact de polluants qui peuvent être nocifs pour l'homme ou les cultures agricoles. L'ozone, par exemple, espèce chimique secondaire (non émise) est nocif pour les voies respiratoires. Cette espèce est essentiellement produite en présence de dioxyde d'azote et de composés organiques volatils (principalement issus de l'activité humaine).

Ces modèles font parfois intervenir plusieurs centaines de champs tridimensionnels, soumis à des phénomènes atmosphériques complexes à des temps et distances caractéristiques très dispersés. Actuellement, seule la simulation numérique permet d'estimer l'évolution de telles espèces dans l'atmosphère.

Les simulations numériques permettent de remplir différents objectifs. Un objectif est de prévoir, sur une échéance de quelques jours, les concentrations de différentes espèces au voisinage du sol. Le système PREV'AIR [1] par exemple, fournit des prévisions opérationnelles de polluants gazeux et de particules (PM10, PM2.5) de manière quotidienne pour une échéance de quelques jours en France et en Europe. S'ajoutent à cela les réglementations françaises et européennes sur différents polluants : ozone, SO_2, NO_2, PM10. La prévision de dépassement de seuil de concentration peut alors être un enjeu sanitaire et économique.

Un second objectif concerne la possibilité de réaliser des études d'impact. L'étude d'impact en qualité de l'air consiste à simuler puis étudier la pollution suivant plusieurs scénarios d'émission.

1. http ://www.prevair.org/

On peut s'intéresser à l'étude d'une implantation d'une usine, par exemple. L'impact des émissions de cette dernière sera alors étudié dans la région d'installation. L'étude d'impact peut aussi être appliquée aux scénarios de réduction des émissions du trafic.

Un autre objectif est la modélisation inverse. On se propose dans ce cas de d'affiner les données d'entrée grâce à des méthodes d'assimilation de données. Cela a un double objectif : (1) améliorer les simulations en ayant des données d'entrée de meilleure qualité et (2) retrouver des données d'entrée mal estimées ou inconnues. Le deuxième cas peut être rencontré lors d'un rejet accidentel d'espèces dans le but de retrouver une bonne estimation de la quantité de polluant émise.

Ces modèles ont néanmoins des limitations non négligeables du fait des nombreuses incertitudes dues à leurs données d'entrée, leurs paramétrisations physiques et leurs approximations numériques.

Incertitudes

La première source d'incertitude concerne les données nécessaires en entrée d'un modèle de chimie-transport : champs météorologiques, émissions, conditions aux limites, ... Ces données sont évidemment entachées d'incertitudes. À titre d'exemple, l'évaluation, par des experts, de la qualité de ces données d'entrée [Hanna *et al.*, 1998, 2001] révèle des incertitudes de l'ordre de 50%, et jusqu'à 100% pour les données d'émissions.

La seconde source a son origine dans la formulation physico-chimique d'un modèle. Les mécanismes chimiques en sont un bon exemple. Il en existe plusieurs — CB05, RACM, RACM 2, RADM2 — et chacun propose des hypothèses et un nombre d'espèces et de réactions chimiques différents. Toutes les espèces et toutes les réactions chimiques ne peuvent être intégrées dans les mécanismes dédiés aux modèles opérationnels. En conséquence, des simplifications et des regroupements d'espèces sont réalisés afin de prendre en compte au mieux les principales espèces et réactions intervenant dans les processus concernant les cibles modélisées.

Une autre source d'incertitude non négligeable réside dans les approximations numériques : pas de temps, résolutions horizontale et verticale, schémas numériques.

Dans ce contexte de fortes incertitudes, on considère que la modélisation doit approcher la densité de probabilité des concentrations, vues comme variables aléatoires, plutôt que les concentrations seules. Au lieu de considérer un modèle déterministe

$$y = M(x),$$

où la cible y et les entrées x sont des vecteurs fixés, on considère un modèle stochastique

$$Y = M(X),$$

où Y et X sont des vecteurs aléatoires. S'il s'agit de proposer une unique valeur ou prévision pour la cible, on estime alors l'espérance $\mathrm{E}(Y)$. Cette espérance doit être accompagnée d'une mesure de l'incertitude, c'est-à-dire classiquement la variance $\mathrm{Var}(Y) = \mathrm{E}\left((Y - \mathrm{E}(Y))(Y - \mathrm{E}(Y))^T\right)$.

En pratique, cette approche est déclinée par la génération d'ensemble de simulations déterministes. Dans une approche Monte Carlo, on repose sur une série x_i de tirages des entrées (selon certaines distributions probabilistes) et on obtient l'ensemble des $(y_i)_i$ tels que

$$y_i = M(x_i).$$

Dans cette thèse, je m'intéresse particulièrement à une extension de cette approche où le modèle lui-même est considéré comme incertain. On repose alors sur plusieurs modèles déterministes $(M_k)_k$ pour constituer l'ensemble des $(y_{ik})_{i,k}$ tels que

$$y_{ik} = M_k(x_i).$$

Un travail essentiel, objet d'une partie de cette thèse, est de s'assurer que cet ensemble est capable d'estimer la variance associée à Y. De plus, on souhaite que l'ensemble apporte un moyen de calculer des probabilités de dépassement de seuil, soit que $\mathrm{P}(Y \geq S)$ où S est typiquement un seuil de concentration réglementaire.

Plan de la thèse

Le chapitre 1 est un chapitre d'introduction à la physique et à la chimie atmosphériques. On se concentre essentiellement sur les phénomènes liés à la production d'ozone. Je décris ensuite les sources d'incertitude qui interviennent dans le cadre de la simulation numérique de la qualité de l'air : (1) les données d'entrée, (2) les paramétrisations physiques et (3) les approximations numériques. J'introduis ensuite les ensembles de prévisions comme étant un moyen de prendre en compte ces trois sources d'incertitude. Je décris enfin un certain nombre de scores d'ensemble qui permettent d'évaluer la fiabilité et la résolution d'un ensemble de prévisions probabilistes.

Le chapitre 2 traite d'une méthode originale pour générer de manière automatique des ensembles de simulations numériques de la qualité de l'air. Cette méthode prend en compte toutes les incertitudes. Un ensemble d'une centaine de simulations photochimiques à l'échelle européenne est lancé sur toute l'année 2001. Les performances statistiques de différentes simulations d'ozone sont calculées. L'ensemble présente des membres assez différents et une grande variabilité.

Au chapitre 3, on s'intéresse au problème de calibration d'un ensemble pour l'estimation de l'incertitude et la prévision des risques de dépassement de seuil. La méthode de calibration se fonde sur l'ensemble de simulations, précédemment généré, un score d'ensemble tel que la variance du diagramme de rang ou du diagramme de fiabilité, et un algorithme d'optimisation combinatoire. La calibration consiste à sélectionner, via l'algorithme d'optimisation, un sous-ensemble qui est très performant pour le score d'ensemble sélectionné. Des champs d'incertitude, fournis par l'écart type empirique des sous-ensembles calibrés pour le diagramme de rang, sont ensuite calculés. Les robustesses spatiale et temporelle de la méthode sont évaluées. Les ensembles calibrés permettent de produire des prévisions probabilistes plus précises et de prévoir les incertitudes.

Dans le chapitre 4, on compare deux ensembles : l'un généré avec l'approche décrite dans le deuxième chapitre (prenant en compte toutes les sources d'incertitudes), et l'autre avec la méthode Monte Carlo. Des calibrations sont réalisées pour les deux ensembles de simulations photochimiques. Des champs d'incertitude (variances et co-variances) sont analysés. Ensuite, une régression linéaire est réalisée entre les champs d'entrée perturbés et les simulations d'ozone dans le but de quantifier l'impact des incertitudes des champs d'entrée sur les simulations d'ozone. Enfin, j'estime la part de l'erreur de mesure (liée aux instruments de mesure), l'erreur de modélisation et l'erreur de représentativité dans l'écart entre les observations et les simulations.

Enfin, dans le chapitre 5, une génération d'ensembles a été réalisée pour deux domaines régionaux (Île-de-France et Pays de la Loire) dans le but d'effectuer une étude d'impact autour des centrales thermiques de Porcheville et Cordemais. L'ensemble de simulations photochimiques a été lancé sur l'année 2007 avec et sans les émissions des centrales. La méthode de calibration pour l'estimation d'incertitude et la prévision des risques de dépassement de seuil est appliquée pour ces deux régions et différentes espèces chimiques : O_3, NO_2, SO_2.

Chapitre 1

Physique et modélisation de la qualité de l'air, incertitudes et ensemble de prévisions

Ce chapitre présente succinctement les phénomènes physiques liés à la qualité de l'air et le modèle de chimie-transport qui en découle, plus particulièrement pour des prévisions photochimiques en phase gazeuse. Les différentes sources d'incertitude dans les simulations de la qualité de l'air sont ensuite introduites. La deuxième partie traite de l'utilisation des ensembles de prévisions et de leur utilité dans l'amélioration des prévisions ou dans l'estimation des incertitudes. La dernière partie est consacrée aux différents scores existants pour évaluer un ensemble de prévisions dans le cadre d'une approche probabiliste.

Sommaire

1.1 Simulation numérique pour la qualité de l'air

1.1.1 Phénomènes physiques et chimie de l'atmosphère

L'atmosphère, qui s'étend jusqu'à plus de 500 km, est composée de plusieurs couches. Ces couches, de différentes épaisseurs, sont définies par le profil vertical de la température. Les profils verticaux de température dans les différentes couches de l'atmosphère sont donnés sur la figure 1.1. Ces couches sont :

- la troposphère, dont l'épaisseur varie de 8 à 18 km, selon la saison et la latitude. La température y décroît avec l'altitude ;
- la stratosphère qui se situe au-dessus de la troposphère et s'étend jusqu'à une cinquantaine de kilomètres. La température y augmente avec l'altitude. Ce phénomène s'explique par l'absorption du rayonnement solaire par l'ozone ;
- la mésosphère qui est au-dessus de la stratosphère et est épaisse de 30 km environ ;
- la thermosphère qui s'étend à plus de 500 km.

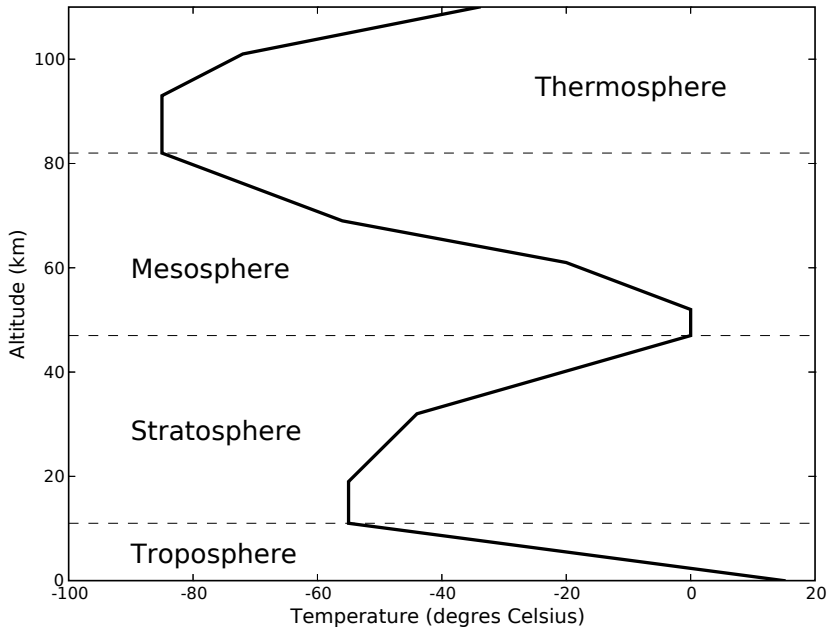


FIGURE 1.1 – Profils verticaux typiques de la température et principales couches atmosphériques. Source : Mallet [2005].

En qualité de l'air, on s'intéresse aux concentrations atmosphériques des polluants qui peuvent avoir un impact sur les populations ou la végétation. Ainsi, seules les concentrations au voisinage du sol sont retenues, c'est-à-dire à quelques mètres d'altitude. Les échanges d'espèces chimiques entre la troposphère et la stratosphère sont assez lents pour qu'on puisse, aux échelles régionales, considérer uniquement l'étude du transport des polluants dans la troposphère. On considère alors que les espèces chimiques évoluent dans la première couche de l'atmosphère — la troposphère — et plus particulièrement dans une basse couche de cette dernière : la couche limite atmosphérique.

Cette couche, appelée plus communément « couche limite », se définit selon Stull [1988] comme la partie de la troposphère qui est influencée en une heure ou moins par des changement au voisi-

nage du sol — émissions de polluants, réchauffement, évaporation, ... L'épaisseur de la couche limite peut varier entre quelques mètres et quelques kilomètres. Elle dépend essentiellement du cisaillement du vent pendant la nuit et du réchauffement au sol par rayonnement solaire pendant la journée.

Le transport des espèces chimiques évoluant dans la couche limite se fait de deux façons : (1) un transport par le vent, appelé vent « moyen », dont la vitesse atteint en moyenne quelques mètres par seconde et qui est la source principale du transport horizontal — les vents verticaux étant très faibles, ils sont négligeables par rapport à la turbulence — et (2) la turbulence qui est la source du transport vertical. La turbulence représente des mouvements turbulents de masse d'air qui sont principalement générés par le cisaillement du vent, le réchauffement du sol par rayonnement solaire et la convection nuageuse. Ces déplacements turbulents, importants durant la journée, brassent et mélangent les polluants se situant dans la couche limite.

Parmi toutes les espèces chimiques, on distingue les espèces primaires des espèces secondaires. Les premières sont (1) d'origine anthropique, c'est-à-dire dues aux activités humaines telles que le trafic, les émissions de cheminées industrielles, ou (2) issues de la biomasse — la végétation, les volcans, les feux de forêt, ... Au contraire, les espèces secondaires se distinguent du fait qu'elles ne sont pas émises mais sont issues de réactions chimiques, comme l'ozone par exemple.

Ces espèces transportées et mélangées dans la couche limite peuvent réagir entre elles pour disparaître et créer d'autres espèces. La durée de vie d'une espèce peut être de l'ordre de quelques secondes à plusieurs mois. La figure 1.2 représente des ordres de grandeur de temps de résidence de quelques espèces chimiques atmosphériques. Ces temps de résidence peuvent dépendre :

- de la vitesse avec laquelle l'espèce considérée va réagir avec d'autres espèces ;
- de leur capacité à déposer rapidement sur le sol ou dans la mer ;
- de leur solubilité dans les gouttes d'eau et nuages.

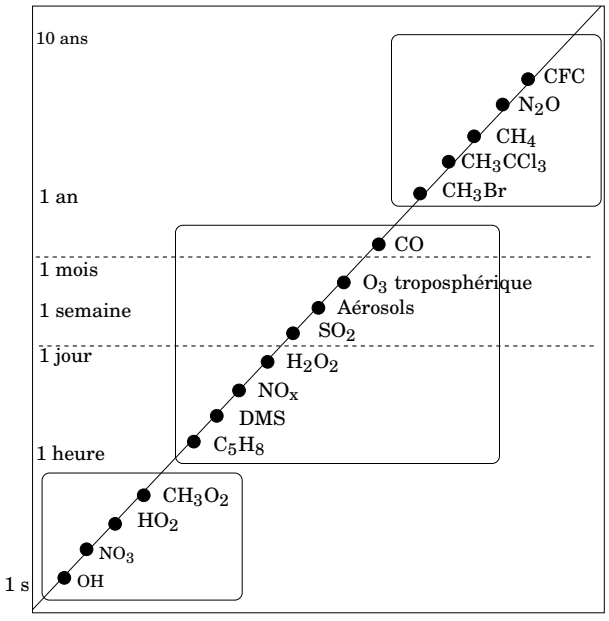


FIGURE 1.2 – Ordre de grandeur du temps de résidence des principales espèces atmosphériques. Source : Sportisse [2008].

Réaction chimique

Les réactions entre les espèces troposphériques sont multiples et se déroulent à des échelles de temps très différentes. La cinétique de certaines réactions chimiques dépend parfois de l'ensoleillement, ce sont les réactions photolytiques. Dans le cycle de production de l'ozone par exemple, l'oxygène O intervenant dans la réaction R 1.1 est issu de la réaction photolytique R 1.2,

$$O_2 + O + M \longrightarrow O_3 + M \tag{R 1.1}$$

où M est un tiers corps — en général N_2 ou O_2.

$$NO_2 + h\nu \xrightarrow{J_{NO_2}} NO + O \tag{R 1.2}$$

Le terme $h\nu$ correspond à la quantité de lumière venant dissocier la molécule NO_2. Au contraire, d'autres réactions viennent consommer l'ozone. La réaction R 1.3 correspond à la titration de l'ozone par le monoxyde d'azote, et la réaction R 1.4 correspond à la photolyse de l'ozone, où O^{1D} est un atome d'oxygène excité.

$$O_3 + NO \longrightarrow NO_2 + O_2 \tag{R 1.3}$$

$$O_3 + h\nu \longrightarrow O_2 + O^{1D} \tag{R 1.4}$$

D'autres espèces entrent en jeu dans le cycle de production de l'ozone, notamment les composés organiques volatils (COV), qui sont des composés de carbone et d'hydrogène notamment émis par les transports et l'industrie.

Dépôts sec et humide

Le phénomène d'absorption de polluants par le sol, la végétation ou la mer est appelé « dépôt sec ». Il dépend de l'espèce considérée, des conditions météorologiques, du type de sol, de la saison et de la solubilité de cette espèce dans le cas de dépôt au-dessus des masses d'eau. Il constitue le principal terme de perte des polluants troposphériques.

Un autre phénomène de dépôt existe et dépend essentiellement de la solubilité des espèces. Une espèce très soluble pourra pénétrer plus facilement dans les gouttes de pluie ou dans les nuages. Lors de pluies, les polluants piégés dans les gouttes sont précipités au sol. Ce transfert de masse vers la phase aqueuse est qualifié de « dépôt humide » ou de « lessivage ».

La compréhension de ces phénomènes complexes est nécessaire pour modéliser et prévoir le transport des polluants et leurs concentrations au voisinage du sol.

1.1.2 Modèles de chimie-transport

Le modèle mathématique de chimie-transport découle d'un système d'équations aux dérivées partielles décrit dans 1.1 et appelé plus communément équation d'advection–diffusion–réaction.

$$\frac{\partial c_i}{\partial t} = -\underbrace{\operatorname{div}(V c_i)}_{\text{advection}} + \underbrace{\operatorname{div}\left(\rho K \nabla \frac{c_i}{\rho}\right)}_{\text{diffusion}} + \underbrace{\chi_i(c,t)}_{\text{réaction}} + S_i - P_i\,, \tag{1.1}$$

où c_i est la concentration de l'espèce i, V le champs de vent, K la matrice de diffusion, χ_i les réactions chimiques associées à l'espèce considérée, S_i le terme source et P_i le terme de perte.

Cette équation décrit le cycle de vie d'une espèce chimique i et représente les phénomènes précédemment décrits. On suit l'évolution spatio-temporelle de la concentration de cette espèce émise par la végétation ou par une activité humaine, terme source S_i, ou issue de réactions chimiques, le terme non-linéaire χ_i. Il vient ensuite les termes de transports horizontal et vertical : le terme d'advection $\operatorname{div}(V c_i)$ constitue le terme du transport horizontal par le vent, le terme de diffusion $\operatorname{div}(\rho K \nabla \frac{c_i}{\rho})$ régit le transport vertical et enfin, le terme de perte P_i concerne le dépôt humide. Pour un souci de lisibilité, certaines dépendances ont été omises. Notamment la dépendance en temps du vent V, de la matrice de diffusion K, des termes de source/perte ou de dépendance aux champs météorologiques des termes χ_i et P_i.

L'équation est accompagnée d'une condition aux limites au sol :

$$\rho K \nabla \frac{c_i}{\rho} \cdot n = D_i - E_i\,, \tag{1.2}$$

où n est la normale au sol orientée dans le sens des altitudes croissantes, D_i est un flux dû au dépôt sec et E_i est un flux dû aux émissions de surface.

Terme de transport

Le terme de transport advection découle de l'équation 1.3 où $\widetilde{c}$ représente un champ tridimensionnel de concentration pour une espèce chimique donnée et $\widetilde{V}$ est le vent.

$$\frac{\partial \widetilde{c}}{\partial t} = -\operatorname{div}(\widetilde{V}\widetilde{c}) \tag{1.3}$$

Sachant qu'il est impossible de décrire toutes les échelles numériquement, il est nécessaire de travailler avec des quantités moyennes. Le vent peut être décomposé en un vent moyen plus des fluctuations de moyenne nulle (décomposition de Reynolds). Soit V le vent moyen et V' les fluctuations. On a alors $\widetilde{V} = V + V'$. Dans un souci de simplicité, on appelle simplement « vent » le vent moyen V. Cette convention est valable pour toute la suite. Une décomposition équivalente est effectuée sur le champ de concentration $\widetilde{c}$: $\widetilde{c} = c + c'$. En moyennant l'équation 1.3, il vient

$$\frac{\partial c}{\partial t} = -\operatorname{div}(V c) - \operatorname{div}(\overline{V'c'})\,, \tag{1.4}$$

où $\overline{V'c'}$ est la moyenne de $V'c'$. Afin de fermer cette équation, on approche classiquement $\overline{V'c'}$ ainsi :

$$\overline{V'c'} \simeq -K \nabla c\,, \tag{1.5}$$

où K représente la matrice de diffusion turbulente. Cette fermeture d'équation est souvent qualifiée « théorie K ». Il est toutefois nécessaire de respecter l'équation de continuité moyenne qui s'écrit

$$\frac{\partial \rho}{\partial t} = -\mathrm{div}(\rho V), \tag{1.6}$$

où ρ représente la densité de l'air. On souhaite naturellement décrire l'évolution de c de la même manière que pour la densité du fluide porteur ρ. Dans le cas présent, si l'on pose $c = \rho$ et que l'on injecte l'équation 1.5 dans l'équation 1.4, on ne retrouve pas l'équation 1.6. Il est alors nécessaire de modifier l'approximation 1.5 par

$$\overline{V'c'} \simeq -\rho K \nabla \frac{c}{\rho}. \tag{1.7}$$

Au final, l'équation de transport s'écrit

$$\frac{\partial c}{\partial t} = -\mathrm{div}(Vc) + \mathrm{div}\left(\rho K \nabla \frac{c}{\rho}\right). \tag{1.8}$$

Mécanisme chimique

Le terme χ_i se définit comme étant un mécanisme chimique. C'est un terme non-linéaire qui permet de prendre en compte un certain nombre d'espèces chimiques ou de groupe d'espèces chimiques avec leurs réactions associées. On s'intéresse particulièrement aux mécanismes qui prennent en compte le cycle d'ozone. Deux mécanismes chimiques ont été utilisés durant cette thèse. Il s'agit de RADM2 [Stockwell *et al.*, 1990] et RACM [Stockwell *et al.*, 1997]. Le premier prend en compte 61 espèces et 157 réactions. Quand au second, il y a 72 espèces et 237 réactions. On notera les travaux de Gross et Stockwell [2003] qui comparent ces deux mécanismes.

Paramétrisations

Plusieurs coefficients intervenant dans l'équation de transport réactif tels que la matrice de diffusion K ou le terme de perte P_i sont décrits par des paramétrisations physiques. On peut citer par exemple le calcul du coefficient de diffusion verticale K_z par les paramétrisations de Louis [1979] ou de Troen et Mahrt [1986]. D'autres paramétrisations physiques n'apparaissant pas directement dans l'équation (1.1) sont nécessaires, notamment pour le calcul des taux de photolyse — paramétrisations de l'humidité relative critique et de l'atténuation nuageuse. Ces différentes paramétrisations physiques, largement utilisées dans le domaine de la qualité de l'air et plus particulièrement en photochimie, sont décrites plus précisément dans la section 2.2.

Données d'entrée

Les derniers coefficients apparaissant dans l'équation 1.1 concernent les données d'entrée : champs météorologiques, données d'émission, conditions aux limites, occupation des sols, ...

La prise en compte de tous ces éléments — nombreuses espèces chimiques, champs météorologiques, non linéarité — font du modèle de chimie-transport un système complexe de grande dimension. Il n'est pas rare, comme pour d'autres modèles environnementaux, d'avoir plusieurs centaines de milliers de variables à traiter à des échelles de temps et à des résolutions spatiales très différentes. Il est alors indispensable de faire des choix, en ce qui concerne les approximations

numériques et les paramétrisations physiques, dans le but d'avoir un modèle de simulation opérationnel.

Polyphemus

Tous les résultats de cette thèse ont été réalisés avec Polyphemus [1]. C'est une plate-forme logicielle dédiée à la qualité de l'air, sous licence libre GNU GPL [2] et développée en commun par l'École des Ponts ParisTech, l'INRIA [3] et EDF R&D, au sein du CEREA [4]. Elle inclut des modèles de chimie-transport eulérien, gaussien ou lagrangien permettant de calculer la dispersion de différentes espèces comme les radionucléides, des traceurs inertes, des polluants en phase gazeuse ou encore des aérosols. Ces différents modèles permettent de prévoir la dispersion de polluants dans l'atmosphère à différentes échelles : de quelques centaines de mètres dans le cas d'études près d'une cheminée industrielle ou d'un rejet accidentel, de quelques kilomètres à l'échelle d'une ville et de plusieurs dizaines et centaines de kilomètres pour les échelles régionale et continentale.

S'ajoutent aux différents modèles existants des méthodes qui permettent d'améliorer les prévisions — à l'aide de méthodes d'assimilations de données [Wu *et al.*, 2008] ou de prévisions d'ensemble [Mallet *et al.*, 2009] — de retrouver les sources d'émission [Quélo *et al.*, 2005] ou d'estimer les incertitudes [Mallet et Sportisse, 2006b].

Le choix de ces différents modèles auxquels il est possible de « brancher » les méthodes citées précédemment ne se fait pas sans une conception orientée objet de la plate-forme, écrite en C++, Fortran et Python. Cette conception offre une forte flexibilité tant aux développeurs qu'aux utilisateurs durant toute la chaînes de calcul. On pourra se rapporter à Mallet *et al.* [2007b] et plus brièvement à la section 2.3.2 pour la description technique de cette plate-forme de modélisation.

Durant cette thèse et sauf mention contraire, nous avons utilisé le solveur numérique Polair3D [Boutahar *et al.*, 2004] pour simuler la dispersion de polluants en phase gazeuse à l'échelle continentale. On pourra se référer à Verwer *et al.* [1998] et Sportisse et Mallet [2006] pour une description plus détaillée des schémas numériques utilisés pour l'advection et la chimie dans Polair3D.

Objectifs des simulations

La plupart des polluants sont réglementés : des niveaux de concentration (plus ou moins moyennés) sont définis comme des limites au-delà desquelles des mesures doivent être prises. Pour l'ozone par exemple, les pouvoirs publics sont tenus de prévenir la population lorsque le seuil dit « d'information » (appelé aussi « protection »), fixé à $180\,\mu g\,m^{-3}$, est dépassé. Ils doivent prendre des mesures pour réduire la pollution lorsque le seuil d'alerte, fixé à $240\,\mu g\,m^{-3}$, est dépassé. Ces seuils sont issus de la réglementation européenne. Afin d'anticiper des prises de décision, il est nécessaire de se donner les moyens de prévoir les concentrations de polluants, en ayant recours à la modélisation numérique notamment.

Des simulations photochimiques à l'échelle européenne ou nationale peuvent être réalisées afin de prévoir les concentrations de polluants tels que l'ozone, le SO_2 ou les NO_x à des échéances

1. http://cerea.enpc.fr/polyphemus/
2. http://www.gnu.org/licenses/gpl.html
3. Institut national de recherche en informatique et en automatique
4. Centre d'enseignement et de recherche en environnement atmosphérique

de deux à trois jours. Au-delà de cette échéance, les prévisions météorologiques ne sont pas assez fiables pour avoir des prévisions pertinentes.

Des exemples de résultats pour la prévision d'ozone sont donnés sur la figure 1.3. La figure 1.3(a) correspond à une prévision d'ozone sur l'Europe pour la date du premier avril 2001 à 14 heures TU (Temps Universel — en anglais, *Universal Time Coordinate*, UTC). Elle apporte des informations précises et utiles aux prévisionnistes : la localisation de forte pollution, les concentrations prévues, ... On constate qu'elle contient des structures spatiales non triviales. Il est aussi intéressant de noter que la concentration d'ozone près des zones urbaines est moins élevée à cause de la titration de l'ozone par le monoxyde d'azote émis — voir la réaction R 1.3. La figure 1.3(b) est un profil journalier typique de concentration d'ozone. Le minimum est atteint pendant la nuit, lorsque la hauteur de la couche limite est au plus bas et qu'il n'y a pas de transport vertical. Le maximum est quant à lui atteint aux heures d'ensoleillement les plus importantes. Le rayonnement solaire favorise la production de précurseurs d'ozone par photolyse et augmente la concentration d'ozone au voisinage du sol grâce à un mélange vertical plus important dû à la turbulence. En effet, dans les premiers mètres de l'atmosphère, le profil vertical de concentration d'ozone est croissant avec l'altitude. Une forte turbulence engendre donc une concentration d'ozone plus importante au voisinage du sol. Enfin, la figure 1.3(c) compare les résultats d'une simulation de pics journalier d'ozone avec les données observées à une station de mesure. Même si la simulation n'est pas parfaite, les deux courbes ont des tendances assez proches. L'écart entre les observations et les concentrations simulées est principalement dû aux incertitudes liées aux modèles et aux données d'entrée utilisées.

Il est important de tenir compte des incertitudes liées aux modèles déterministes que l'on utilise et aux données d'entrée associées. Il est donc nécessaire de connaître la source de ces incertitudes et de les quantifier.

1.1.3 Incertitudes

Il existe différentes sources d'incertitude dans un modèle de chimie-transport :

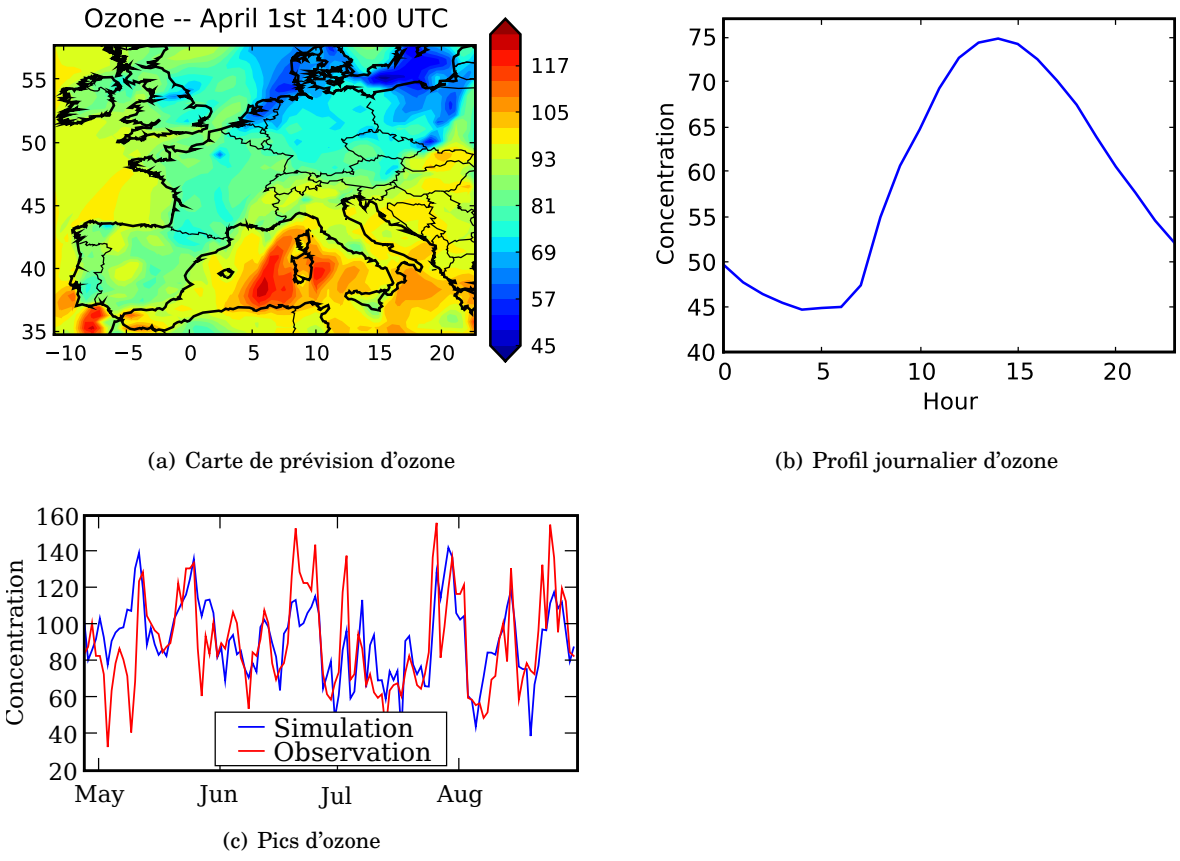
- les incertitudes liées aux données d'entrée tels que les champs météorologiques, les données d'émissions ou les conditions aux limites ;
- les incertitudes liées aux paramétrisations physiques tels que le calcul de la composante verticale de la matrice de diffusion K nécessaire pour la description du transport vertical des espèces chimiques ; le calcul de la vitesse de dépôt ou le mécanisme chimique ;
- les approximations numériques tels que les schémas numériques utilisés pour la résolution numérique de l'équation 1.1, la résolution verticale, le pas de temps.

Données d'entrée

Plusieurs données d'entrée sont indispensables pour lancer une simulation photochimique. Certaines données comme les champs météorologiques ou les conditions aux limites sont issues de modèles. Ces modèles, qui ont parfois des résolutions spatiales très différentes de celles utilisées pour nos études, sont entachés de leurs propres incertitudes qui ont un impact direct sur les résultats.

Les données d'émission peuvent être fournies par des modèles ou des campagnes de mesure. EMEP[5], par exemple, décompose les quantités émises de polluants par secteurs d'activité ap-

5. *European Monitoring and Evaluation Program* – http://www.emep.int/



(a) Carte de prévision d'ozone

(b) Profil journalier d'ozone

(c) Pics d'ozone

FIGURE 1.3 – Carte de prévision d'ozone (a), profil journalier moyen (b) et pics d'ozone à la station de mesure de Saint-Nazaire (c) ($\mu g m^{-3}$).

pelés SNAP[6] — cela concerne le type d'industrie comme les centrales nucléaires, les raffineries de pétrole, ou simplement les émissions dues aux transports routiers et à l'agriculture. Ces émissions sont données en quantité totale annuelle par polluant (CO, NH_3, SO_x), par SNAP et pour chaque grille du domaine considéré. Des hypothèses sont alors faites pour estimer la quantité des polluants émis à chaque heure de la journée en fonction du jour de la semaine — jour de travail ou week-end — et du mois de l'année. Des cadastres d'émission plus précis existent pour des résolutions spatiales plus fines : sources ponctuelles, prise en compte du trafic, ... Il est très difficile d'obtenir des données précises sur toutes les sources d'émission. Un autre point essentiel concerne l'altitude à laquelle sont émis les polluants, surtout dans le cas de cheminées industrielles. En effet, en fonction de la direction du vent et de la hauteur d'émission, le panache de polluants peut se disperser de manière très différente. L'estimation des données d'émission est très importante dans le modèle de chimie-transport puisque c'est elle qui va orienter la dynamique de la chimie.

D'autres données, comme la couverture au sol, donne le type d'occupation du sol par classification — zone urbaine, champs, forêt, mer, montagne — ainsi que leur localisation. Ces données interviennent entre autre dans le calcul de la vitesse de dépôt.

Russell et Dennis [2000] présente une liste de données d'entrée utilisées dans les modèles de chimie-transport avec leurs incertitudes. Ils tirent ces données de plusieurs articles traitant de l'estimation des incertitudes dans les données d'entrée. À titre d'exemple, Hanna *et al.* [1998] donne un intervalle d'incertitude à 95% de $\pm 30^\circ$ pour la direction du vent ou encore $\pm 80\%$ pour les émissions de COV — pour une résolution horizontale de 5×5 km dans la région près de New-York.

En plus d'être incertaines de manière intrinsèque, les données d'entrée des modèles de chimie-transport sont à la fois multiples et parfois interdépendantes : (1) quelles données d'émission choisir à quelle résolution, quels modèles météorologiques, quelles conditions aux limites ; (2) les émissions biogéniques vont dépendre de la météo et de la couverture au sol par exemple.

Paramétrisations physiques

Les paramétrisations physiques interviennent pour deux raisons : (1) une méconnaissance des phénomènes physiques qui sont modélisés à partir de mesures et de formulations empiriques et (2) par la nécessité de représenter des phénomènes physiques qui ont une résolution plus fine que la résolution spatiale du modèle choisie et qui ne peuvent pas être représentés directement dans le modèle lui-même. Les incertitudes liées à ces paramétrisations interviennent lors du calcul du coefficient de diffusion verticale, du mécanisme chimique, du calcul de la vitesse de dépôt ou encore dans le calcul de la couverture nuageuse. Ces incertitudes ne sont *a priori* pas connues. Il est néanmoins possible d'estimer les incertitudes liées aux paramétrisations physiques. Dans Mallet et Sportisse [2006b], il est montré qu'il y a des fortes incertitudes liées au calcul du coefficient de diffusion verticale et au mécanisme chimique.

Rentrons plus en détail dans un exemple concret. Il a été mentionné plus tôt que la composante verticale K_z de la matrice de diffusion turbulente est essentielle dans le processus de modélisation — la diffusion horizontale étant quant à elle faible par rapport aux transports dus aux vents. Cette matrice intervient dans le terme de diffusion $\mathrm{div}(\rho K \nabla \frac{c}{\rho})$ de l'équation 1.1 et découle de la fermeture de l'équation 1.4 décrite dans la section 1.1.2. Deux paramétrisations physiques pour l'estimation de ce coefficient de diffusion verticale (en $m^2 s^{-1}$) sont proposées dans Polyphemus et largement utilisées dans les modèles de chimie-transport :

6. *Selected Nomenclature for Air Pollution*

– la première est de Louis [1979] et se décrit comme suit :

$$K_z = L^2 F(R_i) \left[\left(\frac{\Delta U}{\Delta z} \right)^2 + \left(\frac{\Delta V}{\Delta z} \right)^2 \right], \tag{1.9}$$

où L est la longueur de mélange, R_i le nombre de Richardson, F la fonction de stabilité et le couple (U,V) sont les composantes horizontales du champ de vent ;

– la deuxième est donnée par Troen et Mahrt [1986] :

$$K_z = u_* \kappa z \Phi_m^{-1} \left(1 - \frac{z}{PBLH} \right)^p, \tag{1.10}$$

où u_* correspond à la vitesse de friction, κ est la constante de von Kármán, Φ_m est un nombre sans dimension et correspond au cisaillement, et $PBLH$ signifie *Planetary Boundary Layer Height*, donc la hauteur de la couche limite atmosphérique décrite dans la section 1.1.1. L'exposant p vaut la plupart du temps entre 2 ou 3.

Il est à noter que les deux paramétrisations dépendent essentiellement de données météorologiques : stabilité de l'atmosphère, champs de vent, hauteur de la couche limite, ... La paramétrisation de Troen&Mahrt est plus paramétrique et plus robuste que celle de Louis. Les incertitudes liées aux données météorologiques peuvent avoir un impact plus important sur la paramétrisation de Louis.

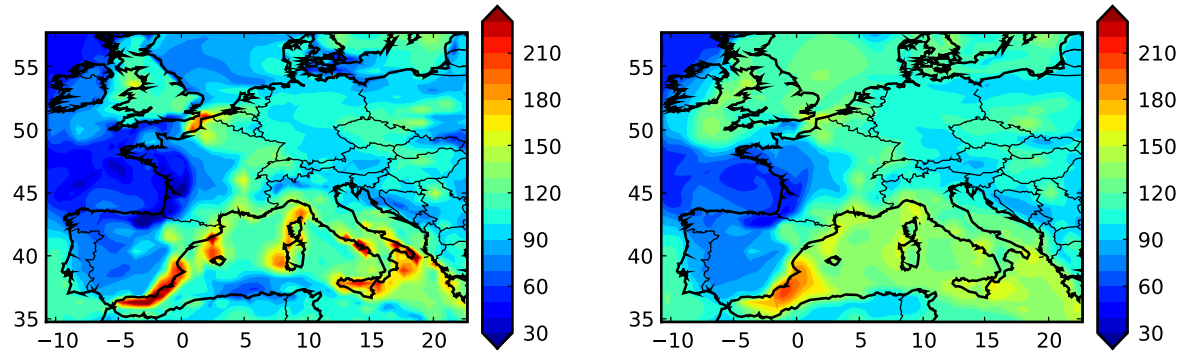
Le calcul de ce coefficient a un impact direct sur la distribution verticale des polluants puisqu'il va quantifier la transition des polluants d'un niveau vertical à l'autre. Les concentrations des polluants au voisinage du sol sont très sensibles à ce coefficient. Afin de montrer l'impact du calcul du coefficient de diffusion verticale sur la concentration d'un polluant tel que l'ozone, deux simulations photochimiques à l'échelle européenne ont été lancées sur toute l'année 2001.

Le premier résultat concerne les différences entre les moyennes globales de K_z pour les deux paramétrisations différentes ainsi que pour les concentrations d'ozone. La moyenne du K_z calculé via la paramétrisation de Louis vaut $2.68\,\mathrm{m^2 s^{-1}}$ tandis que le K_z calculé avec Troen&Mahrt est égal à $7.17\,\mathrm{m^2 s^{-1}}$. Ces deux moyennes ont été calculées pour les valeurs du coefficient de diffusion verticale à la première interface, donc à une hauteur de 50 m. Les concentrations moyennes d'ozone au voisinage du sol valent $66.9\,\mu\mathrm{g\,m^{-3}}$ pour la première et $83.4\,\mu\mathrm{g\,m^{-3}}$ pour la seconde. De plus, la variabilité de l'ozone est plus importante dans le premier cas que dans le second puisque l'écart type spatio-temporel est égal $31.6\,\mu\mathrm{g\,m^{-3}}$ contre $26\,\mu\mathrm{g\,m^{-3}}$.

Le second résultat concerne les champs de prévision d'ozone. La figure 1.4 montre deux cartes de prévision pour le 4 juillet 2001 à 14 heures UTC. Les champs d'ozone dont le coefficient de diffusion verticale a été calculé avec Louis puis Troen&Mahrt sont respectivement les cartes 1.4(a) et 1.4(b). Même si les minima et maxima se trouvent approximativement aux mêmes endroits, c'est-à-dire respectivement à l'ouest et au sud de l'Europe, les champs montrent des différences notables. Les concentrations d'ozone sont bien plus élevées le long des côtes de la Méditerranée et au nord de la France pour K_z Louis même si la moyenne du champ est plus faible de $10\,\mu\mathrm{g\,m^{-3}}$ que pour le champ d'ozone avec K_z Troen&Mahrt.

Par ailleurs, la figure 1.5 présente les coefficients de diffusion verticale à la première interface, c'est-à-dire 50 m, pour la même date que précédemment. Le champ fourni par Troen&Mahrt est globalement plus élevé que celui de Louis. On notera une structure spatiale un peu différente entre les deux champs. Les valeurs plus élevées du champ K_z par Troen&Mahrt à la première interface permettent un meilleur mélange des espèces chimiques et augmentent ainsi la concentration d'ozone au voisinage du sol.

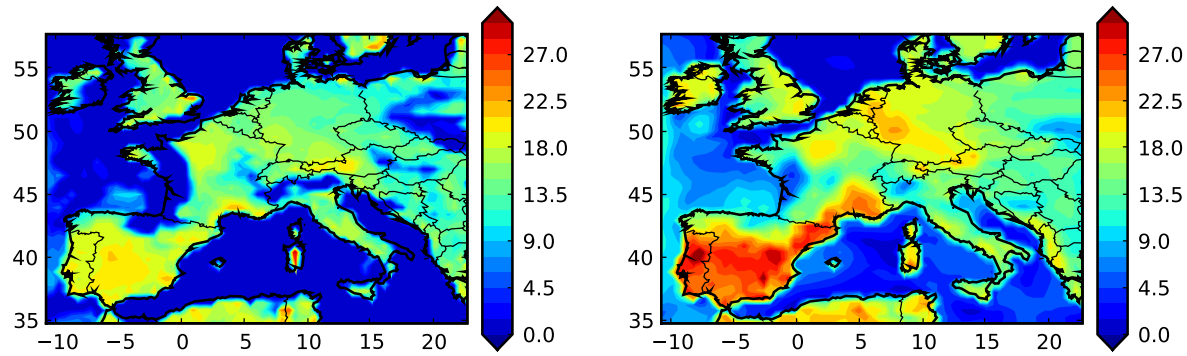
L'écart d'environ $15\,\mu\mathrm{g\,m^{-3}}$ entre les moyennes globales d'ozone et la différence des champs d'ozone montrent l'impact que peut avoir le choix de l'une ou l'autre des paramétrisations physiques



(a) Ozone avec K_z Louis

(b) Ozone avec K_z Troen&Mahrt

FIGURE 1.4 – Cartes de prévision d'ozone ($\mu g\,m^{-3}$) pour le 4 juillet 2001 14 heures UTC avec deux paramétrisations physique de K_z différentes : Louis (a) et Troen&Mahrt (b).



(a) K_z Louis

(b) K_z Troen&Mahrt

FIGURE 1.5 – Champs de K_z à la première interface (50 m) issues de deux paramétrisations différentes : Louis et Troen&Mahrt (m^2s^{-1}).

du calcul du coefficient de diffusion verticale. Cela met en évidence l'incertitude liée à cette paramétrisation physique.

Il en va de même pour toutes les autres paramétrisations physiques avec un degré d'impact différent sur les concentrations de sortie. On peut citer par exemple les paramétrisations pour le calcul de l'atténuation nuageuse ou de la vitesse de dépôt de chaque espèce. Tout comme la paramétrisation du coefficient de diffusion verticale, (1) ces dernières peuvent dépendre les unes des autres — le calcul des taux de photolyse dépend de l'atténuation nuageuse qui dépend elle-même du calcul de l'humidité relative critique qui va enfin dépendre des différents champs météorologiques ; (2) il existe plusieurs choix possibles pour calculer ces paramètres. Il en existe deux pour le calcul de l'atténuation et deux autres pour le calcul de l'humidité relative critique. Un tableau récapitulatif de toutes les alternatives est donné à la section 2.2.3 dans le cadre de simulations photochimiques.

En plus de dépendre de certaines données d'entrée, le calcul de ces multiples paramètres peut changer en fonction des choix sur les approximations numériques. On peut citer principalement le choix de la résolution verticale ou la hauteur de première couche.

La dernière source d'incertitude concerne justement les approximations numériques. De la même manière que les deux sources d'incertitude précédentes, les différentes approximations utilisées dans un modèle de chimie-transport résultent de choix, imposés par les problèmes liés aux différentes échelles spatio-temporelles, à la grande dimension du problème et par la nécessité d'avoir un modèle opérationnel.

Approximations numériques

La résolution numérique de l'équation 1.1 découle de choix de modélisation et de différents schémas numériques. On distingue alors le choix d'une discrétisation — pas de temps, résolutions verticale et horizontale — et le choix des schémas numériques. Le choix de ces approximations numériques est contraint par le temps de calcul et les moyens mis à disposition — rappelons que les choix de modélisation doivent être faits dans un souci de prévision — et par la précision voulue des résultats.

Des développements et des évaluations de schémas numériques ont donc été effectués dans le but de résoudre numériquement l'équation de chimie-transport de manière à la fois raisonnablement précise et efficace [Verwer *et al.*, 1998]. D'autres études sont plus spécifiques et concernent la séparation d'opérateurs [Sportisse, 2000; Lanser et Verwer, 1999] ou les problèmes raides [Verwer *et al.*, 1999; Sandu *et al.*, 1997b,a].

Il est très difficile de connaître l'incertitude liée aux schémas numériques. Une étude de sensibilité peut être réalisée en comparant le choix de différents schémas numériques sur la sortie du modèle. Mallet [2005] et Mallet *et al.* [2007a] réalisent de telles études avec des simulations à l'échelle européenne et présentent des résultats concernant la séparation d'opérateur, l'intégration de la chimie, l'advection ou encore le pas de temps d'intégration et leur impact sur plusieurs polluants : O_3, NO, NO_2, SO_2 et HO qui ont chacun des temps de réaction et une répartition spatiale très différents — certaines espèces sont localisées près sources d'émission comme le NO. Les résultats de cette étude mettent en évidence l'influence du pas de temps et du schéma d'advection sachant que les problèmes liés à la séparation d'opérateur sont moins importants. Il est à noter que les espèces ne sont pas toutes sensibles aux mêmes schémas. Néanmoins, il est important de comparer l'impact de ces approximations numériques avec les autres incertitudes mentionnées précédemment. Mallet et Sportisse [2006b] reviennent justement sur les incertitudes liées aux paramétrisations physiques et aux approximations numériques et modèrent l'impact des schémas numériques tant l'incertitude liée aux paramétrisations physiques est grande.

Malgré les efforts pour l'amélioration des modèles de chimie-transport et la compréhension des phénomènes physiques dans le but de produire des simulations et des prévisions plus performantes, les incertitudes liées aux modèles eux-mêmes sont encore peu quantifiées. Plusieurs questions se posent : (1) comment prendre en compte toutes les sources d'incertitude dans les simulations de la qualité de l'air, (2) quelle confiance peut-on avoir dans les prévisions produites par nos systèmes, (3) quelles évaluations peuvent être mises en place pour mesurer la qualité des estimations d'incertitude ?

Une approche entièrement stochastique consisterait à introduire des variables aléatoires en lieu et place des variables et paramètres incertains dans l'équation 1.1. La résolution d'une telle équation reviendrait à estimer la distribution de probabilité des concentrations de polluants résultantes. Ceci nécessiterait une très bonne connaissances des incertitudes sur les champs d'entrée, les paramétrisations physiques et les approximations numériques. Or, si l'on peut obtenir des hypothèses sur les distributions de probabilités de quelques champs météorologiques ou des sources d'émission, il devient pour l'heure impossible d'obtenir de telles informations sur les paramétrisations physiques et autres approximations numériques. De plus, rappelons que nous sommes face à un problème non-linéaire de grande dimension, avec plus de 10^6 concentrations calculées, ce qui exclut la résolution de l'équation stochastique associée.

Une autre approche consiste à construire un ensemble de modèles déterministes afin de mieux prendre en compte ces sources d'incertitude, d'améliorer les prévisions et d'estimer les incertitudes liées aux modèles.

1.2 Ensemble de prévisions

Au vu des fortes incertitudes qui entrent en jeu dans les modèles de chimie-transport, il paraît indispensable de les prendre en compte lors des simulations et de ne pas reposer sur les simulations d'un unique modèle. De plus, à l'instar des travaux menés pour la prévision météorologique, il est utile de connaître et de quantifier la confiance que l'on a dans les prévisions. Ces objectifs peuvent reposer soit sur un unique modèle déterministe dont on perturbe les données d'entrée, soit sur un ensemble de plusieurs modèles différents.

1.2.1 Construction d'ensemble

L'utilisation d'ensemble de prévisions a d'abord été développée pour des prévisions météorologiques [Lorenz, 1963; Epstein, 1969b]. Cette démarche découle de la volonté de prendre en compte l'aspect chaotique des phénomènes atmosphériques modélisés de manière déterministe. Les ensembles sont construits à partir de perturbations des conditions initiales afin d'obtenir une série de trajectoires probables. Elle a ensuite été étendue à la dispersion atmosphérique [Straume *et al.*, 1998; Warner *et al.*, 2002]. L'approche pour la génération d'un ensemble dépend essentiellement des sources d'incertitudes que l'on souhaite prendre en compte, de la faisabilité et du coût de calcul qu'une telle opération peut engendrer.

Données d'entrée

Les données d'entrée sont multiples et entachées d'incertitudes — voir la section 1.1.3. L'idée la plus simple consiste donc à utiliser différentes sources de données d'entrée — différents champs météorologiques issus de plusieurs modèles, idem pour les conditions aux limites ou les sources d'émission — avec leur propres incertitudes. Chaque combinaison de données d'entrée donne lieu

à une simulation différente, le but étant de construire un ensemble de prévisions. La mise en place d'une telle méthode peut s'avérer lourde et contraignante. En effet, il est nécessaire d'avoir à disposition toutes les données et leurs modèles associés. La plupart des modèles tiers, comme les modèles météorologiques et les modèles globaux pour l'obtention des conditions aux limites par exemple, ne reposent pas sur les mêmes systèmes de coordonnées et ne calculent pas les mêmes champs. Un pré-traitement conséquent des données d'entrée est donc nécessaire et peut devenir fastidieux lorsque plusieurs modèles différents entrent en jeu.

Une autre solution est de considérer les données entrée comme étant des variables aléatoires dont les distributions de probabilité sont supposées connues. Un grand nombre de tirages aléatoires sont effectués sur les différentes données d'entrée et permettent de construire un ensemble de simulations dites de *Monte Carlo*. Une estimation de la distribution de probabilité des concentrations de polluant en sortie ainsi qu'une étude de sensibilité des incertitudes liées aux données d'entrée peuvent être ainsi réalisées.

Des simulations d'ensemble utilisant cette méthode ont déjà été réalisées en perturbant les champs météorologiques [Straume, 2001; Draxler, 2003; Warner *et al.*, 2002] dans le cadre d'études de dispersion. Pour les études avec simulations photochimiques, les principaux travaux sont Hanna *et al.* [1998], Hanna *et al.* [2001] et Beekmann et Derognat [2003]. Dans tous les cas, un grand nombre de données d'entrée sont perturbées. De plus, il faut que le nombre de simulations d'ensemble soit suffisamment important afin d'obtenir une convergence de plusieurs variables cibles. Ainsi, Hanna *et al.* [1998] perturbe plus d'une vingtaine de champs d'entrée et produit une cinquantaine de simulations, ce qui peut difficilement garantir la convergence. Beekmann et Derognat [2003] produisent quant à eux 500 simulations. Un exemple d'incertitude sur les données d'entrée, issue de Mallet [2005] et adapté de Hanna *et al.* [1998, 2001], est donné dans le tableau 1.1.

Une difficulté réside dans le coût de calcul d'un grand nombre de simulations. En conséquence, la plupart des études se concentrent sur une courte période de temps. Mallet [2005] lance 800 simulations photochimiques à l'échelle européenne sur 12 jours — période du 1er juillet au 11 juillet 2001. Le nombre de simulations est suffisant pour estimer les moyennes et les écart types de concentrations de polluants mais insuffisant pour estimer les densités de probabilités de ces derniers. De plus, il montre aussi que l'incertitude due aux données d'entrée est importante, mais moins forte que celle due aux paramétrisations physiques.

TABLE 1.1 – Exemple d'incertitude sur les données d'entrée dont la distribution est supposée lognormale. L'incertitude est donnée pour un intervalle de confiance à ~ 95%. L'écart type est donné sur le logarithme de la variable. Source : Mallet [2005]

Donnée	**Incertitude**	**Écart type**
Atténuation nuageuse	±30%	0.131
Vitesse de dépôt	±30%	0.131
Conditions aux limites	±20%	0.091
Émissions anthropogéniques	±50%	0.203
Émissions biogéniques	±100%	0.347
Constantes photolytiques	±30%	0.131

Cela montre qu'il est important de prendre en compte les incertitudes liées aux données d'entrée sans négliger les incertitudes liées aux paramétrisations et aux approximations numériques. Par ailleurs, les simulations Monte Carlo ne permettent pas de prendre en compte les incertitudes dues aux paramétrisations physiques et aux approximations numériques. C'est pourquoi

une approche multi-modèles telle que décrite dans la section suivante propose une approche plus complète du problème.

Multi-modèles

L'approche multi-modèles, comme son nom l'indique, consite à prendre en compte plusieurs modèles de chimie-transport dans le but de construire un ensemble de simulations. Cette approche est particulièrement intéressante puisque chaque modèle utilise sa propre combinaison de données d'entrée, d'approximations numériques et de paramétrisations physiques — ce qui considère toutes les sources d'incertitude identifiées précédemment. Russell et Dennis [2000] donne une liste non exhaustive de quelques modèles pour des simulations photochimiques à l'échelle régionale :

- EMEP [Simpson, 1992] ;
- LOTOS [Builtjes, 1992] ;
- un modèle de la NOOA [McKeen *et al.*, 1991] ;
- REM-CALGRID [Stern, 1994; Stern *et al.*, 2003] ;

auxquels on peut encore ajouter :

- CHIMERE [Vautard *et al.*, 2001; Schmidt *et al.*, 2001] ;
- TM5 [Krol *et al.*, 2005] ;
- MATCH [Andersson *et al.*, 2007].

Des études regroupant plusieurs de ces modèles permettent à la fois de réaliser des comparaisons entre ceux-ci [van Loon *et al.*, 2007] ou d'améliorer les prévisions — via le calcul de la moyenne d'ensemble ou des méthodes statistiques comme nous les verrons par la suite dans la partie 1.2.2 [Galmarini *et al.*, 2004b,a; Pagowski *et al.*, 2005; Delle Monache et Stull, 2003].

Peu d'articles proposent de construire de tels ensembles afin d'étudier et d'estimer les incertitudes [Vautard *et al.*, 2009]. La principale difficulté réside dans le fait de réunir un grand nombre de modèles variés. En effet, chacun travaille sur ses propres grilles de calculs, se base sur des modèles météorologiques différents. De plus, les formats des données en entrée et en sortie sont dissemblables et le fait d'homogénéiser le tout pour réaliser des comparaisons ou produire de meilleures prévisions peut devenir une tâche longue et difficile. Cette difficulté s'amplifie au fur et à mesure que l'on souhaite agrandir l'ensemble de modèles. De plus, il paraît impossible à l'heure actuelle de construire de tels ensembles avec plusieurs centaines de modèles comme le fait la méthode Monte Carlo.

Une autre solution pour la construction d'un ensemble multi-modèles consiste à générer un grand nombre de simulations à partir d'une même plate-forme de développement. Grâce à la flexibilité et à la modularité de Polyphemus, il est possible de construire un ensemble de modèles très différents en perturbant les données d'entrée et en choisissant différentes paramétrisations physiques et approximations numériques. Mallet et Sportisse [2006b,a] construisent ainsi un ensemble de 48 modèles dans le but d'étudier les incertitudes liées aux paramétrisations physiques et aux approximations numériques mais aussi afin de fournir de meilleures prévisions. La figure 1.6 représente le profil journalier moyen d'ozone de 48 modèles. La moyenne du profil est calculée sur quatre mois et sur tout le domaine européen. On remarque que la dispersion de l'ensemble est importante avec des profils très variés qui dénotent de la grande diversité des modèles.

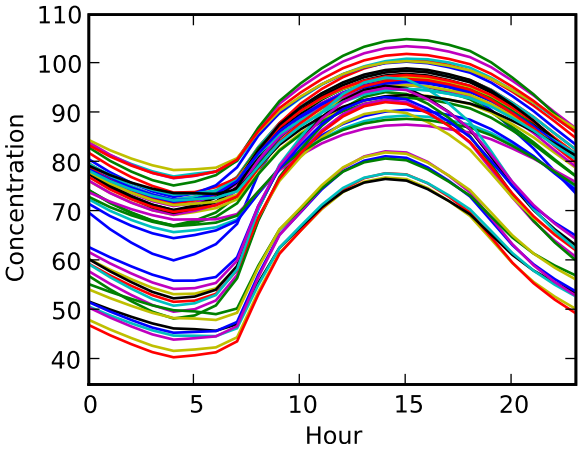


FIGURE 1.6 – Profil journalier moyen d'ozone pour 48 modèles ($\mu g m^{-3}$). Source : Mallet et Sportisse [2006a].

Malgré la très grande flexibilité de la plate-forme logicielle Polyphemus, cette étude a nécessité de modifier puis de lancer « à la main » une chaîne de calculs correspondant à chaque modèle. L'ajout de modèles dans l'ensemble en suivant cette méthodologie peut s'avérer fastidieux.

Une nouvelle méthode de génération automatique d'ensemble a donc été réalisée pendant cette thèse. Elle permet de construire plus d'une centaine de membres en perturbant plusieurs champs d'entrée — conditions aux limites, sources d'émission, champs de vent, température, coefficient de diffusion verticale — et en sélectionnant aléatoirement différentes paramétrisations et approximations numériques disponibles pour chaque simulation. Une fois l'ensemble construit, un script Python se charge de lancer toute la chaîne de calcul et peut profiter du calcul distribué si plusieurs machines sont disponibles. Le chapitre 2 décrit largement cette méthode et l'illustre avec un ensemble d'une centaine de simulations photochimiques à l'échelle européenne sur toute l'année 2001.

Au vu de l'importance des incertitudes dans les modèles de chimie-transport, il est indispensable de les prendre en compte via une approche Monte Carlo ou multi-modèles. Ces ensembles permettent à la fois de produire de meilleures prévisions et d'estimer les incertitudes sur les concentrations des polluants. Ces deux aspects sont discutés dans les deux prochaines sections.

1.2.2 Prévision d'ensemble

Dans le cadre de prévisions opérationnelles, il est indispensable de comparer les données simulées aux observations dans le but de « valider » ou plutôt d'évaluer ces prévisions. Des stations d'observation mesurant des concentrations de polluants — O_3, NO, NO_2, SO_2, particules — sont disponibles dans toute l'Europe, ou plus localement par pays et par région, et servent à mesurer la performance des modèles de chimie-transport. À l'échelle européenne, on peut citer par exemple les réseaux d'observation EMEP ou Airbase qui fournissent au mieux des concentrations horaires sur plus d'une centaine de stations de mesure. La figure 1.3(c) à la fin de la section 1.1.2 présente une comparaison de pics d'ozone simulés et observés à la station de mesure de Saint-Nazaire sur une période de 4 mois.

Les outils statistiques utilisés pour mesurer les performances de ces modèles sont généralement la RMSE, le facteur de biais, la corrélation, ... Le tableau 1.2 donne les principaux indica-

teurs statistiques évaluant les performances d'un modèle sur des séries temporelles simulation–observation.

TABLE 1.2 – Indicateurs statistiques évaluant les performances d'un modèle. $(x_i)_i$ est la série temporelle simulée. $(o_i)_i$ est la série correspondante observée. n est le nombre d'éléments dans chaque série. $\bar{x}$ et $\bar{o}$ sont respectivement la moyenne de $(x_i)_i$ et $(o_i)_i$. Source : Mallet [2005]

Indicateur *Nom anglais*	Notation(s)	Formule
Racine de l'erreur quadratique moyenne *Root mean square error*	RMS ou RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-o_i)^2}$
Biais *Bias*	Bias	$\frac{1}{n}\sum_{i=1}^{n}(x_i-o_i)$
Facteur de biais *Bias factor*	BF	$\frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{o_i}$
Corrélation *Correlation*	Corr	$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(o_i-\bar{o})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(o_i-\bar{o})^2}}$
Bias normalisé moyen *Mean normalized bias error*	MNBE	$\frac{1}{n}\sum_{i=1}^{n}\frac{x_i-o_i}{o_i}$
Mean normalized gross error	MNGE	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert x_i-o_i\rvert}{o_i}$

Lors des deux dernières décennies, les performances des modèles de chimie-transport n'ont cessé de s'améliorer grâce à une compréhension des phénomènes physiques plus approfondie et de meilleures approximations numériques. Néanmoins, depuis quelques années, il est difficile d'améliorer de manière drastique les performances de tels modèles sans prendre en compte l'effet des incertitudes ou l'apport non négligeable d'information que sont les observations. C'est ainsi que s'est développée la prévision d'ensemble.

Il existe plusieurs méthodes qui permettent de fournir des prévisions plus performantes à partir d'un ensemble de simulations. L'une d'entre elle est basique et consiste à considérer la moyenne de l'ensemble des prévisions. Plusieurs études ont montré que la moyenne d'ensemble pouvait parfois donner une prévision meilleure comparée aux prévisions des modèles pris individuellement [Delle Monache et Stull, 2003; Delle Monache *et al.*, 2006a; McKeen *et al.*, 2005]. Cette amélioration n'est cependant pas toujours observée et reste toujours faible.

D'autres méthodes plus performantes utilisent les observations pour donner des poids à chacun des modèles de l'ensemble. Une combinaison linéaire optimale de tous les modèles est ensuite réalisée pour produire une prévision. L'équation 1.11 montre la construction d'une telle prévision où $x^m_{t,i}$ correspond à la concentration simulée du modèle m au temps t et à la station i et α^m_t correspond au poids accordé au modèle m et au temps t :

$$\tilde{x}_{t,i} = \sum_m \alpha^m_t x^m_{t,i}\,. \tag{1.11}$$

Les poids peuvent par exemple être calculés de telle sorte qu'ils minimisent la quantité

$$C_t(\alpha) = \sum_{t-T \leq t' < t, i} \left[o_{t',i} - \sum_m \alpha^m x^m_{t',i} \right]^2 , \tag{1.12}$$

où $o_{t',i}$ est la concentration observée à la date t' pour la station i, et T est la largeur de la période d'apprentissage. En pratique, cette méthode permet d'améliorer la performance des prévisions comparée à la moyenne d'ensemble ou à la prévision du meilleur modèle que ce soit en concentration horaire ou en pics d'ozone. Ces différents poids peuvent être calculés à chaque pas de temps ou pour chaque station. Mallet et Sportisse [2006a] montrent que sur une période de 4 mois et sur le réseaux européen EMEP, cette méthode permet de réduire de manière significative la RMSE et d'améliorer la corrélation.

D'autres algorithmes ont été développés, sur des bases statistiques [Pagowski *et al.*, 2006] et sur la base des méthodes d'apprentissage automatique [Mallet *et al.*, 2009]. Les méthodes d'apprentissage ont l'avantage d'être robustes et fournissent de meilleures prévisions que le meilleur des modèles ou la moyenne d'ensemble, et elles garantissent des performances au moins comparables à celles de la meilleure combinaison linéaire à poids constants en temps. De telles méthodes sont d'ailleurs utilisées sur le système PREV'AIR[7] de manière opérationnelle. Ce système fournit quotidiennement des prévisions pour la qualité de l'air en l'Europe et en France, pour les principaux polluants atmosphériques.

Récemment, ces méthodes de prévision d'ensemble ont été couplées aux techniques classiques d'assimilation de données [Mallet, 2010]. Le couplage permet notamment d'introduire des poids dépendant de la position et de tenir compte de l'erreur des observations.

L'apport d'un ensemble de modèles dans un cadre prévisionnel n'est pas négligeable puisqu'il permet, à l'aide de méthode mathématiques robustes, d'utiliser les observations afin de fournir des prévisions avec des performances fortement améliorées (jusqu'à 30% de réduction de RMSE). Cela nécessite néanmoins d'avoir la capacité de calcul et une plate-forme suffisamment modulaire afin d'avoir un ensemble de prévisions opérationnel.

1.2.3 Approche probabiliste

Nous avons vu qu'une approche stochastique de l'équation de chimie-transport était rendue impossible par la complexité et la taille du problème. On suppose dans cette section que la cible observée — concentration d'une espèce chimique quelconque — est un vecteur aléatoire dont on ne connaît pas la distribution de probabilité a priori. Grâce à un ensemble, on ne fournit plus une seule prévision déterministe, mais plusieurs prévisions plus ou moins probables sur l'état du système.

Soit Y une variable aléatoire correspondant à la cible observée : concentration d'ozone à un endroit donné et à une date donnée. Y représente la meilleure connaissance que l'on a de la cible : il s'agit de la vraie valeur de la cible plus une erreur inconnue et considérée comme aléatoire. L'ensemble fournit N réalisations $(x_1, x_2, \ldots, x_i, \ldots, x_N)$ possibles de Y. Cet échantillon de concentrations d'ozone produit par l'ensemble peut permettre d'estimer l'espérance, la variance et pourquoi pas la distribution de probabilité de Y qui n'est a priori pas connue. Par la suite et pour une variable aléatoire quelconque, on utilisera les notations suivantes :

– l'espérance mathématique : $\mathrm{E}[X]$;

7. http://www.prevair.org/fr/

– la variance : Var[X]. Rappelons que la variance est donnée par :

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 . \tag{1.13}$$

Une estimation possible de l'incertitude liée à une variable aléatoire peut être donnée par le calcul de son écart type σ (1.14) :

$$\sigma_X^2 = \mathrm{Var}[X]. \tag{1.14}$$

Prenons l'exemple d'une variable aléatoire X suivant une loi normale quelconque $\mathcal{N}(\mu, \sigma^2)$ où μ et σ^2 correspondent respectivement à l'espérance et la variance de X. On peut montrer que 95% des réalisations de X seront dans l'intervalle $[\mu - 2\sigma, \mu + 2\sigma]$. On peut aisément prendre σ comme étant une mesure de l'incertitude puisqu'il fournit un intervalle de confiance dans les prévisions.

Un des objectifs de cette thèse sera, après avoir construit un ensemble de simulations photochimiques, d'estimer l'incertitude liée au modèle de qualité l'air via le calcul de l'écart type de l'ensemble en question.

Pour un ensemble de N membres, l'estimation de l'incertitude peut être donnée par

$$\tilde{\sigma}_X = \sqrt{\frac{1}{N-1}\sum_i^N (x_i - \bar{x})^2}, \tag{1.15}$$

où $\bar{x}$ correspond à l'espérance empirique de X. Néanmoins, rien ne nous garantit que cette estimation de l'incertitude est proche de la véritable incertitude liée à notre modélisation numérique — c'est-à-dire de la variance de Y. C'est pourquoi, il est nécessaire de « calibrer » l'ensemble afin de d'obtenir un ensemble représentatif de l'incertitude. Pour ce faire, les performances de l'ensemble en tant que système de prévision probabiliste doivent être calculées. Une méthode automatique de calibration d'ensemble est largement décrite et discutée dans le chapitre 3.

De plus, contrairement à une prévision déterministe, un ensemble de prévisions permet de produire des probabilités pour un certain type d'évènements. Dans le cadre de la qualité de l'air, il est très important de prévoir les dépassements de seuils réglementaires. Dans le cas de l'ozone, deux seuils existent :
– seuil d'information fixé à 180 $\mu g\,m^{-3}$;
– seuil d'alerte fixé à 240 $\mu g\,m^{-3}$.

Ainsi, on souhaite pouvoir calculer la probabilité $P([O_3] \geq 120\ \mu g\,m^{-3})$ par exemple. Dans le cas où l'ensemble fournit une bonne estimation de la distribution de probabilité, il est facile de calculer une telle probabilité. Dans le cas d'une distribution normale par exemple, il suffit de calculer l'aire sous la courbe comme le montre la figure 1.7.

Dans le cas où la distribution de probabilité n'est pas connue, il est encore possible d'approcher cette probabilité en comptant le nombre de simulations qui dépassent effectivement le seuil (1.16) [Hamill et Colucci, 1997] :

$$P([O_3] \geq 180\ \mu g\,m^{-3}) = \frac{|m/x_m \geq 180\ \mu g\,m^{-3}|}{N}. \tag{1.16}$$

De la même manière que pour l'estimation de l'incertitude, il est indispensable d'évaluer les performances de l'ensemble dans le cas où celui-ci produit des probabilités pour un évènement donné.

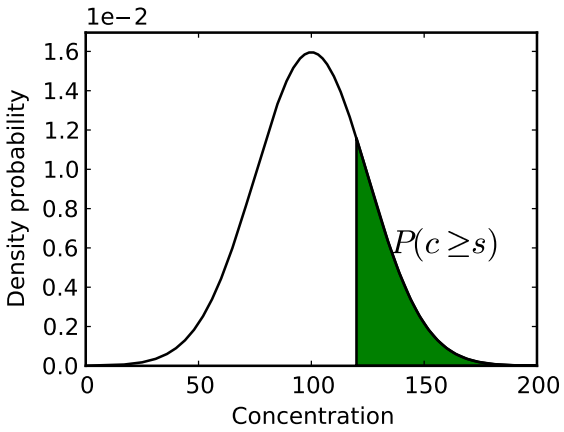


FIGURE 1.7 – Probabilité de dépassement de seuil pour une variable aléatoire suivant une distribution normale. La probabilité correspond à l'aire sous la courbe. Le seuil de concentration s en $\mu g\,m^{-3}$ est ici arbitrairement fixé et correspond à 120 $\mu g\,m^{-3}$.

Nous avons vu que la construction d'un ensemble de modèles de qualité de l'air est primordiale pour prendre en compte toutes les sources d'incertitude et qu'elle permet à la fois de fournir de meilleures prévisions qu'un unique modèle déterministe, d'estimer les incertitudes et aussi de calculer des probabilités pour un évènement donné.

Dans tous les cas, l'évaluation d'un ensemble repose sur des comparaisons aux observations. La section suivante présente les principaux scores mesurant les performances d'un ensemble de simulations.

1.3 Évaluation d'ensemble

Dans cette partie, on s'intéresse à un « système de prévision » qui a pour objectif de fournir une estimation d'une densité de probabilité ou de donner un ensemble de probabilités pour l'occurrence d'un évènement incertain. On présente ici des scores qui permettent de mesurer la qualité d'un tel système de prévision.

Par la suite, on appelle « climatologie » une variable ou une fonction de distribution qui ne dépend pas du temps et issue ou déduite, la plupart du temps, d'une longue série d'observations. À titre d'exemple, on peut calculer la moyenne d'occurrence d'un évènement — par exemple, un dépassement de seuil — en se basant sur un très grand nombre d'observations.

1.3.1 Fiabilité

La fiabilité — *reliability* en anglais — est la capacité du système de prévision à prévoir de manière cohérente la réalité. Il faut qu'il y ait un accord entre les éléments prévus et les éléments effectivement observés.

Le système fournit des probabilités pour un évènement $\mathscr{E}$ donné. Par exemple, le système prévoit m fois la probabilité p_0 que $\mathscr{E}$ se réalise. On compare ensuite cette probabilité prévue p_0 aux observations. Pour ce faire, on compte le nombre d'occurrences de l'évènement les m fois où la probabilité p_0 a été produite (par le système). Le système est fiable si la fréquence d'occurence de $\mathscr{E}$, quand p_0 a été produite, est égale à p_0.

1.3.2 Résolution

La résolution est une seconde qualité requise pour un système de prévision probabiliste. La résolution d'un système de prévision correspond à sa capacité à mesurer la variabilité des éléments prévisibles. Autrement dit, c'est la capacité à classer ou discriminer un ensemble d'évènements en sous-ensembles d'évènements. Elle mesure la variabilité des éléments observés en fonction des probabilités produites par le système. Plus un système aura la capacité à fournir un grand nombre de différentes probabilités, plus la résolution du système de prévision sera susceptible d'être grande.

À titre d'exemple, une prévision climatologique n'a aucune résolution étant donné qu'il n'y a qu'une seule probabilité prévue, quelle que soit l'échéance prévue. Elle est incapable de distinguer des groupes parmi les différentes réalisations d'un évènement.

1.3.3 Diagramme de fiabilité

On peut noter $O(p)$ la fréquence d'occurence de $\mathscr{E}$ observée quand p a été produite (par le système). Le système est parfait si pour tout évènement et pour tout p :

$$O(p) = p\,. \tag{1.17}$$

Pour un système parfaitement fiable, la courbe $O(p)$ en fonction de p est confondue avec la première diagonale sur $[0,1]$. Le diagramme de fiabilité se propose de comparer $O(p)$ à p, généralement en les intégrant sur certains intervalles. Les probabilités d'occurrence d'un évènement produites par le système sont découpées en plusieurs catégories. Celles-ci peuvent être des probabilités égales à 0.1, 0.2, 0.3, ... ou bien des intervalles tels que $[0,0.1[$, $[0.1,0.2[$, $[0.2,0.3[$, ... On trace les fréquences relatives d'occurrence $O(p)$ de $\mathscr{E}$ en fonction des probabilités produites par le système. La figure 1.8 représente un diagramme de fiabilité. Il arrive que la prévision climatologique ou la fréquence climatologique soit représentée sur le diagramme ; cette dernière est souvent la fréquence d'occurrence de l'évènement observée sur une longue période.

Lorsque le diagramme de fiabilité est au-dessous de la diagonale, cela signifie que le système surévalue la probabilité de prévision — *overforecasting*. Au contraire, si le diagramme est au-dessus de la diagonale, le système sous-évalue la probabilité de prévision — *underforecasting*. Plus la courbe est plate — on entend par plate le fait que l'inclinaison de la courbe soit faible — plus la résolution du système est faible. En effet, la courbe aura tendance à approcher la valeur de la probabilité d'occurrence climatologique qui a une résolution nulle.

1.3.4 Diagramme d'acuité

On associe souvent le diagramme d'acuité — *sharpness* en anglais — au diagramme de fiabilité. Il correspond au nombre d'occurrences de chaque catégorie de probabilité produite par le système. Un système de prévision a une bonne acuité si les probabilités fournies par le système ont des valeurs extrêmes — c'est-à-dire 0 ou 1. Au contraire, un système produisant un grand nombre de probabilités proches de 0.5 a une mauvaise acuité. Un système qui fournit de telles probabilités ne permet pas en effet de donner une information utile sur la réalisation ou la non réalisation d'un évènement. L'acuité peut être importante pour le prévisionniste et pour l'aide à la décision.

Comme pour le critère de résolution :

– un système peut avoir une bonne acuité mais une mauvaise fiabilité ;

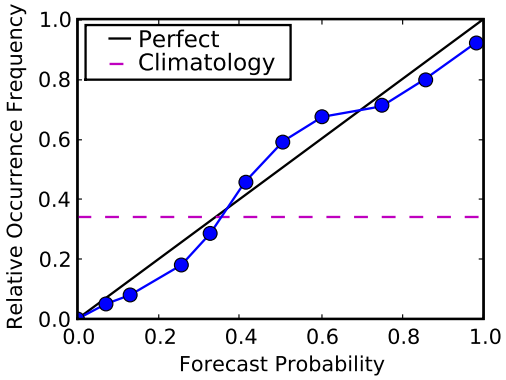


FIGURE 1.8 – Exemple d'un diagramme de fiabilité.

– la climatologie n'a pas d'acuité.

La figure 1.9 montre un diagramme d'acuité d'un système de prévision pour un évènement donné.

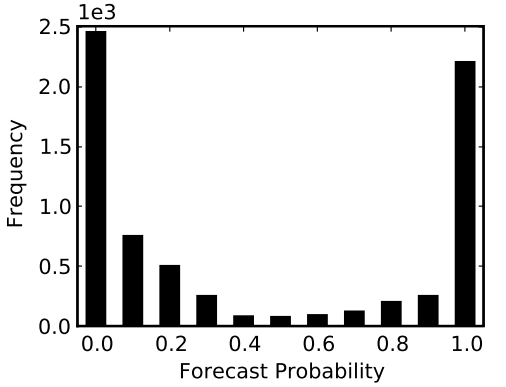


FIGURE 1.9 – Exemple d'un diagramme d'acuité.

Les diagrammes de fiabilité et d'acuité sont très utiles voire indispensables quand il s'agit d'évaluer un système qui a pour objectif de fournir des probabilités pour l'occurrence d'un évènement.

1.3.5 Diagramme de rang

Le diagramme de rang, développé par Anderson [1996], Talagrand *et al.* [1999] et Hamill et Colucci [1997], est un moyen visuel qui permet de comparer la distribution de deux variables aléatoires : la première est estimée par le système de prévision, la seconde correspond à la cible observée.

Généralités

Soit $X = (X_1, X_2, \ldots, X_i, \ldots, X_{n-1}, X_n)$ une suite de n variables aléatoires indépendantes et identiquement distribuées. Soit Y la variable aléatoire correspondant à la cible considérée. On suppose ici que les X_i et Y suivent la même loi de probabilité et que l'ensemble des X_i est rangé

— $X_1 < X_2 < X_i < X_n$ par exemple.

On peut alors montrer (voir la démonstration plus loin) que :

$$\mathrm{E}[P(Y \leq X_i)] = \frac{i}{n+1}. \tag{1.18}$$

On peut aussi montrer que :

$$\mathrm{E}[P(X_{i-1} < Y \leq X_i)] = \frac{1}{n+1}. \tag{1.19}$$

On introduit les valeurs y et x_i qui sont des réalisations des variables aléatoires Y et X_i. Ces valeurs sont données par une observation et l'ensemble des prévisions — des concentrations en pic d'ozone pour une date et une station données, par exemple. La construction d'un diagramme de rang se fait en comptant le nombre d'observations d'un rang donné. Un rang correspond à la position de l'observation y parmi les x_i rangés, ou encore au nombre de x_i inférieurs à y. On construit le diagramme de rang ainsi :

1. Un ensemble de prévisions fournit n valeurs x_i pour une observation y. Ces valeurs correspondent aux réalisations des n variables aléatoires X_i et à la réalisation de Y.
2. On classe les x_i par ordre de valeur croissante.
3. On situe l'observation y parmi les x_i triés et on retient le rang — le rang étant la position de y. Par exemple, le rang 0 correspond à $y \leq x_1$, tandis que le rang k correspond à $x_k < y \leq x_{k+1}$
4. On réitère l'opération pour chaque nouvelle observation.
5. On trace ensuite un histogramme correspondant aux nombres d'observations de rang k, où $k \in [0, n+1]$.

La figure 1.10 montre un exemple de diagramme de rang.

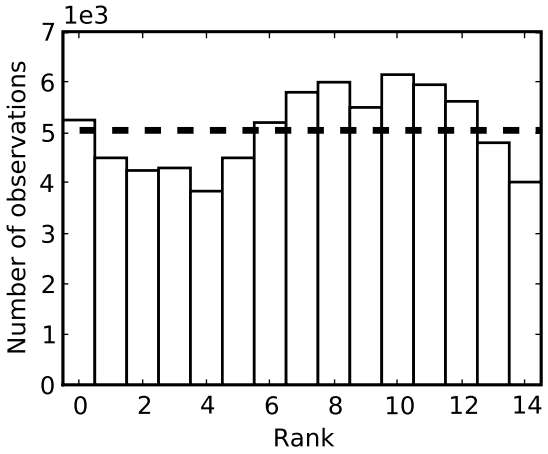
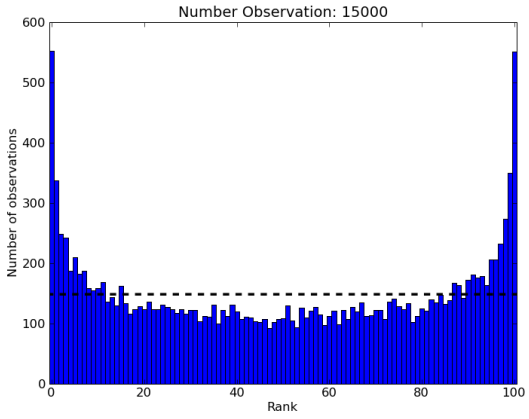


FIGURE 1.10 – Exemple d'un diagramme de rang. Il représente le nombre d'observations ayant un rang donné.

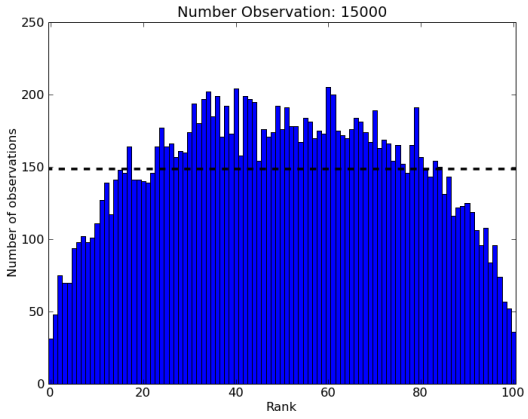
La forme du diagramme de rang permet d'évaluer la cohérence entre les distributions des variables X_i et Y, autrement dit entre la distribution fournie par l'ensemble et celle liée à la variable observée.

– Un diagramme de rang plat correspond à un diagramme de rang parfait et vérifie l'équation 1.18 ;
– un diagramme de rang en forme de « U » — *U-shape* en anglais — correspond à un ensemble trop peu dispersé. En effet, la variable observée est à la fois trop souvent supérieure à X_n et inférieure à X_1 ;
– un diagramme en forme de cloche ou de dôme — *dome-shape* en anglais — correspond à un ensemble trop dispersé ;
– un diagramme asymétrique correspond à un ensemble qui a un biais.

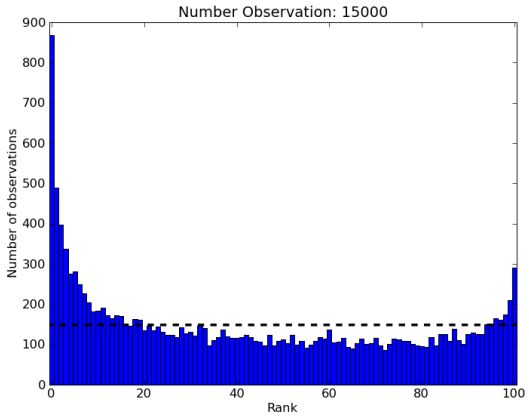
Quatre diagrammes de rang avec des formes différentes sont présentés à la figure 1.11. Ils donnent la forme d'un diagramme dans les cas sur-dispersés et sous-dispersés d'un ensemble avec un biais et sans biais. Le biais, présenté ici à titre d'exemple, est un biais positif. C'est-à-dire que $\mathrm{E}[X] > \mathrm{E}[Y]$.



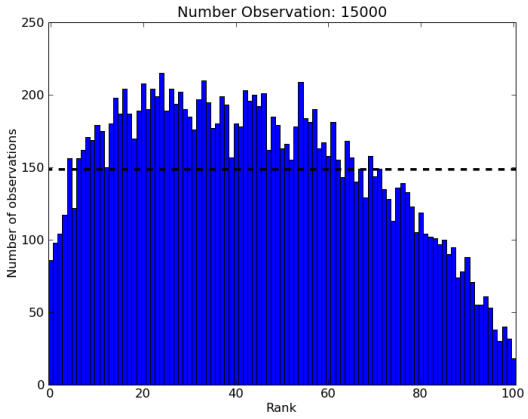
(a) Forme de « U »



(b) Forme de cloche



(c) Forme de « U » avec biais



(d) Forme de cloche avec biais

FIGURE 1.11 – Exemple de quatre diagrammes rang de 100 membres pour 15000 observations. Les diagrammes (a) et (b) correspondent respectivement à 2 ensembles non biaisés sous-dispersé et sur-dispersé. Les diagrammes (c) et (d) correspondent quand à eux à des ensembles biaisés sous-dispersé et sur-dispersés.

La forme d'un diagramme de rang peut donc être très caractéristique du comportement d'un ensemble de prévisions comparé aux observations. Dans l'exemple de la figure 1.11, on remarque

très bien les formes en « U » et en cloche dans le cas où l'ensemble n'est pas assez dispersé ou est trop dispersé. Un ensemble biaisé se caractérise bien par une asymétrie de son diagramme de rang. Dans le cas des diagrammes 1.11(c) et 1.11(d), le nombre d'observations est plus important à gauche, c'est-à-dire vers le rang 0, ce qui dénote d'une surestimation globale des prévisions par rapport aux observations.

Les exemples donnés figure 1.11 sont issus d'expériences contrôlées, avec des distributions choisies pour générer aléatoirement des observations et les ensembles de prévisions. Le tableau 1.3 donne les moyennes et les écarts types des distributions normales choisies pour cette expérience. Le nombre d'observations est fixé à 15000.

TABLE 1.3 – Moyennes et écarts type des distributions normales utilisées pour générer les observations et les ensembles de 100 membres dont les diagrammes de rang sont présentés sur la figure 1.10.

	Y	**1.11(a)**	**1.11(b)**	**1.11(c)**	**1.11(d)**
Moyenne	0.	0.	0.	0.25	0.25
Écart type	1.	0.75	1.25	0.75	1.25
Biais		aucun	aucun	positif	positif
Dispersion		sous	sur	sous	sur

On se propose ensuite de démontrer les équations 1.18 et 1.19 pour 2 puis n variables.

Démonstration pour 2 variables

Soient 2 variables aléatoires X et Y de densité de probabilité notées f_X et f_Y — on en suppose ici l'existence. On note F_X et F_Y les fonctions de répartition. La probabilité

$$P(Y \leq X) = \int_{-\infty}^{X} f_Y(t)\,\mathrm{d}t = F_Y(X) \tag{1.20}$$

est une variables aléatoire. On peut en calculer l'espérance :

$$\mathrm{E}[P(Y \leq X)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{x} f_Y(t)\,\mathrm{d}t\, f_X(x)\,\mathrm{d}x\,. \tag{1.21}$$

On suppose que X et Y suivent la même loi de probabilité : $f_Y = f_X = f$. On veut calculer l'espérance $\mathrm{E}[P(Y \leq X)]$. Cela revient à montrer l'équation 1.18 pour $n = 1$, soit $\mathrm{E}[P(Y \leq X)] = \frac{1}{2}$.

$$\begin{aligned}
\mathrm{E}[P(Y \leq X)] &= \int_{-\infty}^{+\infty} \int_{-\infty}^{x} f(t)\,\mathrm{d}t\, f(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{+\infty} F(x) f(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{+\infty} F(x) F'(x)\,\mathrm{d}x \\
&= \frac{1}{2}\left[F^2(x)\right]_{-\infty}^{+\infty} \\
&= \frac{1}{2}(1-0)
\end{aligned}$$

$$\mathrm{E}[P(Y \leq X)] = \frac{1}{2} \tag{1.22}$$

Démonstration pour n variables

On souhaite démontrer maintenant l'équation 1.19 avec n variables X_i. On utilise ici non plus la définition de la fonction de répartition mais la statistique d'ordre.

On définit les $n+1$ variables suivantes $(U_1,\ldots,U_i,\ldots,U_n,U_{n+1})$ comme étant les $(X_1,\ldots,X_i,\ldots,X_n,Y)$. On introduit en plus les variables S_i qui correspondent aux U_i rangées. Soit $S_1 < S_2 < \ldots S_i < \ldots S_{n+1}$ les U_i triés. On se propose de calculer l'espérance de la probabilité d'avoir les U_i à la $j^{\text{ième}}$ place.

$$\begin{aligned}\sum_i \mathrm{E}\left[P(U_i = S_j)\right] &= \mathrm{E}\left[\sum_i P(U_i = S_j)\right] \\ &= \mathrm{E}[1] \\ &= 1\end{aligned}$$

De plus, pour tout i et tout k, on a :

$$\mathrm{E}[P(U_k = S_j)] = \mathrm{E}[P(U_i = S_j)]\,.$$

On a donc $\sum_{i=1}^{n+1} \mathrm{E}[P(U_i = S_j)] = 1$, soit $\mathrm{E}[P(U_i = S_j)] = \frac{1}{n+1}$. Autrement dit, chaque U_i suivant la même loi de probabilité a autant de chance que les autres d'être à une $j^{\text{ième}}$ place quelconque. Il est possible d'écrire autrement :

$$\mathrm{E}\left[P(U_i = S_j)\right] = \mathrm{E}\left[P(U_i \in]S_{j-1}, S_{j+1}[)\right]\,.$$

En posant $i = n+1$ pour avoir Y, il vient :

$$\begin{aligned}\mathrm{E}\left[P(Y \in]S_{j-1}, S_{j+1}[)\right] &= \frac{1}{n+1} \\ \mathrm{E}\left[P(S_{j-1} < Y < S_{j+1})\right] &= \frac{1}{n+1}\,.\end{aligned}$$

Limites du diagramme de rang

- Le diagramme de rang n'évalue que la fiabilité d'un système. La résolution n'est donc pas prise en compte ;
- le diagramme de rang est nécessaire mais non suffisant pour estimer la qualité d'un système de prévision ;
- l'interprétation de la qualité d'un système à partir d'un diagramme de rang est à prendre avec précaution.

Quantification numérique

Un diagramme de rang est avant tout un indicateur visuel. On a décrit dans un paragraphe précédent qu'un diagramme donne des informations sur le biais et la dispersion d'un ensemble. Une manière de comparer l'adéquation de deux ensembles est d'identifier l'ensemble dont le diagramme de rang est le plus « plat ». Il est judicieux d'introduire un score quantitatif qui le mesure. Ce score devient même indispensable si l'on souhaite comparer plusieurs diagrammes de rang avec un nombre de rangs différents. Cette partie décrit le score normalisé d'un diagramme

de rang. Par la suite, on fera référence à ce score normalisé comme étant le « score » du diagramme de rang.

On sait que $\mathrm{E}[P(Y \leq X_i)] = \frac{i}{n+1}$. Soit b_i la valeur du nombre d'observations de rang i — nombre d'observations dans le i^{e} intervalle, c'est-à-dire entre le membre i de l'ensemble et le membre $i+1$. Le système est fiable si $b_i = \frac{M}{n+1}$, avec M le nombre total d'observations. On définit le score comme étant la variance du diagramme de rang soit :

$$\mathscr{S} = \sum_{i=0}^{n} \left(b_i - \frac{M}{n+1} \right)^2 . \tag{1.23}$$

On cherche à calculer l'espérance de $\mathscr{S}$. On décide de poser :

$$b_i = \sum_{j=1}^{M} q_j , \tag{1.24}$$

où

$$q_j = \begin{cases} 1 & \text{avec probabilité } p = \frac{1}{n+1} \\ 0 & \text{avec probabilité } q = 1-p = \frac{n}{n+1} \end{cases} \tag{1.25}$$

q_j suit donc une loi de Bernoulli avec la probabilité p. On peut aisément vérifier $\mathrm{E}[b_i] = \frac{M}{n+1}$. En effet,

$$\begin{aligned} \mathrm{E}[b_i] &= \mathrm{E}\left[\sum_{j=1}^{M} q_j \right] \\ &= \sum_{j=1}^{M} \mathrm{E}[q_j] \\ &= \frac{M}{n+1} . \end{aligned}$$

b_i est une somme de variables suivant une loi de Bernoulli. Ainsi, b_i suit une loi binomiale de paramètres M et p — M tirages de la variable q suivant la probabilité p. On a donc les propriétés suivantes :

- $\mathrm{E}[b_i] = Mp$;
- $\mathrm{Var}[b_i] = Mp(1-p)$.

De plus, $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$. L'espérance de b_i^2 vaut alors

$$\begin{aligned} \mathrm{E}[b_i^2] &= \mathrm{Var}[b_i] + \mathrm{E}[b_i]^2 \\ &= Mp(1-p) + (Mp)^2 \\ &= \frac{M}{n+1}\left(1 - \frac{1}{n+1}\right) + \left(\frac{M}{n+1}\right)^2 \\ &= \frac{M(n+M)}{(n+1)^2} . \end{aligned}$$

On veut maintenant calculer l'espérance de $\left(b_i - \frac{M}{n+1}\right)^2$.

$$\begin{aligned}\mathrm{E}\left[\left(b_i - \frac{M}{n+1}\right)^2\right] &= \mathrm{E}[b_i^2] - \frac{2M}{n+1}\mathrm{E}[b_i] + \left(\frac{M}{n+1}\right)^2 \\ &= \frac{M(n+M)}{(n+1)^2} - \frac{2M^2}{(n+1)^2} + \left(\frac{M}{n+1}\right)^2 \\ &= \frac{nM}{(n+1)^2}.\end{aligned}$$

L'espérance du score $\mathscr{S}$, noté $\mathscr{S}_0$, est donc

$$\mathscr{S}_0 = \mathrm{E}[\mathscr{S}] = \frac{nM}{n+1}. \tag{1.26}$$

On définit alors le score normalisé du diagramme de rang comme

$$\delta = \frac{\mathscr{S}}{\mathscr{S}_0}. \tag{1.27}$$

Un diagramme de rang plat revient à avoir $\delta = 1$.

Ce score permet d'apporter une quantification numérique relative aux diagrammes de rang, en plus de l'information visuelle qu'apportent ces derniers. De plus, il est aussi un moyen de vérifier la qualité des diagrammes de rang qui ont une forme semblable, et qu'il est donc difficile de comparer.

La figure 1.12 présente deux diagrammes de rang de deux ensembles assez semblables — léger biais positif et une sous-dispersion — qui ont un nombre de membres différents (40 et 30). Malgré leur forme assez proche, leurs scores normalisés sont assez différents : 27.3 pour le diagramme 1.12(a) et 58.4 pour le diagramme 1.12(b). Comme précédemment, les observations ont été générées aléatoirement avec une loi normale centrée réduite. Quant aux deux ensembles, ils ont été générés avec les lois normales $\mathscr{N}(0.2, 0.85)$ et $\mathscr{N}(0.23, 0.8)$. Dans cet exemple, il paraît délicat d'affirmer lequel de ces deux diagrammes de rang est le plus « plat ». Cependant, la distribution de l'ensemble de 40 membres, qui a un meilleur score que son homologue, est plus proche de la distribution des observations. L'intérêt du score normalisé présenté dans cette partie est de s'affranchir d'une comparaison uniquement visuelle qui peut s'avérer trompeuse.

1.3.6 Score de Brier

Général

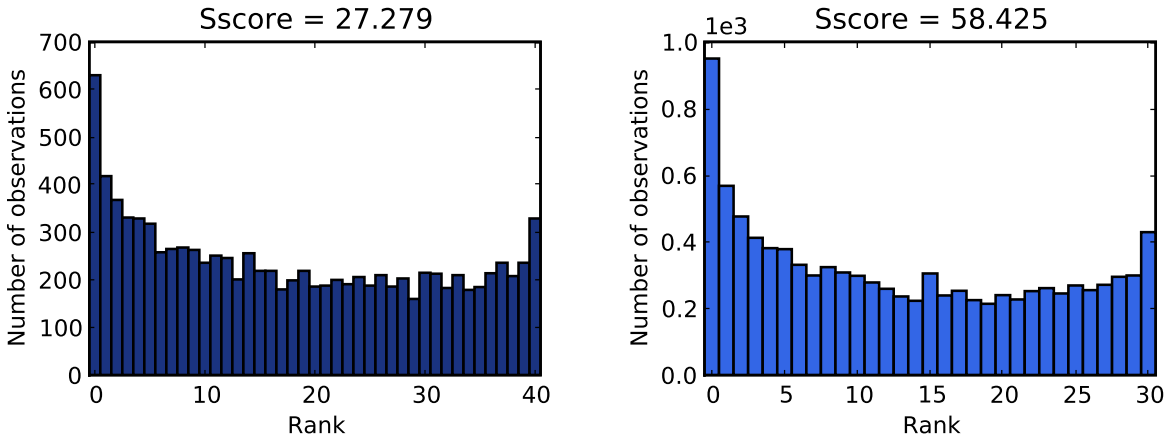
Le score de Brier, Brier [1950] et Wilks [2005], est une mesure de la performance d'un système de prévision pour la probabilité d'occurrence d'un évènement. Soit un évènement $\mathscr{E}$. Soit M le nombre d'observations, p_i les probabilités produites par le système de prévision que l'évènement se réalise et o_i la probabilité observée. La probabilité observée vaut 1 si l'évènement s'est réalisé, 0 sinon. Le score s'écrit sous forme continue :

$$\mathscr{B} = \mathrm{E}[(p-o)^2], \tag{1.28}$$

ou sous forme discrète :

$$\mathscr{B} = \frac{1}{M}\sum_{i=1}^{M}(p_i - o_i)^2. \tag{1.29}$$

Le score varie entre $[0,1]$. Le système est performant pour la prévision de l'évènement $\mathscr{E}$ si son score de Brier est proche de 0.



(a) Ensemble de 40 membres

(b) Ensemble de 30 membres

FIGURE 1.12 – Comparaison de deux diagrammes de rang. Malgré leur forme assez semblable, ces deux diagrammes n'ont pas le même score $\mathscr{S}$.

Ce score mesure à la fois la fiabilité et la résolution d'un système de prévision. Il est d'ailleurs possible de le décomposer afin d'accéder aux termes de « fiabilité » et de « résolution ».

Décomposition

Le score de Brier peut se décomposer en trois termes. Un terme qui mesure la fiabilité du système. Un deuxième terme mesurant la résolution du système. Et un troisième qui ne concerne que l'occurrence de l'évènement en lui-même. La première étape de la décomposition consiste à calculer l'espérance de $(p-o)^2$ pour une probabilité particulière. Ensuite, il suffit de généraliser l'expression obtenue pour l'ensemble des probabilités produites par le système.

Soit p_0 une probabilité prévue par le système, O_0 la fréquence d'occurrence de l'évènement $\mathscr{E}$ quand p_0 a été produite. La probabilité observée de $\mathscr{E}$, quand p_0 a été produite, est noté o et suit une distribution de Bernoulli (1.30).

$$o = \begin{cases} 1 & \text{à la fréquence } O_0 \\ 0 & \text{à la fréquence } 1-O_0 \end{cases} \tag{1.30}$$

L'espérance de $(p_0-o)^2$ vaut :

$$\begin{aligned} \mathrm{E}[(p_0-o)^2] &= (p_0-1)^2 O_0 + p_0^2(1-O_0) \\ &= O_0(p_0^2-2p_0+1) - O_0 p_0^2 + p_0^2 \\ &= -2p_0 O_0 + O_0 + p_0^2 + O_0^2 - O_0^2 \\ &= (p_0-O_0)^2 + O_0(1-O_0). \end{aligned} \tag{1.31}$$

La deuxième et dernière étape de la décomposition consiste à généraliser l'équation 1.31 pour toutes les probabilités. On pose $f(p)$ la fonction de distribution de p. On a :

$$\int_0^1 f(p)\mathrm{d}p = 1, \tag{1.32}$$

et la fréquence climatologique s'écrit :

$$o_c = \int_0^1 O(p) f(p) \mathrm{d}p \,. \tag{1.33}$$

On intègre 1.31 sur $[0,1]$:

$$\begin{aligned}\mathscr{B} &= \int_0^1 (p - O(p))^2 f(p)\mathrm{d}p + \int_0^1 O(p)(1-O(p)) f(p)\mathrm{d}p \\ &= \int_0^1 (p - O(p))^2 f(p)\mathrm{d}p + \int_0^1 O(p) f(p)\mathrm{d}p - \int_0^1 O(p)^2 f(p)\mathrm{d}p \,.\end{aligned}$$

On introduit le terme $2o_c^2 - 2o_c^2$, qui peut s'écrire $o_c^2 + o_c^2 \int_{[0,1]} f - 2o_c \int_{[0,1]} Of$:

$$\begin{aligned}\mathscr{B} &= \int_0^1 (p - O(p))^2 f(p)\,\mathrm{d}p + o_c - \left[\int_{[0,1]} O^2 f \overbrace{-2o_c^2 + 2o_c^2}^{} \underbrace{-2o_c \int_{[0,1]} Of + o_c^2 \int_{[0,1]} f}_{} + o_c^2 \right] \\ &= \int_0^1 (p - O(p))^2 f(p)\,\mathrm{d}p - \int_{[0,1]} (O - o_c)^2 f + o_c(1-o_c) \\ \mathscr{B} &= \underbrace{\int_0^1 (p - O(p))^2 f(p)\,\mathrm{d}p}_{\text{fiabilité}} - \underbrace{\int_0^1 (O(p) - o_c)^2 f(p)\,\mathrm{d}p}_{\text{résolution}} + \underbrace{o_c(1-o_c)}_{\text{incertitude}} \,.\end{aligned} \tag{1.34}$$

Cette décomposition du score de Brier (1.34), introduite par Murphy [1973], contient trois termes :

- le premier terme est clairement une mesure de la fiabilité puisqu'il calcule l'écart entre p et $O(p)$ — se référer à la notion de fiabilité à la section 1.3.1 ;
- le deuxième est une mesure de la résolution du système de prévision. Effectivement, lorsque $\int_{[0,1]} (O - o_c)^2 f$ est proche de 0, cela signifie que l'occurrence de $\mathscr{E}$, quand différentes probabilités sont produites par le système, est proche de la fréquence climatologique. Or, la fréquence climatologique n'a aucune résolution. Ce terme peut être représenté sur le diagramme de fiabilité comme étant la moyenne de la somme de l'écart au carré entre les points de la courbe et la droite correspondant à la fréquence climatologique. Ce terme est orienté positivement. Un exemple est donné sur la figure 1.13. Plus la courbe de fiabilité sera « plate », moins le système aura une bonne résolution et plus grand sera le score de Brier ;
- le troisième terme est appelé « incertitude » dans la littérature et est indépendant du système de prévision puisqu'il dépend uniquement de $\mathscr{E}$. Il correspond au score de Brier de la prévision climatologique. Ce terme sera faible pour une prévision climatologique très proche de 1 ou au contraire très proche de 0. On peut aussi voir ce terme comme l'erreur quadratique moyenne — noté aussi MSE [8] — de l'occurrence de l'évènement. En reprenant la définition du score de Brier (1.29) et en remplaçant les p_i par la prévision climatologique o_c, on a bien :

$$\mathscr{B}_{\text{climatologie}} = \frac{1}{M} \sum_{i=1}^{M} (o_i - o_c)^2 = o_c(1-o_c) \,, \tag{1.35}$$

8. *MSE : Mean Square Error*

où o_c n'est autre que $\frac{1}{M}\sum o_i$, la moyenne de l'occurrence de l'évènement. La figure 1.14 représente la valeur de ce terme en fonction de la prévision climatologique.

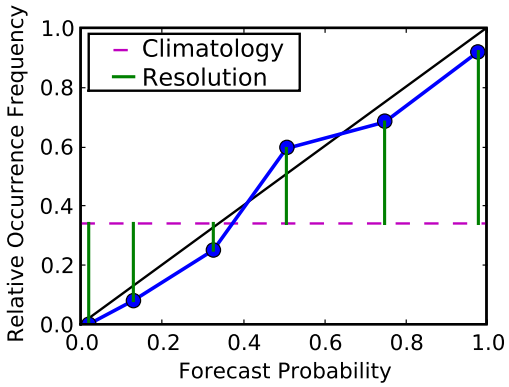


FIGURE 1.13 – Exemple d'un diagramme de fiabilité avec le terme de résolution (en vert) issu de la décomposition du score de Brier.

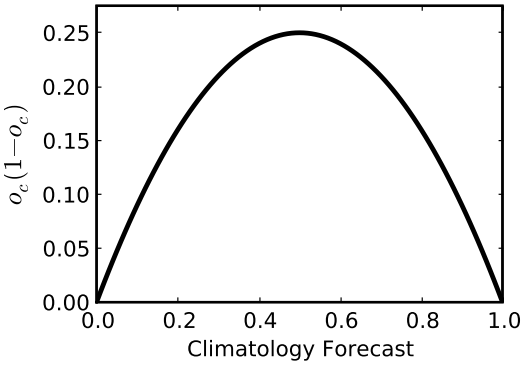


FIGURE 1.14 – Évolution du troisième terme de la décomposition du score de Brier en fonction de la prévision climatologique — aussi appelée *uncertainty* dans la littérature. Il est maximum pour une prévision climatologique égale à 0.5. Dans ce cas, on peut dire que l'occurrence de l'évènement est très incertaine.

Un système de prévision performant doit minimiser le premier terme tout en maximisant le second. Le score de Brier sera plus faible pour un évènement très courant ou au contraire très rare puisque le troisième terme sera plus proche de 0.

Skill score

Il peut être intéressant de comparer le score de Brier d'un système de prévision avec le score de Brier de la prévision climatologique (1.35). Le *Brier Skill Score* mesure l'écart relatif entre ces deux scores de Brier. On appelle donc *skill* l'aptitude d'un système de prévision à avoir un meilleur score que le score donné par la prévision climatologique :

$$\mathscr{B}_S = 1 - \frac{\mathscr{B}}{o_c(1-o_c)}, \tag{1.36}$$

où o_c est la climatologie. Le score de Brier, quand la probabilité de prévision est égale à la prévision climatologique, vaut $o_c(1-o_c)$ — voir la décomposition du score de Brier dans la section 1.3.6 et plus précisément les équations 1.31 et 1.35. Ce *Brier Skill Score* varie entre $[-1,1]$. Contrairement au score de Brier, ce score est orienté positivement, c'est-à-dire que le système est d'autant meilleur que la prévision climatologique qu'il s'approche de 1. Il est supérieur à 0 quand le système est plus performant que la prévision climatologique. Au contraire, il est négatif quand le système est moins performant que la prévision climatologique.

Le *Brier Skill Score* permet de normaliser le score de Brier lorsque l'on veut comparer plusieurs scores pour différents évènements.

La décomposition du score de Brier (1.34) peut être aussi normalisée par $o_c(1-o_c)$:

$$\begin{aligned}
\mathscr{B}_s &= 1 - \frac{\mathscr{B}}{o_c(1-o_c)} \\
&= 1 - \left[\frac{1}{o_c(1-o_c)} \int_0^1 (p-O)^2 f(p)\,\mathrm{d}p + \frac{1}{o_c(1-o_c)} \int_0^1 O(1-O) f(p)\,\mathrm{d}p \right] \\
&= 1 - \left[\underbrace{\frac{1}{o_c(1-o_c)} \int_0^1 (p-O)^2 f(p)\,\mathrm{d}p}_{\mathscr{B}_{sf}} + \underbrace{1 - \frac{1}{o_c(1-o_c)} \int_0^1 (O-o_c)^2 f(p)\,\mathrm{d}p}_{\mathscr{B}_{sr}} \right] \\
\mathscr{B}_s &= 1 - (\mathscr{B}_{sf} + \mathscr{B}_{sr}).
\end{aligned} \tag{1.37}$$

$\mathscr{B}_{sf}$ et $\mathscr{B}_{sr}$ sont tous deux orientés négativement et correspondent respectivement au terme de fiabilité et de résolution de la décomposition du *Brier Skill Score*.

Le score de Brier ne s'intéresse qu'à un unique évènement. Il est intéressant de connaître la qualité et les performances d'un système de prévisions pour une série d'évènements comme par exemple la capacité à bien prévoir différents dépassements de seuil dans le cadre de la qualité de l'air.

1.3.7 *Discrete Ranked Probability Score*

Le *Discrete Ranked Probability Score*, nommé par la suite par son acronyme $DRPS$ et introduit par Epstein [1969a] et Murphy [1971], permet de mesurer la fiabilité et la résolution d'un système de prévision pour une catégorie d'évènements.

On suppose que l'on a L évènements définis pour L valeurs seuils classées par ordre croissant. Il existe alors L catégories de prévisions pour chaque échéance i. Le $DRPS$ vaut :

$$DRPS = \frac{1}{ML} \sum_{i=1}^{M} \sum_{l=1}^{L} (p_{il} - o_{il})^2, \tag{1.38}$$

où p_{il} est donc la probabilité calculée à l'échéance i pour la prévision de catégorie l.

On peut décomposer le $DRPS$ de la même manière que le score de Brier (1.34). Soit P une séquence de probabilités $P = (p_1, p_2, \ldots, p_l, \ldots, p_L)$ produites par l'ensemble pour chaque catégorie d'évènement. Sous forme continue :

$$
\begin{aligned}
DRPS = & \frac{1}{L}\sum_{l=1}^{L}\int (p_l - O_l)^2 f(P)\mathrm{d}p \\
& - \frac{1}{L}\sum_{l=1}^{L}\int (O_l - o_{lc})^2 f(P)\mathrm{d}p \\
& + \frac{1}{L}\sum_{l=1}^{L} o_{lc}(1-o_{lc}),
\end{aligned} \tag{1.39}
$$

où o_{lc} correspond à la fréquence climatologique d'occurrence de l'évènement l. Le premier et deuxième terme correspondent clairement à un terme de fiabilité et un terme de résolution respectivement. Le troisième est quant à lui associé à la fréquence d'occurrence des différents évènements et est appelé *uncertainty*.

Comme pour le *Brier Skill score* (1.36), nous pouvons normaliser l'équation (1.39) par le troisième terme. Ainsi, il vient

$$
DRPS_{skill} = 1 - \frac{DRPS}{\frac{1}{L}\sum_l o_{lc}(1-o_{lc})}. \tag{1.40}
$$

Comme le *Brier Skill score*, le $DRPS_{skill}$ est orienté positivement et varie entre $[-1,1]$. Des valeurs négatives signifient que le système de prévision est moins bon que la climatologie.

1.4 Conclusions

Après une large partie consacrée à la présentation des phénomènes physiques et aux mécanismes chimiques liés à la qualité de l'air, ce chapitre a mis en exergue les incertitudes liées au modèle de chimie-transport qui ont pour origines (1) les données d'entrée, (2) la formulation physico-chimique et (3) les approximations numériques.

Ces incertitudes étant importantes (sources d'émissions, calcul du coefficient de diffusion verticale, ...), il est nécessaire de les prendre en compte dans les prévisions de la qualité de l'air. Un moyen, éprouvé par la communauté météorologique, est de construire un ensemble de prévisions. Un tel ensemble permet (1) de prendre en compte ces incertitudes — via la perturbation des données d'entrée par exemple — mais aussi (2) d'améliorer les prévisions en calculant la moyenne de l'ensemble ou une prévision agrégée par une méthode mathématique, par exemple d'apprentissage statistique.

L'ensemble peut être la base d'un système de prévision probabiliste pour (1) estimer la distribution des concentrations et (2) de fournir des probabilités d'occurrence d'un évènement — dépassement de seuil de concentration réglementaire, en général. Les scores d'ensemble tels que le diagramme de fiabilité, le score de Brier et le diagramme de rang peuvent ensuite mesurer la fiabilité et la résolution du système de prévision. Ces scores serviront à « calibrer » un ensemble de simulations photochimiques dans les chapitres 3, 4 et 5.

Un aspect essentiel reste la construction d'un ensemble suffisamment riche. La plate-forme logicielle Polyphemus donne la possibilité, grâce à sa modularité, de produire plusieurs simulations avec des paramétrisations physiques et des approximations numériques variées, tout en perturbant les données d'entrée. Le chapitre suivant traite d'un large choix de possibilités dans la construction de simulations photochimiques. Une méthode a été mise en œuvre dans le but d'automatiser la génération et le calcul d'un ensemble.

Chapitre 2

Génération automatique d'ensemble

Ce chapitre décrit une méthode qui permet de générer de manière automatique un ensemble dit « multi-modèles » avec de nombreuses simulations pour la qualité de l'air. L'objectif est de prendre en compte toutes les sources d'incertitude : les données d'entrée, les formulations physique et numérique. La principale idée est de construire des modèles de chimie-transport à partir d'une même plate-forme logicielle. Ainsi, la génération de l'ensemble sera totalement contrôlée. Un ensemble contenant de nombreux modèles peut être généré avec des simulations Monte Carlo qui abordent à la fois les incertitudes dans les données d'entrée and dans la formulation du modèle. La génération automatique de l'ensemble réalisée dans ce chapitre est effectuées à l'aide de la plate-forme logicielle Polyphemus. Cette plate-forme est assez flexible pour construire des modèles très différents. Elle offre un très large choix d'options dans la construction d'un modèle : plusieurs paramétrisations physiques, de nombreux schémas numériques et différentes données d'entrée peuvent être combinés. De plus, les données d'entrée peuvent être perturbées. Dans ce chapitre, une trentaine d'alternatives est disponible pour la génération d'un modèle unique. Pour chaque alternative, des poids sont donnés aux options de manière arbitraire en fonction de leur fiabilité supposée. Chaque modèle de l'ensemble est construit en sélectionnant aléatoirement une option par alternative. Dans le but de diminuer la charge de calcul, un certain nombre de calculs sont partagés par les modèles de l'ensemble. Un exemple d'un ensemble d'une centaine de simulations photochimiques à l'échelle européenne pour l'année 2001 est présenté. Les performances de chaque modèle sont brièvement examinés et la structure de l'ensemble est analysée. Nous trouvons un ensemble généré très dispersé et des résultats très différents dans la performance des modèles. Il est à noter que beaucoup de modèles se révèlent être le meilleur modèle dans une région et à une date données.

Sommaire

Ce chapite est constitué de Garaud et Mallet [2010].

2.1 Introduction

Due to the great uncertainties that arise in air quality modeling, ensembles of simulations are now considered in a wide range of applications. They are primarily built for uncertainty estimation. They can therefore evaluate the reliability of exposure studies based on model simulations. In the context of short-term forecasts, they can be used to evaluate risks, with probabilistic forecasts (e.g., threshold exceedence). Uncertainty estimation may also be useful for screening studies in which the impact of emission abatement, as predicted by numerical simulations, should be compared with the uncertainties. Data assimilation is another application where ensembles are often used: e.g., in the popular ensemble Kalman filter, the background-error covariance matrix is derived from them. For operational forecasts, an ensemble simulation may be sequentially aggregated so as to form forecasts better than the individual models.

A key step is the generation of the ensembles. They may be built (1) with perturbations in the input data to a single model or with an ensemble of input data [Straume, 2001], (2) with models that share little or no computer code [Galmarini *et al.*, 2004b; McKeen *et al.*, 2005], or with models built on the same modeling platform. Uncertainty estimation, for instance, has been conducted with Monte Carlo simulations, thus with perturbations in the input data to a given model [Hanna *et al.*, 1998; Beekmann et Derognat, 2003], and with different models built on the same platform [Mallet et Sportisse, 2006b]. In data assimilation, the ensemble Kalman filter [Evensen, 1994] approximates background-error covariance matrices using an ensemble of simulations generated with perturbations in the input data [in air quality, e.g., Segers, 2002]. A few studies make use of ensembles composed of models developed in different teams [for long-term simulations, see van Loon *et al.*, 2007]. For operational forecasting, a weighted linear combination of models can form an improved forecast, as has been shown with an ensemble of models from different teams [Pagowski *et al.*, 2006] and with an ensemble built on the same modeling platform [Mallet et Sportisse, 2006a; Mallet *et al.*, 2009].

Whatever the application may be, a key step is the generation of the ensemble. In an ideal setting, one should take into account all uncertainty sources based on the best description available. Essentially, this would mean relying on Monte Carlo perturbations for uncertain input data like emissions, on the alternative descriptions available for data like land use cover, on calibrated ensemble weather forecasts, on different formulations for the subgrid parameterizations in the chemistry-transport models, on different numerical schemes in the chemistry-transport models. In this paper, we tend to this ideal setting with a simplified approach: we do not use a meteorological ensemble (the meteorological inputs are treated like other input data), and we rely on an alternative sampling approach to full Monte Carlo simulations. Nevertheless all uncertainty sources can be considered, and they are all taken into account at the numerical-simulation stage: no statistical correction is applied in a postprocessing. The approach described in this paper may be seen as a three-fold extension that of Mallet et Sportisse [2006b]: new uncertainty sources are included, the uncertainty in input data is specifically taken into account, and the ensemble generation is entirely automatic.

From a technical point of view, building an ensemble of simulations is rather straightforward in the case of Monte Carlo simulations: one simply applies random perturbations to the input data of a single model. The perturbation scheme may be complex if it takes into account spatial and temporal correlations in the input fields and if advanced Monte Carlo variants are implemented. However this involves little complexity compared to building an ensemble composed of different models, e.g. of models based on various chemical mechanisms. There are essentially two ways (that may be combined) to form an ensemble with different models. One is to use existing models, usually developed in research groups. The resulting ensemble then includes a small number of models, say about ten. Another way is to generate different models within the

same modeling platform: the models are assembled using basic components such as the chemical mechanism or the deposition module. Building such a platform is a tedious task, but it makes the generation of ensembles, even very large ones, practicable. In addition, the structure of the ensemble is fully controlled, which eases the scientific interpretation. This approach has been implemented in the modeling system Polyphemus [Mallet *et al.*, 2007b], and it is described in this paper.

All the models considered in the platform assume that the concentrations of pollutants satisfy a system of partial differential equations and they approximate their solutions by discretizing the equations in an Eulerian framework. Each equation of the system is an advection–diffusion–reaction equation of the form:

$$\frac{\partial c_i}{\partial t} = -\mathrm{div}(V c_i) + \mathrm{div}\left(\rho K \nabla \frac{c_i}{\rho}\right) + \chi_i(c,t) + E_i - \Lambda c_i \,, \tag{2.1}$$

where c_i is the concentration of the ith species, $c = (c_1, \ldots, c_S)$ is the vector of all concentrations, V the wind vector, K is the turbulent diffusion matrix, ρ the air density, χ_i the production term due to chemical reactions involving species i, E_i represents the emissions and Λc_i accounts for losses due to scavenging. The boundary conditions at ground level involve the surface emissions S_i and the deposition velocity v_i:

$$K \nabla c_i \cdot n = S_i - v_i c_i \,, \tag{2.2}$$

if n is the upward-oriented normal to the ground.

All models solve a system of reactive transport equations like equation (2.1), but they rely on different coefficients in the equations (e.g., in the chemistry χ_i) and on different numerical schemes. The coefficients in the equations are estimated according to data from many sources (emission inventory, meteorological model, ...) and many physical parameterizations (vertical diffusion, photolysis attenuation, ...). We therefore uniquely define a model with (1) the input data and the physical parameterizations it uses and (2) its numerical schemes. Many alternative parameterizations, data sources and numerical schemes are available in Polyphemus—this flexibility is part of Polyphemus design principles. Most options are described in section 2.2 which identifies the models that can be built on the platform.

In section 2.3, the actual generation of the ensemble is addressed. This means selecting the models, also called ensemble members, which in turn means selecting the components (input data, physical parameterization, numerical options) for every model. One model is actually a set of programs that are launched in a given order. The simulation chain should be properly established to take into account the dependencies (e.g., the deposition velocities depend on the land use cover). It should also share the common computations among groups of models and distribute the computations over several computer processors in order to minimize the overall computational time. In addition to the changes in the physical and numerical formulations, several input fields that appear in the reactive transport equation are perturbed. It is assumed that the fields have a normal or a log-normal distribution, and they are perturbed accordingly.

In section 2.4, the method is illustrated with a 101-member ensemble with gas-phase chemistry only.

2.2 Building One Model

In this section, many options available in Polyphemus 1.5 (released 20 May 2009) for photochemical simulations are introduced. A summary of these options is given in table 2.1.

2.2.1 Physical Formulation (Parameterizations and Input Data)

Land Use Cover

The land use cover (LUC) describes the material covering the ground with a few categories. Polyphemus supports the USGS (U.S. Geological Survey) LUC with its 24 categories and the GLCF (Global Land Cover Facility) LUC that includes 14 categories. Both LUC have a $1 \times 1\text{km}^2$ resolution, with categories such as grassland, cropland, deciduous forest, urban areas, ...

Chemistry

The chemical mechanism is a simplified representation of atmospheric chemistry, here related to photochemical activity. The mechanism includes species that may or may not exist as such, since many (real) chemical species are lumped into a few (model) species (e.g., the terminal alkenes are lumped into "OLT" in RACM). The mechanism describes the chemical reactions between these species. Here, we consider two chemical mechanisms: RADM 2 [Stockwell *et al.*, 1990] with 61 species and 157 reactions, and RACM [Stockwell *et al.*, 1997] with 72 species and 237 reactions.

Critical Relative Humidity

The critical relative humidity is used to compute the cloud fraction, the cloudiness and the attenuation. One option is to compute the critical relative humidity q_c as a function of σ:

$$q_c = 1 - \alpha\sigma^a(1-\sigma)^b\left(1+\beta(\sigma-\frac{1}{2})\right), \tag{2.3}$$

where $\sigma = \frac{P}{P_s}$, P is the pressure, P_s is the surface pressure, $\alpha = 1.1$, $\beta = \sqrt{1.3}$, $a = 0$ and $b = 1.1$. In another option (*two layers*), the critical relative humidity is simply constant in two distinct layers: $q_c = 0.75$ below 700 hPa and $q_c = 0.95$ above.

Photolysis

Two options are considered. Clear sky photolysis rates J_{clear} can be those computed by the JPROC software which is part of the Community Multiscale Air Quality (CMAQ) Modeling System [Byun et Ching, 1999], or they can be computed based on the zenith angle alone. The photolysis rates are of the form $J = \mathscr{A} J_{clear}$ where $\mathscr{A}$ is the attenuation.

Attenuation

The cloud attenuation $\mathscr{A}$ measures the decrease in the rates of photolysis reactions when solar radiation is partially absorbed or reflected by clouds. It can be computed using the RADM method [Madronich, 1987; Chang *et al.*, 1987]:

$$\begin{cases} \mathscr{A}_b = 1 - \min(1, \mathscr{N}_m + \mathscr{N}_h)(1 - 1.6\,Tr\cos Z) \\ \mathscr{A}_a = 1 + \min(1, \mathscr{N}_m + \mathscr{N}_h)(1 + (1-Tr)\cos Z) \end{cases} \tag{2.4}$$

where $\mathscr{A}_b$ and $\mathscr{A}_a$ are the attenuations below and above the clouds, $\mathscr{N}_m$ and $\mathscr{N}_h$ are the medium cloudiness and the high cloudiness, Tr is the cloud transmissivity and Z is the zenith angle. The photolysis rates below and above the clouds are respectively $J_b = \mathscr{A}_b J_{clear}$ and $J_a = \mathscr{A}_a J_{clear}$.

A second parameterization was developed after the ESQUIF campaign [ESQUIF, 2001], using measurements of the photolysis rates for NO_2. The attenuation is approximated by

$$\mathscr{A} = (1 - a\,\mathscr{N}_h)(1 - b\,\mathscr{N}_m)e^{-cB}\,, \tag{2.5}$$

where a, b, c and B are constants.

Vertical Diffusion

The vertical diffusion coefficient K_z ($\mathrm{m^2\,s^{-1}}$) is the third diagonal term of the turbulent diffusion matrix K (equation 2.1). This coefficient is computed at the interfaces of the model layers and can be estimated with two parameterizations. K_z may be computed with the Louis parameterization [Louis, 1979] at interface k:

$$K_{z,k} = L_k^2 F(R_{ik}) \left[\left(\frac{\Delta U_k}{\Delta z_k} \right)^2 + \left(\frac{\Delta V_k}{\Delta z_k} \right)^2 \right], \tag{2.6}$$

where L_k is the mixing length at level k, R_i is the Richardson number and F is the stability function. Alternatively, K_z can be computed with the Troen&Mahrt parameterization [Troen et Mahrt, 1986]:

$$K_{z,k} = u_* \kappa z_k \Phi_{m,k}^{-1} \left(1 - \frac{z_k}{PBLH} \right)^p, \tag{2.7}$$

where u_* is the friction velocity, κ is the von Kármán constant, $\Phi_{m,k}$ is the non-dimensional shear and $PBLH$ is the planetary boundary layer height. This parameterization is more parametric and more robust than the Louis parameterization. A third option is a combination of both parameterizations: the Louis parameterization used in stable conditions and the Troen&Mahrt parameterization in unstable conditions.

In the Troen&Mahrt parameterization (2.7), the exponent p may be 2 or 3. In the ensemble generation, the boundary layer height $PBLH$ may be perturbed at that stage.

In addition to the selected parameterization, a few options remain with the minimum value for K_z, the minimum value of K_z over urban areas, and whether the minimum values for K_z are applied only in the first layer or in all layers.

Deposition Velocities

The deposition velocities ($\mathrm{m\,s^{-1}}$) are assumed to be in the form

$$V_d = \frac{1}{R_a + R_b + R_c}, \tag{2.8}$$

where R_a is the aerodynamic resistance, R_b is the quasi-laminar sublayer resistance and R_c is the canopy resistance. R_a can be computed with the heat flux or the momentum flux. R_c can be computed by the Zhang parameterization [Zhang *et al.*, 2003] or the Wesely parameterization [Wesely, 1989]. It depends on the LUC.

Emissions

Pollutant emissions are usually divided into two parts: biogenic emissions emitted by vegetation and anthropogenic emissions originating from human activities (transport, industries, ...). The biogenic emissions are surface emissions computed following Simpson *et al.* [1999]. They depend on LUC. At the European scale, anthropogenic emissions are estimated by EMEP (European Monitoring and Evaluation Programme). EMEP provides annual quantities for a few pollutants (NO_x, VOC, SO_2, CO and aerosols) and for 10 different sectors called SNAP (Selected Nomenclature for Air Pollution). These annual emissions are multiplied by monthly, daily (Saturday, Sunday, week days) and hourly factors which depend on the country and SNAP. Finally the emissions are split into surface and volume emissions, according to SNAP. The vertical distribution of the volume emissions is subject to a choice; here, we consider two options: a *low* distribution and a *medium* distribution—the former distribution assumes that the pollutants are released closer to the ground than with the latter distribution. Table 2.8 describes the 10 different SNAP and the emission vertical distribution for the options *low* and *medium*.

2.2.2 Numerical Issues

In Polyphemus, three numerical schemes (for advection, diffusion and chemistry) are assembled to form a numerical integrator, called Polair3D [Boutahar *et al.*, 2004], whatever schemes are used. The numerical integrators share the coordinate system: regular horizontal grid in latitude/longitude, vertical levels with fixed altitudes (in meters from the ground). The integration makes use of operator splitting: in one time step, the advection is integrated first, then the diffusion and finally the chemistry.

Very few numerical options are considered here, because the uncertainty sources were mainly found in the physical formulation and in the input data [Mallet et Sportisse, 2006b]. In Mallet *et al.* [2007a], a detailed study of many numerical options shows that the splitting time step and the advection scheme may have a significant impact. In the present study, the advection scheme is not an option: a third-order direct-space-time scheme with flux limiting [Verwer *et al.*, 2002] is used in all the models. On the other hand, the splitting time step is an option (see below). Both diffusion and chemistry are integrated using a second-order Rosenbrock scheme [Verwer *et al.*, 2002].

Time Step

The (splitting) time step is set to 600 s (the usual time step) or 1200 s.

Simulation Grid

The horizontal resolution is set to 0.5° in all simulations.

Along the vertical, the grid is made up of 5 layers or 9 layers, up to 3000 m. The height of the first layer may be 40 m or 50 m. Consequently, there are 4 possible vertical grids. Note that a change in the vertical grid has consequences in almost all computations.

Vertical-Wind Diagnosis

The vertical wind may be reconstructed from the horizontal-wind components by solving the equation $\mathrm{div}(\rho V) = 0$ where ρ is the air density and V the wind vector.

It may also be estimated with the simplified equation $\mathrm{div}(V) = 0$. In this case, the diffusion term in equation 2.1 is changed, for consistency, to $\mathrm{div}(K\nabla c)$.

This diagnosis is carried out after the horizontal winds have been perturbed.

2.2.3 Other Options

The options previously mentioned are summarized in table 2.1. Other options are available in Polyphemus. They are not reported in this paper because they are not used in the illustrative example (section 2.4).

Many of the other options are related to aerosols. Polyphemus includes a size-resolved aerosol module called SIREAM [Debry *et al.*, 2007] and related preprocessing (anthropogenic emissions, sea salt emissions, deposition, boundary conditions). The aerosol module offers numerous options: choice of the aqueous module, nucleation model (binary, ternary), heterogeneous reactions, calculation of the wet diameter, aerosol density, thermodynamics module, ... This ability was used in the sensitivity study by Sartelet *et al.* [2008], and it should be used in the generation of an ensemble. In the preprocessing steps, several options also relate to aerosols, e.g., the parameterization for estimating the emissions of sea salt could that of Smith et Harrison [1998] or that of Monahan *et al.* [1986].

Table 2.1: Alternatives for the physical parameterizations and numerical options. The numbers enclosed in brackets correspond to the occurrence probability of an option.

#	**Parameterization**	**First option**	**Other option(s)**	**Comment**
		Physical parameterizations		
1.	Land use cover	USGS (0.5)	GLCF (0.5)	
2.	Chemistry	RACM (0.6)	RADM 2 (0.4)	
3.	Cloud attenuation	RADM method (0.6)	ESQUIF (0.4)	
4.	Critical relative humidity	Depends on σ (0.7)	Two layers (0.3)	Used in the RADM method to compute cloud attenuation
5.	Vertical diffusion (K_z)	Troen & Mahrt (0.35)	Louis (0.3)	
			Louis stable (0.35)	Troen & Mahrt kept in unstable conditions
6.	Deposition velocity	Zhang (0.5)	Wesely (0.5)	
7.	Coefficient Ra	Heat flux (0.7)	Moment flux (0.3)	For aerodynamic resistance (in deposition velocities)
8.	Emissions vertical distribution	Low (0.5)	Medium (0.5)	
9.	Photolysis rates	JPROC (0.7)	Zenith angle (0.3)	
		Numerical issues		
10.	Time step	600 s (0.9)	1200 s (0.1)	
11.	Vertical resolution	5 layers (0.5)	9 layers (0.5)	The first layer height can be 50 m or 40 m
12.	First layer height	50 m (0.5)	40 m (0.5)	The top of every other layer does not change
13.	Vertical-wind diagnosis	$\mathrm{div}\left(\rho V\right) = 0$ (0.5)	$\mathrm{div}(V) = 0$ (0.5)	
14.	Minimal K_z	0.2 $\mathrm{m^2 s^{-1}}$ (0.7)	0.5 $\mathrm{m^2 s^{-1}}$ (0.3)	
15.	Minimal K_z in urban area	0.2 $\mathrm{m^2\,s^{-1}}$ (0.3)	0.5 $\mathrm{m^2\,s^{-1}}$ (0.3)	
			1. $\mathrm{m^2\,s^{-1}}$ (0.4)	
16.	Vertical application of minimal K_z	Yes (0.8)	No (0.2)	If *no*, the lowest threshold is applied only to the top of the first layer, otherwise it is applied to all levels
17.	Exponent p	2 (0.7)	3 (0.3)	The value of the p exponent to compute the vertical diffusion coefficient (T&M only)
18.	Boundary layer height	raw value (0.6)	+10% (0.2)	Used to compute Kz (T&M only)
			-10% (0.1)	
			+20% (0.1)	

2.3 Ensemble Generation

In order to build a large ensemble and to take into account all possible options, an automatic procedure is necessary. In addition to the changes in the model formulation, the procedure includes a perturbation step (section 2.3.1): the input data of the numerical model are perturbed so as to take into account additional uncertainty sources. After that step, all simulations are completely defined and launched (sections 2.3.1 and 2.3.2).

2.3.1 Input Data Perturbation

In the final stage of a simulation, the numerical integration of equation (2.1) is carried out with the selected numerical scheme. At this stage, the fields that appear in the equation are also perturbed.

Estimations of the uncertainties were established by experts and reported in Hanna *et al.* [1998, 2001], for 5 km and 12 km resolutions, in regions of eastern U.S.A., and for a few days. These estimations should be seen as guidelines to be adapted to the simulation region, to the resolution of the simulation, to the time span, and to other considerations on the quality of the fields. For instance, the uncertainty in the values of a field should decrease when the resolution gets higher. In addition, a few ensembles were generated in order to roughly calibrate the uncertainty parameters, based on comparisons with observations (not reported here). Several fields are given a distribution, normal or log-normal, and an uncertainty range determined by a parameter α. It is assumed that any value of the field is the random variable $\hat{p}$ that satisfies

- $\hat{p} = p + \frac{\gamma}{2}\alpha$ for a normal distribution,
- $\hat{p} = p\sqrt{\alpha}^{\gamma}$ for a log-normal distribution,

where γ is distributed according to $\mathcal{N}(0,1)$, and p is a (deterministic and known) value which is assumed to be the median of $\hat{p}$.

For a normal distribution, $\hat{p} \in [p-\alpha, p+\alpha]$ has a probability of 95%. Thus $\pm\alpha$ is an uncertainty range, around the mean (or median) p, associated with a probability of 0.95. α is twice the standard deviation of $\hat{p}$. For a log-normal distribution, the same applies to $\ln\hat{p}$, with an uncertainty range of width $\pm\frac{1}{2}\ln\alpha$. The probability that $\hat{p} \in [p/\alpha, \alpha p]$ is 0.95.

Note that the perturbation will not depend on the date or on the position. We simply assume that $\hat{p}(t,x) = p(t,x)\frac{\gamma}{2}\alpha$ (or $\hat{p}(t,x) = p(t,x)\sqrt{\alpha}^{\gamma}$) for any date t and position x. Since γ does not depend on t and x, two values $\hat{p}(t,x)$ and $\hat{p}(t',x')$ are fully correlated (correlation equals 1) for a normal distribution. In the log-normal case, $\ln\hat{p}(t,x)$ and $\ln\hat{p}(t',x')$ are fully correlated.

The list of the perturbed fields and the corresponding values of α are shown in table 2.2. These values were used in the example (section 2.4). The list of input fields includes meteorological variables, the boundary conditions, the emissions for different species and variables related to chemical species such as the deposition velocities or the photolysis rates. These input data come from different models (ECMWF, Mozart 2, EMEP) or are processed during the preprocessing.

Once the distribution and the parameter α are determined, the actual perturbation is not given by a random sampling of γ. The actual perturbed value is randomly and uniformly selected in a set of three values: the unperturbed value (i.e., the median) $\tilde{p}_0 = p$, and two other points $\tilde{p}_1$ and $\tilde{p}_2$ defined below. The points are chosen so that the empirical mean and the empirical standard deviation are the same as the mean and the standard deviation of $\hat{p}$, thus

- $E(\hat{p}) = \dfrac{\tilde{p}_0 + \tilde{p}_1 + \tilde{p}_2}{3}$;
- $\mathrm{Var}(\hat{p}) = \dfrac{(\tilde{p}_0 - \bar{p})^2 + (\tilde{p}_1 - \bar{p})^2 + (\tilde{p}_2 - \bar{p})^2}{2}$.

Table 2.2: Perturbation of input data. The last column is the parameter α that defines the uncertainty range. If the median value of a normally-distributed random variable $\hat{p}$ is p, the probability that $\hat{p} \in [p-2\alpha, p+2\alpha]$ is 0.95. If $\hat{p}$ is log-normally distributed, the probability that $\hat{p} \in [p/\alpha, \alpha p]$ is 0.95.

#	**Field**	**Source**	**Distribution**	**Uncertainty range**
1.	Horizontal-wind module	ECMWF	Log-normal	1.5
2.	Horizontal-wind angle	ECMWF	Normal	±40 degrees
3.	Temperature	ECMWF	Normal	±3 K
4.	O_3 boundary conditions	Mozart 2	Log-normal	2.0
5.	NO_x boundary conditions	Mozart 2	Log-normal	3.0
6.	VOCs (Volatile Organic Compounds) boundary conditions	Mozart 2	Log-normal	2.0
7.	NO_x anthropogenic emissions	EMEP	Log-normal	1.5
8.	VOCs anthropogenic emissions	EMEP	Log-normal	1.5
9.	Biogenic emissions	Computed	Log-normal	2.
10.	Vertical diffusion	Computed	Log-normal	1.7
11.	Deposition velocities	Computed	Log-normal	1.5
12.	Photolysis rates	Computed	Log-normal	1.4

In the case where $\hat{p} \sim \mathcal{N}(p, \frac{1}{4}\alpha^2)$:

$$\begin{aligned} \tilde{p}_0 &= p\,; \\ \tilde{p}_1 &= p - \tfrac{1}{2}\alpha\,; \\ \tilde{p}_2 &= p + \tfrac{1}{2}\alpha\,. \end{aligned}$$

When $\hat{p}$ is log-normally distributed:

$$\begin{aligned} \tilde{p}_0 &= p\,; \\ \tilde{p}_1 &= p\sqrt{\alpha}^{\gamma_1}\,; \\ \tilde{p}_2 &= p\sqrt{\alpha}^{\gamma_2}\,, \end{aligned}$$

with

$$\begin{aligned} \gamma_1 &= 2\log(\tfrac{3\beta-1-\sqrt{\Delta}}{2})/\log\alpha\,; \\ \gamma_2 &= 2\log(\tfrac{3\beta-1+\sqrt{\Delta}}{2})/\log\alpha\,, \end{aligned}$$

and

$$\begin{aligned} \beta &= \exp(\tfrac{1}{8}\log^2\alpha)\,; \\ \Delta &= 4\beta^4 - 7\beta^2 + 6\beta - 3\,. \end{aligned}$$

Models Automatic Selection

The selection of the models to be included in the ensemble is carried out randomly. A probability is given to each option. The numbers between brackets in table 2.1 are these probabilities. In any alternative, the sum of the probabilities equals 1. Each perturbation in the input data (table 2.2) is uniformly selected from three possible values (section 2.3.1). A model is defined once an option has been selected for any alternative (18 alternatives are shown in table 2.1, 12 perturbations are listed in table 2.2).

Except for the perturbations in the input data, the probabilities are chosen according to the confidence put in each option. There is no direct indicator to determine these probabilities. If two parameterizations are available for a given option, the choice lies between giving a probability one to a parameterization (no uncertainty), and giving 0.5 to both parameterizations (which leads

to the largest uncertainty). If one option is supposed to be more accurate (a priori quality of a parameterization, finer grid resolution, ...) or if it is usually associated with better model results (comparison with observations), its weight should be higher than that of alternative choices. For example, a time step equal to 600 s is supposed to give more accurate results than 1200 s—the numerical solution converges to the exact solution as the time step tends to 0. Therefore, a higher probability is associated with the time step fixed to 600 s. Another example is the chemical mechanism RACM which is more detailed than RADM 2, and which has shown slightly better results in several studies [Gross et Stockwell, 2003].

2.3.2 Technical Aspects

The structure of the Polyphemus system contains four (mostly) independent levels: data management, physical parameterizations, numerical solvers and high-level methods such as data assimilation. Figure 2.1 illustrates the structure of the modeling platform.

During the first stage, several C++ programs carry out the preprocessing. This is the most important part of the simulation process, both in terms of simulation definition (the physics is set there) and computer code. Almost all terms of the reactive transport equation (2.1) are computed at this stage. The computations are split into several programs to ensure flexibility. For instance, there is one program to process land use cover (actually two programs: one for USGS data and another for GLCF data), one program for the main meteorological fields, one program to compute biogenic emissions, another program for anthropogenic emissions, ... Another example is the vertical diffusion coefficient: one program computes it with Louis parameterization and another with the Troen&Mahrt parameterization. In addition, these programs have several options (e.g., the parameter p in the Troen&Mahrt parameterization, see equation (2.7)). The use of multiple programs makes it an efficient system to build an ensemble. Adding new options is easy since one may simply add a new program (or add the option into an existing program). Moreover the computations are well managed. For example, if two models have the same options except the deposition velocity, all computations except those depending on deposition (i.e., the computation of the deposition velocities, and the numerical integration of the reactive transport equation) will be shared.

In the second stage, the numerical solver carries out the time integration of the reactive transport equation. The numerical solver is actually embedded in a structure called the "driver". The driver is primarily in charge of perturbing the input data as detailed in section 2.3.1.

At a postprocessing stage, the ensemble is completely generated and the results are analyzed. At all stages, a few libraries, mainly in C++ and Python, offer support, especially for data manipulation.

Disk space usage is optimized since the models can share part of their preprocessing. Moreover, the perturbed input fields (table 2.2) are not stored; only the unperturbed fields (medians) are stored, and the driver applies the perturbations during the simulation.

Python scripts generate the identities (i.e., the set of options and perturbations) of all models to be launched. The corresponding configuration files are created. The scripts then launch the preprocessing programs and the simulations. The simulations can obviously be run in parallel, so the scripts can launch the programs over SSH on different machines and processors. The only constraint lies in the dependencies between the programs: e.g., the deposition velocities must be computed after the meteorological fields because they depend on winds (among other fields). Groups of programs are defined with different priorities, and the scripts launch one group after the other. It is possible to relaunch parts of the ensemble computations. It is also possible to add new models (new simulations) after an initial ensemble has been generated. The Python code is available in the module EnsembleGeneration, from Polyphemus 1.5.

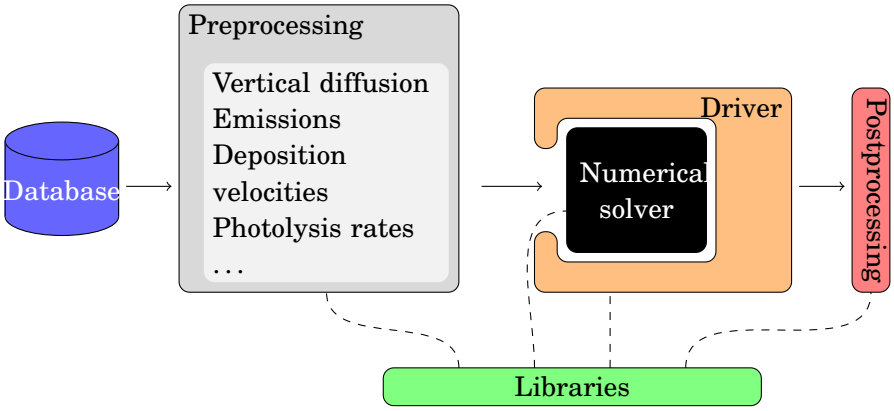


Figure 2.1: A view on Polyphemus design: database storing all raw data, preprocessing stages for most physical computations, drivers in which the numerical solver is embedded, postprocessing and libraries that may be called at any time.

The same approach may be applied to another modeling system providing enough options (in the model formulation) are available. This requires that significant diversity is maintained in the system. In particular, when a new formulation (e.g., a more accurate chemistry) is developed, the previous formulation should remain available to the user. The rationale is that, while a formulation may seem better from a deterministic point of view (based on a priori considerations or on performance analysis), the previous formulation still has a significant probability (though lower than that of the new formulation) from a stochastic point of view.

2.4 An Example of 101-Member Ensemble

With the previous method, about 620 billion models can be generated. An ensemble of 101 models is built and run throughout the year 2001 over Europe ($[10.75°W, 22.75°E]\times[34.75°N, 57.75°N]$). The models are not simplified to reduce the computational costs. All models have a 0.5 ° horizontal resolution, which is a usual resolution. Because the total computational cost is high, the ensemble size is limited to 101 simulations. This size is enough at least for the spatio-temporal empirical mean (of ozone peaks) to converge, as shown in figure 2.2.

Six reference models are included in the ensemble. These models are not generated automatically, but each of them corresponds to a possible combination of options in that they could have been selected by the automatic procedure.

Aerosols are not taken into account in these simulations. The output stored on disk are the hourly concentrations in the first layer for O_3, NO, NO_2 and SO_2—which already amounts to 45 Gb of data.

Section 2.4.1 briefly summarizes which members are included in the ensemble. Although this paper is a technical description of the ensemble generation procedure, we aim to provide insight into the ensemble structure. We review the performance of the models, compared to ground observations, in section 2.4.2. We analyze the spread of the ensemble in section 2.4.3. We do not address more complex issues like probabilistic forecasts, uncertainty estimation or sequential aggregation.

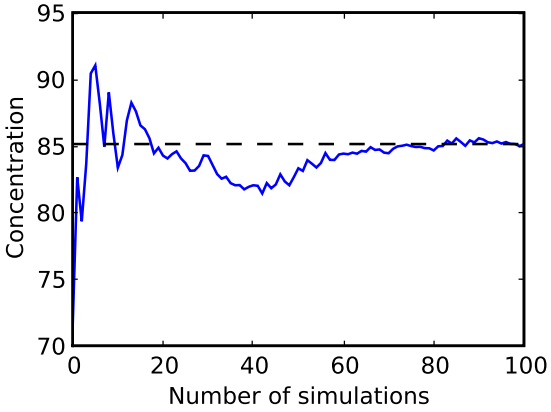


Figure 2.2: The empirical mean of the ozone peaks averaged over all stations of network 1 and during 2001. It seems to have converged after about 70 simulations.

2.4.1 Experiment Setup

In table 2.1, a probability is associated with every option. The models are built according to these probabilities, but the actual frequency of an option in the 101-member ensemble may differ slightly because of the random sampling. The occurrence frequency (in percentages) of each parameterization, numerical option and field perturbation in the 101-member ensemble are shown in table 2.3. For the field perturbations, there are three options: no perturbation (raw data), increased values in the field ($p\alpha$ if $p \geq 0$, or $p + \alpha$) and decreased values (section 2.3.1).

The six additional models can be seen as reference models. They are built with the parameterizations that we trust the most, and without any perturbation in the input field. If we had to build a model for forecast, we would a priori choose one of them. They are formed with the parameterizations and numerical options from the first column of table 2.1 but for the vertical diffusion parameterization and the mass conservation. Considering the three options for vertical diffusion (line 5) and the two options for vertical-wind diagnosis (line 13), six models may be constructed. These are listed in table 2.4.

2.4.2 Evaluation of the Ensemble Members

Performance Measures

In order to evaluate a model performance, n available observations o_i from different ground stations are compared with the corresponding simulated concentrations y_i, using

1. the root mean square error:
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - o_i)^2}\,;$$
2. the correlation:
$$\mathrm{corr} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(o_i - \bar{o})^2}}\,,$$
where $\bar{o} = \sum_{i=1}^{n} o_i$ and $\bar{y} = \sum_{i=1}^{n} y_i$;
3. the bias factor:
$$\mathrm{BF} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{o_i}\,.$$

Table 2.3: Occurrence frequency of each parameterization, numerical option and field perturbation for the 101-member ensemble. As for the perturbations, "raw" means no perturbation, "raw$^-$" means lower value after perturbation (p/α or $p-\alpha$) and "raw$^+$" means higher value after perturbation ($p\alpha$ or $p+\alpha$).

#	**Option**				
		Physical parameterizations			
1.	Land use cover	USGS (50)	GLCF (50)		
2.	Chemistry	RACM (61)	RADM 2 (39)		
3.	Attenuation	RADM method (50)	ESQUIF (50)		
4.	CRH	Depends on σ (75)	Two layers (25)		
5.	K_z	T&M (44)	Louis (30)	Louis stable (26)	
6.	Deposition velocity	Zhang (55)	Wesely (45)		
7.	Coefficient Ra	Heat flux (74)	Moment flux (36)		
8.	Emissions vertical distribution	Low (54)	Medium (46)		
9.	Photolysis rates	JPROC (88)	Zenith angle (12)		
		Numerical issues			
10.	Time step	600 s (91)	1200 s (9)		
11.	Vertical resolution	5 layers (48)	9 layers (52)		
12.	First layer height	50 m (40)	40 m (60)		
13.	Vertical-wind diagnosis	$\mathrm{div}(\rho V)=0$ (52)	$\mathrm{div}(V)=0$ (48)		
14.	Minimal K_z ($\mathrm{m^2 s^{-1}}$)	0.2 (66)	0.5 (34)		
15.	Minimal K_z in urban area ($\mathrm{m^2 s^{-1}}$)	0.2 (30)	0.5 (35)	1.0 (35)	
16.	Vertical application of minimal Kz	Yes (84)	No (16)		
17.	Exponent p	2 (75)	3 (25)		
18.	Boundary layer height	raw value (61)	+10% (18)	-10% (7)	+20% (14)
		Input data			
19.	Temperature (K)	raw (39)	raw$^-$ (34)	raw$^+$ (27)	
20.	Horizontal-wind angle (degrees)	raw (35)	raw$^-$ (31)	raw$^+$ (34)	
21.	Horizontal-wind velocity ($\mathrm{m\,s^{-1}}$)	raw (36)	raw$^-$ (40)	raw$^+$ (24)	
22.	K_z ($\mathrm{m^2\,s^{-1}}$)	raw (33)	raw$^-$ (32)	raw$^+$ (35)	
23.	O_3 boundary conditions ($\mu\mathrm{g\,m^{-3}}$)	raw (33)	raw$^-$ (36)	raw$^+$ (31)	
24.	NO_x boundary conditions ($\mu\mathrm{g\,m^{-3}}$)	raw (29)	raw$^-$ (35)	raw$^+$ (36)	
25.	VOCs boundary conditions ($\mu\mathrm{g\,m^{-3}}$)	raw (35)	raw$^-$ (37)	raw$^+$ (28)	
26.	Biogenic emissions	raw (34)	raw$^-$ (28)	raw$^+$ (38)	
27.	NO_x emissions	raw (34)	raw$^-$ (35)	raw$^+$ (31)	
28.	VOCs emissions	raw (27)	raw$^-$ (25)	raw$^+$ (48)	
29.	Deposition velocities ($\mathrm{cm\,s^{-1}}$)	raw (35)	raw$^-$ (32)	raw$^+$ (33)	
30.	Photolysis rates	raw (34)	raw$^-$ (37)	raw$^+$ (29)	

Table 2.4: Description of the 6 reference models.

#	**Vertical diffusion**	**Vertical-wind diagnosis**
R0.	T&M	$\mathrm{div}(\rho V)=0$
R1.	T&M	$\mathrm{div}(V)=0$
R2.	Louis stable - T&M unstable	$\mathrm{div}(\rho V)=0$
R3.	Louis stable - T&M unstable	$\mathrm{div}(V)=0$
R4.	Louis	$\mathrm{div}(\rho V)=0$
R5.	Louis	$\mathrm{div}(V)=0$

In practice, not all observations are retained. Stations that fail to provide observations at over 10% of all considered dates are discarded as these stations may not be reliable.

For ozone, the observations from three networks are considered:

– Network 1 is composed of 243 urban and regional stations, primarily in France and Germany (116 and 81 stations respectively). It provides about 1 365 000 hourly concentrations and 61 000 peaks.

– Network 2 includes 96 EMEP stations (regional stations distributed over Europe), with about 776 700 hourly observations and 33 300 peaks.

– Network 3 includes 371 urban and regional stations in France. It provides 2 800 000 hourly measurements and 122 000 peaks. Note that it includes most of the French stations of network 1.

The Models' Performance on Ozone

Table 2.5 shows the performance of the six reference models for ozone and of the best model in the ensemble. The best model is selected with respect to the RMSE for the considered network and target (ozone peaks or ozone hourly concentrations). It is noteworthy that, except for network 2 and for hourly concentrations, there is always at least one model in the 101-member ensemble which is better than the six reference models (according to the RMSE and the correlation). The automatic generation of 101 models therefore created models that are as good as or better than the models derived from experience.

It also generated models with poor performance. Figure 2.3 shows the performance of the 101 models sorted according to the mean, bias factor, correlation and RMSE. The performance can obviously vary greatly.

The Best Model

In this section, we define the "best model" as the model with the lowest RMSE. Of course, it depends on the target (the network, the pollutant, the time period), and considering the RMSE only is not enough to identify the best model, if any can be identified, as a modeler would do. Still this gives insights on the performance of models automatically generated.

Model 98 in the 101-member ensemble is the best model for ozone peaks on network 1, for ozone hourly concentrations and ozone peaks on network 2 (table 2.5). For these targets, it beats the reference models. Several parameterizations and numerical options of model 98 are the same as those of the reference models (photolysis rates, deposition velocities, time step, ...), but several selected options are unexpected. For instance, its chemical mechanism is RADM 2, and four fields are perturbed. See table 2.6 for the complete description of model 98. It is interesting to note that (1) the random sampling generates several models with good performance (compared to the observations, with the RMSE), (2) the random sampling generates a model with lower square errors (over a long time period) than the models tuned by the modelers.

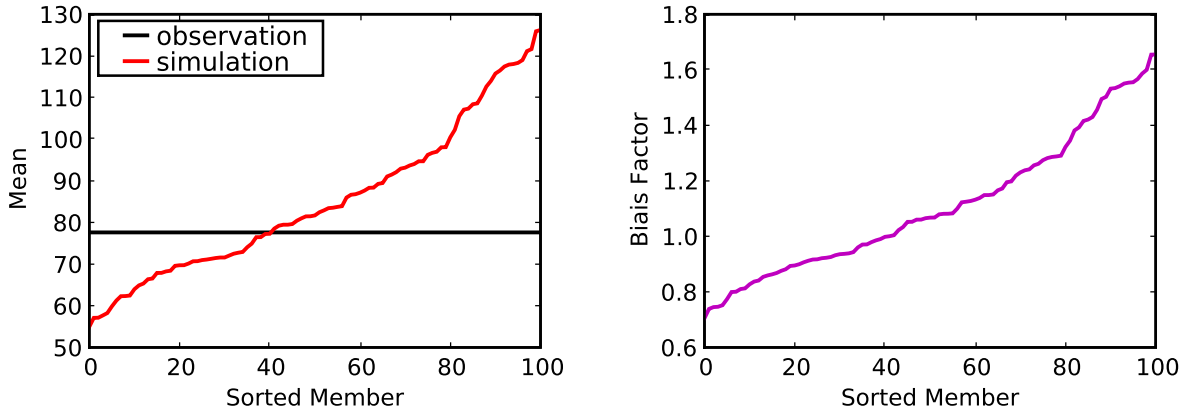
Among the 101 simulations, the median RMSE is about 27 $\mu g\,m^{-3}$ and the median correlation is close to 0.73.

2.4.3 Ensemble Variability

Every model in the ensemble is unique, but one may ask whether the ensemble contains enough information and has a rich structure. For example, the ensemble should not be clustered into distinct groups of similar models. One measure of the difference between two models is the number of options that differ between them. Interestingly enough, two models with a similar RMSE can be made with many different options: for example models 98 and 58, which have close

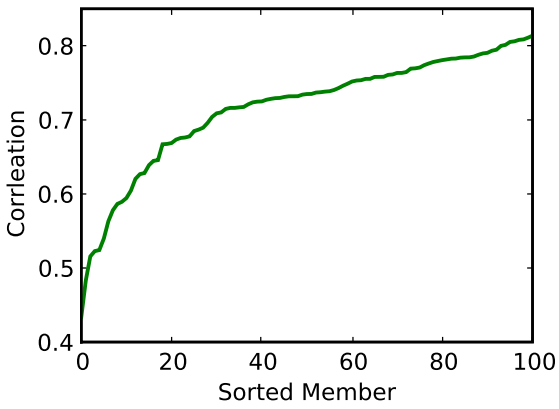
Table 2.5: Statistical measures for the 6 reference models and the best model from the 101-member ensemble, for hourly ozone concentrations and hourly ozone peaks. R0–5 refer to the 6 reference models.

#	BF	mean	corr	RMSE	#	BF	mean	corr	RMSE
Network 1 – Hourly					*Network 1 – Peak*				
R0.	1.06	62.0	0.67	28.09	R0.	1.10	85.1	0.76	24.54
R1.	0.96	55.5	0.68	25.55	R1.	1.00	76.9	0.77	23.19
R2.	1.21	72.3	0.68	31.19	R2.	1.13	87.1	0.78	24.49
R3.	1.10	65.1	0.68	26.85	R3.	1.03	78.8	0.78	22.82
R4.	0.89	51.1	0.69	25.87	R4.	0.98	75.6	0.79	23.30
R5.	0.82	46.9	0.70	25.74	R5.	0.91	70.1	0.78	23.95
48.	0.9	51.81	0.73	**22.42**	98.	1.08	83.6	0.80	**22.54**
Network 2 – Hourly					*Network 2 – Peak*				
R0.	0.99	65.2	0.64	25.28	R0.	1.06	84.2	0.73	21.66
R1.	0.90	59.0	0.64	24.90	R1.	0.97	76.7	0.73	21.51
R2.	1.12	74.1	0.65	25.74	R2.	1.09	86.0	0.74	21.22
R3.	1.02	67.3	0.65	**23.52**	R3.	0.99	78.4	0.74	20.74
R4.	0.83	54.2	0.66	26.47	R4.	0.93	74.4	0.74	23.36
R5.	0.77	50.1	0.66	27.75	R5.	0.87	69.5	0.73	24.80
98.	1.05	69.1	0.67	24.02	98.	1.04	82.6	0.76	**20.24**
Network 3 – Hourly					*Network 3 – Peak*				
R0.	1.12	65.1	0.64	31.18	R0.	1.15	86.0	0.76	26.59
R1.	1.01	58.0	0.66	27.21	R1.	1.04	77.3	0.76	23.98
R2.	1.27	75.2	0.66	35.98	R2.	1.18	88.0	0.77	27.09
R3.	1.15	67.4	0.67	30.44	R3.	1.06	79.2	0.77	24.15
R4.	0.96	54.3	0.68	26.34	R4.	1.02	76.9	0.79	23.09
R5.	0.89	49.7	0.69	25.13	R5.	0.95	71.1	0.79	22.79
48.	0.93	52.8	0.72	**23.29**	99.	0.91	68.5	0.81	**22.41**

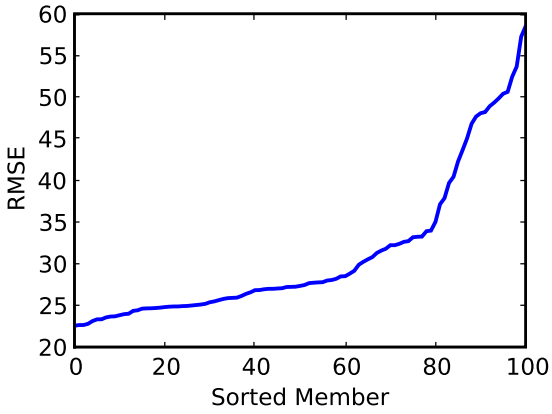


(a) Mean

(b) Bias Factor



(c) Correlation



(d) RMSE

Figure 2.3: Mean ($\mu g m^{-3}$), bias factor, correlation and RMSE ($\mu g m^{-3}$) for ozone peaks on network 1. In each plot, the models are sorted according to the indicator.

Table 2.6: Description of the best model.

#	**Name**	
	Physical parameterizations	
1.	Land use cover	GLCF
2.	Chemistry	RADM 2
3.	Attenuation	ESQUIF
4.	CRH	Two layers
5.	K_z	T&M unstable – Louis stable
6.	Deposition velocity	Zhang
7.	Coefficient Ra	Heat flux
8.	Emissions vertical distribution	Low
9.	Photolysis rates	JPROC
	Numerical issues	
10.	Time step	600 s
11.	Vertical resolution	5 layers
12.	First layer height	40 m
13.	Mass conservation	$\mathrm{div}\left(\rho V\right) = 0$
14.	Minimal K_z ($\mathrm{m^2 s^{-1}}$)	0.2
15.	Minimal K_z in urban area ($\mathrm{m^2 s^{-1}}$)	1.0
16.	Vertical application for minimal Kz	Yes
17.	Exponent p	3
18.	Boundary layer height	raw value
	Input data	
19.	Temperature (K)	raw^+
20.	Horizontal-wind angle (degrees)	raw
21.	Horizontal-wind velocity ($\mathrm{m{\cdot}s^{-1}}$)	raw
22.	K_z ($\mathrm{m^2 s^{-1}}$)	raw^-
23.	O_3 boundary conditions ($\mu\mathrm{g\,m^{-3}}$)	raw
24.	NO_x boundary conditions ($\mu\mathrm{g\,m^{-3}}$)	raw
25.	VOCs boundary conditions ($\mu\mathrm{g\,m^{-3}}$)	raw
26.	Biogenic emissions	raw^+
27.	NO_x emissions	raw
28.	VOCs emissions	raw^-
29.	Deposition velocities ($\mathrm{m{\cdot}s^{-1}}$)	raw
30.	Photolysis rates	raw

RMSEs (22.54 and 23.65 respectively, ozone peak, network 1), are generated with 17 different options (out of 30) shown in table 2.7. This fact can be observed with the whole ensemble. In figure 2.4, the models are sorted according to their RMSE for ozone peaks on network 1 (model 0 has the lowest RMSE, and model 100 has the highest RMSE), and the matrix of the differences between the models (measured with the number of differing options) is shown. No overall structure can be identified. This tends to show that quite different models can achieve similar performance. The RMSE, seen as a function of the parameters, seems to have many local minima.

On the other hand, the output of the best models are correlated. This is shown in figure 2.5 with the correlation computed with all ozone peaks observed in network 1. Two skillful models therefore have a similar spatio-temporal variability.

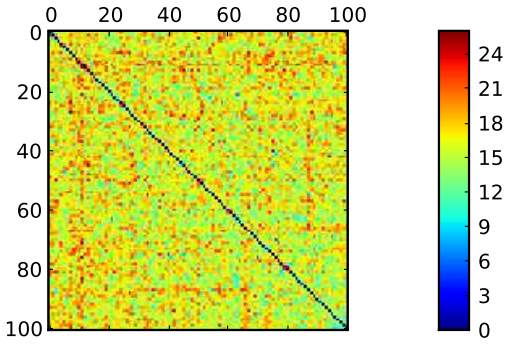


Figure 2.4: Matrix of the number of different options between two models. The models are sorted according to the RMSE (from the best to the worst value).

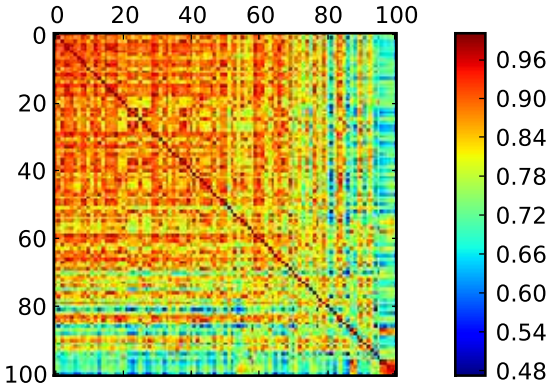
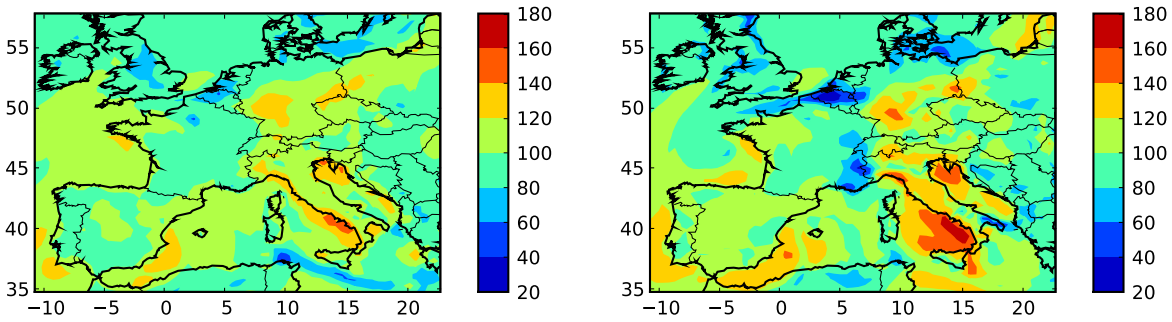


Figure 2.5: Matrix of correlation between all observed ozone peaks (on network 1) and the corresponding model-concentrations. The models are sorted according to the RMSE (from the best to the worst value).

These high correlations are partly due to the structure of ozone fields. Because of the physical constraints, two reasonable ozone fields necessarily share a set of common features, such as higher concentrations in the south compared to the north, or low concentrations at high NO

emission sources. However, two skillful models can show significant differences in their spatial patterns, as figure 2.6 demonstrates.



(a) Model 98 (b) Model 58

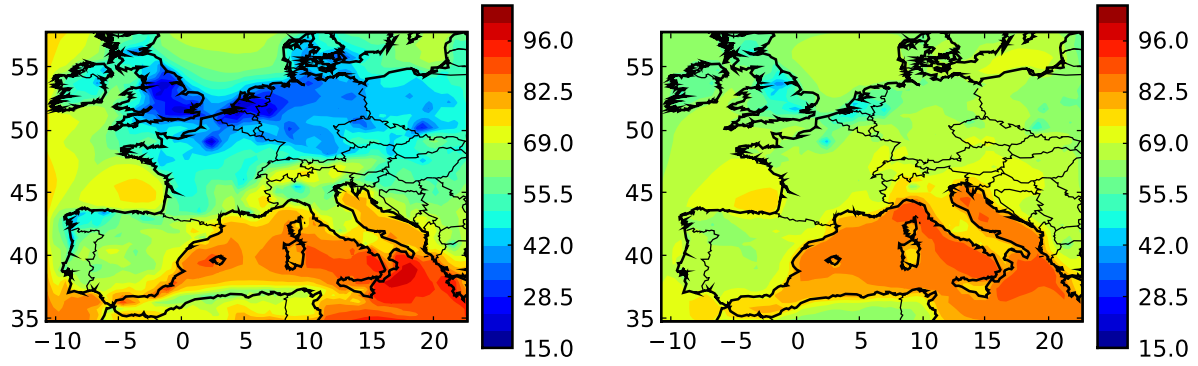
Figure 2.6: Ozone map of model 98 (left) and model 58 (right), on 5 May 2001 at 17:00 UT. Both models show good performance, but they can produce ozone fields that differ significantly.

Table 2.7: A comparison between the model 98 and 58.

Name	**Model 98**	**Model 58**
Chemical Mechanism	RADM 2	RACM
Cloud attenuation	ESQUIF	RADM
Critical relative humidity	on 2 layers	with σ
Vertical diffusion	Troen & Mahrt unstable – Louis stable	Louis
Coefficient Ra	Heat flux	Moment flux
Vertical resolution	5 levels	9 levels
Time step	600 s	1200 s
Exponent p for K_z	3	2
First layer height	40 m	50 m
Minimal K_z in urban area	1.0	0.5
Temperature	raw$^+$	raw
NO_x boundary conditions	raw	raw$^+$
VOCs boundary conditions	raw	raw$^+$
Biogenic emissions	raw$^+$	raw
NO_x emissions	raw	raw$^+$
VOCs emissions	raw$^-$	raw
Deposition velocities	raw	raw$^+$

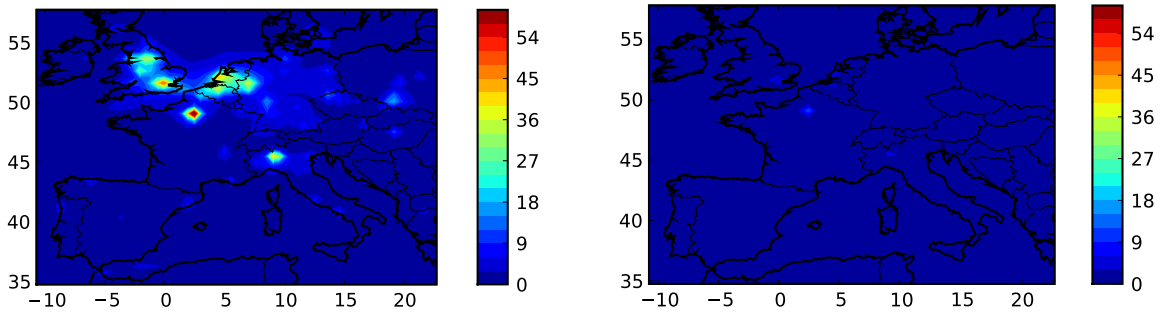
Figures 2.7, 2.8, 2.9 and 2.10 show the temporal mean of the concentration map of the fifth reference model and of a model from the 101-member ensemble, for O_3, NO, NO_2 and SO_2 respectively. Again, the physical constraints make the models reproduce specific features, like high NO concentrations only at emission locations, but significant differences are found.

Figures 2.11 shows the mean daily profiles of all models from the 101-member ensemble, for O_3, NO, NO_2 and SO_2 respectively. For the species O_3 and NO_2, the daily profiles are computed on network 3 whereas the daily profiles for the species NO and SO_2 are computed with all cells. All models produce a similar profile shape, which is due to the physical phenomena accounted for in every model and the fact that these profiles are highly averaged (whole year, and full domain



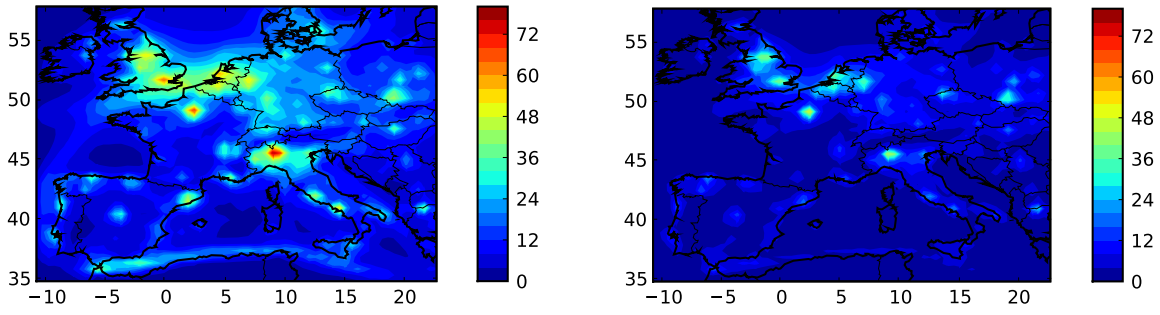
(a) Reference 5 (b) Model 76

Figure 2.7: Temporal average of ozone map for reference model 5 (left) and for model 76 of the 101-member ensemble (right).



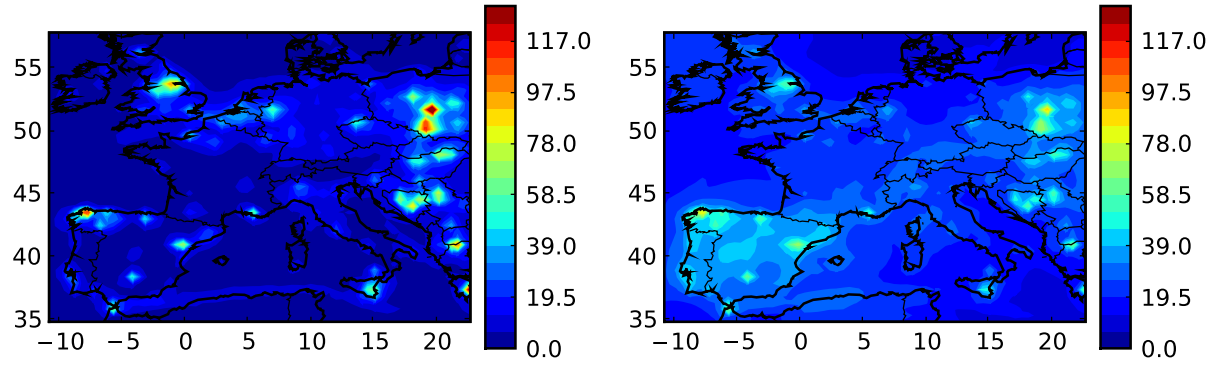
(a) Reference 5 (b) Model 52

Figure 2.8: Temporal average of NO map for reference model 5 (left) and for model 52 of the 101-member ensemble (right).



(a) Reference 5 (b) Model 90

Figure 2.9: Temporal average of NO_2 map for reference model 5 (left) and for model 90 of the 101-member ensemble (right).



(a) Reference 5

(b) Model 88

Figure 2.10: Temporal average of SO_2 map for reference model 5 (left) and for model 88 of the 101-member ensemble (right).

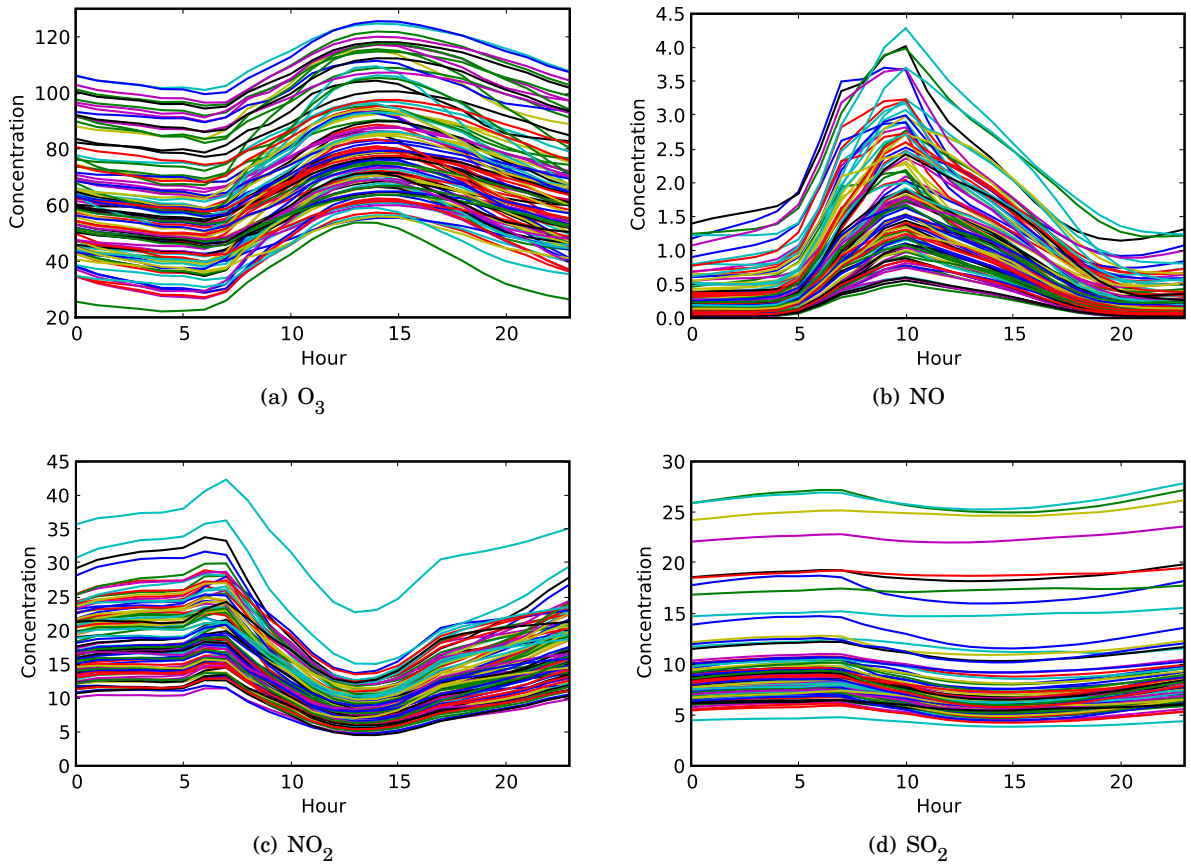
or all stations). The means can differ a lot, and, obviously, not all models are equally likely.

Nevertheless, even if the average performance of a model is very low, it may produce the best forecast at some location or some date. In other words, from a stochastic viewpoint, the model may have a very low probability, but it is still likely to produce the best forecast. This can be verified with a "map of the best-model index". At a given date, the best model in each grid cell is determined as follows. The concentrations of the models and the observed concentration at the closest station (to the grid cell) are compared. The model that produces the closest concentration to the observed concentration is considered as the best model in the grid cell. Hence, in every grid cell, one "best model" is determined. A color is associated to each model (actually each model index) to generate the maps in figure 2.12. These maps show the best model for three different dates in June 2001. The best model varies frequently from one grid cell to another, and from one date to another. This shows that many models bring useful information, at least in some regions or on given dates.

2.5 Conclusions

This paper describes how a large ensemble may be automatically generated using the Polyphemus system. Contrary to most traditional approaches, which are based on perturbations of input data only, or on small ensembles of models from different teams, our approach takes into account all sources of uncertainties at once: input data, physical formulation and numerical formulation. Each member of the ensemble is a complete chemistry-transport model whose contents are clearly defined within the modeling platform. In this context, the ensemble and the differences between its members can be rigorously analyzed, and also controlled through the probabilities associated with every option. Our approach tries to combine the flexibility of Monte Carlo simulations (large ensembles of simulations with perturbed input data) and the completeness of a multimodel ensemble (models with alternative physical parameterizations, like in ensembles made of a few models from different teams).

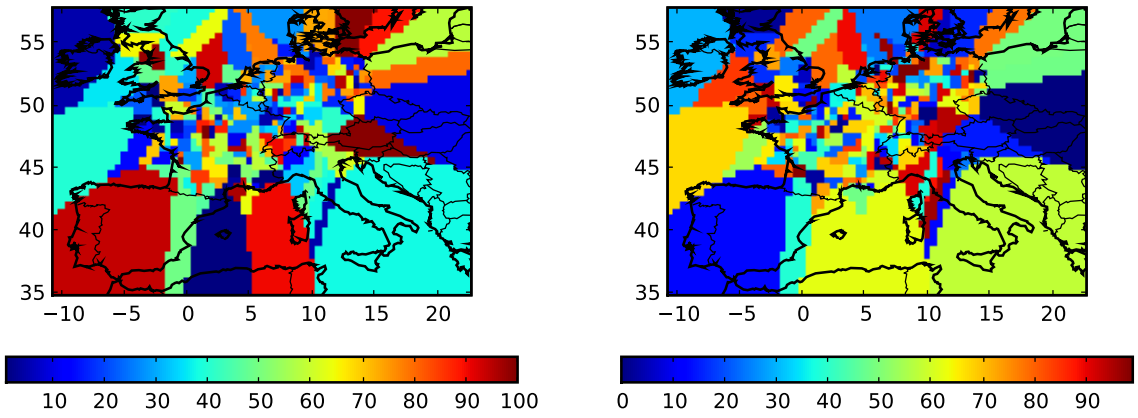
In the ensemble, each model is defined by a unique set of physical parameterizations, numerical schemes and input data. Hence building a model means picking an option for every alternative that the system provides. The options are associated with probabilities—depending



(a) O_3 (b) NO

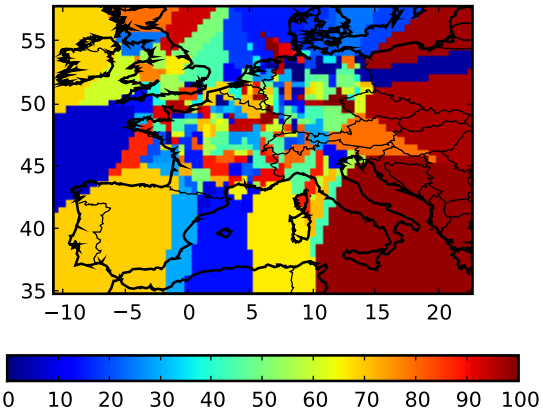
(c) NO_2 (d) SO_2

Figure 2.11: Daily profile for ozone (*network 3*), NO, NO_2 (*network 3*) and SO_2. The profile is computed at observation stations for O_3 and NO_2. It is computed with all computed values (that is, from all grid cells) for NO and SO_2.



(a) June 1^{st}

(b) June 11^{th}



(c) June 13^{th}

Figure 2.12: Maps of best-model indexes. In each grid cell of the domain, the color shows which model (marked with its index, in $[\![0,100]\!]$) gives the best ozone peak forecast on 1 June, 11 June and 13 June 2001 at the closest station to the cell center. It shows that many models can deliver the best forecast at some point. Stations of network 1 are used. Of course, the colors are only reliable in regions that contain stations.

on how reliable the option is supposed to be—and they are randomly selected. In addition, input data is sampled from normal or log-normal distributions.

The computations are carried out, from the preprocessing to the actual simulation, using small programs whose output results may be shared by different models. This minimizes computational costs and increases flexibility. Thanks to the automatic procedure, the configuration and the generation of an arbitrarily-large ensemble is straightforward. The method can be applied to any simulation with Eulerian models in Polyphemus, such as simulations over a smaller region, or simulations with aerosols.

The ensemble given as example includes 101 photochemical models generated and run for the year 2001, over Europe. The ensemble has a wide spread for all chemical species. The models show a strong diversity both in their formulation and their performance. Many of them appear to be the best in many different regions and periods.

Many research issues are related to this procedure. One relates to the choice of the models to be included in the ensemble. How many models should be included for the ensemble to properly represent the uncertainties? Which models should be included? What probabilities should be associated with the options, and what distributions should be given to the input data? How should meteorological ensembles be integrated? Other research issues may deal with the best structure for an ensemble. How does our procedure compare with other approaches, such as Monte Carlo simulations or small ensembles based on models from different teams? How much information on the uncertainties is provided by the different approaches?

2.6 Appendix

2.6.1 Emissions from EMEP

As described in section 2.2, anthropogenic emissions are provided by EMEP. The vertical distribution of the pollutants depends on SNAP category. Two vertical distributions are used in this paper: a "*low* distribution" and a "*medium* distribution"—see table 2.8.

Table 2.8: Emission distribution in percentages for each level and for each SNAP category. Combustion in energy and transformation industries (S1) ; non-industrial combustion plants (S2) ; combustion in manufacturing industry (S3) ; production processes (S4) ; extraction and distribution of fossil fuels and geothermal (S5) ; solvent use and other product use (S6) ; road transport (S7) ; other mobile sources machinery (S8) ; waste treatment and disposal (S9) and agriculture (S10).

SNAP	**Low**					**Medium**				
	ground	0–50 m	50–150 m	150–300 m	>300 m	ground	0–50 m	50–150 m	150–300 m	>300 m
S1	0	28.6	71.4	0	0	0	13.8	34.5	51.7	0
S2	12.5	50	37.5	0	0	6.6	26.7	66.7	0	0
S3	0	28.6	71.4	0	0	0	13.8	34.5	51.7	0
S4	25	75	0	0	0	22.2	77.8	0	0	0
S5	25	75	0	0	0	22.2	77.8	0	0	0
S6	100	0	0	0	0	100	0	0	0	0
S7	100	0	0	0	0	100	0	0	0	0
S8	100	0	0	0	0	100	0	0	0	0
S9	0	28.6	71.4	0	0	0	13.8	34.5	51.7	0
S10	100	0	0	0	0	100	0	0	0	0

Acknowledgements
We thank Richard James for proofreading the paper.

Chapitre 3

Calibration d'ensemble, estimation de l'incertitude et prévision probabiliste

On s'intéresse dans ce chapitre au problème de calibration d'un ensemble pour l'estimation de l'incertitude. La méthode de calibration comprend (1) un ensemble généré de manière automatique, (2) un score d'ensemble tel que la variance du diagramme de rang et (3) la sélection d'un sous-ensemble basée sur un algorithme d'optimisation combinatoire et qui minimise le score d'ensemble choisi. Les scores d'ensemble sont le score de Brier (dédié aux prévisions probabilistes) ou sont issus du diagramme de rang ou du diagramme de fiabilité. Ces scores permettent de mesurer la qualité d'une estimation d'incertitude, la fiabilité et la résolution d'un ensemble. L'ensemble est généré à l'aide de la méthode « multi-modèles » décrite précédemment. C'est le même ensemble, d'une centaine de simulations d'ozone au voisinage du sol lancé sur toute l'Europe durant l'année 2001, qui est utilisé dans ce chapitre. Nous avons évalué cet ensemble avec les scores précédemment mentionnés. Plusieurs calibrations d'ensemble sont réalisées selon différents scores d'ensemble. La calibration permet de sélectionner des ensembles de 20–30 membres. Ces ensembles calibrés améliorent grandement les scores d'ensemble : essentiellement la fiabilité, alors que la résolution reste inchangée. La robustesse spatiale des cartes d'incertitude est effectuée grâce à une validation croisée. L'impact du nombre d'observations et des erreurs d'observations est aussi vérifié. Finalement, les ensembles calibrés permettent de produire des prévisions probabilistes plus précises et de prévoir les incertitudes, bien que ces incertitudes soient fortement dépendantes du temps.

Sommaire

Ce chapitre est constitué de Garaud et Mallet [2011].

3.1 Introduction

Air quality simulation involves complex numerical models that rely on large amounts of data from different sources. Most of the input data is provided with high uncertainties in their time evolution, spatial distribution and even average values. Chemistry-transport models are themselves subject to uncertainties in both their physical formulation and their numerical formulation. The multi-scale nature of the problem leads to the introduction of subgrid parameterizations that are an important source of errors. The dimensionality of the numerical system, involving up to hundreds of pollutants in a three-dimensional mesh, is much higher than the number of observations, which also leads to high uncertainties in non-observed variables.

In order to quantify the uncertainties, classical approaches rely on Monte Carlo simulations. The input fields and parameters of the chemistry-transport model are viewed as random vectors or random variables. These are sampled according to their assumed probability distribution, and a model run is carried out with each element of the sample. The set of model outputs constitutes a sample of the probability distribution function of the output concentrations. Typically, the empirical standard deviation of the output concentrations measures the simulations uncertainties. This approach has been applied for air quality simulations [Hanna *et al.*, 1998, 2001; Beekmann et Derognat, 2003].

Another approach is the use of models which differ by their numerical formulation or physical formulation. The models can originate from different research groups [e.g., van Loon *et al.*, 2007; Delle Monache et Stull, 2003; McKeen *et al.*, 2005; Vautard *et al.*, 2009] or from the same modular platform Mallet et Sportisse [2006b]. In addition to this multimodel strategy, the input data can also be perturbed so that all uncertain sources are taken into account. It is also possible to choose between different emission scenarios and meteorological forecasts as Delle Monache *et al.* [2006a,b] did. Pinder *et al.* [2009] split the uncertainty into a structural uncertainty due to the weaknesses in the physical formulation and a parametric uncertainty due to the errors in the input data. In Garaud et Mallet [2010], the ensemble is built with several models randomly generated within the same platform and with perturbed input data.

Whatever the strategy for the generation of an ensemble, several assumptions are made by the modelers. One needs to associate probability density functions to every input field or parameter to be perturbed. Under the usual assumption that the distribution of a field or parameter is either normal—or log-normal, one has to estimate a median and a standard deviation. For a field, providing a standard deviation is complex as it should take into account spatial correlations, and possibly time correlations. As for multimodel ensembles, one has little control over the composition of the models when they are provided by different teams. When the models are derived within the same platform, the key points are the amount of choice in the generation of an individual model, and the probability associated to each choice. Once all the assumptions and choices have been made, it is technically possible to generate an ensemble. However, it is quite difficult to determine the proper medians and standard deviations of the perturbed fields, and to design a multimodel ensemble that properly takes into account all formulation uncertainties.

In order to evaluate the quality of an ensemble, several a posteriori scores compare the ensemble simulations with observations. These scores, such as rank histograms, reliability diagrams or Brier scores, assess the reliability, the resolution or the sharpness of an ensemble. For instance, a reliable ensemble gives a well estimated probability for a given event in comparison to the frequency of occurrence of this event, whereas the resolution describes the capacity of an ensemble to give different probabilities for a given event.

Improving the quality of an ensemble should lead to improved scores, e.g., to a flat rank diagram or low Brier score. One strategy could be tuning the perturbations of the input fields or optimizing the design of the multimodel ensemble (that is, choosing or developing physical

parameterizations or numerical schemes, and better weighting each design option), so as to minimize or maximize some score. This is a complex and computationally expensive task that would require the generation of many ensembles.

In this paper, we adopt a strategy based on a single, but large, ensemble. Out of a large ensemble, a combinatorial optimization algorithm extracts a sub-ensemble that minimizes (or maximizes) a given score such as the variance of a rank diagram. This process is referred to as (a posteriori) calibration of the ensemble. Section 3.2 describes it in detail. It is applied in Section 3.3 to a 101-member ensemble of ground-ozone simulations with full chemistry-transport models run across Europe during the year 2001. The scores of the full ensemble and the optimized sub-ensemble (i.e., the calibrated ensemble) are studied, based on observations at ground stations. In Section 3.4, the uncertainty estimation given by the calibrated ensemble is analyzed. In Section 3.5, probabilistic forecasts for threshold exceedance are studied.

3.2 Calibration Method

Hamill et Colucci [1997] use rank histograms to calibrate precipitation probabilistic forecasts. When the ensemble is not reliable enough, the probabilistic forecasts cannot be derived directly from the ensemble relative frequencies. Assuming the shape of the rank histogram remains the same in the forecast period, the authors propose to rely on the past rank distribution to compute the probabilistic forecasts. Hopson et Webster [2010] calibrate an ensemble prediction to improve floods forecasting. An empirical cumulative distribution function is provided by ensemble predictions of precipitation. Then, it is calibrated with observations, using a quantile-to-quantile mapping technique.

In this paper, by "ensemble calibration" we mean extracting a sub-ensemble from a large ensemble so that a certain criterion is satisfied. A preliminary step is therefore to generate a large ensemble, composed of simulations that are sufficiently different from each other to provide substantial information. A criterion is defined to assess the quality of an ensemble, and a corresponding score measures how well the criterion is satisfied. An automatic selection of a sub-ensemble is finally carried out to minimize the score. The criterion usually assesses the uncertainty representation of an ensemble, based on the additional information brought by the observations. This section details the method employed to generate a large ensemble and to carry out an automatic calibration.

3.2.1 Generation of a Large Ensemble

The method employed for the automatic generation of a large ensemble is described in Garaud et Mallet [2010]. A wide range of options should be available for the design of a single model: several physical parameterizations, several numerical discretizations, different sources for the input data and random perturbations in the input fields. In the paper referred to, thirty alternatives are available for the generation of a single model. Each member of the ensemble is defined after the random selection of one option per alternative.

In this paper, we rely on the same ensemble as in Garaud et Mallet [2010]. It includes 101 members run throughout the year 2001 over Europe. This ensemble will be used and calibrated in Section 3.3.

3.2.2 Automatic Selection

Suppose a base ensemble with N members. There are $\sum_{k=1}^{N}\binom{N}{k}$ possible sub-ensembles [1]. If $N = 100$, there are over 10^{30} sub-ensembles. It is obviously impossible to consider all possible combinations in order to select the best combination with respect to the given criterion. Consequently a combinatorial optimization algorithm is required to minimize the score associated with the criterion.

Let $\mathscr{E}$ be the full ensemble and $\mathscr{S}$ be a sub-ensemble of $\mathscr{E}$. $\mathscr{S} \subseteq \mathscr{E}$ is supposed to be non-empty. Let $J(\mathscr{S})$ be the score of $\mathscr{S}$. The following sections describe different scores and algorithms which may be used in the ensemble calibration.

Criterion and Score

The main reasons for generating an ensemble are to improve forecasts with the so-called ensemble forecasts, and to estimate the uncertainty in the model's output. In this paper, we focus on the second objective. The criterion typically measures the quality of an uncertainty estimation or of the prediction of exceeding a threshold. It can be based on two desirable features of an ensemble:

1. Reliability: an ensemble has high reliability when its probabilistic forecasts for a given event match, on average, the observed frequency of this event.
2. Resolution: the capacity of the prediction system to distinguish the outcomes for a given event.

Rank Histogram A rank histogram measures the reliability of an ensemble. Let $\{x_1,\ldots,x_j,\ldots,x_N\}$ be the output of a N-member ensemble at a given time, sorted in increasing order. This ensemble is considered as a sample of a random variable X with some probability distribution, which means that all x_j are supposed to follow the same probability distribution. Let Y be a random variable representing the true state. At a given point, if Y has the same probability distribution as X, then $\mathrm{E}_X[P_Y(y \leq x_j)] = \frac{j}{N+1}$, where $\mathrm{E}_X[.]$ denotes the expectation related to X, P_Y the probability associated with Y and y a realization of the true state, i.e., an exact measured ozone concentration for instance. The rank histogram, developed by Anderson [1996]; Talagrand *et al.* [1999]; Hamill et Colucci [1997], is computed by counting the rank of the true state to an actual sorted ensemble of forecasts. A perfect diagram is flat, whereas a U-shaped rank histogram means a lack of variability in the ensemble.

Let r_j be the number of observations of rank j. An observation of rank j is an observation which is higher than the concentrations of exactly j members of the ensemble. Suppose we have M observations. The expectation of r_j is $\bar{r} = \mathrm{E}[\sum_{m=1}^{M} P_Y(x_j < y_m \leq x_{j+1})] = \frac{M}{N+1}$. The score related to the rank histogram flatness is based on the squared error

$$\mathscr{S} = \sum_{j=0}^{N} (r_j - \bar{r})^2 . \tag{3.1}$$

The score $\mathscr{S}$ gets lower as the histogram gets flatter, since $\bar{r}$ corresponds to the height of a flat histogram. Obviously, this measure depends on the number of members. It can be normalized by $\mathscr{S}_0 = \mathrm{E}[\mathscr{S}] = \frac{NM}{N+1}$ because $\mathrm{E}[(r_j - \bar{r})^2] = \frac{NM}{(N+1)^2}$. Finally the following score is used to measure the flatness of the rank histogram:

$$\delta = \frac{N+1}{NM} \sum_{j=0}^{N} (r_j - \bar{r})^2 , \tag{3.2}$$

which should ideally be close to 1.

1. Also noted $C(N,k)$ or C_k^N: number of k-element subsets of an N-element set.

Reliability Diagram Instead of simply predicting whether an event will occur or not, an ensemble can provide a probabilistic forecast. This is especially useful for the prediction of a threshold exceedance. A basic probabilistic forecast may be given by the number of models which exceed the threshold over the total number of models [Anderson, 1996]. In order to construct a reliability diagram, the range of forecast probabilities, $[0,1]$, is divided into $K+1$ bins $[p_0,p_1]$, $\ldots$, $[p_k,p_{k+1}]$, $\ldots$, $[p_{K-1},p_K]$ where $p_0=0$, $p_K=1$, and the sequence $(p_k)_k$ is increasing. Let O_k be the (observed) relative occurrence frequency of the event when the ensemble predicts in $[p_k,p_{k+1}]$. A reliable ensemble should give $O_k \in [p_k,p_{k+1}]$. The reliability diagram [Wilks, 2005] plots O_k against p_k or $\frac{1}{2}(p_k+p_{k+1})$. A perfect reliability diagram should follow the diagonal.

Brier score The Brier score measures the mean squared probability error for a specific event [Brier, 1950; Wilks, 2005]. Let M be the total number of observations. Let p_i be the forecast probability and o_i be the observed probability at a date i. The observed probability o_i is equal to 1 if the event occurred, and 0 otherwise. The Brier score is given by:

$$\mathscr{B} = \frac{1}{M}\sum_{i=1}^{M}(p_i-o_i)^2. \tag{3.3}$$

A Brier score for an ensemble can be compared with the Brier score of the climatological forecast. The climatological forecast is given by a single occurrence frequency o_c, observed in the past. If o_i follows the Bernoulli distribution and is equal to 1 with the frequency o_c and to 0 with the frequency $1-o_c$, the expectation of the Brier score $\mathscr{B}_{cl}$ of the climatological forecast is given by

$$\mathscr{B}_{cl} = \frac{1}{M}\sum_{i=1}^{M}[o_c(o_c-1)^2+(1-o_c)o_c^2] = o_c(1-o_c). \tag{3.4}$$

The so-called Brier skill score is defined by

$$\mathscr{B}_s = 1-\frac{\mathscr{B}}{o_c(1-o_c)}. \tag{3.5}$$

It ranges between $[-1,1]$ and is greater than 0 when the ensemble prediction gives a better forecast than the climatological forecast.

Discrete ranked probability score Suppose a set of L events, and let p_{li} be the forecast probability for the l-th event at the date i. The total number of observations M is the same for each event. The discrete ranked probability score (DRPS), which is a variant of RPS (ranked probability score) [Epstein, 1969a; Murphy, 1971], is given by:

$$\begin{aligned}\mathrm{DRPS} &= \frac{1}{LM}\sum_{i=1}^{M}\sum_{l=1}^{L}(p_{il}-o_{il})^2 \\ \mathrm{DRPS} &= \frac{1}{L}\sum_{l=1}^{L}\mathscr{B}(\mathscr{E}_l).\end{aligned} \tag{3.6}$$

This score is a generalization of the Brier score from a single event to a set of events.

While the rank histogram and the reliability diagram measure the reliability of a prediction system, the Brier score, and thus the DRPS, can measure the reliability and the resolution of an ensemble as shown in Murphy [1973]. The latter scores can be broken down into three terms: reliability, resolution and uncertainty. For instance, the Brier score is an estimation of $\mathrm{E}[(p-o)^2]$. Let p_0 be the specific probability for a given event $\mathscr{E}$ and O_0 be the occurrence frequency of $\mathscr{E}$ when

p_0 is provided. The occurrence of $\mathscr{E}$ denoted o follows Bernoulli's distribution. Thus, o takes value 1 with frequency O_0 and takes value 0 with frequency $1-O_0$. The expected value of $(p_0-o)^2$ is

$$\begin{aligned}\mathrm{E}[(p_0-o)^2] &= (p_0-1)^2 O_0 + p_0^2(1-O_0)\\ &= (p_0-O_0)^2 + O_0(1-O_0).\end{aligned} \tag{3.7}$$

Then, we compute (3.7) for many probabilities. In our case, the prediction system provides discrete probabilities for a given event. Suppose the system provides $K+1$ different probabilities denoted p_k, ranging in $[0,1]$. Let n_k be the number times p_k is computed with the ensemble. Thus, the frequency distribution of p_k is given by $\frac{n_k}{M}$ with M the total number of considered dates, i.e., the total number of observations. We have $\frac{1}{M}\sum_{k=0}^{K} n_k = 1$. Let O_k be the observed occurrence frequency of the event when the ensemble predicts p_k. The climatological occurrence frequency is $o_c = \frac{1}{M}\sum_{k=0}^{K} n_k O_k$.

$$\begin{aligned}\mathscr{B} &= \mathrm{E}[(p-o)^2]\\ &= \frac{1}{M}\sum_{k=0}^{K} n_k (p_k - O_k)^2 + \sum_{k=0}^{K} n_k O_k (1-O_k)\\ &= \underbrace{\frac{1}{M}\sum_{k=0}^{K} n_k (p_k - O_k)^2}_{\text{reliability}} - \underbrace{\sum_{k=0}^{K} n_k (O_k - o_c)^2}_{\text{resolution}} + \underbrace{o_c(1-o_c)}_{\text{uncertainty}}.\end{aligned} \tag{3.8}$$

The first term is a reliability term since it compares the probability provided by the forecast system with the occurrence frequency of the event. The second term is called "resolution" and is equivalent to the variance of O_k. The third one is the "uncertainty" term which corresponds to the score of the climatological forecast. It is constant for a specific event and is maximum when the climatological forecast is equal to 0.5. This means that the climatological forecast has the worst Brier score when it provides the most uncertain occurrence probability, i.e., 0.5. The same decomposition can be carried out for the Brier skill score and the DRPS (3.9).

$$\begin{aligned}DRPS = &\frac{1}{LM}\sum_{l=1}^{L}\sum_{k=0}^{K} n_k^l (p_{lk} - O_{lk})^2\\ &- \frac{1}{LM}\sum_{l=1}^{L}\sum_{k=0}^{K} n_k^l (O_{lk} - o_{lc})^2\\ &+ \frac{1}{L}\sum_{l=1}^{L} o_{lc}(1-o_{lc}).\end{aligned} \tag{3.9}$$

The choice of a criterion, i.e., an ensemble score, is the first step of the ensemble calibration. The second step is the choice of a combinatorial optimization algorithm.

Combinatorial optimization algorithm

Two combinatorial optimizations are employed in order to minimize the scores previously introduced: a genetic algorithm and simulated annealing.

Genetic algorithm The genetic algorithm, described in Fraser et Burnell [1970] and Crosby [1973], takes evolutionary biology as its basis, with the selection, crossover and mutation of a population of individuals. Let $\mathscr{S}_i$ be an individual, that is, a sub-ensemble, and let $\mathscr{P} = \{\mathscr{S}_1,\ldots,\mathscr{S}_i,\ldots,\mathscr{S}_{N_{pop}}\}$ be a population of N_{pop} individuals. The first step of the genetic algorithm is the random generation of the first population (denoted $\mathscr{P}^0$). Each $\mathscr{S}_i$ randomly collects an arbitrary number of models of the ensemble $\mathscr{E}$. Then, three important steps generate the population $\mathscr{P}^{k+1}$ based on $\mathscr{P}^k$:

1. Selection: a few individuals are selected according to some method. In practice, we select half the best individuals with respect to the score.
2. Crossover: among the selected individuals, a crossover is carried out. Two parents $\mathscr{S}_a$ and $\mathscr{S}_b$ create two new children $\mathscr{S}_c$ and $\mathscr{S}_d$. All the models of $\mathscr{S}_a$ and $\mathscr{S}_b$ are randomly dispatched into $\mathscr{S}_c$ and $\mathscr{S}_d$. The list of models in an individual can be seen as its genetic print. A new population denoted $\tilde{\mathscr{P}}^{k+1}$ is generated with $N_{pop}/2$ parents and $N_{pop}/2$ children.
3. Mutation: each individual of the previous population $\tilde{\mathscr{P}}^{k+1}$ can mutate. In our case, a model can be replaced by another one, removed from an individual or added to an individual. These mutations constitute the new population $\mathscr{P}^{k+1}$.

The operation is repeated until some stopping criterion has been satisfied, e.g., when a given number of iterations is reached. The final population contains many individuals that are better (with respect to the cost function) than those of the initial population. It is the best individual of the final population that is considered as the calibrated ensemble.

Simulated annealing Simulated annealing, described in Kirkpatrick *et al.* [1983], is a basic optimization method inspired by a thermodynamic process. Each sub-ensemble of the search space is analogous to a state of some physical system.

In our case, the first state is just a random generation of a sub-ensemble. The current state has a lot of neighbor states which correspond to the current state with a unit change, that is, a removed, added or replaced model in the sub-ensemble. Let $\mathscr{S}$ be the current sub-ensemble and $\mathscr{S}'$ be a neighbor sub-ensemble. $\mathscr{S}'$ is a new sub-ensemble which is randomly built from the current sub-ensemble with one removed, added or replaced model. In order to minimize (*resp.* maximize) a score J, two transitions to the neighbor are possible:

1. If the score $J(\mathscr{S}')$ is lower (*resp.* higher) than $J(\mathscr{S})$, then the current sub-ensemble moves to the neighbor sub-ensemble. $\mathscr{S}'$ becomes the current sub-ensemble and another neighbor is generated.
2. If the $J(\mathscr{S}')$ is greater (*resp.* lower) than $J(\mathscr{S})$, moving to $\mathscr{S}'$ is allowed to occur with an acceptance probability. This acceptance probability is equal to $\exp(-\frac{J(\mathscr{S}')-J(\mathscr{S})}{T})$ (*resp.* $\exp(\frac{J(\mathscr{S}')-J(\mathscr{S})}{T})$) where T is called temperature and is decreased after each iteration. A state movement is carried out if $u < \exp(-\frac{J(\mathscr{S}')-J(\mathscr{S})}{T})$ where u is a random number uniformly drawn from $[0,1]$. At the beginning of the algorithm, the acceptance probability is high. Thus, the probability of switching to neighbor is higher than at the end of the algorithm.

At the end of the process, the best state encountered in all the iterations, i.e., the best sub-ensemble, is taken as the calibrated ensemble.

3.3 Application to a 101-Member Ensemble

We consider the 101-member ensemble, launched throughout the year 2001 over Europe and described in detail in Garaud et Mallet [2010]. The ensemble was automatically generated for the simulation of ground-level ozone, with a horizontal resolution of half a degree. Each member

of the ensemble is a unique combination of physical parameterizations, numerical schemes and input data. For instance, the members can differ in the chemical mechanism (RACM or RADM2), the computation of the vertical diffusion coefficient (Louis' or Troen&Mahrt's parameterizations), the vertical resolution (5 or 9 levels) or the perturbation of the meteorological fields (wind, temperature, etc.) and emission sources. About 30 alternatives are available for the generation of a member. The generated ensemble contains very different members and has a wide spread. The following subsections deal with the assessment of this ensemble and its calibration according to ensemble scores previously mentioned.

3.3.1 Evaluation of the Ensemble

In this sub-section, we quickly review the performance of the models and then of the ensemble.

The ensemble evaluation is carried out using the observation network Airbase [2]. This database, managed by the European Environment Agency, provides ground-level ozone observations at 210 rural background, 702 rural, 647 suburban and 1324 urban stations across Europe.

Stations that fail to provide observations at over 10% of all the dates considered are discarded as the scores at these stations may not reliable. In order to have stations which are representative of the ozone peak concentration at the model scale (half a degree in the horizontal), only rural and background stations are kept. There are about 123 000 observations for ozone peaks during the year 2001. Following usual recommendations Russell et Dennis [2000]; Hogrefe *et al.* [2001]; US EPA [1991], a cut-off is applied to the observations. Observations below 40 $\mu g\,m^{-3}$ are discarded so as to focus on the most harmful concentrations.

Models Skills

The different models show quite different skills and performances. The spatio-temporal mean of ground-level ozone peaks ranges from 60 to 130 $\mu g\,m^{-3}$. Their variability is also quite different because the global standard deviation of ozone peak simulations ranges between 17 and 44 $\mu g\,m^{-3}$.

Figure 3.1 shows the performance, compared to the observations, of the 101 simulations in a single diagram. This Taylor diagram [Taylor, 2001] takes into account the standard deviation of the observations and the correlation between each simulation and the observations. The radial coordinate of the Taylor diagram corresponds to $\frac{\sigma_x}{\sigma_y}$ where σ_x is the empirical standard deviation $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j)^2}$ of the simulated sequence $(x_i)_{i=1,\dots,n}$, and σ_y is the empirical standard deviation of the observed sequence $(y_i)_{i=1,\dots,n}$. The azimuth is the arccosine of the correlation between $(x_i)_{i=1,\dots,n}$ and $(y_i)_{i=1,\dots,n}$. The lower azimuth, the higher correlation between a simulation and the observations. A Taylor diagram shows the performance of an ensemble of simulations in term of correlation, the variability of each simulation compared with the observed variability, and the spread of these performances. Although a large number of simulations show less variability than the observations, a number of members still show good variability. The correlations range between 0.3 and 0.77.

This shows that the ensemble has a strong variety and that the models can have very different statistical measures and performance. A few models have weak skill, i.e., a high RMSE (up to 29.6 $\mu g\,m^{-3}$) and a low correlation (down to 0.3). However these models should not be discarded because they can bring useful information. Figure 3.2 shows the number of times each model is closer to an observation than any other model. Most of the bars are close to the mean

2. http://air-climate.eionet.europa.eu/databases/airbase/airbasexml/index_html

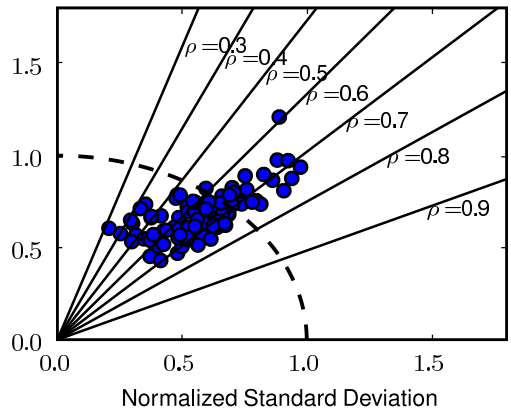


Figure 3.1: Taylor plots of ozone peak averaged over stations. The radial coordinate is the standard deviation normalized by the standard deviation of observations. The angles between the abscissa axis and the lines correspond to the arcosine of the correlation ρ between each simulation and observations.

(1091 observations). Figure 3.2 shows that all the members give the closest concentrations to the observations for a significant number of times. In the worst case, the count is about half the mean count. The worst model in terms of RMSE and correlation gives the closest concentrations to 1061 observations, which is about the average performance. This means that even if a member shows a bad performance on average, it still brings useful information in some regions and at some dates.

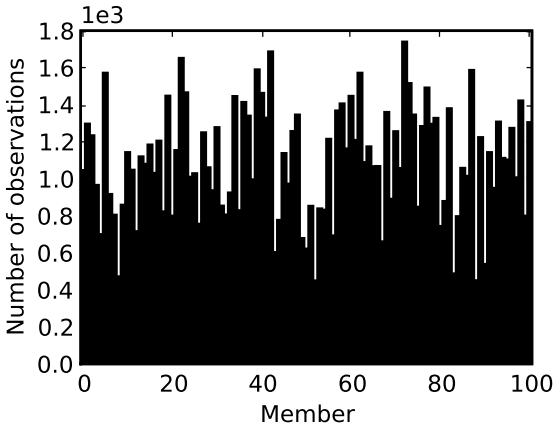


Figure 3.2: Best models count for ozone peaks on the network Airbase. A model is counted "best" when the discrepancy between the simulated concentration and the observation is minimal. The count is carried out for all observations.

Ensemble Scores

Reliability Diagram Figure 3.3 shows the reliability diagram for the event $[O_3] \geq 120\ \mu g\,m^{-3}$. The ensemble shows a reasonable performance since the diagram roughly follows the diagonal. Below the forecast probability 0.4, the ensemble overforecasts the event occurrence since the reliability curve is below the diagonal. On the other hand, the ensemble underforecasts the

event occurrence when the forecast probabilities are greater than 0.4. The diagram shows that the ensemble has an acceptable resolution. An ensemble with lower resolution would have a flatter reliability diagram which would be close to the climatological forecast. Unfortunately, for an event based on a higher concentration, such as $[O_3] \geq 180\,\mu g\,m^{-3}$, the ensemble leads to a poor reliability diagram. This can be explained by the very low occurrence of the event – about 0.6% of all cases – and by the sharpness histogram. Two sharpness histograms are shown in figure 3.4 and represent the frequency of the forecast probabilities for the two previous events. The sharpness indicates the tendency of an ensemble to provide probabilities near 0 or 1. The forecast probabilities provided for the first event ($120\,\mu g\,m^{-3}$) are quite frequent and close to 0. Thus, most of the time, no simulation exceeds the threshold, so that the ensemble gives a null probability of event occurring. For the threshold $180\,\mu g\,m^{-3}$, the sharpness histogram is even worse since over 98% of forecast probabilities are less than 0.1. As the number of forecast probabilities greater than 0.1 is so low, it seems difficult to correctly build a reliability diagram. Hence for the threshold $180\,\mu g\,m^{-3}$, the calibration cannot be carried out using the reliability diagram.

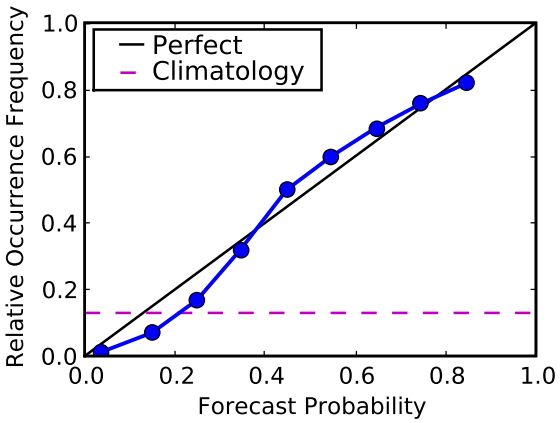
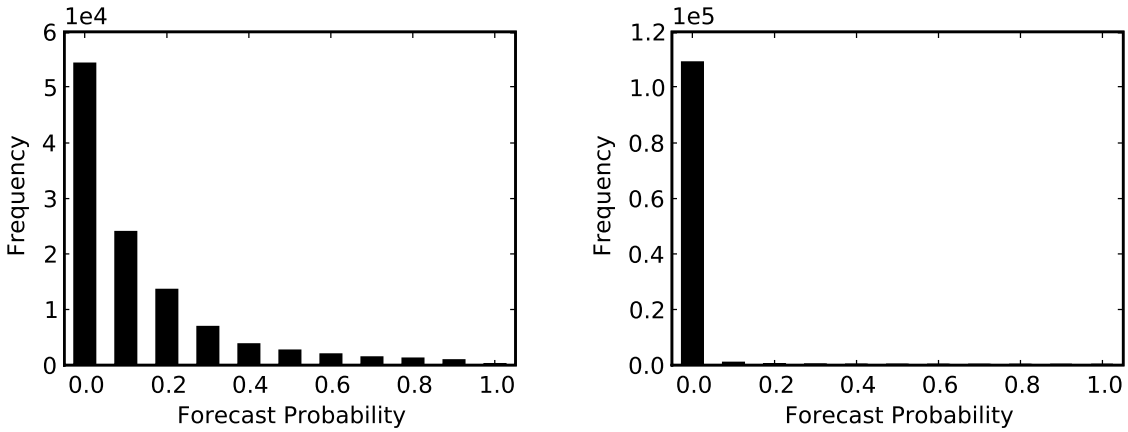


Figure 3.3: Reliability diagram of the ensemble for ozone peaks. The ozone concentration threshold is $120\,\mu g\,m^{-3}$. The black line corresponds to a perfect reliability diagram. The dashed horizontal line is the value of the climatological forecast.

Rank Histogram Figure 3.5 is the rank histogram of the 101-member ensemble for ozone peaks. The histogram does not show any extremely low or extremely high bar, but several bars have half the height they should have and several others are significantly higher than expected. The first bar, which corresponds to the number of observations below the lower envelope of ensemble, is especially high. It means that, at certain locations and dates, the spread of the ensemble is insufficient to cover the observations. The measure of the flatness described in the section 3.2.2 is 148.

Brier Score and DRPS The Brier score, Brier skill score and discrete ranked probability score are computed with the full ensemble, with the "best" model alone and with the climatological forecast. The "best" model will be the member from the full ensemble that minimizes or maximizes the given score. The climatological forecast is given by the all-year relative (observed) occurrence frequency of the event. These different scores are reported in table 3.1. The DRPS is computed with the threshold exceedances for 80, 100, 120, 140 and $160\mu g\,m^{-3}$.



(a) $120\,\mu g m^{-3}$ (b) $180\,\mu g m^{-3}$

Figure 3.4: Sharpness histograms for two ozone concentration thresholds: $120\,\mu g m^{-3}$ (left) and $180\,\mu g m^{-3}$ (right).

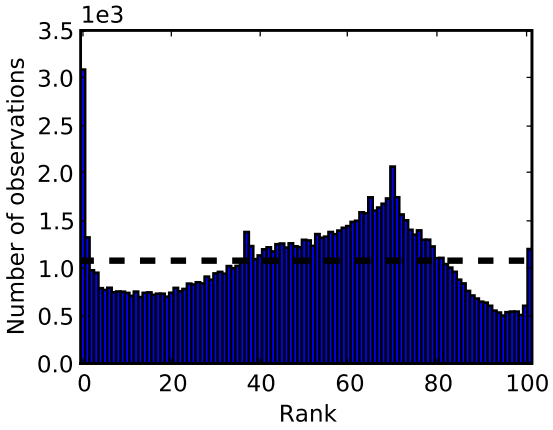


Figure 3.5: Rank histogram of the 101-member ensemble on network Airbase for ozone peaks. The horizontal dashed line corresponds to the ideal value for a flat rank histogram with respect to the number of members. The large number of observations on the left means there are many observations below the lower envelope of the ensemble.

Table 3.1: Brier scores and Brier skill scores for the event $[O_3] \geq 120\,\mu g m^{-3}$ for the ensemble, the "best" model (with respect to the score) and the climatological forecast. The DRPS is computed with the threshold exceedances for 80, 100, 120, 140 and $160\,\mu g m^{-3}$.

	Full Ensemble	Best Model	Climatology
Brier	$76\ 10^{-3}$	$95\ 10^{-3}$	$113\ 10^{-3}$
Brier skill	$32.7\ 10^{-2}$	$15.6\ 10^{-2}$	0.0
DRPS	$90.3\ 10^{-3}$	$124\ 10^{-3}$	$130\ 10^{-3}$

It is interesting to notice that the "best" model is always the same for all scores and corresponds to the model which has the smallest RMSE (20.5 $\mu g\,m^{-3}$). This "best" model is always better than the climatological forecast. It should, however, be noted that, firstly, one model can only provide probabilities equal to 0 or 1 and secondly, a large majority of the models have worse scores than the climatological forecast. For instance, over 77% of the models have a negative Brier skill score for the 120 $\mu g\,m^{-3}$ threshold exceedance. Whatever the score, the full ensemble always performs better than the "best" model. Consequently it seems that an ensemble is necessary to provide forecast probabilities which are more accurate than probabilities provided by a single model.

3.3.2 Calibration

Reliability Diagram

We introduce the average probability $\overline{p}_k$ of all forecast probabilities lying in the interval $[p_{k-1}, p_k]$. As described in the section 3.2.2, a perfect reliability leads to $\overline{p}_k = O_k$. In order to have an optimized reliability diagram, the calibration method is therefore carried out with the mean squared error of the diagram. The score to minimize can be written as

$$\mathscr{C}_{rel} = \frac{1}{K} \sum_{k=1}^{K} (\overline{p}_k - O_k)^2 . \tag{3.10}$$

We consider the event $[O_3] \geq 120\,\mu g\,m^{-3}$, and we apply the genetic algorithm and the simulated annealing. Figure 3.6 shows the two resulting reliability diagrams. The calibrated diagrams are better than the reliability diagram of the full ensemble since they are closer to the diagonal. The 35-member calibrated ensemble from the genetic algorithm is very reliable and has a mean squared error lower than 10^{-5}. As the reliability is improved, the Brier skill score of the two calibrated sub-ensembles are equal to 34 10^{-2} and 35 10^{-2}, which represents slight improvements compared with the full ensemble. The Brier score decomposition shows that the reliability term is better after calibration whereas the resolution term is slightly worse. For the best calibrated sub-ensemble (genetic algorithm), the reliability term decreases by about 93% while the resolution term decreases by about 1%. Candille et Talagrand [2005] show that there is a compromise between reliability and resolution. Thus, resolution can be degraded when reliability is improved. Nevertheless, this calibration dedicated to improving reliability degrades resolution very slightly.

Rank Histogram

We now apply the calibration with criterion (3.2) so as to get a flat rank histogram. Note that it is desirable to obtain a sub-ensemble with the largest number of models so that an accurate uncertainty estimation can be produced. It is possible to obtain a perfectly flat diagram with just one model, providing half the observations are below the model concentrations and half the observations are above; but one model cannot help in providing an uncertainty estimation.

The calibration results depend on the height of the highest bar (here, the left bar) of the full-ensemble histogram. All observations with rank 0 (left bar) are below the lower envelope of the ensemble. For any sub-ensemble, the height of the left bar cannot be lower than the number r_0 of observations below the lower envelope. In a flat histogram, at best, the height of the left bar is still r_0 and all the bars have the same height. In this case, there cannot be more than 34 members (which is deduced from the total number of observations divided by r_0). Figure 3.7 is the rank histogram of the calibrated sub-ensemble using simulating anneal. There are 33 members and the flatness score is about 6 instead of 148 for the full ensemble score.

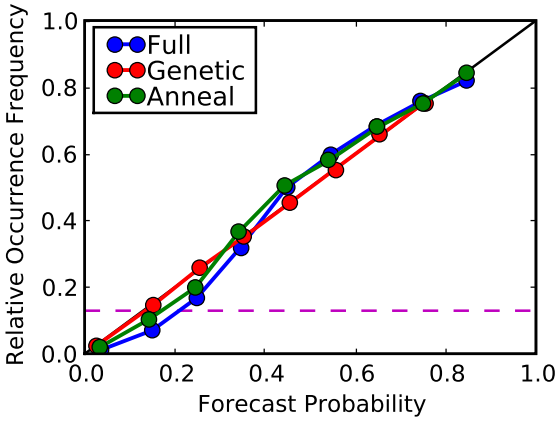


Figure 3.6: Calibrated reliability diagrams for the event $[O_3] \geq 120\ \mu g\,m^{-3}$ from the simulated annealing and the genetic algorithm. The dashed line corresponds to the value of the climatological forecast.

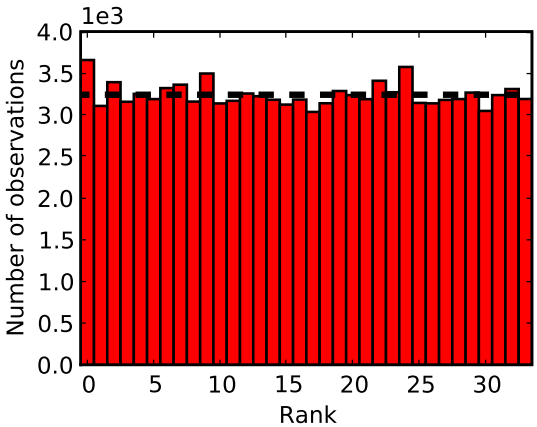


Figure 3.7: Rank histogram of the calibrated ensemble on network Airbase for ozone peaks. The horizontal dotted line corresponds to the ideal value for a flat rank histogram according to the number of members.

This calibrated sub-ensemble also improves the Brier scores and the DRPS. For the same events as before, the Brier skill score and DRPS respectively give 36 10^{-2} and 90 10^{-3}. It is interesting to notice that the reliability (from the DRPS decomposition (3.9)) is decreased by 90%, while the resolution remains unchanged. This is consistent with the fact that the rank histogram is an ensemble score which measures reliability.

DRPS

The calibration according to the DRPS gives $\text{DRPS}_{calib} = 66\,10^{-3}$. The DRPS of the full ensemble is reduced by 15%. The reliability part (see (3.9)) is reduced by 47% and the resolution part by 10%.

For all ensemble scores, the calibration provides well balanced sub-ensembles. They always are better than the full ensemble, the best model or the climatology. The calibrated sub-ensembles also improve the reliability. However, the resolution essentially remains the same. As for the Brier score decomposition (3.8), the resolution term does not depend directly on the agreement between the forecast probability and the event occurrences. The improvement in the resolution depends on the definition of forecast probabilities bins described in paragraph 3.2.2 and Candille et Talagrand [2005]. The ensemble calibration essentially improves the quality of forecast probabilities, i.e., the reliability, rather than the variance of frequency occurrence O_k.

3.4 Uncertainty Estimation

We now analyze the uncertainty estimation based on the sub-ensemble calibrated for the rank histogram. This calibration is chosen because it is related to the probability distribution of ozone concentrations, whereas the other scores are used to assess an ensemble for specific events.

The uncertainty can be estimated with the (empirical) standard deviation of the ensemble. A monthly average of the standard deviation of the calibrated ensemble is computed in each cell of the domain studied. Figure 3.8 shows the corresponding uncertainty map over Europe, averaged over June 2001. A higher ozone uncertainty appears along the south-coasts of Europe. This is consistent with a well-known difficulty of predicting ozone along the coasts, mainly because of poor representation of winds and turbulence in these areas.

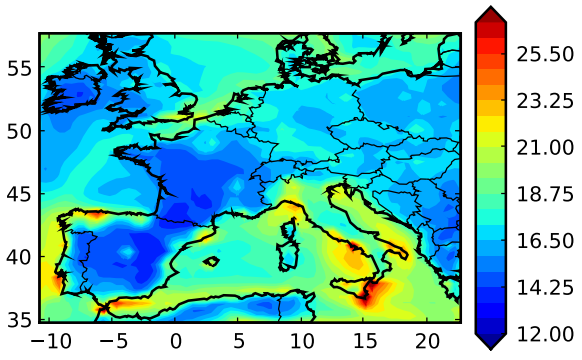


Figure 3.8: Monthly average of ozone uncertainty from a calibrated sub-ensemble for June 2001 across Europe ($\mu g\,m^{-3}$).

Before presenting further results, it is important to assess the robustness of the calibration

method. One question is the spatial robustness. A calibrated sub-ensemble is spatially robust if it is still reliable at non-observed locations. In order to check this robustness, we randomly exclude stations from the calibration, and assess the calibration on the remaining stations.

Figure 3.9 shows all observation stations previously used to compute the ensemble scores and to calibrate the ensemble. This network is randomly split into two sub-networks (cyan and yellow). The rank histogram calibration is then carried out on each sub-network, that is, using only the observations of the sub-network. Figure 3.10 shows four rank histograms for the two calibrated sub-ensembles. At the top of the figure, the calibrated rank histograms are shown, each computed with the observations used for their calibration. At the bottom, the rank histograms are computed using the observations of the other sub-network. The rank histograms are almost flat, which shows that the calibration is robust. It is noteworthy that the two sub-ensembles have a similar number of members (27 and 28 members for the "cyan" and "yellow" sub-ensembles, respectively).

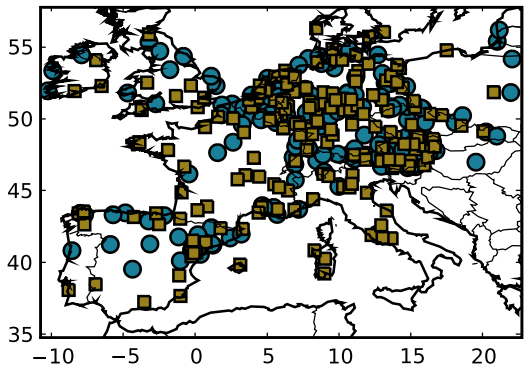
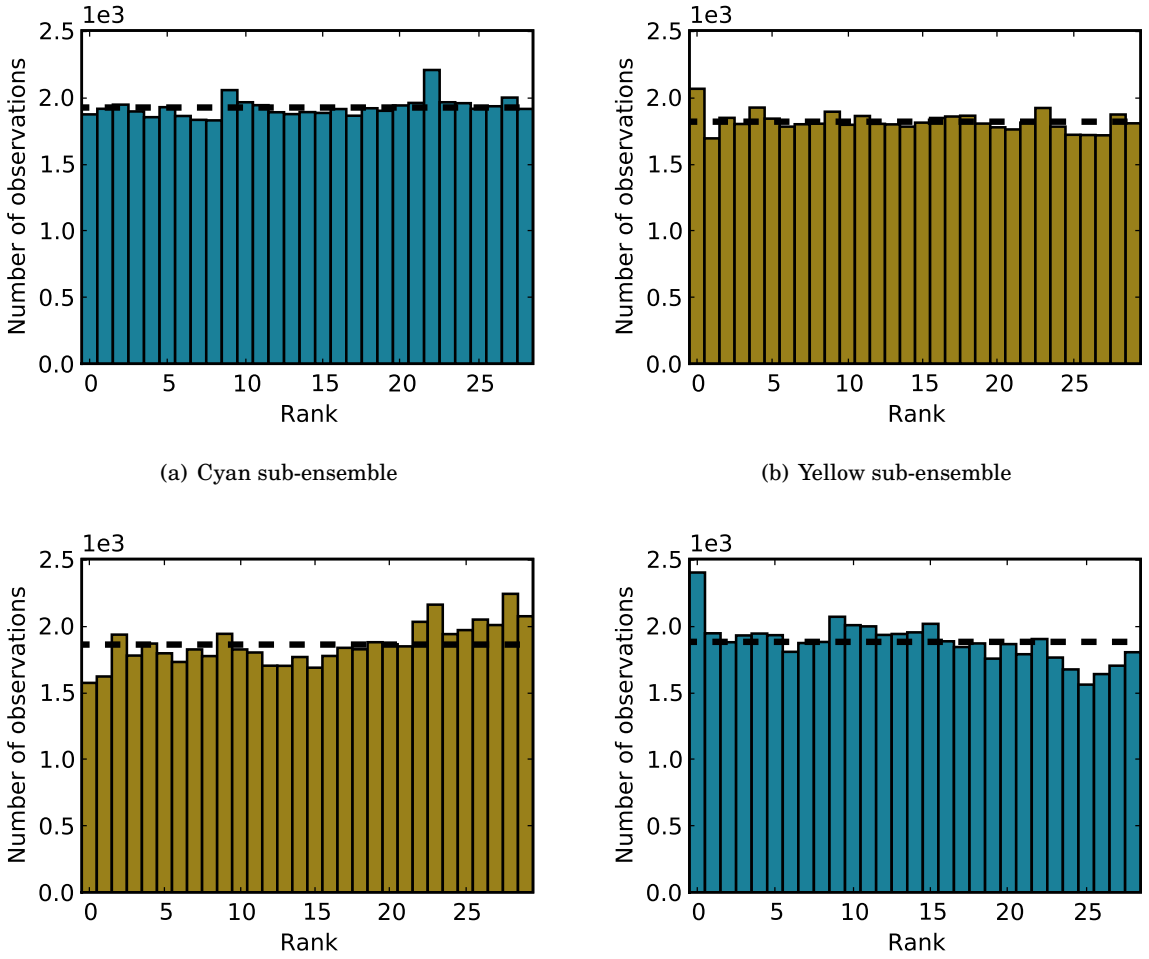


Figure 3.9: The two random sets of stations over Europe. These two sub-networks are used to assess the spatial robustness of the ensemble calibration method. The two sub-networks are a partition of the full network: each station of the full network belongs to one and only one sub-network.

We can now compare the uncertainty estimation maps from the two previous calibrated sub-ensembles. Figure 3.11 shows the uncertainty estimation of the two calibrated sub-ensembles from the two previous random sub-networks. The spatial structures are similar. The high and low uncertainty values are located at the same places. In figure 3.12, these uncertainty maps are also compared with the uncertainty map obtained after calibration with all observations. The relative difference between these maps is about 3% on average, and marginally exceeds 10%. For reference, the figure also shows the relative difference with the uncertainty derived from the full ensemble.

Besides spatial robustness, the previous results also show that here, half observations are sufficient to calibrate an ensemble and estimate uncertainties. This raises the question of how many observations are needed for the calibration. An experiment was carried out to estimate this number. First, a rank histogram is computed for the full ensemble with about 30000 hourly observations. These observations are selected arbitrarily. Then, observations are randomly removed and the rank histogram is computed again. After a few iterations, we can compare several rank histograms with a different number of observations. Figure 3.13 shows two rank histograms of the full ensemble with about 32500 observations and about 10100 observations. Their shapes are very similar. Below 8000 observations, the shape of the rank histogram starts changing. So



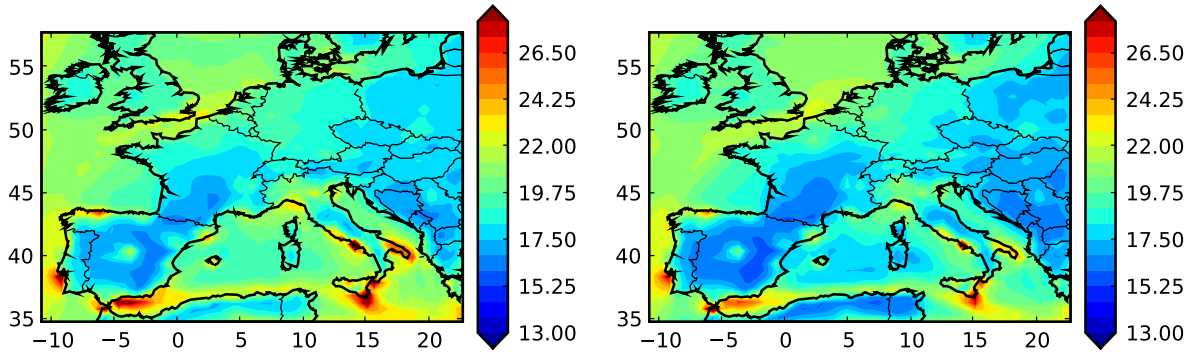
(a) Cyan sub-ensemble

(b) Yellow sub-ensemble

(c) Yellow sub-ensemble on the cyan network

(d) Cyan sub-ensemble on the yellow network

Figure 3.10: Rank histograms of the calibrated sub-ensembles on the two random sub-networks. At the top, the calibrated rank histograms of the cyan and yellow sub-ensembles. At the bottom, the rank histograms computed from the yellow sub-ensemble (*left*) on the cyan sub-network and from the cyan sub-ensemble (*right*) on the yellow sub-network.



(a) From cyan sub-ensemble

(b) From yellow sub-ensemble

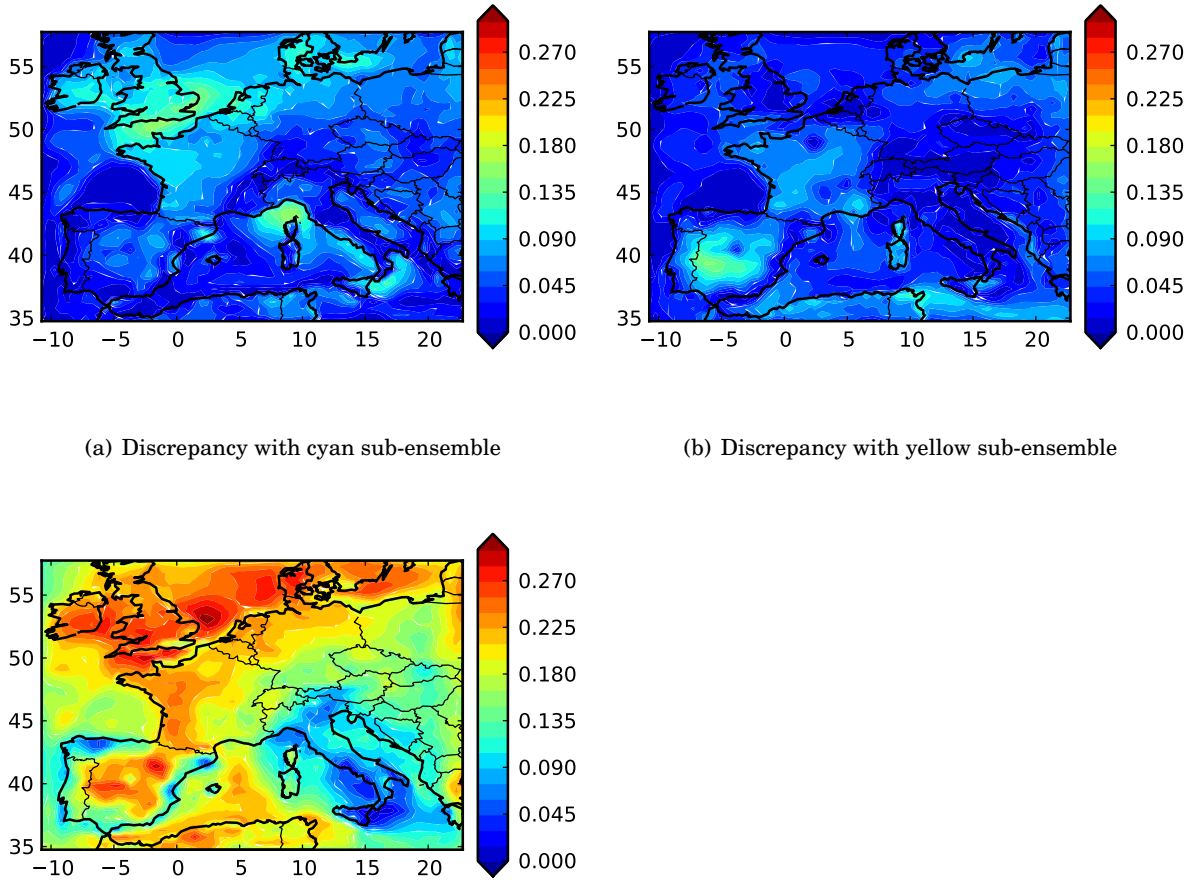
Figure 3.11: Temporal average of uncertainty estimation in $\mu g\,m^{-3}$ from two sub-ensembles which were calibrated with two random sub-networks over Europe. On the left is uncertainty map from the cyan network while the right map corresponds to the yellow network for June 2001.

we conclude that 8000–10000 observations are required to assess the quality of the 101-member ensemble.

A similar experiment was carried out to determine the number of observations needed for the calibration to be reliable. The full ensemble is calibrated a few times with a total number of observations (initially 32500) divided by 2, 3, 5, 8 and 13. The calibrated ensembles contain a similar number of members, ranging from 22 to 27. The rank histograms for the calibrated ensembles are then computed, each time with the observations used in the calibration. The rank histograms remain flat in every case. The uncertainty estimations starts depending on the number of observations when there are fewer than 8000–10000 observations.

Another question is the impact of observational errors on the calibration and on the uncertainty estimation [Anderson, 1996; Hamill, 2001]. The rank histogram checks whether two random variables sample the same distribution. Noise in the observations should therefore be added to ensemble so that we can check the ensemble samples the real uncertainty without observation noise. Let $x_i^m(t)$ be the simulated concentration at station i and date t for the model m. We assume that observational errors do not depend on the station and date. We introduce the perturbed concentrations $\hat{x}_i^m(t) = x_i^m(t)(1+\alpha_i^m)$ where α_i^m follows a uniform distribution on the interval $[-\varepsilon,\varepsilon]$. This form allows us to introduce a noise relative to the concentration, which is a usual feature for ozone observations. Based on Airparif [2007], $\varepsilon \simeq 0.13$ for ozone peak concentrations measured over the year 2009 at about 30 stations from the Airparif monitoring network (in the Paris region). This noise is introduced before the calibration. The calibrated ensemble with perturbation ($\varepsilon = 0.13$) shows a flat rank histogram, and the resulting uncertainty estimations are plotted in figure 3.14. The values and spatial patterns of the standard deviation are very similar to those of the calibration without perturbations. The observation errors therefore seem to have a limited impact on the calibration.

Finally, we investigate the robustness of the calibration over time. A calibration is carried out during a learning period, and the relevance of this calibration is evaluated for a forecast period. The sub-ensemble selected based on the learning period is referred to as an a priori sub-ensemble. The quality of the forecast is measured by comparing the a priori sub-ensemble and

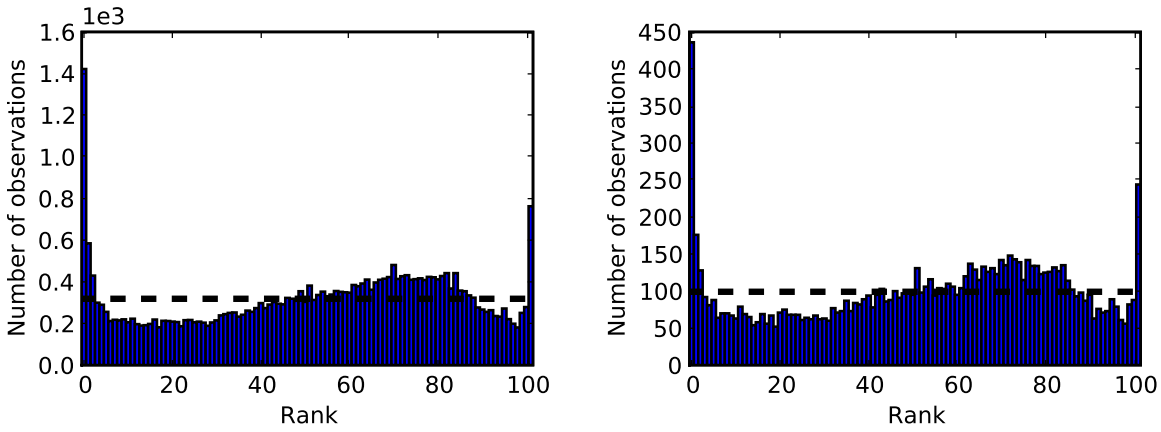


(a) Discrepancy with cyan sub-ensemble

(b) Discrepancy with yellow sub-ensemble

(c) Discrepancy with full ensemble

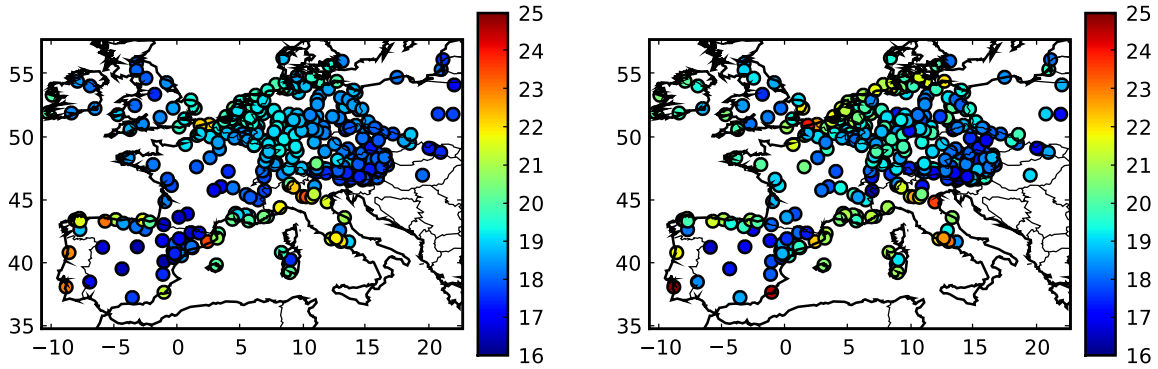
Figure 3.12: Relative discrepancy on uncertainty fields (averaged over June) between the sub-ensemble calibrated with all observations and (a) the sub-ensemble calibrated on the cyan sub-network, (b) the sub-ensemble calibrated on the yellow sub-network, and (c) the full ensemble. For example, the relative discrepancy (c) is defined (pointwise) as the difference between the averaged uncertainty obtained with the full ensemble and the averaged uncertainty obtained with the calibrated sub-ensemble, divided by the latter.



(a) 32500 observations

(b) 10100 observations

Figure 3.13: Rank histograms with a different number of observations — about 32500 observations on the left and 10100 on the right.



(a) Reference calibration

(b) With perturbed simulations

Figure 3.14: Uncertainty estimations at stations, for the reference calibration on left, and for the calibration with perturbed simulations ($\varepsilon = 0.13$) on right.

the a posteriori sub-ensemble that is calibrated over the forecast period.

The learning period is a week, from April 3rd to April 9th, with 50000 hourly ozone observations. It is an arbitrary chosen period. The forecast period ranges from April 10th to April 16th. Figure 3.15 shows the uncertainty map computed during the learning period and the forecast uncertainty map. These maps clearly show different patterns, e.g., with higher forecast uncertainties over the North Sea, over France and Germany, and with lower forecast uncertainties over several parts of the Mediterranean Sea. This, and tests not reported here, show that the uncertainty estimations can vary strongly over time. Figure 3.15 also shows the a posteriori uncertainty map. The forecast and a posteriori maps essentially show the same patterns and uncertainty levels. This means that, despite the significant variation in time, the calibration seems robust over time. Here the calibration can be used to forecast the uncertainties for a few days. The root mean square error between the forecast and a posteriori maps (daily averages), divided by the mean of the a posteriori map, is equal to about 5% over each of the next six days. It is noteworthy that the learning period should be long enough—two-day or four-day periods do not appear to be long enough to ensure a good forecast.

3.5 Risk Assessment and Probabilistic Forecast

In order to check that the calibration can help in risk assessment and in forecasting a given event, the same tests as in the previous section are carried out with the Brier skill score and the reliability diagram instead of the rank histogram.

Figure 3.16 shows, for each sub-network, reliability diagrams for calibrated sub-ensembles and the full ensemble. Any sub-ensemble calibrated on one sub-network performs well on the other sub-network.

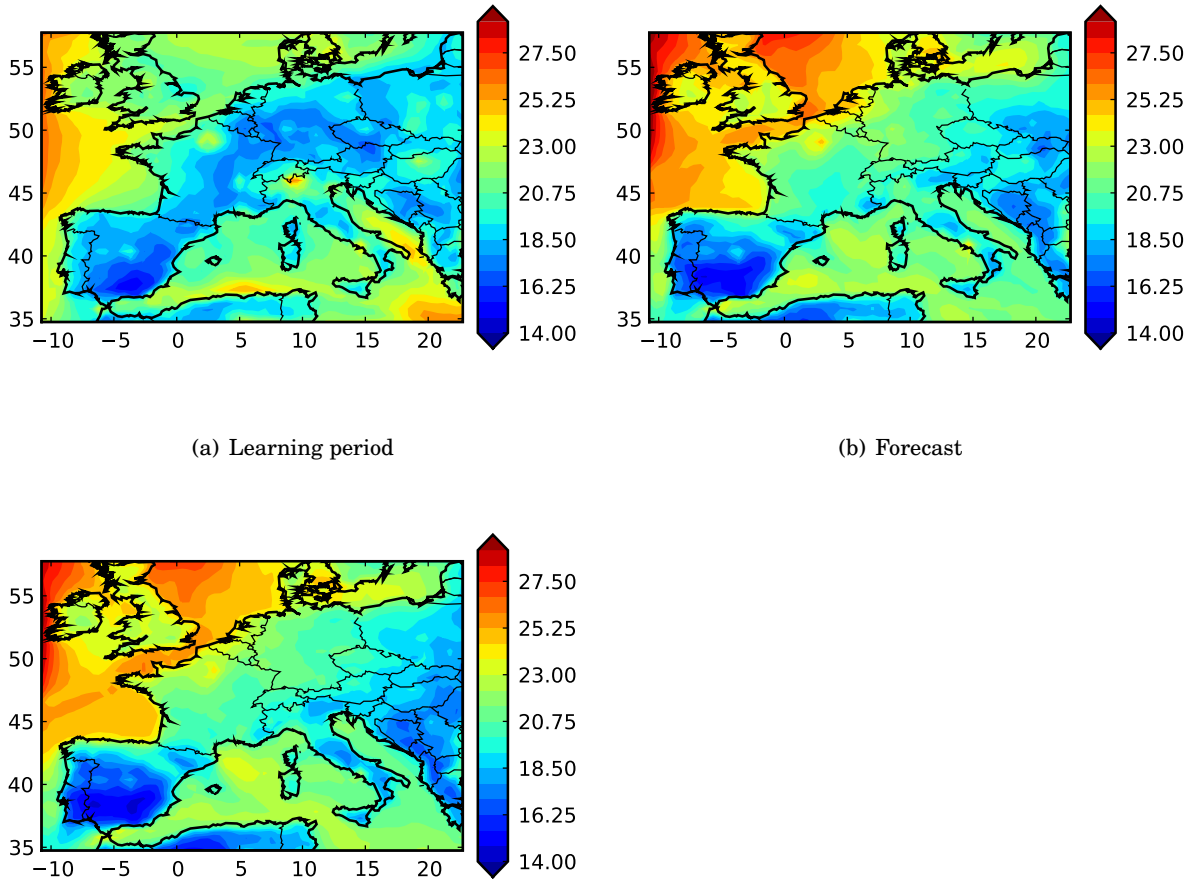
The same conclusion can be drawn from the Brier skill score calibration. Table 3.2 shows the Brier skill scores of the full ensemble and calibrated sub-ensembles computed on the cyan sub-network for three different thresholds — 80, 100 and 120 $\mu g\,m^{-3}$. Whatever the threshold exceedance, the calibrated sub-ensembles perform significantly better than the full ensemble. The sub-network over which the calibration was carried out does not impact the results.

Table 3.2: Brier skill scores of the full ensemble and the calibrated sub-ensembles. The scores are computed using the observations of cyan sub-network.

Threshold exceedance	$[O_3] \geq 80\ \mu g\,m^{-3}$	$[O_3] \geq 100\ \mu g\,m^{-3}$	$[O_3] \geq 120\ \mu g\,m^{-3}$
Full ensemble	0.35	0.37	0.34
Cyan calibrated sub-ensemble	0.40	0.46	0.44
Yellow calibrated sub-ensemble	0.40	0.46	0.44

According to these results, the calibrations based on the reliability diagram and the Brier skill score seem spatially robust.

In order to assess the temporal robustness, we select arbitrarily the learning period from May 31th to June 6th and rely on the corresponding calibrated sub-ensemble to forecast the period from June 7th to June 13th. Figure 3.17 shows the reliability diagrams of the full ensemble, the a priori calibrated sub-ensemble and the a posteriori calibrated sub-ensemble, for the threshold exceedance $[O_3] \geq 100\ \mu g\,m^{-3}$. The a priori sub-ensemble performs better than the full ensemble, but its reliability diagram is deteriorated compared to the a posteriori sub-ensemble. Note that the forecast period is long (7 days) because the reliability diagram requires a significant amount of data to be computed. It is possible that the results would be better if the diagram could be computed with the observations of the very first forecast days only.

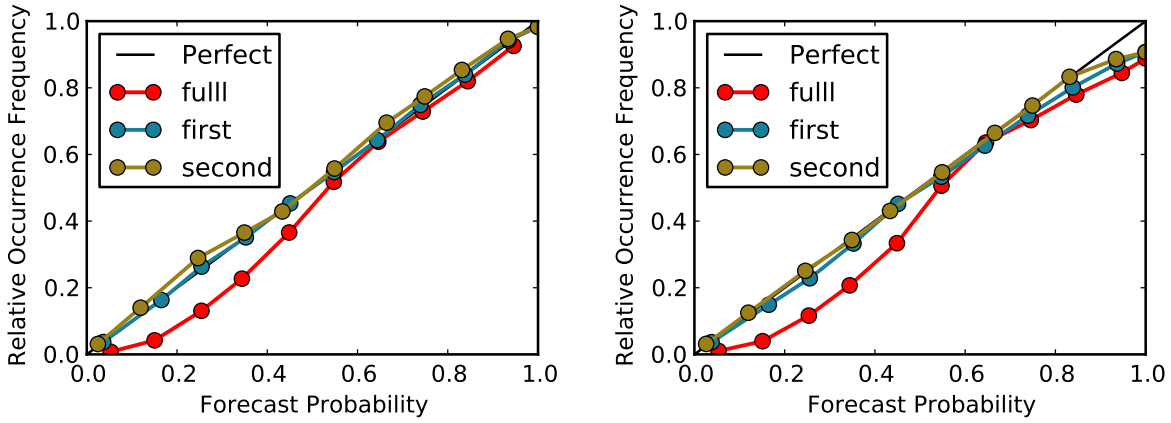


(a) Learning period

(b) Forecast

(c) A posteriori

Figure 3.15: Comparison of ozone uncertainty maps averaged over one week in $\mu g m^{-3}$. The first one is the uncertainty estimation during the learning period (from 3^{rd} to 9^{th} April 2001), the second one is the uncertainty forecast (10^{th} to 16^{th} April 2001), and the last one (bottom) is the a posteriori uncertainty.



(a) Cyan sub-network

(b) Yellow sub-network

Figure 3.16: Reliability diagrams for $[O_3] \geq 100\ \mu g\,m^{-3}$ of the calibrated sub-ensembles and the full ensemble. On the left, the reliability diagrams are computed on the cyan sub-network. On the right, the reliability diagrams are computed on the yellow sub-network.

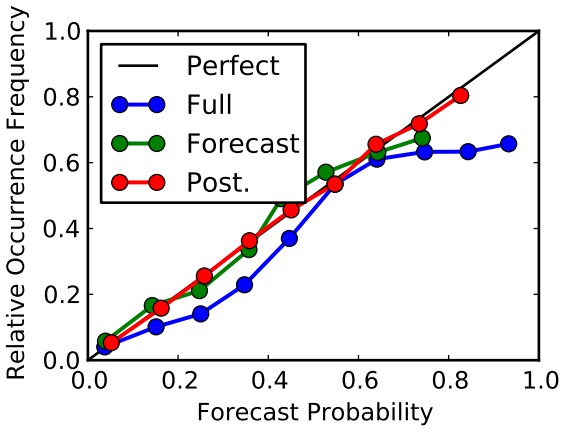


Figure 3.17: Reliability diagrams for $[O_3] \geq 100\ \mu g\,m^{-3}$ of the full ensemble (cyan), the a posteriori calibrated sub-ensemble (red) and the a priori calibrated sub-ensemble (green). This is based on observations from June 7^{th} to June 13^{th}.

The Brier skill scores in the same forecast period are 0.18, 0.27 and 0.25 for the full ensemble, the a posteriori sub-ensemble and the a priori sub-ensemble, respectively. It shows that the calibration can be relevant in the context of probabilistic forecast.

3.6 Conclusion

The work presented in this paper relies on a 101-member ensemble that was automatically generated on the Polyphemus platform. This large ensemble is evaluated for uncertainty estimation and for probabilistic forecasts. The tests show that about 10000 observations are required to properly evaluate the 101-member ensemble. A calibration method is designed to select a sub-ensemble from the full ensemble that better estimates the uncertainties.

Several calibrations for different ensemble scores are carried out and show significant improvements in the ensemble scores. An almost perfect reliability diagram and a very flat rank histogram can result from the calibration. We note that observation errors have a slight impact on calibration, since uncertainty maps with and without observation errors have the same pattern. The quality of the spatial distribution of the uncertainty estimation is assessed by a cross validation. Again, the calibration seems robust as the uncertainty maps are reasonably sensitive to the observation network. Finally, we show that the method can be applied in a forecasting context. The calibration can be carried out on a learning period, and the resulting sub-ensemble is able to estimate the uncertainties in the subsequent period almost as well as the sub-ensemble calibrated on this subsequent period.

It would therefore be a natural next step to apply the method proposed here in operational conditions, including for aerosols for which the number of available observations may be significantly lower. A question is how much the proposed approach can help forecast threshold exceedances. The results show that the scores associated with such forecasts are improved, but the impact in an operational platform for decision making has yet to be assessed.

The complexity of the method mainly lies in the automatic generation of a large ensemble in which many sources of uncertainties are taken into account. An open question is what ensemble design should be considered for uncertainty estimation and probabilistic forecasting. This question is especially important when considering forecasts because the sub-ensemble selected over one period should still represent the right uncertainty sources in another period. Monte Carlo simulations, for instance, are easier to carry out, but they might miss important uncertainty sources coming from the model formulation itself.

Further work should address the partition of the uncertainty sources in order to better identify modeling errors, representativeness errors and measurement errors. Also the spatial and temporal correlations in the errors should be evaluated.

Acknowledgements
We would like to thank to Hélène Marfaing and Christophe Debert from Airparif for their very useful studies and their data about measurement uncertainties. We thank Richard James for proofreading the paper.

Chapitre 4

Estimation et décomposition de l'incertitude à l'aide d'un ensemble Monte Carlo et d'un ensemble multi-modèles

Ce chapitre traite dans un premier temps de l'estimation de l'incertitude pour deux ensembles différents. Le premier est un ensemble Monte Carlo – seuls les données et paramètres d'entrée sont perturbés – tandis que le second est l'ensemble multi-modèles, largement décrit dans le chapitre 2. Les deux ensembles sont composés d'une centaine de simulations photochimiques à l'échelle européenne sur l'année 2001. Des calibrations sont réalisées pour les deux ensembles afin d'estimer puis comparer les champs d'incertitude (variances et co-variances). L'ensemble multi-modèles montre plus de variabilité et semble mieux représenter les incertitudes que l'ensemble Monte Carlo.

Dans un deuxième temps, une régression linéaire est réalisée entre les champs d'entrée perturbés et les concentrations simulées d'ozone dans le but de quantifier l'impact des incertitudes des champs d'entrée sur les simulations d'ozone. Cette régression est réalisée pour les deux ensembles sur différents types de station : urbain, rural et de fond. Il s'avère que les conditions aux limites d'ozone jouent un rôle important – ceci pour les deux ensembles. Plusieurs autres champs d'entrée ont un impact significatif, par exemple, le taux de photolyse de NO_2 ou les émissions de NO_x.

Enfin, on estime la part de l'erreur de mesure (liée aux instruments de mesure), l'erreur de modélisation et l'erreur de représentativité dans l'écart entre les observations et les simulations. Deux méthodes indépendantes permettent d'estimer la variance de l'erreur de représentativité. Les erreurs de mesure sont assez faibles comparées aux deux autres erreurs. L'erreur de représentativité représente en moyenne au moins un tiers de la variance de l'écart entre les observations et les simulations.

Sommaire

Ce chapite est constitué de Garaud et Mallet [2012].

4.1 Introduction

Many studies and decision making are based on modern chemistry-transport models that solve 3D reactive transport equations for the main atmospheric chemical species. The primary output of these models is 3D concentrations of the chemical species against time. A next step in the development of air quality modeling systems should be to provide, along with the concentrations, an estimate of their uncertainty. The uncertainty can be significantly high, to the point of changing or at least softening conclusions drawn from the concentrations alone.

In order to estimate the uncertainty on a model's output, all the sources of this uncertainty should a priori be taken into account and propagated through the reactive transport equations. The ideal target would be the probability density function of the chemical concentrations, and its time evolution. Computing this probability density function is an impossible task, considering a model's state vector with easily one million or ten million components. In practice, the most accurate uncertainty estimations are derived from ensemble simulations. There are two main approaches: Monte Carlo simulations and multimodel simulations. In the former, the input data and parameters of the model are randomly perturbed in each member (i.e., simulation) of a typically 50- or 100-member ensemble [e.g., Hanna *et al.*, 1998, 2001; Beekmann et Derognat, 2003; Boynard *et al.*, 2011]. It requires to have knowledge on the probability distributions of the input data and parameters, and to have enough computational power to reasonably sample these distributions. The other approach relies on different chemistry transport models that are based on different physical, chemical and numerical formulations [e.g., Delle Monache et Stull, 2003; Mallet et Sportisse, 2006a; McKeen *et al.*, 2007]. This approach can be easily combined with Monte Carlo simulations since the input data of the members of a multimodel ensemble can be randomly perturbed. The combined approach can take into account all sources of uncertainties, from the model formulation itself (what Pinder *et al.* [2009] call structural uncertainty) and from the input data.

The Monte Carlo approach is rather easy to implement since only the input data and parameters are modified, within a single model. The multimodel approach has a higher implementation cost since it involves models with varying requirements: different input data, a range of resolutions, different chemical species, ... However this approach, when combined with perturbations in the input data, provides richer ensembles. One may wonder whether the better quality of the multimodel ensembles is worth the additional implementation effort. This raises the question of the main sources of uncertainties and of the merits of each approach with respect to the representation of the uncertainty sources. In this paper, we try to bring some answers by comparing a Monte Carlo ensemble with a multimodel ensemble. In section 4.2, we describe and compare the generation of two 100-member ensembles, one with the Monte Carlo approach and another with the multimodel approach, for the full year 2001. Uncertainty estimations computed with the empirical standard deviations are then analyzed. In order to improve these estimations, we also carry out an a posteriori calibration, using the observations and the method introduced in Garaud et Mallet [2011].

In section 4.3, we identify which input data and parameters are the main sources of the uncertainties in the output concentrations. We carry out the identification for both Monte Carlo and multimodel ensembles. For the main sources of uncertainties, the impact on the output concentrations is estimated. The results point out which input fields or parameters should receive more attention in the generation of an ensemble.

In section 4.4, the discrepancy between observations and models' concentrations are decomposed into three errors: (1) measurement error due to instrument limitations, (2) modeling error originating from the shortcomings and uncertainties in the simulations, and (3) the representativeness error due to the misrepresentation of point observations by model's concentrations that

are averaged over a grid cell. The latter is seldom estimated and yet proves to be quite significant. It is consistently estimated by two independent methods, one solely based on observations, and another based on the observations and any model from the ensembles.

4.2 Comparison of Monte Carlo and Multimodel Ensembles

In this section, we compare the uncertainty estimations that can be derived from Monte Carlo simulations and a multimodel approach. After each generation of an ensemble, a calibration is carried out, in order to select a sub-ensemble that better estimates the uncertainty than the full ensemble. This lets us to produce more accurate uncertainty estimations and to evaluate the estimation quality from both approaches.

4.2.1 Generation of the Ensembles

Air quality models produce estimations of the pollutant concentrations that can be highly uncertain because of (1) shortcoming in their chemical and physical formulations (turbulence modeling, deposition velocities, chemical mechanism, ...), (2) uncertainties in the various input data (meteorological fields, emission sources, boundary conditions, ...), and (3) the numerical approximations (numerical schemes, time step, vertical resolution). Considering all these uncertainty sources, many different models and associated simulations can be carried out with a great variety of results. In order to explore these possible results, two classical approaches are the Monte Carlo simulations that solely rely on perturbations in the uncertain input data or model parameters, and the multimodel ensembles that try to take into account all uncertainty sources.

In the multimodel ensemble approach, each simulation is built with a given combination of chemical/physical parameterizations, numerical schemes and input data. This ensemble can be derived from models developed in different teams or inside a flexible modeling platform. In this paper, we rely on the latter, with the air quality modeling system Polyphemus [Mallet *et al.*, 2007b] and the automatic generation of large ensembles described in Garaud et Mallet [2010]. In fact we use the same 101-member ensemble as described in Garaud et Mallet [2010] for photochemistry simulation over Europe during 2001. Every member of the ensemble has a unique combination of (1) perturbed input fields such as wind field, temperature or boundary conditions, (2) physical parameterizations (e.g., for the vertical diffusion coefficient or the chemical mechanism) and (3) numerical approximations such as time step, vertical resolution or the first layer height, 40 m or 50 m. A total of thirty alternatives are available for the generation of a single model. Each member of the ensemble is defined by a random selection of one option per alternative.

In this paper, the Monte Carlo ensemble generation is inspired by Hanna *et al.* [1998, 2001] and Beekmann et Derognat [2003]. The uncertain input data is perturbed independently for each member of the ensemble, but each member relies on the same model. This model is referred to as the reference model in Garaud et Mallet [2010]. It uses common options of the Polyphemus system for photochemical simulations: RACM chemical mechanism, Troen&Mahrt parameterization to compute vertical diffusion coefficients, deposition velocities computed with the Zhang parameterization [Zhang *et al.*, 2003], a vertical resolution with 5 levels from 50 m to 3000 m, a time step equal to 600 s. The perturbations are randomly sampled assuming that the input fields and parameters are normally or log-normally distributed with a given standard deviation. Table 4.1 lists the perturbed variables together with their distributions (normal or log-normal) and the associated uncertainties. Assume that p is a scalar to be perturbed. p is either a parameter or the point value of a field at a given time and location. In the case of a log-normal distribution, the perturbed value will be $\tilde{p} = p\sqrt{\alpha}^{\gamma}$ where α is given in the column "Uncertainty" of table 4.1

and γ is sampled according to the standard normal distribution. For a normally-distributed field or parameter, the perturbed value will be $\tilde{p} = p + \frac{1}{2}\alpha\gamma$. For temperature, a relative uncertainty is provided, so that the perturbed value will be $\tilde{p} = p(1 + \frac{1}{2}\alpha\gamma)$. Note that, for a given field, the same perturbation (i.e., $\sqrt{\alpha}^{\gamma}$, $\frac{1}{2}\alpha\gamma$ or $1 + \frac{1}{2}\alpha\gamma$) is applied for all times and locations; γ is therefore generated by a pseudo-random number generator just once for each field or parameter (and for each member). The uncertainties for the reaction rates of RACM are adapted from Beekmann et Derognat [2003] who use MELCHIOR chemical mechanism [Lattuati, 1997]. A comparison between the MELCHIOR and RACM reactions was carried out in order to assume the same uncertainty for almost all corresponding reaction rates. In total, we carried out the Monte Carlo runs to generate a 100-member ensemble.

Table 4.1: Input data uncertainties for Monte Carlo simulations. The column "Uncertainty" reports the interval with 95% confidence whose median coincides with the unperturbed data. If the uncertainty is α and the initial parameter is p, every sample will lie in $[p/\alpha, \alpha p]$ for a log-normal distribution and $[p - \alpha, p + \alpha]$ for a normal distribution. For temperature, the interval is $[(1-\alpha)p, (1+\alpha)p]$.

Field	**Uncertainty**	**Distribution**
Meteorological Fields		
Wind speed	1.5	Log-normal
Wind angle	20°	Normal
Temperature	1%	Normal
Vertical diffusion coefficient	1.9	Log-normal
Attenuation	1.3	Log-normal
Emissions		
Anthropogenic NO_x	1.7	Log-normal
Anthropogenic VOCs	1.7	Log-normal
Biogenic VOCs	2.	Log-normal
Chemical rates – RACM [Stockwell et al., 1997]		
Reaction rates 128, 130, 132, 133		
136, 137, 143, 146, 147		
151, 152, 155, 156, 162		
166, 167, 168, 232, 236	1.3	Log-normal
Reaction rates 127, 129	1.2	Log-normal
Others reaction rates	1.1	Log-normal
Others		
Photolysis rates	1.4	Log-normal
Boundary conditions NO_x	3.	Log-normal
Boundary conditions	2.	Log-normal
Deposition velocities	1.7	Log-normal

A multimodel ensemble has a more detailed representation of the uncertainties as it can include perturbations in the input data (just like Monte Carlo simulations), but also takes into account the uncertainties due to the physical, chemical and numerical formulations. For instance, while the Monte Carlo simulations will take into account the shortcomings in the modeling of the turbulence with perturbations in the vertical diffusion coefficients, a multimodel approach will introduce several parameterizations for the vertical diffusion, each of which is physically consistent with the meteorological fields. A change of parameterization leads to much more relevant differences (e.g., increased vertical diffusion only in unstable conditions) than a simple multiplication of a reference diffusion field. In our case, the perturbations in the multimodel ensemble

are mostly applied to the same fields and parameters as in the Monte Carlo simulations, except for the reaction rates that are perturbed only in the Monte Carlo simulations.

4.2.2 Comparison of the Non-Calibrated Ensembles

The results of the two ensembles are studied for ozone peaks. The evaluation is carried out using the observation network Airbase [1]. The database, managed by the European Environment Agency, provides ground-level ozone observations at 210 rural background stations, 702 rural stations, 647 suburban stations and 1324 urban stations over Europe. In order to select stations which are representative of the ozone peak concentration at the model scale (half a degree in the horizontal), only rural and background stations are kept in this section. In order to select the most reliable stations, we exclude the stations that fail to provide observations for over 10% of all dates (i.e., all days in the year 2001). Following usual recommendations [Russell et Dennis, 2000; Hogrefe *et al.*, 2001; US EPA, 1991], a cut-off is applied to the observations. Observations below 40 $\mu g\,m^{-3}$ are discarded. In total, we keep about $1.1\,10^5$ observations for ozone peaks during the year 2001.

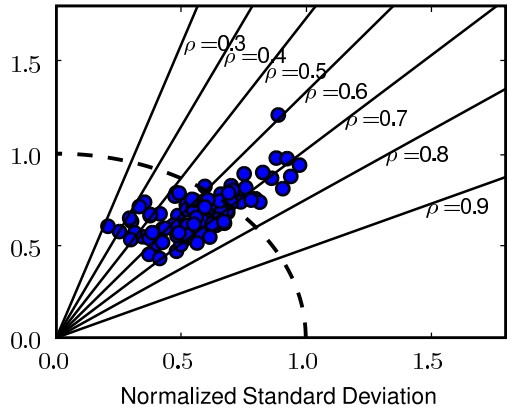
Figure 4.1 shows the models' performances on a single figure, using a Taylor diagram for the two ensembles. A Taylor diagram [Taylor, 2001] takes into account the standard deviation of observations and the correlation between each simulation and the observations. The radial coordinate is the simulation standard deviation normalized by the standard deviation of the observations. The azimuthal coordinate corresponds to the arcosine of the correlation between each simulation and the observations. The performances of the simulations from the multimodel ensemble show higher spread than the performances of the simulations from the Monte Carlo ensemble. In the Monte Carlo case, all the simulations have a standard deviation less than the standard deviation of the observations. Moreover, almost all correlations are between 0.6 and 0.7. On the contrary, the range of correlations in the multimodel ensemble is wide, and several simulations show a higher variability than the observations.

The mean of the ensemble standard deviations (at observation stations) is 22.3 $\mu g\,m^{-3}$ for the multimodel ensemble and 19.9 $\mu g\,m^{-3}$ for the Monte Carlo ensemble. It shows that the simulations from the Monte Carlo experiment have a lower variability than those of the multimodel ensemble. It is noteworthy that the "best" model (in terms of RMSE and correlation) from the multimodel ensemble better compares to the observations than the "best" model from the Monte Carlo ensemble. The "best" simulation in the multimodel ensemble has a RMSE equal to 20.5 $\mu g\,m^{-3}$ and a 0.735 correlation whereas the "best" Monte Carlo simulation has a 21.8 $\mu g\,m^{-3}$ RMSE and a 0.679 correlation.

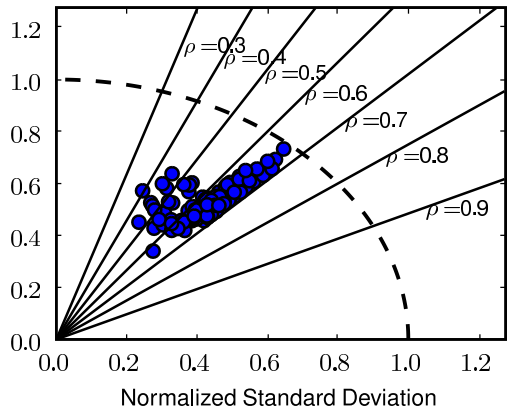
4.2.3 Comparison of the Calibrated Ensembles

The two ensembles take into account many sources of uncertainties, but the actual estimation of uncertainty that can be derived from them may not be reliable. In order to assess the quality of the uncertainty estimation by the ensembles, we compute the rank histogram [Anderson, 1996; Hamill et Colucci, 1997; Talagrand *et al.*, 1999]. Each observation is given a rank which is the number of members that simulate a concentration lower than the observation. The rank is zero if the observation is below the lower envelope of the ensemble, and it is equal to the total number of members if the observation is above the upper envelope of the ensemble. The rank histogram displays, for each rank, the number of observations with that rank. An ensemble properly estimates the uncertainty if the rank histogram is almost flat.

1. http://air-climate.eionet.europa.eu/databases/airbase/airbasexml/index_html



(a) Multimodel



(b) Monte Carlo

Figure 4.1: Taylor diagrams for ozone peaks from Airbase stations, for the multimodel ensemble and the Monte Carlo ensemble.

Figure 4.2 displays the rank histograms for the multimodel and Monte Carlo ensembles. A significant number of observations fall below the lower envelope of the multimodel ensemble. On the contrary, the Monte Carlo ensemble has a lot of observations above its upper envelope. This shows that the ensembles are not spread enough in given time periods or given regions. Besides the first and last bars, the histograms differ from the flat histogram which would coincide with the dotted lines in the figure.

In order to improve the uncertainty estimation that can be derived from the ensembles, we apply independently to both ensembles the calibration procedure that was introduced in Garaud et Mallet [2011]. The main idea is to extract a sub-ensemble with a flat rank histogram so that its uncertainty estimation should be more accurate. The ensemble calibration is therefore a combinatorial optimization problem. We use a genetic algorithm to select a sub-ensemble that minimizes the variance of the rank histogram [Garaud et Mallet, 2011]. Note that the observations outside the envelope of the full ensemble will be outside the envelope of the sub-ensemble. As a consequence, the heights of the first and last bars in the histogram can only increase after the selection of a sub-ensemble. Since the number of bars (i.e., the number of models plus one) in a flat histogram is the total number of observations divided by the height of one bar, only few members can be part of the sub-ensemble if the bars are too high. In order to increase the number of members in the sub-ensemble, we first remove from each member the global (spatio-temporal) bias of the ensemble mean, so as to decrease the height of the extreme bars. Figure 4.3 shows the rank histograms of the two unbiased ensembles. The extreme-bars values are significantly lower than with the bias, which eventually helps increasing the number of members selected by the calibration.

Figure 4.4 displays the rank histograms of the calibrated ensembles. For both ensembles, the rank histograms are almost flat. The calibrated multimodel ensemble includes 34 members while the calibrated Monte Carlo ensemble only has 23 members. This difference can partially be explained by the higher value of the extreme bar in the rank histogram of the full unbiased Monte Carlo ensemble. Also, as was pointed out in section 4.2.2, the Monte Carlo ensemble shows less variability than the multimodel ensemble, which makes it more difficult to properly span the range of the observations.

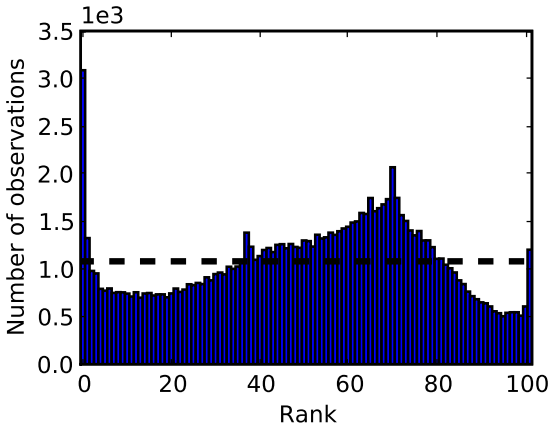
4.2.4 Uncertainty and Covariance Estimation

After calibration, the sub-ensemble is supposed to better sample the distribution of the uncertain ozone concentrations. The empirical standard deviations of the sub-ensembles therefore provide a reliable approximate measure of the uncertainty. If X_i is the state vector of the ith member, in a N-member ensemble, the empirical variance is defined as:

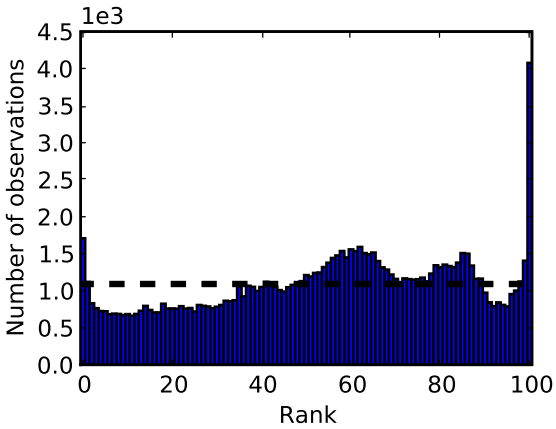
$$\Sigma = \frac{1}{N-1} \sum_{i=1}^{N} \left[X_i - \frac{1}{N} \sum_{j=1}^{N} X_j \right] \left[X_i - \frac{1}{N} \sum_{j=1}^{N} X_j \right]^T . \tag{4.1}$$

Figure 4.5 shows two maps of the empirical standard deviation (i.e., the diagonal of Σ) averaged over May 2001 for the calibrated multimodel and Monte Carlo sub-ensembles. In the Monte Carlo case, there is a high impact of the uncertain boundary conditions in the south of Europe, and the ozone ensemble standard deviation shows moderate variability in the rest of the domain. In the multimodel case, the field shows more gradients and local high values especially along the costs.

Uncertainty estimation is useful in data assimilation where the corrections on the model's state depend on the amplitudes and on the shapes of the state error variance and the observational error variance. A key point lies in the spatial distribution of the covariance between an error on ozone concentration at a given location and the errors at the other locations. This error covariance corresponds to one line of the matrix Σ. Figures 4.6 and 4.7 show two error covariance

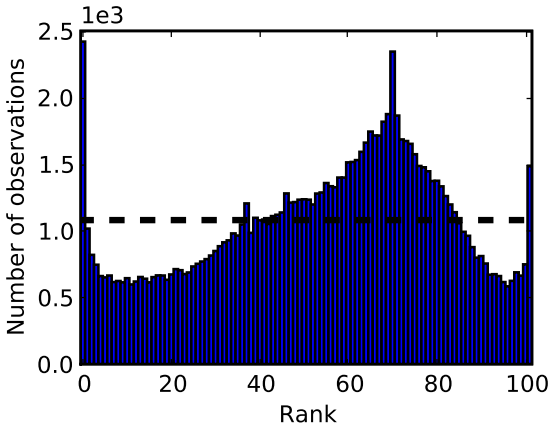


(a) Multimodel

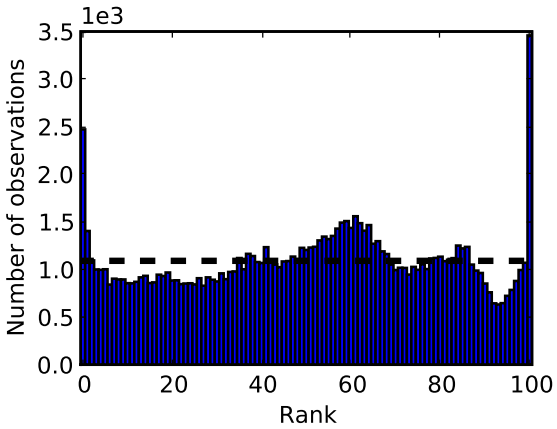


(b) Monte Carlo

Figure 4.2: Rank histograms for the multimodel ensemble and the Monte Carlo ensemble. The dotted line corresponds to the height of the flat rank histogram.

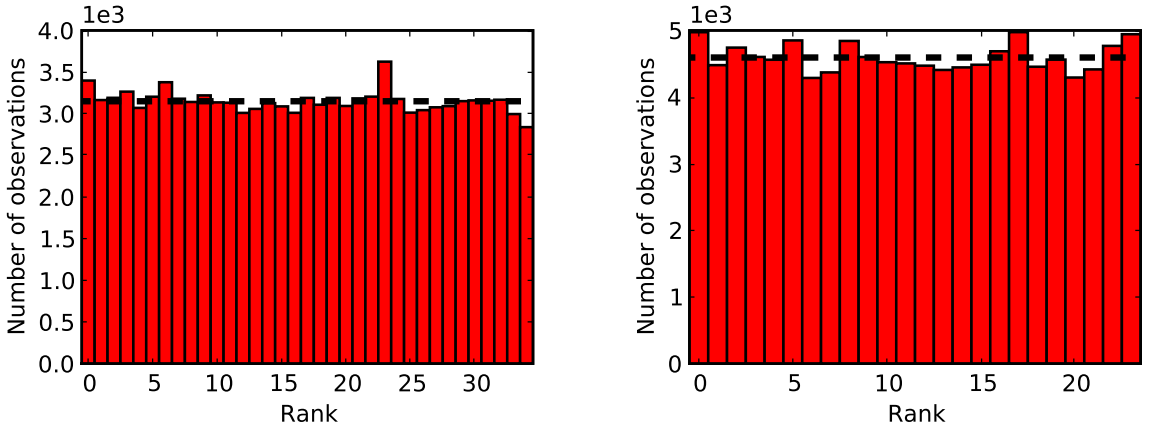


(a) Multimodel



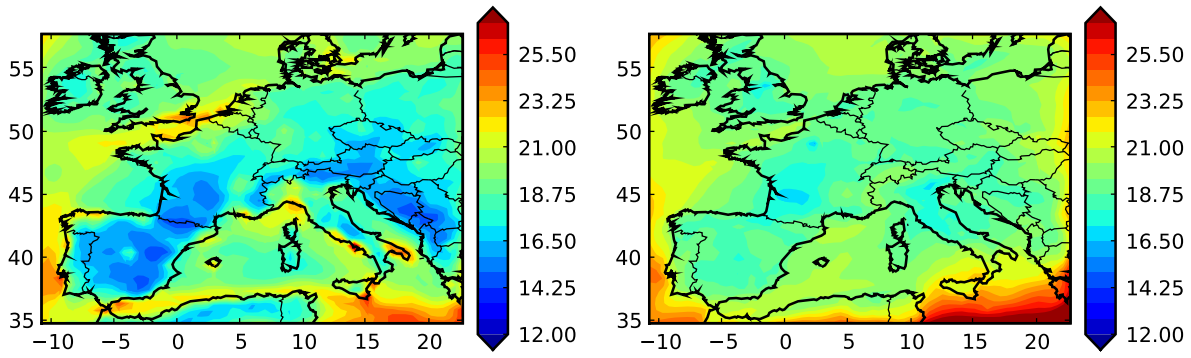
(b) Monte Carlo

Figure 4.3: Rank histograms of the multimodel ensemble and the Monte Carlo ensemble after removing the bias of the ensemble mean. The dotted line corresponds to the height of the flat rank histogram.



(a) Multimodel (b) Monte Carlo

Figure 4.4: Rank histograms of the multimodel sub-ensemble and the Monte Carlo sub-ensemble, after calibration. The dotted line corresponds to the height of the flat rank histogram.



(a) Multimodel (b) Monte Carlo

Figure 4.5: Empirical standard deviation computed from the calibrated multimodel sub-ensemble and calibrated Monte Carlo sub-ensemble. These fields, in $\mu g\,m^{-3}$, are averaged over May 2001.

fields, computed from the calibrated sub-ensembles. The covariances are computed with respect to two points: one location where there are large emissions (Paris), and another location in the center of France (background).

Monte Carlo covariance fields can show very high values along the domain boundaries. These values can be much higher than the covariances with close locations. It means that the perturbations on the boundary conditions travel down to the inner of the domain, and as the perturbations do not depend on time (nor on space), the transported perturbations remain correlated with those on the boundary conditions. Just like for the variance, the covariances show moderate variability (except along the domain boundary). There are however clear patterns mainly due to the emission points. In figure 4.6, the covariances with the background cell are lower at emission locations, while in figure 4.7, the covariance with the point located near emission sources show larger values at all emissions locations (even along the Mediterranean ships route). This raises the so-called localization issue in data assimilation; a data assimilation method that relies on such error covariances would unduly correct the concentrations far away from the observed locations, provided the errors on the model's concentrations at the observed locations are correlated with the errors at these away locations.

The covariances as approximated by the multimodel sub-ensemble show much more variability and the largest values are located in the vicinity of the point with which the covariances are computed. The perturbations in the boundary conditions do not lead to high covariances along the domain boundary, although the relatively high minimum of the covariance (about 200 $\mu g^2\,m^{-6}$) is partially due to the boundary conditions (see section 4.3.2). Finally, as one may expect, the covariance decreases faster (with the distance) for the point located near large emissions than for the background point. As in the Monte Carlo case, most emission locations appear in the maps. This is due to the perturbations on the emissions that do not depend on space and therefore create spurious correlation over long distances.

Compared to the Monte Carlo sub-ensemble, the multimodel sub-ensemble seems to better represent the uncertainty. It includes more models, it is less subject to the covariance localization problem and the patterns in the variance and covariance fields look better (especially along the domain boundaries). In order to improve the results of both approaches, the input data should be perturbed with spatial and temporal decorrelations, instead of with a single space- and time-independent perturbation (multiplicative for log-normal distributions, additive for normal distribution).

4.3 Uncertainty Due To Input Data

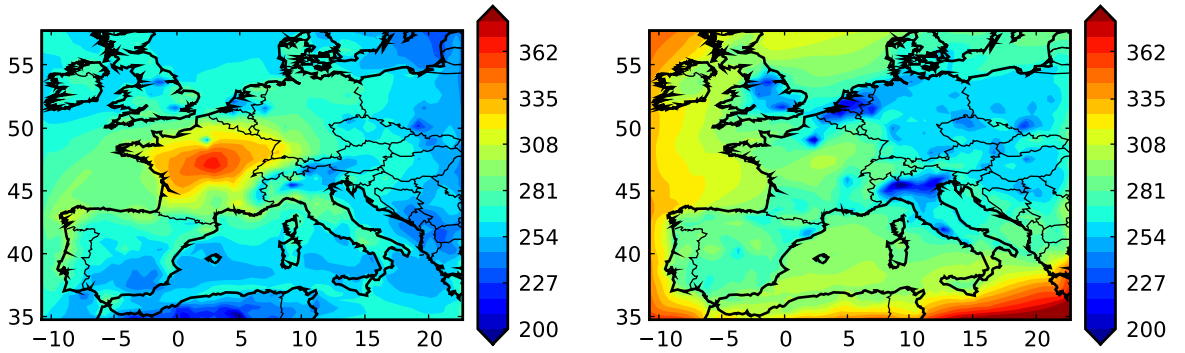
In this section, we evaluate the impact of the uncertainty in the input data on ozone simulated from 1st April to 31th August 2001. We want to identify what are the main uncertainty sources in the temporal mean of ozone peaks at observation stations.

4.3.1 Correlation and Regression

Following Hanna *et al.* [2001], a linear regression is carried out between ozone ensemble simulations and the perturbations applied to the input fields and parameters. It identifies the linear relationship between the ozone concentration x_i of simulation i and the perturbations u_{ik} to the input data:

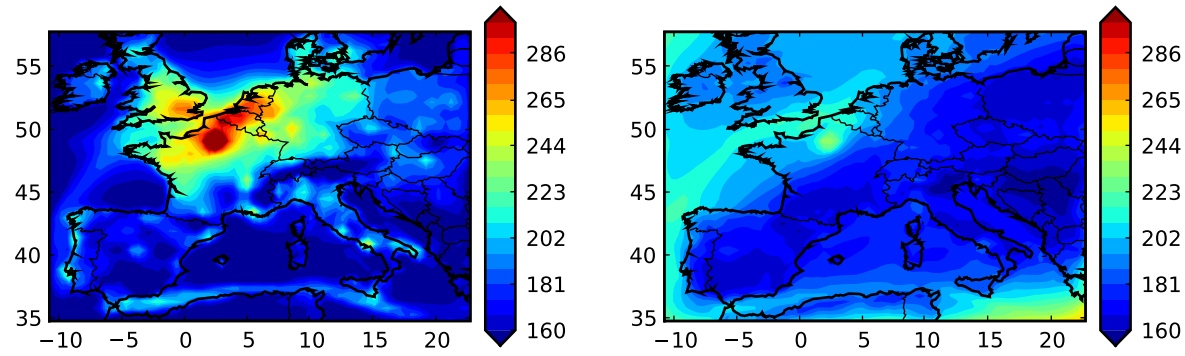
$$x_i = \beta_0 + \beta_1 u_{i1} + \beta_2 u_{i2} + \ldots + \beta_k u_{ik} + \ldots + \beta_K u_{iK} + \epsilon_i \,, \tag{4.2}$$

where β_k is a regression coefficient to be determined and ϵ_i is the error (i.e., the part of x_i that cannot be explained by the linear regression). u_{ik} is (1) the additive random perturbation ($\frac{1}{2}\alpha\gamma$)



(a) Multimodel (b) Monte Carlo

Figure 4.6: Ozone spatial covariance fields for the multimodel sub-ensemble and the Monte Carlo sub-ensemble, in $\mu g^2\,m^{-6}$. The location with respect to which the covariance is computed is a background cell in the center of France.



(a) Multimodel (b) Monte Carlo

Figure 4.7: Ozone spatial covariance fields for the multimodel sub-ensemble and the Monte Carlo sub-ensemble, in $\mu g^2\,m^{-6}$. The location with respect to which the covariance is computed is near Paris (France).

applied to the kth uncertain input field if this field is normally distributed, (2) the multiplicative random perturbation ($\sqrt{\alpha}^{\gamma}$) applied to the kth uncertain input field if the field is log-normally distributed, or (3) the multiplicative perturbation $1+\frac{1}{2}\alpha\gamma$ for temperature. ϵ is supposed to be a vector of white noise with a variance $\text{var}(\epsilon)=\sigma_\epsilon^2 I$ where I is the identity matrix.

If $x=[x_1,\ldots,x_i,\ldots,x_N]^T$ is the vector of the ozone concentrations (from a N-member ensemble), $\beta=[\beta_0,\ldots,\beta_k,\ldots,\beta_K]^T$ and $U=[1,u_1,\ldots,u_k,\ldots,u_K]$ is the matrix of all perturbations (except the first column that is filled with ones), the regression reads

$$x = U\beta+\epsilon\,. \tag{4.3}$$

The regression coefficients are determined as $\beta=(U^TU)^{-1}U^Tx$. If the absolute value of a regression coefficient $|\beta_k|$ is high, the ozone concentration is sensitive to the uncertainty on the field k. However for a regression coefficient β_k to be reliable, we require that its estimated value is higher than twice its standard deviation σ_{β_k}. The variance of β is given by

$$\Sigma_\beta^2=\sigma_\epsilon^2(U^TU)^{-1}. \tag{4.4}$$

$\sigma_{\beta_k}^2$ corresponds to the k^{th} diagonal term of the matrix Σ_β^2. The variance σ_ϵ^2 is computed from the regression residuals. If a coefficient does not satisfy the criterion $|\beta_k|\geq 2\sigma_{\beta_k}$, it is not excluded from the regression, but we ignore it in subsequent analyses.

The Monte Carlo ensemble contains 100 members in which 285 perturbations (i.e., 285 regressors) are applied. In order to decrease the number of regressors, only the regressors k for which the correlation ρ_k with the ozone concentration is greater than a given threshold are selected. The correlation threshold is taken so that the confidence interval on the correlation between the regressor and the ozone concentration does not include zero. The confidence interval depends on the sample size [Fisher, 1921], hence on the number of ensemble members. With normally-distributed perturbations and ozone concentration, it can be shown [Fisher, 1915] that the 95% confidence interval on the correlation ρ_k is

$$\left[\rho_k-\frac{e^{2s}-1}{e^{2s}+1},\rho_k+\frac{e^{2s}-1}{e^{2s}+1}\right] \tag{4.5}$$

where $s=\frac{1.96}{\sqrt{N-3/2}}$. With $N=100$, the confidence interval does not include zero if ρ_k is higher than 0.195. Note that most perturbations are not log-normally distributed, so that we only apply the criterion as a guideline.

The correlations are computed independently at each observation station. We then select the perturbations with a correlation satisfying $\rho_k>0.195$ at over 75% of background, rural or urban stations. It selects 20 parameters for background stations, 19 parameters for rural stations and 18 parameters for urban stations. Among these selected parameters, one finds ozone boundary conditions, NO_2 photolysis rate, the rate of photolysis of O_3 in O^{1D} (and O_2), attenuation, ozone deposition velocity and a few chemical reactions from RACM. A priori important fields are not always selected by this approach, so whatever their correlation with the output concentrations, we systematically include the perturbations of the nine following variables: VOC emissions, NOx emissions, biogenic emissions, NO and NO_2 boundary conditions, vertical diffusion coefficient, temperature, wind module and wind angle.

After the regression, we compute the determination coefficient R^2. It corresponds to the part of the variance that is explained by the linear combination. If we denote $\hat{x}_i=x_i-\epsilon_i$ and $\bar{x}=\frac{1}{N}\sum_{i=1}^N x_i$, then the determination coefficient reads

$$R^2=\frac{\sum_{i=1}^N(\hat{x_i}-\bar{x})^2}{\sum_{i=1}^N(x_i-\bar{x})^2}\,. \tag{4.6}$$

The determination coefficient depends on the number K of regressors and on the number N of samples. In order to compensate, we use the so-called adjusted determination coefficient AR^2:

$$\mathrm{AR}^2 = 1 - (1 - \mathrm{R}^2)\frac{N-1}{N-p-1}. \tag{4.7}$$

4.3.2 Results

The regression is applied independently at each observation station. The regressand x_i is the temporal mean of ozone daily peaks at the station. The regressors depend on the station type (urban, rural or background), since the correlation criterion (see above) selects a slightly different list of parameters for each station type. The regression coefficients and adjusted determination coefficients AR^2 reported in this section are averaged over all stations from one of these types: urban (665 stations), rural (263) or background (92) from Airbase network.

Monte Carlo

Table 4.2 shows the (averaged) regression coefficients and their spread (empirical standard deviation of the regression coefficients), computed for the most important fields and for each station type. Only the average coefficients that are greater than twice their average standard deviation are reported. The average of the adjusted determination coefficients is the same whatever the station type and is equal to 91%.

Table 4.2: Regression coefficients (averaged, in $\mu g\,m^{-3}$) for the three different station types. Output data (regressand) is the temporal average of ozone peaks. Only the average coefficients that are greater than twice their average standard deviation are reported. In brackets, we show the spread among the stations of the coefficients, computed as the empirical standard deviation of all regression coefficients – and not as the mean of the deviations σ_{β_k}.

Field Name	**Background**	**Rural**	**Urban**
O_3 bounday conditions	41.4 (4)	41.8 (4.2)	40.3 (4.3)
NO_2 photolysis	15.8 (10.8)	17.5 (8.8)	18.5 (9.6)
NOx emission	10.7 (5.1)	4.1 (7)	6. (8.8)
Vertical diff. coeff.	4. (2.4)	5.6 (1.8)	5. (3)
ISO emission	2.5 (2.9)	3.9 (2.2)	3.7 (3.1)
Attenutation	11.7 (5.3)	13.6 (5.1)	15 (5.9)
Temperature	162.3 (33.5)	147.7 (32)	160.5 (34)
API boundary conditions	5. (1.7)	–	–
NO_2 boundary conditions	1.8 (0.4)	1.8 (0.4)	1.6 (0.6)
Wind module	–	–	-6.3 (5.7)
Reaction 28	-32.1 (7.6)	-32.3 (8.3)	-32 (10)
Reaction 90	32.8 (9)	26.4 (9.8)	25.7 (11.2)
Reaction 182	-30 (6)	–	–

All regression coefficients shown in table 4.2 are in $\mu g\,m^{-3}$. The magnitude of the regression coefficients together with the typical variation of the perturbation coefficients provide a measure of the impact of input uncertainties on the output ozone concentration. For instance, the temperature regression coefficient is equal to $150\,\mu g\,m^{-3}$ and the associated random perturbation ranges in $[-1\%, 1\%]$ (see table 4.1). Thus, the part uncertainty due to the temperature is up to $\pm 1.5\,\mu g\,m^{-3}$. Another example is ozone boundary conditions whose regression coefficient is

$\sim 41\,\mu g\,m^{-3}$ for all stations, with a perturbation ranging in [1/2,2]. The impact of the uncertainty due to ozone boundary conditions therefore ranges roughly in $[-20.5,+41.0]\mu g\,m^{-3}$. Table 4.3 shows this impact for all input fields which appear in table 4.2.

For any field except NOx emissions, the regression coefficients are essentially the same for all station types. NOx emissions are given a higher coefficient for background stations than rural or urban stations. NOx are the limiting species for ozone production in background areas. Hence the sensitivity of ozone concentrations to NOx emissions is higher in these areas.

Ozone boundary conditions show an especially high positive regression coefficients, which is consistent with the high uncertainties along the domain boundaries that were pointed out in section 4.2.4.

Table 4.3: Regression coefficients and the associated 95% uncertainty range on ozone ($\mu g\,m^{-3}$) for several input fields described in table 4.1. The regression coefficients are averaged over all stations whatever their type, except for NOx emissions and reaction 182 whose regression coefficients are averaged at background stations only. The uncertainty range represents the deviation of the output concentration with respect to its median.

Field Name	**Averaged Coeff.**	**Uncertainty Range**
O_3 boundary conditions	41.0	-20.5, +41.0
NO_2 photolysis	17.0	-4.9, +6.8
NOx emissions	10.7	-4.4, +7.5
Vertical diff. coeff.	5.0	-2.4, +4.5
ISO emissions	3.0	-1.5, +3.0
Attenutation	13.0	-3.0, +3.9
Temperature	150.0	-1.5, +1.5
API bounday conditions	5.0	-2.5, +5.0
NO_2 boundary conditions	1.8	-1.2, +3.6
Wind module	-6.0	-3.0, +2.0
Reaction no. 28	-32.0	-3.2, +2.9
Reaction no. 90	30.0	-2.7, +3.0
Reaction no. 182	-30.0	-3.0, +2.7

The RACM chemical reactions 28 and 90 (see R 4.5 and R 4.6) have a significant regression coefficient for all station types. The former has a negative impact with a coefficient at $-32\,\mu g\,m^{-3}$, i.e., an impact on ozone ranging in $[-3.0,+2.9]\mu g\,m^{-3}$. On the contrary, chemical reaction 90 has a positive impact on ozone.

$$O^{1D} + H_2O \longrightarrow HO + HO \tag{R 4.5}$$

$$ONIT + HO \longrightarrow HC_3P + NO_2 \tag{R 4.6}$$

Chemical reaction 28 has a negative impact, which is difficult to explain since increased concentrations of HO should lead to more ozone production. The latter is confirmed by the sensitivities computed by Martien et Harley [2006] in the context of 3D photochemical simulations.

It is easier to explain why reaction 90 (R 4.6) has a positive regression coefficient, since it tends to indirectly produce ozone through the production of NO_2.

RACM chemical reaction 182 (R 4.7) has a negative regression coefficient which is significant only for background station type.

$$\mathrm{TOLP} + \mathrm{MO_2} \longrightarrow \mathrm{HCHO} + \mathrm{HO_2} + 0{\cdot}35\,\mathrm{MGLY} + 0{\cdot}65\,\mathrm{GLY} + \mathrm{DCB} \qquad \text{(R 4.7)}$$

Species TOLP tends to produce ozone via RACM chemical reactions 41, 142 or 202 [Stockwell *et al.*, 1997]. Thus, if the reaction rate of R 4.7 increases, TOLP concentration decreases and indirectly tends to produce less ozone.

Note that, even if we can explain the sign of the regression coefficients associated to the previous chemical reactions, it is however difficult to explain why these particular reactions among many others lead to higher uncertainties in ozone concentrations.

Multimodel Ensemble

In the multimodel case, all perturbations are included in the regression, since fewer input fields are perturbed. In particular, the rates of the chemical reactions are not perturbed. The perturbed fields are:

- VOC boundary conditions and VOC emissions;
- NO_x boundary conditions and NO_x emissions;
- O_3 boundary conditions;
- biogenic emissions.
- vertical diffusion, temperature and wind;
- deposition velocities;
- photolysis rates.

Note that the same perturbation is applied to all deposition velocities, whatever the chemical species. It was assumed that these velocities where highly correlated since they are based on the same aerodynamic resistance, quasilaminar sublayer resistance and since they are based on a bulk surface resistance derived from those of SO_2 and O_3 [Wesely, 1989; Zhang *et al.*, 2003]. The perturbations of the photolysis rates are also the same for all species. The perturbation scheme is described in Garaud et Mallet [2010].

Table 4.4 shows the eight significant averaged regression coefficients. Ozone boundary conditions have the highest regression coefficient, as in Monte Carlo case. It is noteworthy that ozone boundary conditions and photolysis rates are associated with similar regression coefficients as with the Monte Carlo ensemble: $\sim 44\,\mu g\,m^{-3}$ and $\sim 18\,\mu g\,m^{-3}$ respectively. The adjusted determination coefficient is equal to 72% — instead of 91% for the Monte Carlo case. This is consistent with the fact that part of uncertainty in the multimodel ensemble is represented by changes in the model formulation itself.

The average regression coefficient for deposition velocities is negative. When all deposition velocities increase, ozone deposition should play a key role and ozone concentrations should therefore decrease. The average regression coefficient for wind velocity is also negative, which is probably due to emitted pollutants (especially ozone precursors) being more diluted when the wind velocity increases.

As in the Monte Carlo study, the average regression coefficient for NOx emissions is higher at background stations than at urban and rural stations.

Whatever the ensemble, the part of the uncertainty due to ozone boundary conditions is high. This is consistent with the previous uncertainty maps, figures 4.5 and 4.6. It indicates that a multimodel ensemble, or at least a Monte Carlo ensemble, should be beneficial at higher scale, in

Table 4.4: Regression coefficients (averaged, in $\mu g\,m^{-3}$) for the three different station types. Output data (regressand) is the temporal average of ozone peaks from multimodel ensemble. Only the average coefficients that are greater than twice their average standard deviation are reported. In brackets, we show the spread among the stations of the coefficients, computed as the empirical standard deviation of all regression coefficients – and not as the mean of the deviations σ_{β_k}.

Field Name	**Background**	**Rural**	**Urban**
O3 Boundary Conditions	42 (4.6)	45 (5)	43.8 (5)
NOx Emissions	21 (7)	13.2 (9.8)	15.3 (10.9)
Photolysis	17.9 (5.8)	17.8 (4.8)	18.8 (6.5)
Biogenic emissions	7.8 (3.)	7.8 (2.2)	8.3 (3.2)
Wind module	-14 (3.5)	-14.4 (5.5)	-14 (6)
Deposition	-14 (4.4)	-12.8 (3.6)	-11.8 (4.2)
NOx Boundary Conditions	3 (0.8)	–	2.7 (0.9)

order to provide an ensemble of boundary conditions. The description of the uncertainty in the boundary conditions would then be much more accurate than the perturbation scheme we use in this work.

Now that we have analyzed the main sources of uncertainties due to the input fields and parameters, we investigate in the next section how the discrepancies between the observations and the simulations can be decomposed and what part is due to the shortcomings of the modeling.

4.4 Error Decomposition

In this section, we investigate the observational errors, we try to evaluate the representativeness error, and we derive from them the modeling error. Consider a model' state vector X and a vector of observations Y at the same time. Both can be compared using the observation operator that maps the state space into observation space, so that HX can be compared to Y. If the true concentration vector at the observed locations is Y^t and the true state vector is X^t, the discrepancy between the observation vector and its simulated counterpart can be decomposed as follows

$$\begin{aligned} e &= Y - HX \\ &= \underbrace{Y - Y^t}_{\text{measurement}} + \underbrace{Y^t - HX^t}_{\text{representativeness}} + H\,\underbrace{(X^t - X)}_{\text{modeling}} \\ &= e_o + e_r + He_m\,. \end{aligned} \tag{4.8}$$

The measurement error $e_o = Y - Y^t$ is notably due to limitations of the observation instrument, errors in the calibration of the instrument, possible mistakes in the retrieval and errors in the postprocessing of the raw measurements. The modeling error $e_m = X^t - X$ is due to the shortcomings of the model that computed the state X based on approximate physical, chemical and numerical formulation and erroneous input data. Even if X^t is known, it is not possible to compute the exact concentrations at the observation locations. The exact state vector X^t contains concentrations averaged in the model's grid cells, from which one cannot compute a point concentration inside a grid cell because of the sub-grid variability. We refer to a representativeness error $e_r = Y^t - HX^t$. Note that the relation between the state vector and the observations is provided by H which can be another source of errors. In our case, H is simply a linear operator (the 2D concentration field X being bilinearly interpolated at observed locations), but in general, H can be a complex operator based on approximations.

Our objective is to estimate the variance of each error. First, we assume that the errors e_o, e_r and e_m have zero mean and that they are mutually uncorrelated, e.g., $\mathrm{E}[H(X^t - X)(Y - Y^t)^T] = 0$. If H_i is the ith row of H, the covariance between two error components i and j of e is

$$\begin{aligned}
\mathrm{Cov}(e)_{ij} &= \mathrm{E}([(Y_i - Y_i^t) + (Y_i^t - H_i X^t) + (H_i X^t - H_i X)] \\
&\quad [(Y_j - Y_j^t) + (Y_j^t - H_j X^t) + (H_j X^t - H_j X)]^T) \\
&= \underbrace{\mathrm{E}[(Y_i - Y_i^t)(Y_j - Y_j^t)^T]}_{\text{measurement error variance}} \\
&\quad + \underbrace{\mathrm{E}[(Y_i^t - H_i X^t)(Y_j^t - H_j X^t)^T]}_{\text{representativeness error variance}} \\
&\quad + H_i \underbrace{\mathrm{E}[(X^t - X)(X^t - X)^T]}_{\text{modeling error variance}} H_j^T
\end{aligned} \tag{4.9}$$

4.4.1 Measure Error

In this study, we use ground stations from the European Airbase network. Ozone is measured by spectrometry, using its absorption in ultraviolet. A sample of ambient air is taken. A beam at wavelength 254 nm is emitted. Ozone molecules absorb a part of the radiation. A sensor turns the measured radiation into an electrical signal which is proportional with the sampling ozone concentration. Airparif, the organization responsible for monitoring air quality in the Paris region, produces upper bounds on the measurement uncertainties, along with the measurements themselves [Airparif, 2007]. The uncertainty takes into account many error sources from the different stages of the measurement chain: air sampling, data capture, electronic device, calibration, ...

Hourly ozone measurements and their uncertainty upper bounds, provided by Airparif for year 2009 and 30 monitoring stations, are clustered in a concentration intervals from $[0,20]\mu\mathrm{g\,m^{-3}}$ to $[80,\infty[\mu\mathrm{g\,m^{-3}}$. The uncertainty and the relative uncertainty (i.e., the uncertainty divided by the concentration) are reported in table 4.5 and in figure 4.8. The uncertainty corresponds to the 95% confidence interval, so that, in case of Gaussian errors, it is equal to twice the standard deviation. The uncertainty increases with the concentration, while the relative uncertainty decreases. The relative uncertainty can be higher than 50% when the measured ozone concentration is about $8\ \mu\mathrm{g\,m^{-3}}$. For high concentrations, the relative uncertainty ranges from $\sim 12\%$ to $\sim 9\%$ for ozone concentrations between ~ 80 and $\sim 150\ \mu\mathrm{g\,m^{-3}}$.

Table 4.5: For five concentration intervals, the table shows the average ($\mu\mathrm{g\,m^{-3}}$) of all measurements in the interval, the corresponding average of the uncertainty ($\mu\mathrm{g\,m^{-3}}$) and the corresponding relative uncertainty. The uncertainty corresponds to a 95% confidence interval; so if the error is Gaussian, it is twice the standard deviation.

Range ($\mu\mathrm{g\,m^{-3}}$)	**Av. Measure**	**Uncertainty**	**Rel. Unc.**
0 – 20	8.8	6.8	0.77
20 – 40	30.4	7.4	0.24
40 – 60	49.7	8.1	0.16
60 – 80	68.5	8.9	0.13
≥ 80	97.9	10.4	0.11

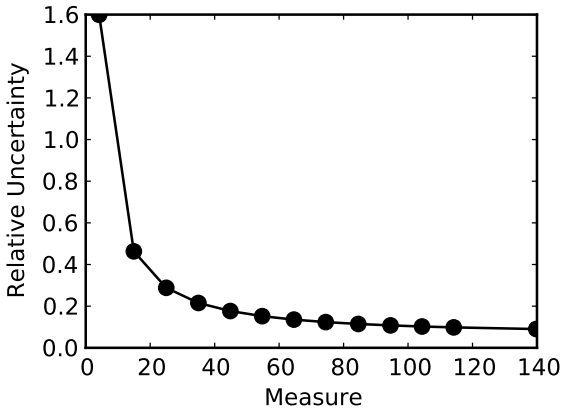


Figure 4.8: Relative uncertainty according to the observed ozone concentration in $\mu g\,m^{-3}$.

4.4.2 Modeling and Representativeness Errors

In this section, we carry out two independent methods to estimate the variance of the representativeness error. We consider all ozone hourly observations (not only the ozone peaks) and, in the second method, the multimodel-ensemble mean.

The first method is solely based on the observations. We assume that the mean concentration in a grid cell can be approximated by the mean of the observed concentrations. This assumption is reasonable only if there are enough observation stations in one model grid cell, and if they are spread all over the grid cell. In grid cell k, we denote J_k the set of the indexes of stations that are inside the grid cell. We approximate the mean concentration in the grid cell by

$$A_k = \frac{1}{|J_k|} \sum_{j \in J_k} Y_j \,, \tag{4.10}$$

where $|J_k|$ is the number of stations inside the grid cell k.

We select six grid cells that contain between 8 and 10 Airbase stations: two cells close to Marseille, one close to Paris, Barcelona, Valencia and London. All selected stations are urban, since in rural regions, the monitoring network is not dense enough to have so many stations in one grid cell.

We first compute $A_k - Y_j$ for the eight grid cells of interest and for all corresponding station $j \in J_k$, which amounts to 300,000 discrepancies. This quantity measures how much observations can deviate from the approximate average in the cell. Figure 4.9 shows the relative occurrence frequency of $A_k - Y_j$. By definition of A_k, the mean of the distribution is zero. The empirical standard deviation, which provides an estimate of the standard deviation of the representativeness error, is equal to $12.9\,\mu g\,m^{-3}$.

The second method is the sometimes referred to as the observational method or the Hollingsworth-Lönnberg method [Hollingsworth et Lönnberg, 1986]. The variance of the modeling error and the sum of measurement and representativeness variances are estimated based on a variogram of the discrepancies $Y - HX$. The variogram plots the empirical covariance between all pairs $(Y_i - H_i X, Y_j - H_j X)$ against the distance between the locations i and j. The first bar of the diagram, which corresponds to variances (because the distance is zero), is due to all three errors e_o, e_r and e_m. If the observational errors are assumed to be uncorrelated, the height of the next bars is only due to the modeling error e_m. If one extrapolates from these bars to the origin, the difference between the ordinate at the origin and the height of the first bar is due to the observational error $(e_o + e_r)$ — see figure 4.10 for an illustration.

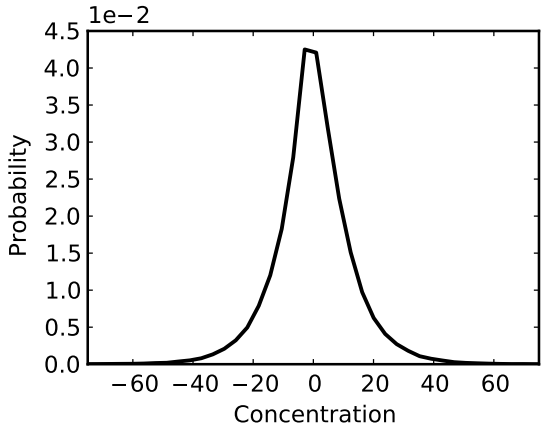


Figure 4.9: Approximate relative frequency occurrence of the representativeness error. The empirical standard deviation is 12.9 $\mu g m^{-3}$.

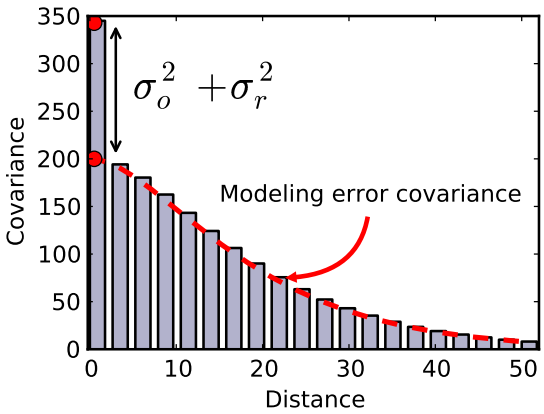


Figure 4.10: Illustration of the method by Hollingsworth et Lönnberg [1986] to estimate observational error variance with a variogram. This figure is inspired by Bouttier et Courtier [1999].

In our case, we consider all pairs (i, j) of urban stations, and all times at which observations are available. The error is computed using the mean of the calibrated multimodel sub-ensemble.

In total, ~ 190,000 covariances are computed. Figure 4.11 shows all covariances that decrease with the distance as the errors become uncorrelated.

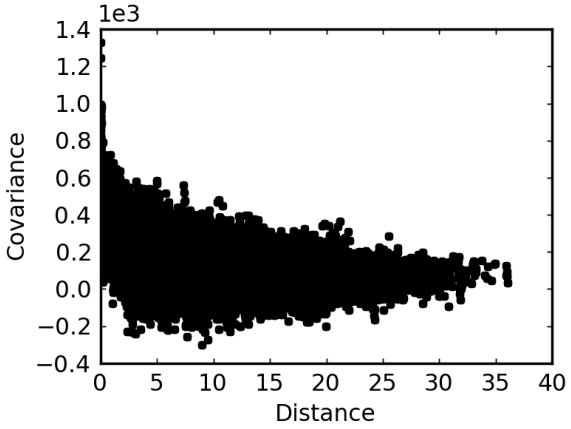


Figure 4.11: Variogram of the error between observed and simulated ozone concentration. The covariance is in $\mu g^2 m^{-6}$ and the distance in degrees (latitude/longitude).

We collect all covariance values from points within a distance in $]0, 0.5°]$. The mean of these covariances is equal to $\sigma_m^2 \simeq 336\,\mu g^2 m^{-6}$. The mean of the variance (computed with all pairs with null distance) is $\sigma^2 = 489\,\mu g^2 m^{-6}$. Hence $\sigma_o^2 + \sigma_r^2 \simeq \sigma^2 - \sigma_m^2 \simeq 153\,\mu g^2 m^{-6}$. The mean of observed concentrations is about $43\,\mu g m^{-3}$. According to section 4.4.1, it means that the variance of measurement errors is about $16\,\mu g^2 m^{-6}$. Consequently, the variance of the representativeness error can be estimated by $137\,\mu g^2 m^{-6}$, hence a standard deviation at $11.7\,\mu g m^{-3}$ which is a bit less than the estimation from the first method ($12.9\,\mu g m^{-3}$).

Observational error should be independent of the model, provided the same observation operator is used, which is the case in our ensemble where all models have the same horizontal resolution. We carry out the method with seven models randomly selected from the multimodel ensemble. The estimated variance of the observational error varies between $146\,\mu g^2 m^{-6}$ and $156\,\mu g^2 m^{-6}$. The standard deviation of the representativeness error is then estimated between 11.7 and $12.5\,\mu g m^{-3}$.

Note that the measurement and representativeness errors are not entirely uncorrelated between two points at close distance. For instance, the measurement errors can depend on the atmospheric conditions, which are obviously correlated at short distance. Therefore the estimation of the modeling error variance is overestimated, and the representativeness errors is underestimated. According to this method, the standard deviation of the representativeness error is likely to be greater than $12.5\,\mu g m^{-3}$, which is consistent with the first method giving $12.9\,\mu g m^{-3}$.

The variance of the discrepancies between hourly ground-level ozone observations and the mean of the calibrated sub-ensemble is $489\,\mu g^2 m^{-6}$. Following (4.9), it can be decomposed in less than ~ 3.2% for measurement errors ($\sigma_o^2 \lesssim 16\,\mu g^2 m^{-6}$), in about 34% for representativeness error ($\sigma_r^2 \simeq 166\,\mu g^2 m^{-6}$) and in about 63% for modeling error ($\sigma_m^2 \simeq 489 - 16 - 166 = 307\,\mu g^2 m^{-6}$).

4.5 Discussion and Conclusions

Two methods for the generation of ensembles are compared in this paper: (1) the Monte Carlo method where several input fields and parameters are perturbed and (2) a multimodel ensem-

ble generation which takes into account uncertainties in input data, numerical approximations and chemical/physical parameterizations. Monte Carlo simulations show less variety than the simulations from the multimodel ensemble. A posteriori calibrations are carried out on both ensembles in order to get more accurate uncertainty estimations. Even after the calibration, the multimodel sub-ensemble shows better features than the Monte Carlo sub-ensemble. More members are selected, and the variances and covariances patterns look better (especially along the domain boundaries). These results seem to justify the use of a multimodel ensemble in place of Monte Carlo simulation whenever it is possible. Note that in this paper, the multimodel ensemble also includes perturbations in the input data and it has a large number of members. A multimodel ensemble with just a few members might not show the same favorable features since the estimation of a variance may require more than a dozen members. It would however be interesting to extend this study with a multimodel ensemble based on models from different teams.

The regressions show that the uncertainties on the ozone boundary conditions, among all other input fields and parameters, have the largest effect on the uncertainties on output ozone concentrations. However, many other parameters and fields have a significant impact on ozone concentrations, which leads to a rather high total uncertainty. The mean uncertainty on ozone hourly concentrations, computed as twice the mean standard deviation divided by the mean concentration, is as high as 57.5%, according to the calibrated multimodel sub-ensemble over the full year 2001.

In the last section, we split the discrepancy between observations and simulations into a measurement error, a representativeness error and a modeling error. The measurement error is comparatively low. The representativeness error appears to be rather high, since it explains about 34% (in terms of variance) of the discrepancy with the observations. Consequently, the errors in the model outputs are significantly lower than the raw comparison with the observations suggests.

In future work, the generation of the ensembles could be improved. For all input fields, the perturbations should depend on time and space, with decorrelation length and time to be determined. The multimodel ensemble should be based on even more alternatives in the models generation. It is important that modern modeling systems get more flexibility so that they can include a wider range of model formulations. This especially applies to aerosol simulations, which were not addressed in this paper. There should be an extended range of options in the multimodel ensemble since many uncertainties lie in the formulation of the aerosol dynamics.

Finally, it would be interesting to evaluate if the uncertainty estimations computed in this study could help improving the performance of data assimilation. For instance, optimal interpolation could benefit from the patterns of the empirical state error variances. Another example is the ensemble Kalman filter that could rely on a multimodel ensemble instead of the usual Monte Carlo ensemble.

Acknowledgments
We would like to thank to Hélène Marfaing and Christophe Debert from Airparif for their very useful studies and their data about measurement uncertainties. We thank Christian Seigneur for his insights on the chemical processes.

Chapitre 5

Application pour l'impact de centrales thermiques

Une génération d'ensembles a été effectuée pour l'année 2007 et pour deux domaines régionaux, l'un en Île-de-France et l'autre dans les Pays de la Loire. Deux centrales thermiques se situent dans ces régions. Les ensembles ont été lancés suivant deux scénarios d'émission : l'un avec les émissions des centrales de Porcheville et de Cordemais, l'autre sans les émissions des centrales. On étudie les concentrations d'ozone, de dioxyde d'azote et de dioxyde de soufre.

Dans un premier temps, ce sont les performances des simulations de chaque polluant dans chacune des régions qui sont évaluées. Les performances pour O_3 et NO_2 sont assez satisfaisantes, mais SO_2 n'est pas correctement simulé. Les scores d'ensemble présentés dans les autres chapitres sont calculés, et des calibrations sont réalisées afin d'estimer l'incertitude et de rendre plus fiables la prévision des risques de dépassement de seuil. La calibration est généralement effective pour l'ozone et NO_2, mais elle ne peut être appliquée pour SO_2 en raison des performances des modèles individuels. Même après calibration, les évènements rares restent difficiles à prévoir. L'étude révèle que l'incertitude sur les concentrations d'ozone est assez faible près des sources d'émission et fortes près des conditions aux limites et le long des côtes. Au contraire, l'incertitude liée aux concentrations de NO_2 et SO_2, qui sont des espèces primaires, se situe d'abord près des sources d'émission.

La génération des ensembles suivant les deux scénarios d'émission a permis d'étudier l'impact des émissions des centrales. L'incertitude liée à cet impact est ensuite étudiée grâce aux simulations d'ensemble. La dispersion de l'ensemble des différences montre que l'incertitude sur l'impact calculé (avec un seul modèle) est très élevée. Sur la base de l'ensemble, on estime un impact maximal.

Enfin, ce sont les robustesses spatiales et temporelles de la calibration d'ensemble qui sont étudiées. La calibration pour l'estimation des incertitudes et pour la prévision des dépassements de seuil est assez robuste spatialement pour l'ozone et le NO_2. La calibration d'ensemble sur une période suffisamment longue rend possible la prévision de champs d'incertitude et de champs de probabilité de dépassement de seuil. Les prévisions d'incertitude et des dépassements de seuil fréquents se comparent bien aux résultats des ensembles calibrés a posteriori. Néanmoins, lorsque les dépassements sont rares, la calibration devient inefficace.

Sommaire

5.1 Introduction

Ce chapitre traite de l'application de l'estimation de l'incertitude et de la prévision de dépassement de seuil autour de centrales thermiques d'Électricité de France (EDF). On reprend ici la plupart des méthodes des chapitres précédents : génération automatique d'ensemble et calibration automatique d'un ensemble via une optimisation combinatoire.

L'étude menée dans ce chapitre ne concerne plus l'échelle continentale puisqu'on s'intéresse à la génération d'un ensemble de simulations photochimiques autour des centrales thermiques de Porcheville et de Cordemais, situées dans les régions d'Île-de-France et des Pays de la Loire respectivement. De plus, on ne s'intéresse plus seulement aux concentrations d'ozone mais aussi aux concentrations de NO_2 et de SO_2 — sachant que les centrales thermiques sont émettrices de ces deux derniers polluants.

La première partie de ce chapitre est consacrée à l'étude des performances des simulations des ensembles d'ozone, de NO_2 et de SO_2 comparées aux observations dans les régions d'Île-de-France et des Pays de la Loire. La deuxième traite plus spécifiquement des performances des ensembles via le calcul des diagrammes de rang, de fiabilité, du score de Brier ... Ces ensembles sont ensuite calibrés afin d'estimer au mieux l'incertitude des concentrations de polluants et aussi dans le but de rendre les probabilités de dépassement de seuil prévues par l'ensemble plus fiables. La génération des ensembles avec et sans les émissions des centrales permettra d'étudier l'impact des émissions des centrales de Porcheville et de Cordemais et aussi d'estimer l'incertitude liée à cette impact. Enfin, les robustesses spatiale et temporelle — cadre prévisionnel — de la calibration d'ensemble sont étudiées.

5.2 Contexte

Deux centrales thermiques — ou CPT pour Centre de Production Thermique — sont étudiées dans cette partie :

1. la centrale de Porcheville située dans les Yvelines à 80 km à l'ouest de Paris. Cette centrale fonctionne au fioul lourd, combustible à haute viscosité ;
2. la centrale de Cordemais dans les Pays de la Loire située sur la rive nord de l'estuaire de la Loire à 34 km à l'ouest de Nantes. Chacune des tranches de la centrale fonctionne soit fioul lourd soit au charbon. Les tranches 4 et 5 disposent d'un système de désulfuration qui réduit principalement les émissions de SO_2.

Dans tous les cas, ces dernières servent de centrale de soutien, elles ne fonctionnent pas à plein régime toute l'année et n'alimentent pas à elle seule la région concernée. La combustion du fioul et du charbon produit principalement des espèces telles que le SO_2, les NOx, du CO_2, du CO, des COV[1] et des particules (PM10, PM2.5). Dans la présente étude, les particules ne sont pas considérées. Deux scénarios d'émissions sont envisagés :

- avec les émissions des centrales ;
- sans les émissions des centrales.

Un certain nombre de données concernant le fonctionnement des centrales et de leurs émissions sur l'année 2007 ont été fournies par EDF R&D :

- le fonctionnement des différentes tranches des centrales sur toute l'année 2007. Les données concernent la puissance de fonctionnement en Watt par tranche horaire ;

1. Composés Organique Volatils ; VOCs en anglais pour « volatile organic compounds »

– la quantité totale en kg des espèces émises sur toute l'année.

La quantité de polluants émise à une date donnée correspond à la quantité totale multipliée par le ratio de fonctionnement de la centrale.

L'objectif de ce chapitre n'est pas uniquement de faire une étude d'impact — c'est-à-dire la comparaison de champs de concentration suivant les deux scénarios — mais aussi bien d'analyser l'estimation des champs d'incertitude et de calibrer l'ensemble pour la prévision des risques de dépassement de seuils réglementaires dans ces régions. Il est important pour EDF de savoir si les émissions de leurs centrales thermiques sont susceptibles d'être la source de dépassement de seuils réglementaires. Il existe deux types de seuils : (1) le seuil d'information et (2) le seuil d'alerte. Les moyens mis en œuvre par les autorités publics ne sont pas les mêmes en fonction du type de seuil. Le tableau 5.1 rappelle les valeurs de ces seuils pour l'ozone, le dioxyde de soufre et le dioxyde d'azote.

	NO_2	O_3	SO_2
Seuil d'information et de recommandation	200 $\mu g m^{-3}$ niveaux horaires	180 $\mu g m^{-3}$ niveaux horaires	300 $\mu g m^{-3}$ niveaux horaires
Seuil d'alerte	400 ou 200[1] en $\mu g m^{-3}$ niveaux horaires	1er seuil[2] 240 $\mu g m^{-3}$ 2nd seuil[2] 300 $\mu g m^{-3}$ 3e seuil 360 $\mu g m^{-3}$ niveaux horaires	500[2] en $\mu g m^{-3}$ niveaux horaires

[1] 200 dans le cas où la (1) procédure d'information/recommandation a été déclenchée la veille et le jour même et (2) si les prévisions font craindre un nouveau risque de déclenchement pour le lendemain.
[2] Concentration dépassée pendant 3 heures consécutives.

TABLE 5.1 – Seuils réglementaires (norme française) pour le NO_2, O_3 et SO_2 en $\mu g m^{-3}$. Source http://www.airparif.asso.fr/reglementation/normes-francaises.

Domaine

Une génération automatique d'ensemble de simulations photochimiques a été effectuée comme décrite dans le chapitre 2. Deux ensembles identiques d'une centaine de membres ont été lancés suivant les deux scénarios d'émission pour toute l'année 2007, dans un premier temps sur le domaine européen puis pour chaque domaine d'étude.

Les conditions aux limites pour les espèces chimiques des deux domaines régionaux proviennent de simulations européennes, dans le cadre d'une imbrication de domaines (*nesting*). Les données des simulations à l'échelle européenne fournissent les concentrations de chaque espèce, à chaque heure, aux abords des deux domaines régionaux et ce, pour chaque scénario.

En tout, six ensembles d'une centaine de simulations ont été lancés. Il y a d'abord deux ensembles pour le domaine européen, l'un sans émission des centrales et l'autre avec les émissions des deux centrales. Les résultats de ces ensembles ont ensuite servi de conditions aux limites pour les deux sous-domaines d'étude. Ensuite, deux ensembles par domaine régional ont été lancés pour chaque scénario d'émission. Au total, six ensembles, donc environ 600 simulations, ont été lancées sur l'année 2007.

À l'échelle européenne, les conditions aux limites sont fournies par des données issues du modèle global MOZART 2 [Horowitz *et al.*, 2003]. Ces données ne semblent cependant pas perti-

nentes pour forcer les simulations régionales, compte tenu de l'incertitude associée à ces données et de la résolution nettement plus fine utilisée par les domaines régionaux. Ce sont donc les données issues des simulations à l'échelle continentale qui ont servi de conditions aux limites des domaines régionaux. Des tests ont montré qu'imbriquer un domaine intermédiaire à l'échelle de la France (forcé par le domaine européen) n'améliorait pas les simulations régionales. La figure 5.1 représente le domaine européen et la taille des deux domaines correspondant approximativement aux régions Île-de-France et Pays de la Loire. Les tailles et les résolutions horizontales des trois domaines de simulation sont définies comme suit :

- Europe $[35.^{\circ}\text{N}, 56.75^{\circ}\text{N}] \times [10.5^{\circ}\text{W}, 22.8^{\circ}\text{E}]$, une résolution horizontale de 0.3° soit 8547 cellules (à chaque niveau de simulation).
- Île-de-France $[48.281^{\circ}\text{N}, 49.301^{\circ}\text{N}] \times [1.339^{\circ}\text{E}, 3.079^{\circ}\text{E}]$, une résolution horizontale de 0.03°, soit 1972 cellules (à chaque niveau de simulation)
- Pays de la Loire $[46.866^{\circ}\text{N}, 47.706^{\circ}\text{N}] \times [2.651^{\circ}\text{W}, 1.09^{\circ}\text{W}]$, une résolution horizontale de 0.03°, soit 1456 cellules (à chaque niveau de simulation).

Par la suite, on se concentrera sur les résultats des simulations à l'échelle régionale. Sauf mention contraire, le réseau d'observation utilisé pour estimer la performance des modèles, calculer les scores d'ensemble et effectuer les calibrations est le réseau français BDQA [2].

5.3 Performances des modèles

Cette section porte sur la comparaison entre les observations et les différentes simulations photochimiques des deux domaines régionaux. On considère les simulations avec les émissions des centrales, puisque les stations de mesures peuvent être impactées par ces émissions. Les stations pour lesquelles la disponibilité des observations est inférieure à 10% (c'est-à-dire que la station fonctionne moins de 10% du temps) sont exclues. Toutes les autres observations horaires sur l'année 2007 servent à la comparaison. On utilise les indicateurs statistiques du tableau 1.2 de la section 1.2.2. On utilise aussi le diagramme de Taylor [Taylor, 2001], précédemment utilisé pour illustrer les performances des modèles issus de l'ensemble photochimique européen pour l'année 2001 section 3.3.1. Il donne une vision graphique de deux indicateurs : (1) la corrélation entre chaque simulation et les observations ; (2) la comparaison entre l'écart type des simulations et celui des observations.

5.3.1 Porcheville

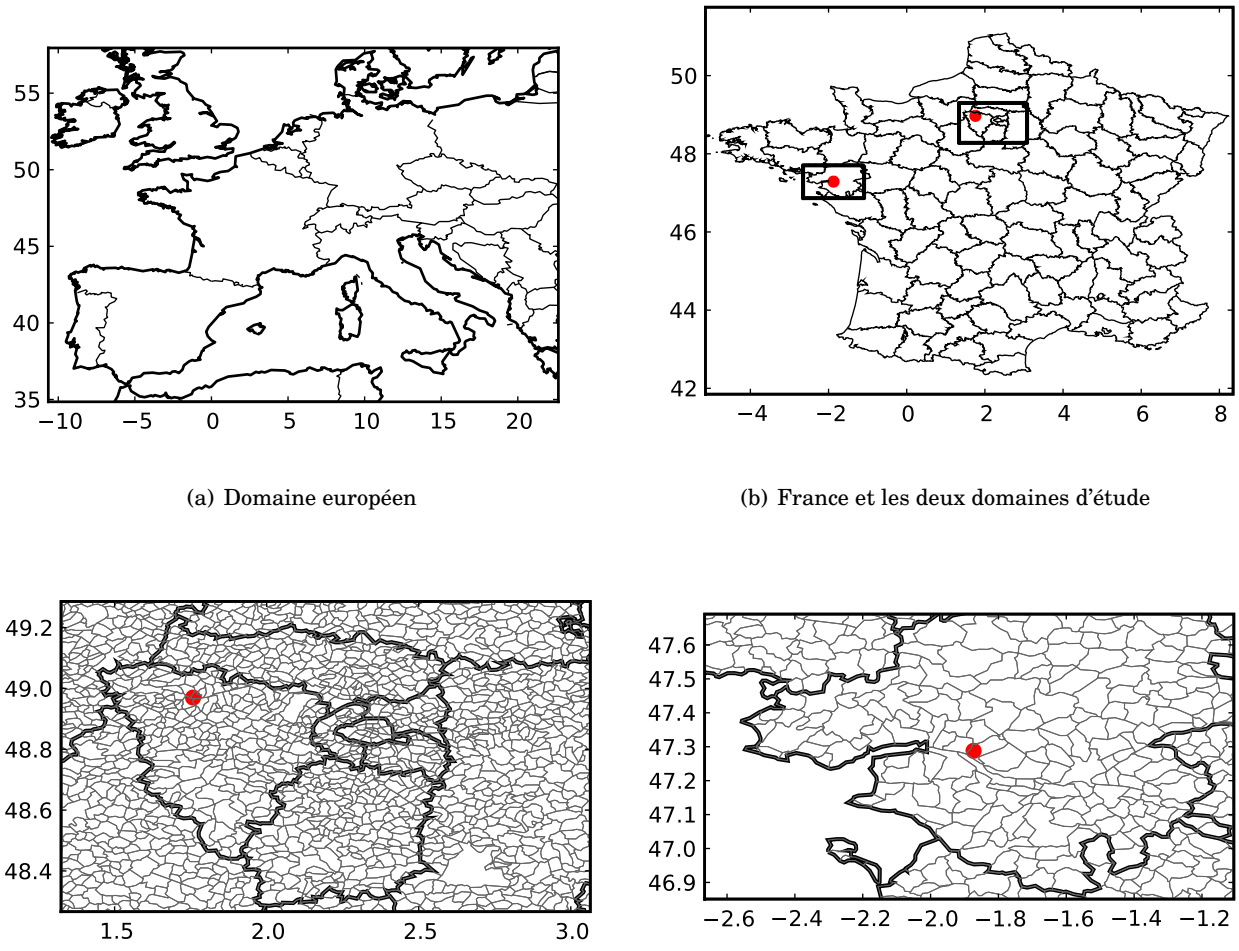
Le tableau 5.2 rassemble les résultats du meilleur modèle pris dans l'ensemble ainsi que de la moyenne de l'ensemble complet. On entend par « meilleur modèle » la simulation dont la RMSE est la plus faible. Généralement, cette simulation admet un bon facteur de biais ainsi qu'une bonne corrélation. Les performances du meilleur modèle et de la moyenne d'ensemble sont données pour chaque polluant. Il faut noter que le meilleur modèle n'est pas le même selon l'espèce considérée.

Le nombre de stations et de données horaires sont les suivants :

- 31 stations pour O_3 avec ~ 253 000 observations ;
- 42 stations pour NO_2 avec ~ 350 000 observations ;
- 12 stations pour SO_2 avec ~ 70 000 observations.

Le meilleur modèle, quel que soit le polluant, présente globalement de meilleures performances que la moyenne de l'ensemble. Cette dernière a néanmoins des performances satis-

2. Base de Données de la Qualité de l'Air — http://ssp.brgm.fr/spip.php



(a) Domaine européen

(b) France et les deux domaines d'étude

(c) Île-de-France et Porcheville

(d) Pays de la Loire et Cordemais

FIGURE 5.1 – Domaine européen qui a servi au *nesting*, carte de France et ses départements avec les limites des deux domaines d'étude. Les centrales thermiques de Porcheville et Cordemais sont localisées par les points rouges. Chacun des domaines régionaux fait apparaître la limite des communes.

faisantes avec 73% de corrélation pour l'ozone par exemple. Par rapport au meilleur modèle, elle admet un écart type plus proche de l'écart type observé pour O_3 et NO_2. Par contre, les résultats concernant le SO_2, que ce soit pour le meilleur modèle ou la moyenne d'ensemble, semblent assez éloignés des observations avec une surestimation de la concentration de SO_2, une RMSE supérieure à la moyenne simulée et des corrélations inférieures à 50%.

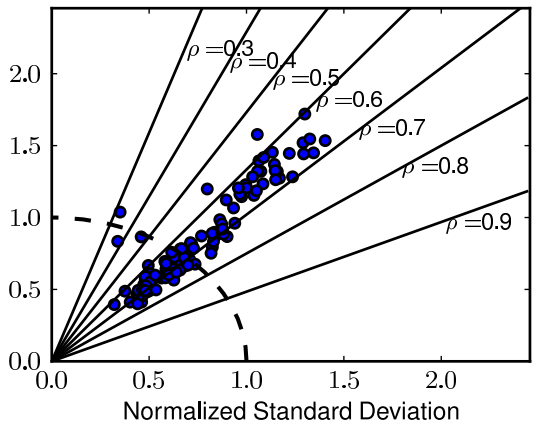
Données	**Moyenne**	**Écart type**	**BF**	**Corrélation**	**RMSE**
Observations O_3	44.	28.	–	–	–
Meilleur modèle O_3	47.	25.6	1.26	0.72	20.4
Moyenne d'ensemble O_3	50.3	28.7	1.33	0.73	21.8
Observations NO_2	38.1	30.8	–	–	–
Meilleur modèle NO_2	35.2	18.5	1.28	0.55	25.9
Moyenne d'ensemble NO_2	46.2	23.7	1.65	0.56	27.6
Observations SO_2	5.3	6.5	–	–	–
Meilleur modèle SO_2	13.9	8.6	4.32	0.42	11.88
Moyenne d'ensemble SO_2	20.7	15.2	6.3	0.41	20.74

TABLE 5.2 – Performances du meilleur modèle (par espèce) en terme de RMSE et de la moyenne d'ensemble pour le domaine de Porcheville en concentration horaire ($\mu g\,m^{-3}$) sur toute l'année 2007 avec le réseau BDQA. Pour le facteur de biais (BF), nous avons exclu les concentrations inférieures à 5 $\mu g\,m^{-3}$ pour le O_3 et NO_2, et à 0.1 $\mu g\,m^{-3}$ pour le SO_2.

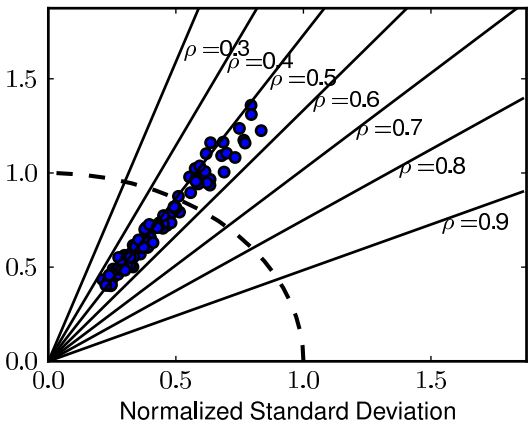
Ce tableau ne donne que des informations concernant un unique modèle et la moyenne de l'ensemble. Dans le but d'afficher les performances de tous les modèles de l'ensemble, des courbes de Taylor sont tracées. La figure 5.2 montre ces courbes pour les trois polluants étudiés. L'ensemble des simulations pour l'ozone admettent des corrélations au voisinage de $60-70$% et des écarts types très variables. Ces derniers peuvent atteindre jusqu'à deux fois l'écart type des observations qui est à 28 $\mu g\,m^{-3}$. Les performances des simulations pour les autres espèces sont globalement moins bonnes que pour l'ozone. Ainsi, les corrélations sont en moyenne égales à 50% et 39% pour le NO_2 et le SO_2 respectivement. Comme pour les écarts types des simulations d'ozone, les simulations de NO_2 présentent des écarts types assez différents comparés aux observations. Par contre, toutes les simulations de SO_2 ont des écarts types bien supérieurs à l'écart type des observations. En moyenne, ils sont égaux à 16 $\mu g\,m^{-3}$ contre un écart type de 6.46 $\mu g\,m^{-3}$ pour les observations. Les simulations de SO_2 surestiment aussi bien l'écart type que la moyenne des observations. En effet, la plupart des simulations ont une moyenne proche de 20 $\mu g\,m^{-3}$ contre 5.3 $\mu g\,m^{-3}$ pour la moyenne des observations.

Ces résultats peuvent être comparés aux performances de la simulation, avec aérosols, effectuée dans la même région par Tombette et Sportisse [2007]. Cette simulation a été lancée avec une résolution horizontale de 0.05° et neuf niveaux verticaux sur la période d'avril à septembre 2001. Comme dans le cas de notre ensemble, les données météorologiques sont issues du centre européen de prévision météorologique[3] et les conditions aux limites proviennent d'un *nesting*. Par contre, les données d'émission ont été fournies par Airparif — contre EMEP dans notre cas. Les performances sont données pour un certain nombre de stations d'observation. Quatre stations pour l'ozone et sept pour le dioxyde d'azote. Les corrélations calculées aux stations pour les concentrations simulées d'ozone sont autour de 70% et une RMSE entre 25 et 26 $\mu g\,m^{-3}$. Les concentrations simulées de NO_2 donnent des corrélations égales à ~ 50% et des RMSE autour

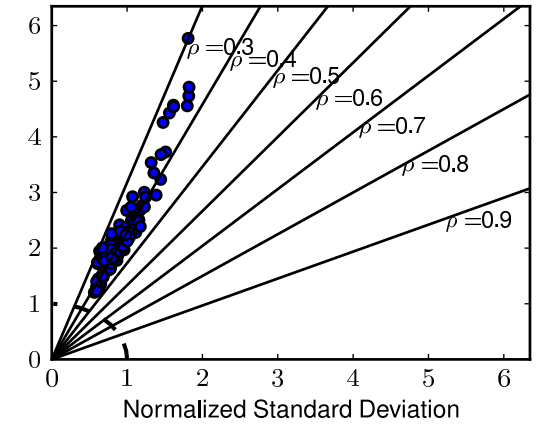
3. ECMWF (sigle anglais) ou CEPMMT (sigle français)



(a) O_3



(b) NO_2



(c) SO_2

FIGURE 5.2 – Courbe de Taylor pour les espèces O_3, NO_2 et SO_2 en concentration horaire pour le domaine de Porcheville. Les observations sont issues du réseau BDQA pour l'année 2007.

de 26 et 28 $\mu g m^{-3}$. On peut remarquer que ces résultats sont proches de ceux décrits dans le tableau 5.2.

5.3.2 Cordemais

De la même manière, nous calculons des scores pour le meilleur modèle et la moyenne d'ensemble, sur les trois polluants, dans la région où se situe la centrale thermique de Cordemais. Le nombre de stations et de données horaires pour cette région sont les suivants :

- 4 stations pour O_3 avec ~ 33 600 observations ;
- 11 stations pour NO_2 avec ~ 86 000 observations ;
- 9 stations pour SO_2 avec ~ 28 000 observations.

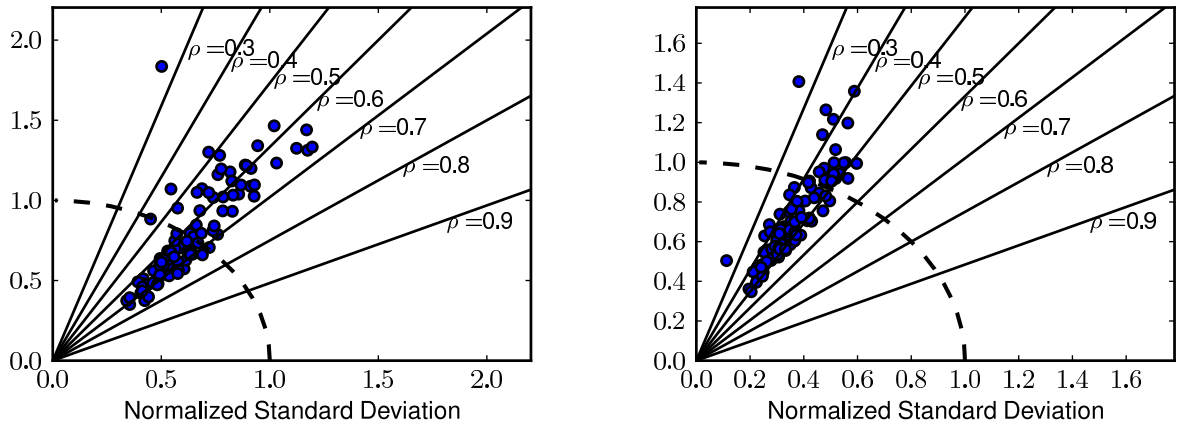
On remarquera que le nombre de stations est beaucoup plus faible que pour la région Île-de-France. Le tableau 5.3 reporte les performances du meilleur modèle — en terme de RMSE — et de la moyenne d'ensemble pour les trois espèces chimiques. Comme précédemment, les meilleurs modèles ont globalement de meilleures performances que la moyenne d'ensemble. Pour l'ozone et le dioxyde d'azote, les corrélations sont équivalentes à celles calculées pour le domaine de Porcheville. La RMSE de l'ozone avoisine les 19 $\mu g m^{-3}$ tandis que la RMSE du NO_2 est bien plus faible que celle calculée en Île-de-France : 13 $\mu g m^{-3}$ avec un assez bon facteur de biais proche de 1. Les concentrations de NO_2 sont cependant beaucoup plus faibles, et la RMSE est élevée, relativement à la moyenne de NO_2. Quant au SO_2, l'évaluation donne des résultats mauvais : une sur-estimation avec un facteur de biais assez élevé — jusqu'à 2.7 pour la moyenne d'ensemble, une très forte sous-estimation de l'écart type — 16.35 $\mu g m^{-3}$ observé contre 2.5 $\mu g m^{-3}$ et 4.7 $\mu g m^{-3}$ pour le meilleur modèle et la moyenne d'ensemble respectivement. De plus, les RMSE sont plus de cinq fois plus élevées que la moyenne simulée et les corrélations sont approximativement nulles.

Données	**Moyenne**	**Écart type**	**BF**	**Corrélation**	**RMSE**
Observations O_3	54.5	26.2	–	–	–
Meilleur modèle O_3	54.9	21	1.16	0.71	18.6
Moyenne d'ensemble O_3	70.8	23.4	1.55	0.71	25.15
Observations NO_2	17.1	14.6	–	–	–
Meilleur modèle NO_2	14.9	10.8	1.05	0.5	13.3
Moyenne d'ensemble NO_2	14.1	10.7	1.	0.5	13.5
Observations SO_2	5.76	16.35	–	–	–
Meilleur modèle SO_2	3.	2.5	1.61	-0.02	16.8
Moyenne d'ensemble SO_2	5.	4.7	2.7	-0.03	17.2

TABLE 5.3 – Performances du meilleur modèle en terme de RMSE et de la moyenne d'ensemble pour le domaine de Cordemais en concentration horaire ($\mu g m^{-3}$) sur toute l'année 2007 avec le réseau BDQA. Pour le facteur de biais (BF), nous avons exclu les concentrations inférieures à 5 $\mu g m^{-3}$ pour le O_3 et NO_2, et à 0.1 $\mu g m^{-3}$ pour le SO_2.

La figure 5.3 représente les courbes de Taylor des espèces O_3 et NO_2 pour le domaine des Pays de la Loire. Aux vues des performances des simulations de SO_2 en terme de corrélation — toutes proches de zéro — et pour des raisons de lisibilité, la courbe de Taylor du SO_2 n'a pas été tracée. Les résultats restent assez similaires aux courbes de Taylor des simulations du domaine de Porcheville avec des corrélations pour l'ozone proches de 60–70% et des écarts types des simulations de part et d'autre de l'écart type des observations. Pour le NO_2, les corrélations oscillent

autour de 50% et les écarts types des concentrations simulés sont globalement inférieurs à celui des observations.



(a) O_3 (b) NO_2

FIGURE 5.3 – Courbes de Taylor pour les polluant O_3 et NO_2 en concentration horaire avec les stations d'observation autour de la centrale de Cordemais.

En étudiant de plus près le type des stations d'observation qui mesurent les concentrations de SO_2 dans la région des Pays de la Loire, 8 stations sur les 9 disponibles sont des stations de type « industriel » tandis que la dernière est une station de type « urbain ». Comme la majorité des stations sont de type industriel, cela signifie qu'elles sont proches de sources d'émission. La concentration simulée du dioxyde de soufre dépend essentiellement des données d'émissions qui s'avèrent être incertaines. De plus, il peut y avoir de fortes incertitudes liées à l'échelle de représentativité, puisque la valeur de la concentration simulée dans la maille n'arrive pas à représenter de manière efficace les concentrations de SO_2 localisées ponctuellement près d'une éventuelle source d'émission. Cela peut expliquer la difficulté à simuler la concentration de SO_2.

On se propose alors de calculer les performances des modèles pour chacune des stations. En moyenne, les corrélations sont proches de zéro, sauf pour deux stations où les corrélations atteignent 33% et 13% avec des RMSE aux alentours de $12\,\mu g\,m^{-3}$ et $7.2\,\mu g\,m^{-3}$ respectivement. La première de ces deux stations est une station industrielle et se situe près de la centrale thermique, au sud-ouest de celle-ci. La deuxième est une station urbaine, la plus éloignée de la côte Atlantique. La figure 5.4 représente les neuf stations BDQA pour le SO_2 autour de la centrale thermique de Cordemais. La station industrielle, qui présente la corrélation la plus élevée, est représentée par le point bleu. La station de type urbain est localisée par le point vert.

Il est à noter que dans le cas du domaine Île-de-France, la majorité des stations d'observation pour le SO_2 est de type urbain et non industriel.

Les performances des modèles, exception faite des simulations de SO_2 du domaine de Cordemais, sont acceptables. Les courbes de Taylor mettent en exergue la diversité des modèles de l'ensemble. Comme dans les chapitres précédents, ces ensembles vont être calibrés dans le but d'estimer l'incertitude et de produire des prévisions probabilistes dans le cadre de prévision de dépassements de seuil. Avant cela, les ensembles, avec et sans calibration, sont évalués dans la section suivante.

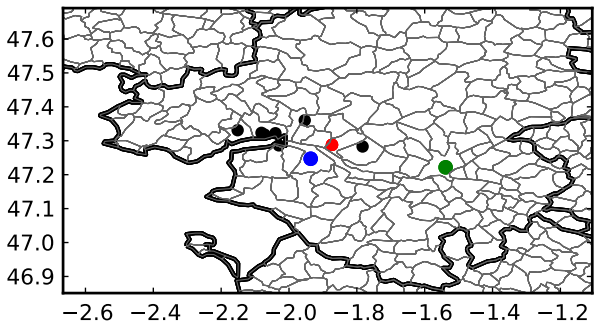


FIGURE 5.4 – Région de Cordemais avec les stations BDQA pour le SO_2. Les stations bleue et verte sont celles où la corrélation entre les simulations et les observations est significativement positive — station industrielle et urbaine respectivement. Le point rouge représente toujours la centrale thermique de Cordemais.

5.4 Score d'ensemble

Cette partie présente pour les deux domaines d'études les différents scores d'ensemble largement décrits et abordés dans les chapitres précédents. Des comparaisons sont effectuées entre les scores de l'ensemble complet et différents ensembles calibrés, pour les trois polluants.

Un autre moyen d'évaluation, non mentionné auparavant et qui concerne la prévision des risques, est le tableau de contingence 5.4. Il compare la prévision de l'occurrence d'un évènement avec des observations. Quatre scénarios a posteriori sont alors possibles :

1. succès : le système a prévu l'occurrence de l'évènement qui a bien eu lieu ;
2. rejet correct ou *correct negative* : le système n'a pas prévu l'évènement et celui-ci n'a effectivement pas eu lieu ;
3. échec : le système n'a pas prévu l'évènement qui s'est produit ;
4. fausse alarme : le système a prévu l'évènement qui n'a pas eu lieu.

	observé	**non observé**
prévu	succès — *hit*	fausse alarme — *false alarm*
non prévu	évènement manqué — *miss*	rejet correct — *correct negative*

TABLE 5.4 – Exemple d'un tableau de contingence pour la prévision d'un évènement.

Il est bien évidemment préférable d'avoir un maximum de succès ainsi que de rejets corrects pour un minimum de fausses alarmes et d'échecs. Contrairement au diagramme de fiabilité et au score de Brier, le système de prévision ne produit pas un ensemble de probabilités comprises entre $[0, 1]$ mais renseigne simplement sur l'occurrence ou la non occurrence d'un évènement particulier — autrement dit le système ne fournit que des probabilités égales à 0 ou à 1. Dans le cadre d'un ensemble de prévisions, on considère que le système prévoit l'occurrence de l'évènement quand au moins la moitié des membres de l'ensemble prévoit l'évènement. Appliqué à la

prévision de dépassement de seuil, le système fournit une probabilité égale à un quand au moins la moitié des simulations de l'ensemble dépassent effectivement le seuil.

Plusieurs ratios sont attachés à ces scores. Quatre scores sont utilisés : *accuracy* (5.1), *hit rate* (5.2), *correct negative ratio* (5.3) et *threat score* (5.4). Le premier permet de mesurer le rapport de réussites sur le total des observations. Le *hit rate* et le *correct negative ratio* mesurent la capacité à prévoir l'occurrence ou la non occurrence d'un évènement. Le dernier ratio, *threat score* proche du *hit rate*, a l'avantage d'indiquer le rapport de succès sur le total des occurrences de l'évènement avec une contrainte pour les fausses alarmes. Au vu des enjeux sanitaires et économiques qu'un dépassement de seuil en qualité de l'air peut engendrer, nous nous intéresserons particulièrement au *threat score*. En effet, il est à la fois nécessaire de mesurer la performance du système quand l'évènement apparaît — succès ou échec — tout en gardant un œil sur les fausses alarmes qui peuvent avoir un coût non négligeable.

$$\text{accuracy} = \frac{\text{hit} + \text{correct negative}}{\text{hit} + \text{correct negative} + \text{miss} + \text{false alarm}} \tag{5.1}$$

$$\text{hit rate} = \frac{\text{hit}}{\text{hit} + \text{miss}} \tag{5.2}$$

$$\text{correct negative ratio} = \frac{\text{correct negative}}{\text{correct negative} + \text{false alarm}} \tag{5.3}$$

$$\text{threat score} = \frac{\text{hit}}{\text{hit} + \text{miss} + \text{false alarm}} \tag{5.4}$$

Par la suite, il sera intéressant de connaître ces scores pour les différents ensembles calibrés selon leur fonction coût (diagramme de rang, fiabilité, score de Brier, ...), et des les comparer aux performances de l'ensemble complet. Par exemple, est-ce qu'un ensemble calibré, selon le diagramme de fiabilité, pour évènement de dépassement de seuil admet un bon *hit rate* ou un bon *threat score* ?

Malheureusement, les mesures de performance et a fortiori la calibration d'ensemble pour la prévision des risques sont très dépendantes des jeux de données d'observations. L'une des principales difficultés pour la prévision des risques en qualité de l'air réside dans la valeur des seuils de concentration réglementaires. Ce sont en effet des évènements rares. Il est donc (1) difficile de fournir des prévisions fiables et (2) délicat d'évaluer les prévisions fournies puisque ces évènements ont une très faible occurrence. À titre d'exemple, les stations d'observation dans la région des Pays de la Loire ne relèvent aucune concentration horaire d'ozone supérieure à $180\,\mu g\,m^{-3}$ sur toute l'année 2007 pour un total de 33600 observations. Il en est de même pour le NO_2 dans la même région. Le seuil réglementaire $200\,\mu g\,m^{-3}$ n'est jamais atteint parmi plus de 80000 observations. En conséquence, en plus d'étudier la prévision des dépassements de seuil réglementaire quand cela est possible, nous nous intéresserons aussi à des valeurs de concentration moins élevées telles que des concentrations d'ozone à $120\,\mu g\,m^{-3}$, utilisée dans la section 3.3.1 par exemple. On se référera au tableau 5.1 pour la valeur des seuils réglementaires de l'ozone, du dioxyde d'azote et du dioxyde de soufre.

Les sections suivantes traitent des différents scores d'ensemble pour l'ensemble complet et les sous-ensembles calibrés, en particulier le diagramme de rang et le diagramme de fiabilité. Ces scores serviront ensuite à la calibration d'ensemble pour (1) l'estimation de l'incertitude et (2)

la prévision des risques de dépassement de seuil pour les trois polluants mentionnés précédemment. Cette étude s'applique aussi bien au domaine d'Île-de-France qu'au domaine des Pays de la Loire.

Nous ne reviendrons pas sur la calibration automatique d'ensemble via l'optimisation combinatoire largement traitée dans le chapitre 3. Rappelons seulement qu'il est nécessaire de définir une fonction coût qui servira à sélectionner un sous-ensemble de l'ensemble complet, sous-ensemble qui minimise ou maximise la dite fonction coût.

5.4.1 Estimation de l'incertitude

Comme expliquée dans la partie 3.4, l'estimation de l'incertitude est mesurée à partir de l'écart type empirique de l'ensemble. Le score d'ensemble utilisé pour la calibration d'un ensemble représentatif de l'incertitude reste le diagramme de rang.

Porcheville

La figure 5.5 présente trois diagrammes de rang de l'ensemble complet pour chacun des polluants. Le diagramme de rang pour l'ozone présente des barres aux extrêmes peu élevés. L'ensemble des simulations d'ozone est d'ailleurs trop dispersé puisque le diagramme de rang à une forme de « cloche ».

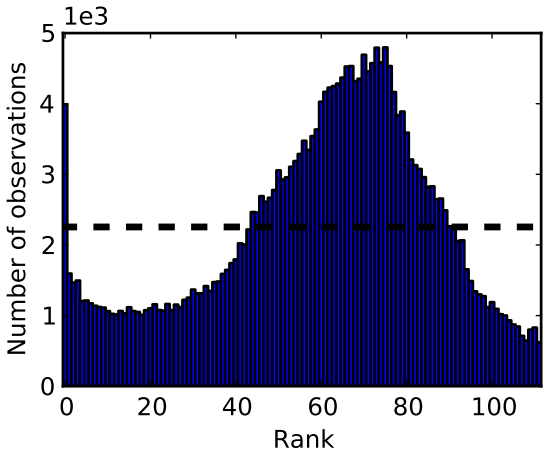
Au contraire, les diagrammes de rang pour le NO_2 et le SO_2 ont une forme de « U » et un biais fort. L'ensemble des simulations de NO_2 et SO_2 n'est (1) pas assez dispersé et (2) surestime globalement les observations — la grande valeur de la barre à l'extrême gauche des diagrammes de rang dénote d'un nombre important d'observations sous l'enveloppe inférieure de l'ensemble. À titre d'exemple, le nombre d'observations du dioxyde de soufre sous l'enveloppe inférieure de l'ensemble représente près de 87% des observations. Quant au diagramme de rang du NO_2, l'ensemble prend en compte plus d'observations à l'intérieur de son enveloppe que pour le cas du SO_2 mais reste très sous-dispersif. Il est difficile de calibrer de tels ensembles pour ces deux espèces. En effet, la calibration du diagramme de rang est contrainte par la valeur des barres les plus élevées et en particulier les barres aux extrémités du diagramme de rang comme expliqué dans la partie 3.3.2. On rappelle que le nombre de membres sélectionnés à l'issue de la calibration dépend de la valeur de la barre la plus haute du diagramme de rang.

Il est néanmoins possible de retirer le biais de la moyenne d'ensemble pour « symétriser » le diagramme de rang. Ceci a été effectué dans 3.4 et 4.

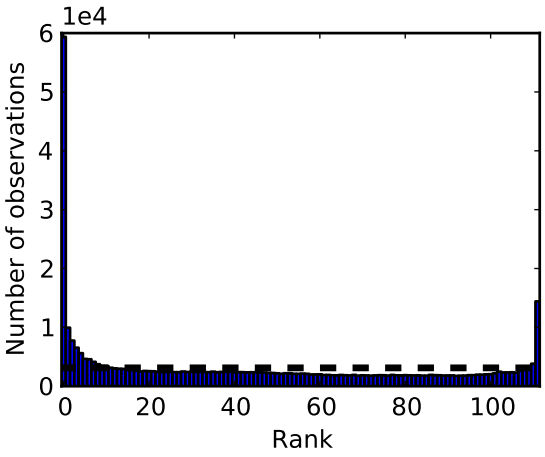
Malheureusement, le minimum de l'enveloppe inférieure de l'ensemble avoisine $0\,\mu g\,m^{-3}$, et ce quel que soit le polluant. Il est alors impossible d'enlever le biais positif sans introduire des valeurs négatives à une ou plusieurs simulations dont les concentrations sont faibles.

Pour le SO_2, l'optimisation combinatoire ne pourra même pas sélectionner un membre de l'ensemble autour duquel les observations sont réparties équitablement au-dessus et au-dessous. Il est donc impossible de construire un sous-ensemble calibré dont le diagramme de rang est plat. Pour NO_2, le nombre idéal de membres est égal à 4 ou 5 (ce qui correspond au nombre total d'observations divisé par la hauteur de la barre la plus haute). Néanmoins, le calcul de l'écart type empirique (estimation de l'incertitude) risque d'être peu fiable s'il est effectué avec si peu de membres.

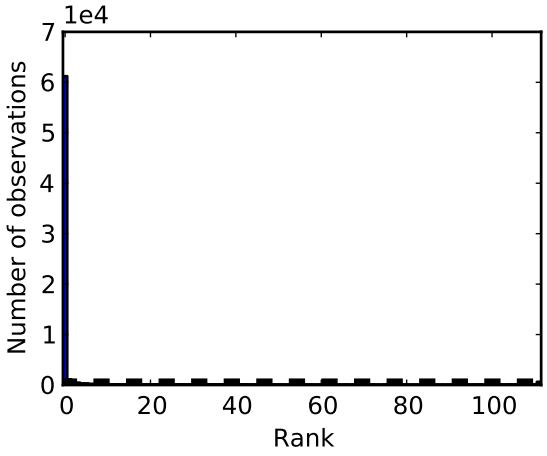
À défaut de retirer le biais aux membres de l'ensemble, il est possible de retirer les faibles concentrations observées tout en gardant un nombre d'observations assez conséquent. Pour le NO_2, nous retirons ainsi toutes les observations au-dessous de $25\,\mu g\,m^{-3}$. Cela permet d'avoir un



(a) O_3



(b) NO_2

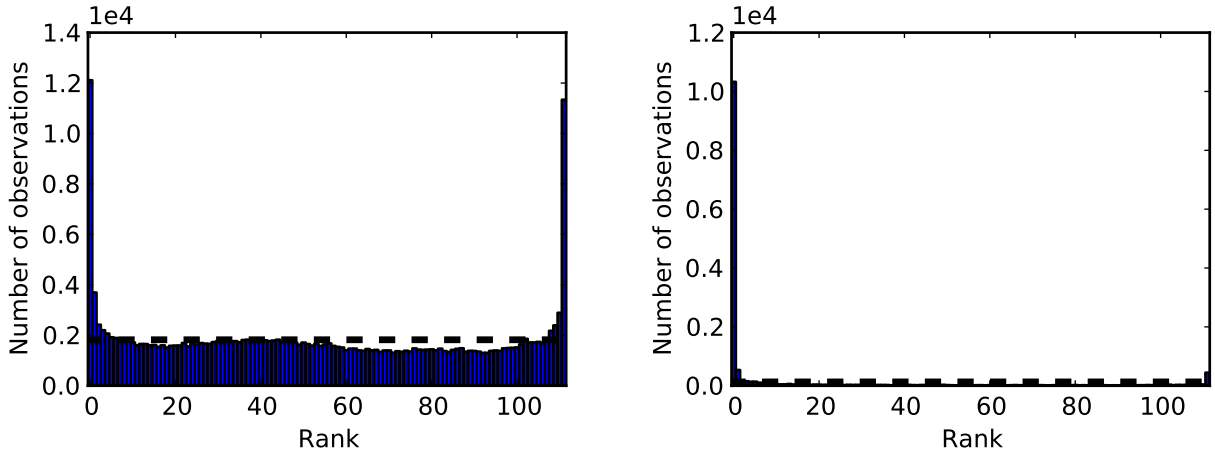


(c) SO_2

FIGURE 5.5 – Diagrammes de rang de l'ensemble complet pour les polluants O_3, NO_2 et SO_2 en concentration horaire pour le domaine Île-de-France avec le réseau BDQA.

diagramme de rang plus symétrique 5.6(a) et un nombre de membres idéal à 15 pour l'ensemble calibré. Le nombre d'observations reste important puisqu'il avoisine les 205000.

Par contre, pour le SO_2, il est impossible de diminuer le nombre d'observations sous l'enveloppe inférieure de l'ensemble de manière significative. En effet, en enlevant toutes les observations au-dessous 8 $\mu g m^{-3}$ — valeur au-dessus de la moyenne des observations, (1) le nombre d'observations passe de 70000 à 10000 environ et (2) ne permet pas d'obtenir une barre extrême gauche raisonnable (figure 5.6(b)). La calibration du diagramme de rang pour le SO_2 s'avère donc impossible ; l'optimisation combinatoire tendra à sélectionner un unique membre qui ne permettra même pas de séparer correctement les observations.



(a) NO_2 filtré à 25 $\mu g m^{-3}$

(b) SO_2 filtré à 8 $\mu g m^{-3}$

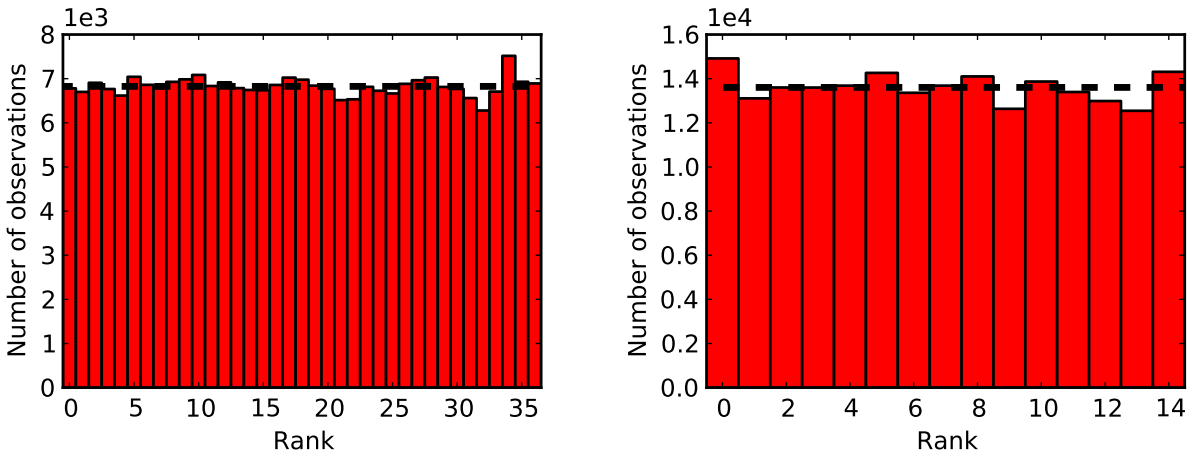
FIGURE 5.6 – Diagrammes de rang de l'ensemble complet pour les polluants NO_2 et SO_2 en concentration horaire pour le domaine Île-de-France. Toutes les observations au-dessous de 25 $\mu g m^{-3}$ et 8 $\mu g m^{-3}$ pour NO_2 et SO_2 respectivement ont été retirées.

Les calibrations du diagramme de rang ont été effectuées sur toute l'année 2007 pour les polluants O_3 et NO_2 en concentration horaire. Pour le NO_2, nous supprimons toutes les observations inférieures à 25 $\mu g m^{-3}$ dans le but de maximiser le nombre de membres dans le sous-ensemble calibré.

Ces diagrammes de rang calibrés sont représentés figure 5.7. Ils admettent des scores normalisés — définis dans l'équation (1.27) — égaux à 6 et 30 pour le O_3 et NO_2 respectivement. Ces scores sont bien meilleurs que ceux de l'ensemble complet puisqu'on a 724 et 1043 pour l'ozone et le dioxyde d'azote respectivement. Le nombre de membres dans ces sous-ensembles calibrés est assez satisfaisant : 36 membres pour l'ensemble O_3 et 14 pour le NO_2.

Après calibration, une estimation de l'incertitude est donnée par l'écart type empirique du sous-ensemble. On obtient ainsi un champ d'incertitude en $\mu g m^{-3}$ qui varie dans le temps. La figure 5.8 présente deux cartes de champs d'incertitude moyennés sur la période d'étude pour l'ozone et le dioxyde d'azote.

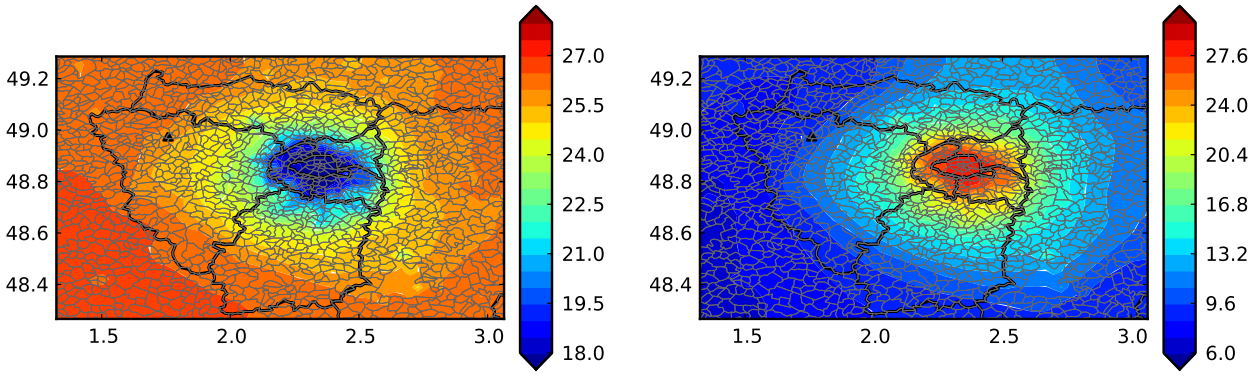
Ces champs d'incertitude sont tout à fait différents. Le premier présente les valeurs minimales près des sources d'émission — principalement autour de Paris et proche banlieue — tandis que le deuxième présente des valeurs maximales près des sources d'émission. Il n'est pas étonnant de noter que les plus grandes incertitudes du NO_2 proviennent des sources d'émission puisque c'est une espèce primaire. L'incertitude liée à l'ozone est toujours élevée près des condi-



(a) O_3 calibré

(b) NO_2 calibré

FIGURE 5.7 – Diagrammes de rang calibrés pour le O_3 et le NO_2 en concentration horaire pour le domaine Île-de-France. Toutes les observations au-dessous de 25 $\mu g m^{-3}$ pour le NO_2 ont été retirées.



(a) Incertitudes O_3 ($\mu g m^{-3}$)

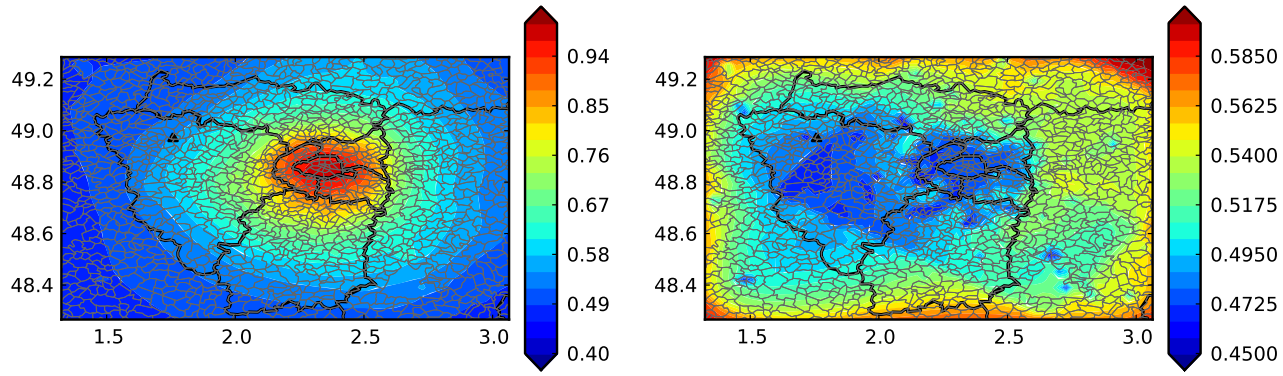
(b) Incertitudes NO_2 ($\mu g m^{-3}$)

FIGURE 5.8 – Cartes d'incertitudes moyennes pour O_3 et NO_2 en $\mu g m^{-3}$ pour le domaine Île-de-France issues des sous-ensembles calibrés via le diagramme de rang.

tions aux limites comme on a pu le constater dans la section 4.2.4.

Il peut être intéressant de tracer l'incertitude relative des champs de concentration. Dans chaque maille et à chaque pas de temps, on calcule l'écart type d'un ensemble calibré divisé par la moyenne de cet ensemble. La figure 5.9 présente les deux champs d'incertitude relative de l'ozone et du dioxyde d'azote moyennés sur l'année. On remarque que la structure spatiale du champ d'incertitude relative d'ozone est à l'opposé du champ d'incertitude absolue. En effet, ce sont au-dessus des zones d'émissions où les valeurs d'incertitude relative sont les plus élevées. Malgré les faibles valeurs de l'incertitude d'ozone à cet endroit, ces dernières ne sont donc pas négligeables par rapport à la valeur moyenne de la concentration d'ozone puisqu'on calcule un écart type relatif qui peut atteindre 1. L'incertitude relative près des conditions aux limites est quant à elle deux fois moins élevée que la valeur moyenne.

Au contraire, l'incertitude relative du dioxyde d'azote révèle des valeurs plus importantes près des conditions aux limites et moins élevées au-dessus des sources d'émission.



(a) Incertitudes relatives O_3 (%)

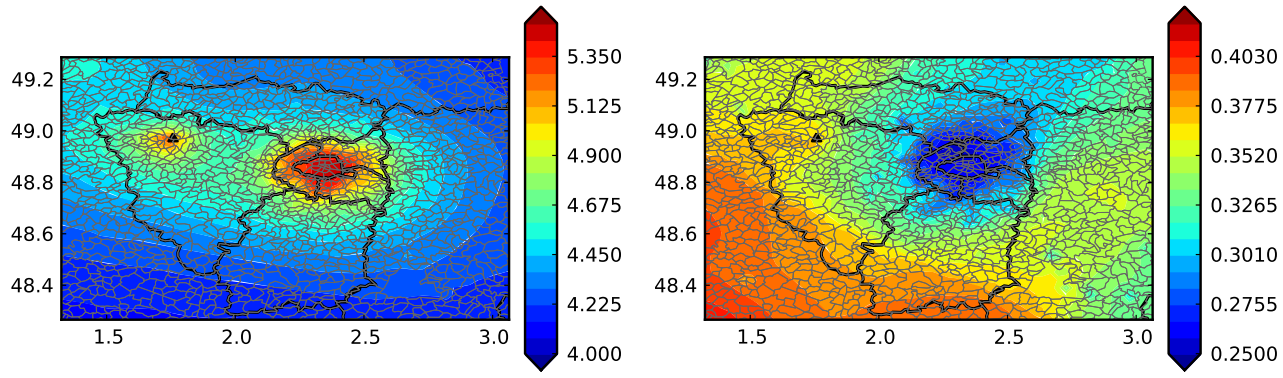
(b) Incertitudes relatives NO_2 (%)

FIGURE 5.9 – Cartes d'incertitudes relatives moyennes de O_3 et NO_2 pour le domaine Île-de-France issues des sous-ensembles calibrés via le diagramme de rang.

Même s'il est impossible pour l'instant de calibrer le diagramme de rang de l'ensemble des simulations de SO_2 de manière efficace, il est néanmoins possible de calculer l'écart type absolu et relatif de l'ensemble complet. Cela peut permettre de se faire une idée de la structure spatiale des incertitudes de SO_2. Attention cependant, l'ensemble n'étant pas calibré, il n'est a priori pas représentatif des incertitudes. Les valeurs données sur la figure 5.10 ne sont là qu'à titre indicatif et ne font pas office de référence.

Il est intéressant de constater que l'incertitude du SO_2, espère primaire, est plus forte près des sources d'émission, tout comme NO_2. On peut même localiser la centrale thermique de Porcheville. L'incertitude relative est plus faible aux abords des sources d'émission, notamment au-dessus de Paris. Néanmoins, la localisation de la centrale dans ce cas n'est pas visible.

La partie suivante traite de la calibration du diagramme de rang et de l'estimation de l'incertitude pour la région des Pays de la Loire.



(a) Incertitudes absolues SO_2 ($\mu g\,m^{-3}$)

(b) Incertitudes relatives SO_2 (%)

FIGURE 5.10 – Cartes d'incertitude absolue ($\mu g\,m^{-3}$) et relative (%) moyennes du SO_2 en Île-de-France.

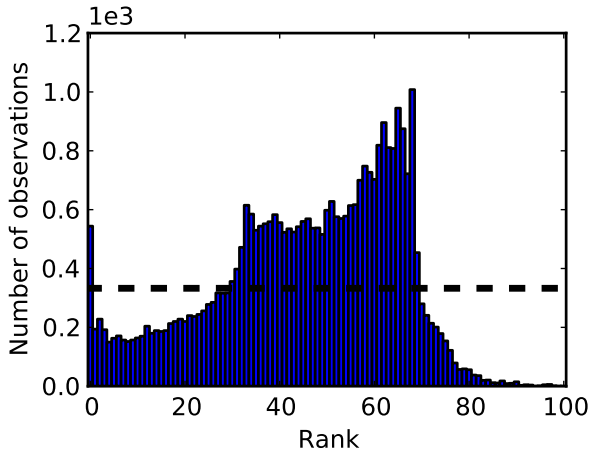
Cordemais

Comme précédemment, nous décidons de tracer les diagrammes de rang des trois polluants de l'ensemble complet, pour ensuite les calibrer et estimer l'incertitude.

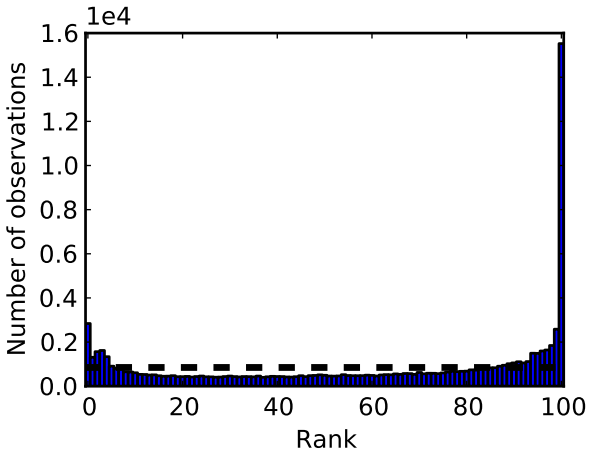
La figure 5.11 présente trois diagrammes de rang de l'ensemble complet pour l'ozone, le dioxyde d'azote et le dioxyde de soufre. Le diagramme de rang de l'ensemble des simulations d'ozone est proche de celui calculé pour le domaine Île-de-France : il a une forme de « cloche », ce qui dénote une sur-dispersion de l'ensemble. Les diagrammes de rang de l'ensemble des simulations de NO_2 et de SO_2 sont sous-dispersifs et encore très fortement biaisés. Néanmoins, pour le NO_2, le biais est cette fois-ci négatif, c'est-à-dire qu'un nombre important d'observations se trouve au-dessus de l'enveloppe supérieure de l'ensemble — représenté par la barre à l'extrême droite du diagramme rang.

Néanmoins, il est possible de retirer ce biais négatif entre la moyenne de l'ensemble et les observations qui est d'environ 3 $\mu g\,m^{-3}$ afin de rendre le diagramme de rang plus « symétrique ». Ainsi, la figure 5.12 fait apparaître le diagramme de rang de l'ensemble NO_2 sans le biais. L'ensemble reste bien évidemment toujours sous-dispersif mais beaucoup plus symétrique. On augmente ainsi légèrement les chances d'avoir un nombre plus élevé de membres dans le sous-ensemble calibré. Dans ce cas, le nombre maximum de membres du sous-ensemble calibré passe de 4 à 6. C'est donc l'ensemble débiaisé qui est utilisé lors de la calibration du diagramme de rang pour le NO_2.

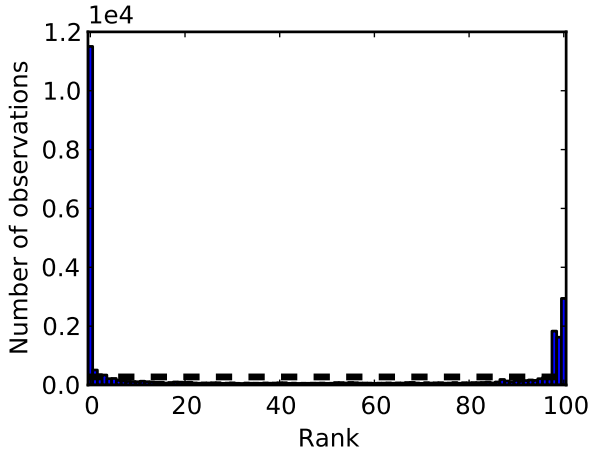
L'ensemble de SO_2 soufre du même symptôme que précédemment, c'est-à-dire qu'il sera impossible (1) de débiaiser l'ensemble sans introduire des concentrations négatives et (2) de retirer des observations sous un seuil donné dans le but de « symétriser » le diagramme de rang afin de garantir un maximum de membres dans le sous-ensemble calibré. Tel qu'il est (au vu de la valeur de la barre la plus haute), l'ensemble de SO_2 ne peut être calibré qu'avec un voire deux membres. Nous nous contenterons donc de calculer l'écart type de l'ensemble complet afin d'estimer les incertitudes du SO_2 près de la centrale thermique de Cordemais.



(a) O_3



(b) NO_2



(c) SO_2

FIGURE 5.11 – Diagrammes de rang de l'ensemble complet pour les polluants O_3, NO_2 et SO_2 en concentration horaire pour le domaine des Pays de la Loire avec le réseau BDQA.

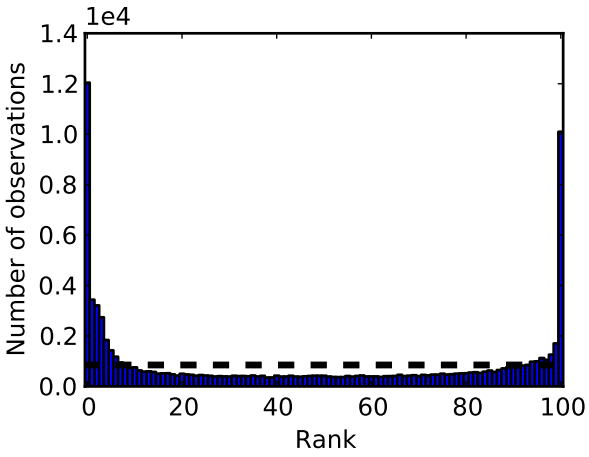
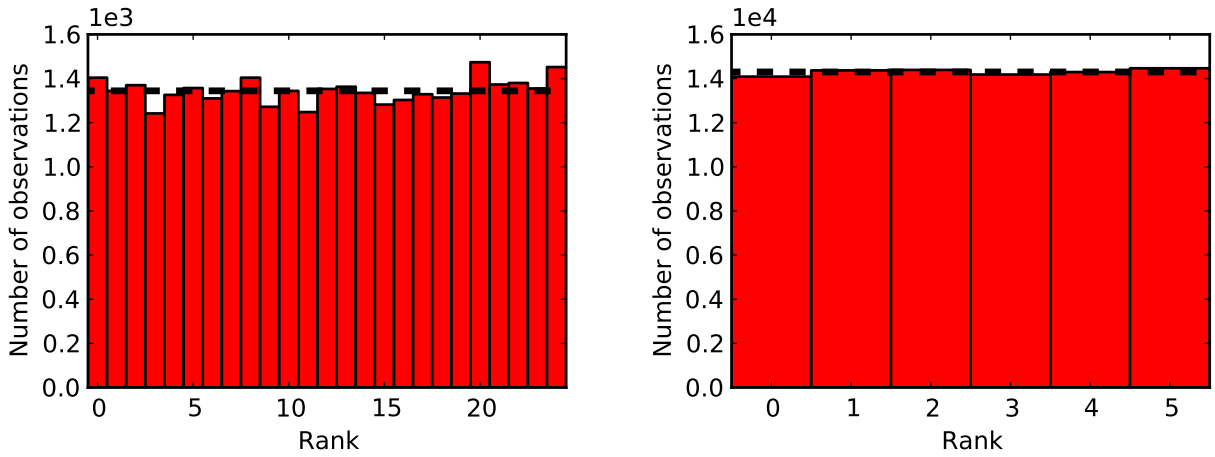


FIGURE 5.12 – Diagramme de rang de l'ensemble complet pour le NO_2 dans la région de Cordemais avec le réseau BDQA. Le biais négatif de la moyenne d'ensemble a été retiré.

Deux calibrations sont effectuées pour l'ozone et le dioxyde d'azote afin d'obtenir des diagrammes de rang plats. La figure 5.13 représente les deux diagrammes de rang calibrés. Le sous-ensemble de l'ozone compte 24 membres avec un score normalisé de 2.25 contre 218.5 pour le diagramme de rang de l'ensemble complet. Comme prévu, le diagramme de rang calibré pour le NO_2 ne compte pas beaucoup de membres mais reste bien plat. Le score normalisé pour ce dernier atteint 1.4 contre plus de 2840 pour le diagramme de rang de l'ensemble complet débiaisé.

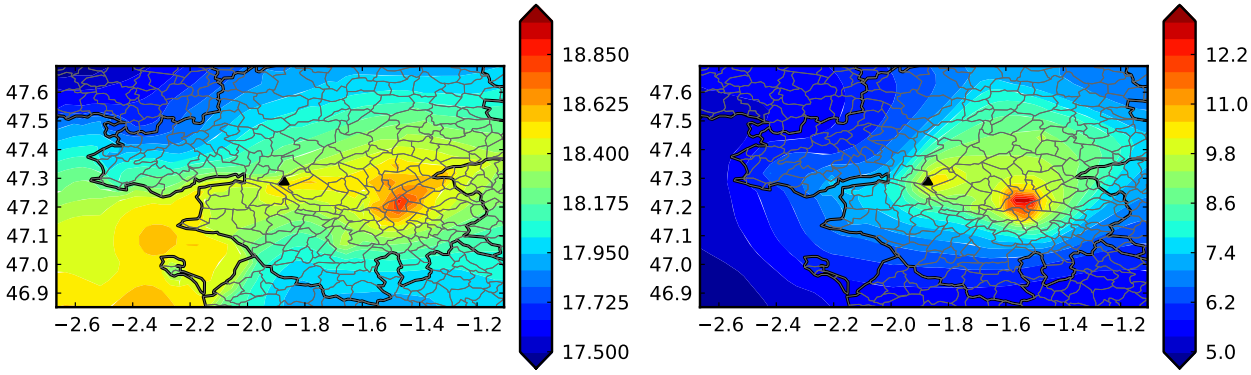


(a) O_3 (b) NO_2

FIGURE 5.13 – Diagrammes de rang calibrés pour les polluants O_3 et NO_2 en concentration horaire pour le domaine des Pays de la Loire avec le réseau BDQA. L'ensemble NO_2 a été débiaisé avant calibration.

Les cartes d'incertitude qui suivent ont été produites à partir du calcul de l'écart type des ensembles calibrés précédents à l'exception du SO_2. Les figures 5.14 et 5.15 représentent respectivement les cartes d'incertitude absolue et relative moyennées sur l'année 2007 de l'ozone et du

NO_2. Dans les deux cas, on aperçoit un maximum sur le continent, à l'est de la centrale. Cela correspond à la ville de Nantes et à sa banlieue, principale zone d'émissions de polluants dans la région. Contrairement à la carte d'incertitude d'ozone dans la région Île-de-France, l'écart type d'ozone est important près des sources d'émission. Ce dernier atteint aussi des valeurs élevées près de la côte Atlantique. Dans les chapitres 3 et 4, nous avons pu remarquer que l'incertitude d'ozone pouvait être importante près du littoral. Outre les fortes incertitudes liées aux sources d'émission près de Nantes pour le NO_2, on peut apercevoir des niveaux d'incertitude un peu plus élevés dans le voisinage de la centrale thermique de Cordemais, elle-même émettrice de dioxyde d'azote.



(a) Incertitudes O_3 ($\mu g m^{-3}$)

(b) Incertitudes NO_2 ($\mu g m^{-3}$)

FIGURE 5.14 – Cartes d'incertitude moyenne de O_3 et NO_2 en $\mu g m^{-3}$ pour le domaine Pays de la Loire issues des sous-ensembles calibrés via le diagramme de rang.

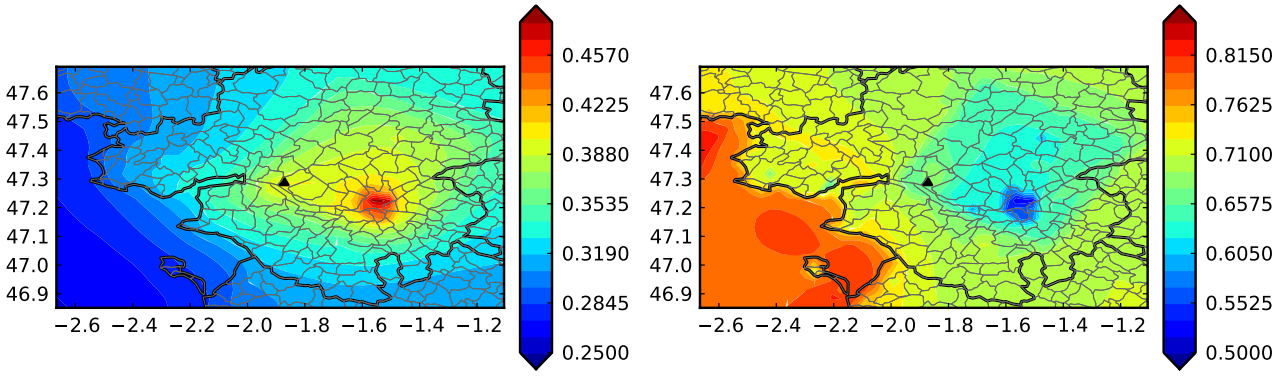
En ce qui concerne l'incertitude relative de ces deux polluants, elle est élevée près des sources d'émission pour l'ozone et faible près des sources d'émission pour le NO_2. L'écart type relatif d'ozone est globalement faible puisque le maximum reste au-dessous de 0.5. L'écart type relatif du NO_2 est quant à lui maximum près des conditions aux limites et près du littoral.

La figure 5.16 montre l'incertitude absolue et relative du SO_2. Nous rappelons que ces cartes ont été produites à titre indicatif et que la valeur des champs calculés à partir de l'ensemble complet, qui surestime fortement la concentration de SO_2, n'est a priori pas représentative de l'incertitude.

Nous n'avons rencontré aucune difficulté particulière pour calibrer et estimer l'incertitude de l'ozone. Pour le NO_2, de forts biais et une sous-dispersion non négligeable de l'ensemble restreignent quelque peu les possibilités d'avoir un nombre de membres conséquent dans le sous-ensemble calibré. Par contre, il est impossible d'estimer de manière fiable l'incertitude sur SO_2 puisque la configuration de l'ensemble complet ne permet aucune calibration.

Les cartes d'estimation d'incertitude montrent que des valeurs importantes sont présentes près des sources d'émission pour le NO_2 et le SO_2 qui sont des espèces primaires. Les concentrations d'ozone restent incertaines près des conditions aux limites et près de la côte Atlantique.

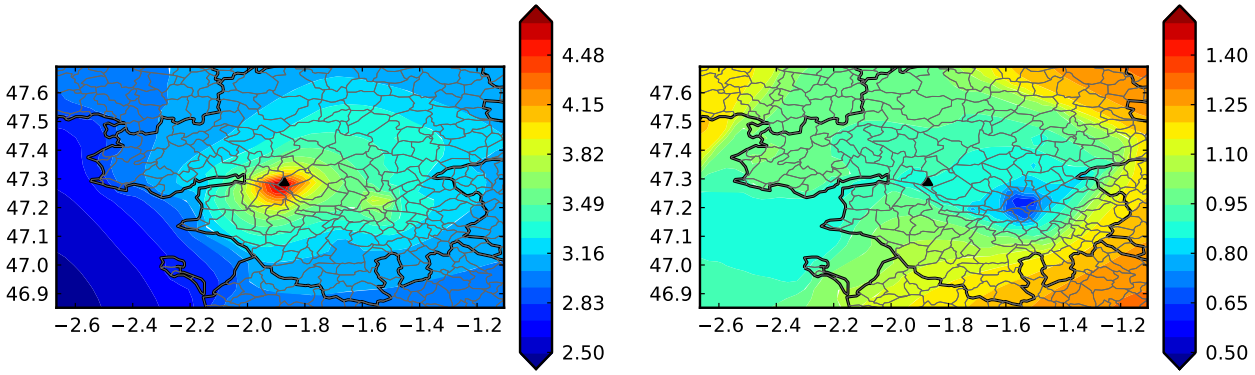
La section suivante traite de la calibration d'ensemble pour la prévision des risques — sélection d'un sous-ensemble qui fournit des probabilités « fiables » pour la prévision d'un dépassement de seuil.



(a) Incertitudes relatives O_3 (%)

(b) Incertitudes relatives NO_2 (%)

FIGURE 5.15 – Cartes d'incertitudes relatives (%) moyennes de O_3 et NO_2 pour le domaine Pays de la Loire issues des sous-ensembles calibrés via le diagramme de rang.



(a) Incertitudes absolues SO_2 ($\mu g m^{-3}$)

(b) Incertitudes relatives SO_2 (%)

FIGURE 5.16 – Cartes d'incertitude absolue ($\mu g\ m^{-3}$) et relative (%) moyennes du SO_2 dans les Pays de la Loire.

5.4.2 Prévision des risques

Les scores d'ensemble dédiés à la prévision des risques sont le score de Brier — nous lui préférerons son homologue *Brier skill score* puisqu'il est normalisé par la fréquence d'occurrence climatologique — le diagramme de fiabilité et le tableau de contingence 5.4. Ce sont ces scores qui seront utilisés pour évaluer la performance, dans le cadre de la prévision des risques, de l'ensemble complet et des sous-ensembles calibrés pour différents évènements.

Nous rappelons que les valeurs des seuils réglementaires, présentés dans le tableau 5.1, sont particulièrement élevées. Ainsi, quels que soit le domaine et le polluant étudié, nous nous intéresserons en premier lieu à des seuils plus faibles.

Contrairement à la section précédente, plusieurs calibrations ont été effectuées selon la fonction coût et l'évènement considérés. Toutes les évaluations des ensembles et des sous-ensembles calibrés se font toujours en concentration horaire à l'aide du réseau d'observations BDQA.

Porcheville

Ozone Dans le cas de l'ozone, trois évènements ont été pris en compte dans l'étude de la prévision : les seuils à 80 $\mu g m^{-3}$, 120 $\mu g m^{-3}$ et 180 $\mu g m^{-3}$. Parmi les données observées, ces évènements apparaissent respectivement pour 9.4%, 1.3% et 0.08% des cas, pour un nombre total d'observations qui s'élève à 252 540. Des calibrations sont effectuées pour les trois scores suivants : *Brier skill score*, le tableau de contingence et le diagramme de fiabilité, et ce pour chaque évènement.

Le *Brier skill score* correspond lui-même à la fonction coût utilisée lors de la sélection du sous-ensemble. La sélection d'un sous-ensemble a pour but de maximiser ce score. Des valeurs négatives de ce dernier sont possibles, cela signifie que le système de prévision, pour un évènement donné, est moins efficace que la fréquence de prévision climatologique. Dans notre cas et sauf mention contraire, la fréquence de prévision climatologique est basée sur la fréquence relative d'occurrence de l'évènement calculée à partir des observations sur toute l'année 2007.

La fonction coût associée au diagramme de fiabilité est celle utilisée dans la section 3.2.2. Elle minimise le carré de la différence entre la courbe et la diagonale.

La fonction coût associée au tableau de contingence est la maximisation du *threat score* (5.4). Il est supposé augmenter le *hit rate* (5.2) avec une contrainte sur les fausses alarmes.

Le tableau 5.5 donne les *Brier skill scores* de l'ensemble complet et des sous-ensembles calibrés (avec le *Brier skill scores*) pour les trois évènements cités. Quel que soit l'évènement, les sous-ensembles calibrés sont meilleurs que l'ensemble complet et que les fréquences climatologiques. L'ensemble complet présente des scores négatifs pour les deux plus hauts seuils. Cela signifie que l'ensemble complet n'est pas un système de prévision efficace pour ces deux évènements puisqu'il est moins bon que la fréquence climatologique. Cela montre aussi l'utilité et la nécessité de la calibration de l'ensemble.

Les tableaux de contingence et les scores associés sont présentés dans le tableau 5.6. Dans le cas du tableau de contingence, on rappelle que l'ensemble fournit une probabilité égale à un, quand au moins la moitié des membres de l'ensemble en question dépassent effectivement le seuil de concentration.

Les sous-ensembles calibrés présentent, quel que soit l'évènement, de meilleurs *threat scores* que l'ensemble complet. Cela paraît normal puisque c'est le score qui a été pris en compte dans la

Évènement	Ensemble complet	Ensemble calibré
$80\,\mu g\,m^{-3}$	0.124	0.459
$120\,\mu g\,m^{-3}$	−0.962	0.289
$180\,\mu g\,m^{-3}$	−17.82	0.021

TABLE 5.5 – *Brier skill score* de l'ensemble complet d'ozone et des sous-ensembles calibrés pour le domaine de Porcheville en fonction de trois évènements.

calibration en tant que fonction coût. Les calibrations permettent aussi d'obtenir de meilleurs *hit rates*. Par contre, le fait d'avoir un meilleur *threat score* n'empêche pas une certaine surestimation de la prévision de l'occurrence de l'évènement puisqu'il y a plus de fausses alarmes dans le cas des sous-ensembles calibrés que dans le cas de l'ensemble complet, et ceci, indépendemment de l'évènement. C'est pourquoi l'ensemble complet a des ratios de précision et de rejet correct légèrement plus élevés que les sous-ensembles calibrés. En ce qui concerne le seuil à $180\,\mu g\,m^{-3}$, seulement 20 occurrences sont observées. L'ensemble complet ne prévoit aucune occurrence — il y a donc 20 échecs — tandis que le sous-ensemble calibré en prévoit 5 — d'où le *hit rate* à 25%. Malgré la contrainte associée aux fausses alarmes dans l'optimisation du *threat score*, le sous-ensemble calibré prévoit 21 fausses alarmes contre zéro pour l'ensemble complet. Le « prix » à payer pour correctement détecter 5 dépassements sur 20 est donc de 21 fausses alarmes.

La capacité de détection de chaque modèle est aussi calculée pour le dernier évènement. L'objectif est de déterminer si, dans ce cas, l'utilisation d'un ensemble apporte plus qu'un unique modèle. Six modèles sur la centaine de membres que comporte l'ensemble arrivent à prévoir les vingts occurrences de l'évènement comparés à l'ensemble complet qui n'en prévoit aucune et à l'ensemble calibré qui en prévoit 5. Ceci s'explique par le fait que ces modèles surestiment fortement la concentration d'ozone. En effet, les moyennes de ces simulation dépassent $100\,\mu g\,m^{-3}$ sachant que la moyenne d'ozone observée est à $\sim 44\,\mu g\,m^{-3}$. Cependant, même si ces modèles prévoient toutes les occurrences de l'évènement, le nombre de fausses alarmes est très important : entre 20000 et 40000 selon les modèles — contre 21 fausses alarmes pour le sous-ensemble calibré. Ces modèles ont donc un *threat score* très faible. Le modèle qui a le meilleur *threat score* (de 0.05 seulement) prévoit l'occurrence de l'évènement seulement deux fois, avec seize fausses alarmes. L'utilisation de la calibration d'ensemble pour ces scores de détection d'occurrence d'évènement semble donc justifiée puisque ni l'ensemble complet, ni le meilleur modèle n'égalent le sous-ensemble calibré en terme de *threat score*. On constate cependant qu'il y a un compromis entre la capacité d'un système à prévoir l'occurrence de l'évènement et le nombre de fausses alarmes, et que le nombre de fausses alertes est encore trop élevé.

Les diagrammes de fiabilité pour deux évènements sont présentés sur la figure 5.17. Pour le seuil à $80\,\mu g\,m^{-3}$, le diagramme de fiabilité issu de la calibration suit parfaitement la diagonale. Le diagramme de fiabilité de l'ensemble complet est sous la diagonale pour des valeurs de probabilités prévues au-dessous de ~ 0.5 et passe au-dessus de la diagonale pour des probabilités prévues supérieures à 0.5. Les probabilités produites par l'ensemble complet semblent mal distribuées. Il y a un trop grand nombre de faibles probabilités produites par le système comparé à la fréquence relative d'occurrence de l'évènement. Au contraire, les probabilités produites qui ont des valeurs proches de 1 ne sont pas assez nombreuses.

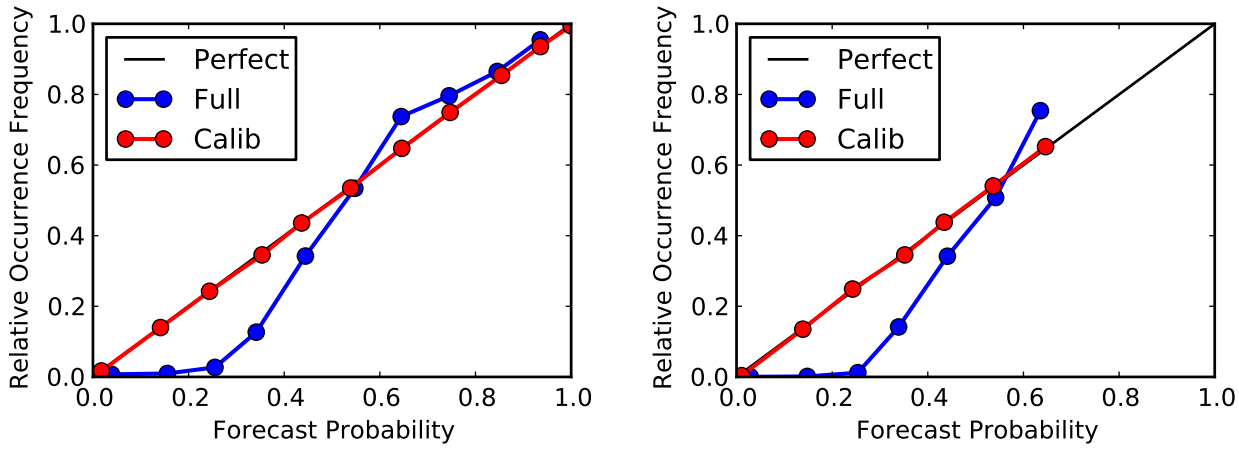
Pour le seuil à $120\,\mu g\,m^{-3}$, le diagramme de fiabilité calibré suit parfaitement la diagonale jusqu'à la probabilité produite égale à $\sim 64\%$. Au-delà, l'ensemble complet et le sous-ensemble calibré ne produisent pas de probabilités supérieures à cette valeur. Cela signifie qu'à aucun moment, plus de 65% des membres vont effectivement dépasser le seuil. Le même problème apparaît

seuil – $\mu g m^{-3}$ / ensemble	**hit**	**miss**	**correct neg.**	**false alarm**
80 / complet	12442	11180	223554	5364
80 / calibré	15452	8170	221394	7524
120 / complet	554	2800	248785	401
120 / calibré	1844	1510	247117	2069
180 / complet	0	20	252520	0
180 / calibré	5	15	252499	21

seuil – $\mu g m^{-3}$ / ensemble	**accuracy**	**correct. neg. ratio**	**hit rate**	**threat score**
80 / complet	0.934	0.977	0.527	0.429
80 / calibré	0.938	0.967	0.654	0.496
120 / complet	0.987	0.998	0.165	0.148
120 / calibré	0.986	0.992	0.550	0.340
180 / complet	0.999	1.000	0.000	0.000
180 / calibré	0.999	0.999	0.250	0.122

TABLE 5.6 – Tableaux de contingence et les ratios associés de trois évènements de dépassement de seuil de concentration d'ozone pour le domaine de Porcheville.

pour le dernier seuil à 180 $\mu g m^{-3}$: jamais plus de 10% des modèles de l'ensemble complet ou du sous-ensemble calibré vont dépasser en même temps le seuil.



(a) $O_3 \geq 80\ \mu g m^{-3}$ (b) $O_3 \geq 120\ \mu g m^{-3}$

FIGURE 5.17 – Diagrammes de fiabilité de l'ensemble complet et du sous-ensemble calibré d'ozone de la région de Porcheville pour les évènements $[O_3] \geq 80\ \mu g m^{-3}$ et $\geq 120\ \mu g m^{-3}$.

On s'intéresse maintenant aux performances des sous-ensembles calibrés sur des scores qui n'ont pas servi à leur calibration. Les sous-ensembles calibrés via le score de Brier normalisé et le diagramme de rang n'ont pas de bons résultats pour le tableau de contingence. L'ensemble complet a globalement de meilleurs scores — *hit rate* et *threat score* — que ces sous-ensembles. L'inverse est aussi vrai. Un sous-ensemble calibré avec un « bon » *threat score* pour un évènement donné a un score de Brier et un diagramme de fiabilité très proche de l'ensemble complet. Par contre, les sous-ensembles qui ont été calibrés pour optimiser le *Brier skill score* ou le diagramme

de fiabilité ont des scores tout à fait acceptables pour l'un et l'autre. À titre d'exemple, le sous-ensemble, calibré pour optimiser le diagramme de fiabilité avec l'évènement $[O_3] \geq 80\,\mu g\,m^{-3}$, a un score de Brier normalisé égal à 0.41, assez proche du score de l'ensemble calibré. Le même constat pour l'opération inverse peut être fait. La figure 5.18 représente le diagramme de fiabilité du sous-ensemble qui a été calibré afin d'optimiser le score de Brier pour le même évènement. Le diagramme est assez proche de la diagonale tout en étant meilleur que le diagramme de l'ensemble complet.

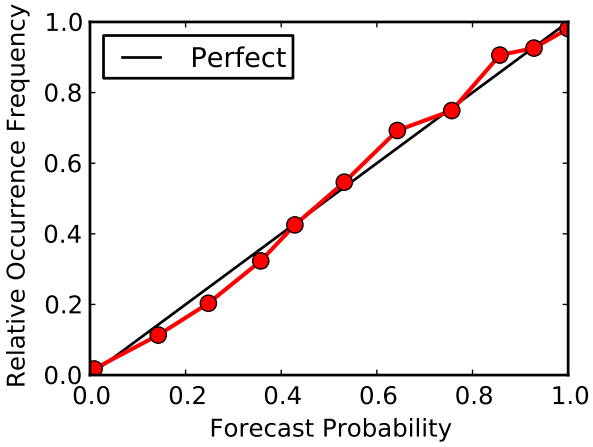


FIGURE 5.18 – Diagramme de fiabilité d'un sous-ensemble d'ozone de la région de Porcheville qui a été calibré pour optimiser le score de Brier normalisé pour l'évènement $[O_3] \geq 80\,\mu g\,m^{-3}$.

Le chapitre 3.2 décrit la décomposition du score de Brier en trois termes : fiabilité, résolution et incertitudes liées aux données d'observations — équation (3.8). Dans la majorité des calibrations effectuées sur l'ensemble de simulations photochimiques à l'échelle européenne pour l'année 2001, la calibration du score de Brier a amélioré de façon plus significative le terme « fiabilité » que l'on retrouve dans le diagramme de fiabilité. Ce n'est donc pas étonnant qu'un sous-ensemble calibré pour optimiser le diagramme de fiabilité ait un bon score de Brier, et vice-versa.

Dioxyde d'azote On considère les mêmes données que dans la section 5.4.1, avec les observations au-dessus de $25\,\mu g\,m^{-3}$. Les évènements pris en compte sont les seuils à $60\,\mu g\,m^{-3}$, $100\,\mu g\,m^{-3}$ et $200\,\mu g\,m^{-3}$ aux fréquences relatives d'occurrence observée suivantes : 32%, 8.7% et 0.2%.

Le tableau 5.7 donne les scores de Brier normalisés de l'ensemble complet et des sous-ensembles calibrés pour les trois évènements cités ci-dessus. L'ensemble complet est moins performant que la fréquence de prévision climatologique pour les deux premiers évènements. Les sous-ensembles calibrés présentent des scores positifs mais qui restent assez proches de zéro.

Les résultats de détection de dépassement de seuil et des ratios associés sont donnés dans le tableau 5.8. Les sous-ensembles calibrés des deux premiers seuils augmentent significativement le taux de réussite. Entre l'ensemble complet et ces sous-ensembles calibrés, les taux passent effectivement de 53% à 70% pour le seuil à $60\,\mu g\,m^{-3}$ et de 5% à 36% pour le seuil à $100\,\mu g\,m^{-3}$. Comme précédemment pour l'ozone, cela implique un plus grand nombre de fausses alarmes. Par exemple, pour le deuxième seuil, le nombre de fausses alarmes passe de 1558 à plus de 21200. Malheureusement, même si la calibration pour le seuil à $100\,\mu g\,m^{-3}$ présente de meilleures résul-

Évènement	**Ensemble complet**	**Ensemble calibré**
$60\,\mu g m^{-3}$	-0.005	0.1
$100\,\mu g m^{-3}$	-0.007	0.06
$200\,\mu g m^{-3}$	0.018	0.024

TABLE 5.7 – *Brier skill score* de l'ensemble complet de dioxyde d'azote et des sous-ensembles calibrés pour le domaine de Porcheville en fonction de trois évènements.

tats que l'ensemble complet, elle n'arrive pas à sélectionner un sous-ensemble capable de prévoir plus de 50% des occurrences de l'évènement. Pour le dernier seuil, dont le nombre d'occurrence est égal à 415, l'ensemble complet et le sous-ensemble calibré ont les mêmes résultats : incapacité à prévoir l'occurrence de l'évènement.

Sur la totalité des membres de l'ensemble, 21 membres dépassent au moins une fois le seuil à $200\,\mu g m^{-3}$. Le modèle qui prévoit le plus de fois l'évènement a un taux de réussite de 38%, un nombre de fausses alarmes égal à 3205 et un *threat score* à 0.04. Ce dernier n'a pas le meilleur *threat score* puisqu'il compte un nombre assez important de fausses alarmes. Le modèle qui a le meilleur *threat score*, égal à 0.07, présente un taux de succès égal à 30% pour 1463 fausses alarmes. La calibration n'arrive donc pas à sélectionner les quelques modèles qui arrivent à détecter l'occurrence de l'évènement. Néanmoins, les *threat scores* de ces derniers ne sont que très légèrement supérieurs à zéro.

seuil – $\mu g m^{-3}$ **/ ensemble**	**hit**	**miss**	**correct neg.**	**false alarm**
60 / complet	35028	30674	103497	34864
60 / calibré	46074	19628	83728	54633
100 / complet	899	17058	184548	1558
100 / calibré	6518	11439	164875	21231
200 / complet	0	415	203648	0
200 / calibré	0	415	203648	0

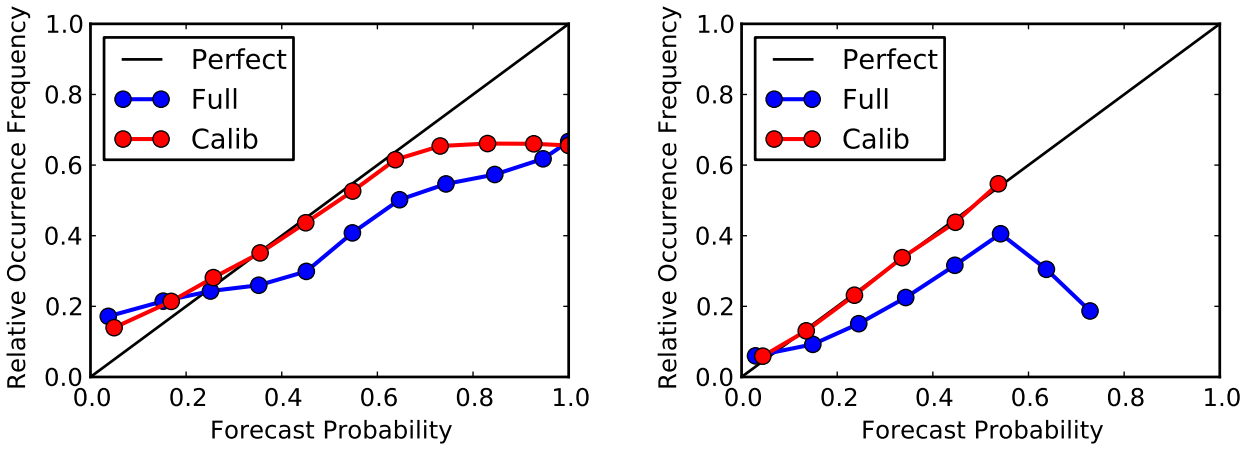
seuil – $\mu g m^{-3}$ **/ ensemble**	**accuracy**	**correct. neg. ratio**	**hit rate**	**threat score**
60 / complet	0.679	0.748	0.533	0.348
60 / calibré	0.636	0.605	0.701	0.383
100 / complet	0.909	0.992	0.050	0.046
100 / calibré	0.840	0.886	0.363	0.166
200 / complet	0.998	1.000	0.000	0.000
200 / calibré	0.998	1.000	0.000	0.000

TABLE 5.8 – Tableaux de contingence et des ratios associés de trois évènements de dépassement de seuil de concentration de NO_2 pour le domaine de Porcheville.

Deux diagrammes de fiabilité calibrés pour le NO_2 sont présentés sur la figure 5.19. Il est plus difficile d'obtenir des diagrammes de fiabilité aussi bons que dans le cas de O_3. Pour l'évènement $[NO_2] \geq 60\,\mu g m^{-3}$, le diagramme de l'ensemble complet est globalement sous la diagonale. Cela signifie une surestimation de l'occurrence de l'évènement. Il en est de même pour le diagramme calibré qui présente le même comportement au-delà de valeur de probabilité produite égale à 0.6. Dans ces cas, la plupart des modèles sont fréquemment au-dessus du seuil. Pour le seuil à $100\,\mu g m^{-3}$, le diagramme calibré a le même comportement que pour le diagramme de fiabilité d'ozone, c'est-à-dire que le sous-ensemble produit des probabilités faibles — au-dessous

de 0.5 environ — mais fiables. Au contraire, le diagramme de fiabilité de l'ensemble complet reste globalement bien au-dessous de la diagonale.

Pour les mêmes raisons que précédemment, aux vues de la fréquence d'occurrence extrêmement faible de l'évènement 200 $\mu g m^{-3}$, les diagrammes de fiabilité n'ont pas été tracés. Que ce soit pour l'ensemble complet ou le sous-ensemble calibré, il n'y a jamais plus de 10% des modèles qui dépassent effectivement le seuil à un moment donné.



(a) $NO_2 \geq 60\,\mu g m^{-3}$ (b) $NO_2 \geq 100\,\mu g m^{-3}$

FIGURE 5.19 – Diagrammes de fiabilité de l'ensemble complet et du sous-ensemble calibré du NO_2 de la région de Porcheville pour les évènements $[NO_2] \geq 60\,\mu g m^{-3}$ et $\geq 100\,\mu g m^{-3}$.

Dioxyde de soufre Malgré la réelle difficulté à calibrer le diagramme de rang de l'ensemble de SO_2, on étudie la capacité de l'ensemble ou d'un sous-ensemble calibré à prévoir un dépassement de seuil.

Plusieurs évènements sont étudiés : dépassements de seuil à 20 $\mu g m^{-3}$, 60 $\mu g m^{-3}$, 100 $\mu g m^{-3}$ et 300 $\mu g m^{-3}$. Les fréquences relatives d'occurrence de ces derniers sont respectivement 2.78%, 0.11%, 0.021% et 4.10^{-4}%. Même pour un seuil de concentration à 20 $\mu g m^{-3}$, qui ne semble pas très élevé, moins de 3% du total des observations sont au-dessus de cette valeur. Le nombre de dépassements du dernier seuil — seuil d'information à 300 $\mu g m^{-3}$ — s'élève à 3 pour plus de 70 000 observations.

Nous n'indiquons pas les *Brier skill score* ni pour l'ensemble complet, ni pour les sous-ensembles calibrés puisqu'ils sont tous négatifs, quel que soit l'évènement. Autrement dit, la fréquence de prévision climatologique donne un meilleur score de Brier que n'importe lequel des ensembles, calibrés ou non.

En ce qui concerne le tableau de contingence, les résultats ne sont guère meilleurs. À titre d'exemple, le seuil à 100 $\mu g m^{-3}$ qui n'apparaît que 15 fois n'est pas prévu par l'ensemble complet et n'est prévu qu'une seule fois par le sous-ensemble calibré. Le seuil d'information est quant à lui ni prévu par l'ensemble complet, ni par le sous-ensemble calibré.

Les calibrations sur les diagrammes de fiabilité ne permettent pas non plus d'obtenir un système de prévisions probabilistes dit « fiable ». La figure 5.20 montre les diagrammes de fiabilité

de l'ensemble complet et du sous-ensemble calibré pour l'évènement $[SO_2] \geq 20\,\mu g\,m^{-3}$ dont le seuil est le plus faible parmi ceux mentionnés plus haut. Les courbes sont très nettement au-dessous de la diagonale. Cela signifie que l'ensemble produit des probabilités d'occurrence de l'évènement bien plus élevées que les occurrences observées. En d'autres termes, les probabilités produites par le système de prévision sont globalement surestimées. Le sous-ensemble calibré ne produit malheureusement pas de meilleurs résultats. Le même constat est à apporter pour les autres évènements.

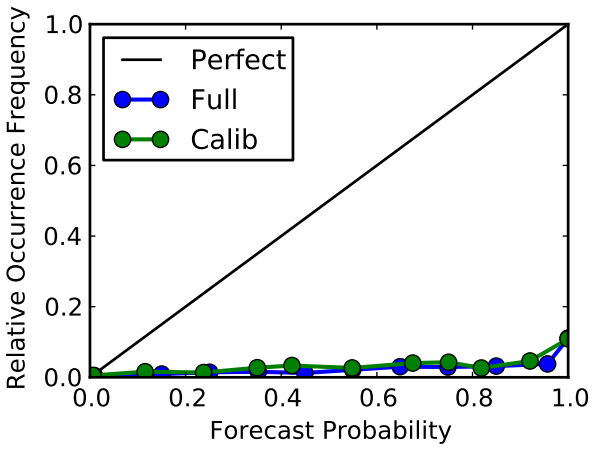


FIGURE 5.20 – Diagrammes de fiabilité de l'ensemble complet et du sous-ensemble calibré du SO_2 de la région de Porcheville pour l'évènement $[SO_2] \geq 20\,\mu g\,m^{-3}$.

L'ensemble de simulations de SO_2, qu'il soit calibré ou non, ne semble pas apporter de solutions satisfaisantes à la prévision des risques.

La calibration des différents scores associés à la prévision des risques dans la région Île-de-France pour l'ensemble de simulations d'ozone donne des résultats assez satisfaisants. Il est plus difficile d'obtenir des résultats équivalents pour l'ensemble de NO_2. En effet, les diagrammes de fiabilité calibrés ne sont pas parfaits. De plus, la calibration du *threat score* pour l'évènement dont le seuil est le plus élevé, c'est-à-dire $200\,\mu g\,m^{-3}$, ne donne pas de meilleurs résultats que l'ensemble complet ou que le meilleur modèle. Dans ce cas-là, la calibration semble incapable de sélectionner un sous-ensemble dont au moins la moitié des modèles dépassent effectivement le seuil, à une date et à une station d'observation donnée. En ce qui concerne l'ensemble de SO_2, les modèles surestiment globalement les observations et la calibration n'arrive pas à rendre fiable l'ensemble.

La sous-section qui suit traite de la calibration des mêmes scores d'ensemble appliqués à l'ensemble de simulations des trois polluants dans la région des Pays de la Loire.

Cordemais

Ozone Il est nécessaire de noter que les données d'observation d'ozone dans la région des Pays de la Loire n'atteignent jamais le seuil d'information fixé à $180\,\mu g\,m^{-3}$. C'est pourquoi les valeurs de seuil sont différentes de celles choisies pour la région Île-de-France. Nous avons donc les seuils de concentration d'ozone suivants : $80\,\mu g\,m^{-3}$, $120\,\mu g\,m^{-3}$ et $140\,\mu g\,m^{-3}$. Ces trois évènements ont été observés respectivement 16.5%, 0.75% et 0.024% sur un total de 33 609 données. Le tableau 5.9 donne les *Brier Skill scores* pour ces évènements. Pour les deux plus hauts seuils, le score de Brier de la fréquence climatologique est meilleure que l'ensemble complet puisque ce dernier a des *Brier skill scores* négatifs. La calibration, dans tous les cas, améliore les scores en

question.

Évènement	Ensemble complet	Ensemble calibré
$80\,\mu g\,m^{-3}$	0.04	0.34
$120\,\mu g\,m^{-3}$	−5.45	0.24
$140\,\mu g\,m^{-3}$	−107.8	0.12

TABLE 5.9 – *Brier skill score* de l'ensemble complet d'ozone et des sous-ensembles calibrés pour le domaine de Cordemais en fonction de trois évènements.

Les scores de détection des trois évènements cités sont présentés dans le tableau 5.10. Le sous-ensemble calibré pour le seuil le plus élevé étudié ($140\,\mu g\,m^{-3}$) présente un *threat score* égal à 75% sur un total de 8 occurrences. Dans la plupart des cas, même si les sous-ensembles calibrés arrivent à mieux détecter les évènements que l'ensemble complet, ces sous-ensembles donnent de plus nombreuses fausses alarmes.

Ces résultats sont comparés avec les scores de détection des modèles pris de manière individuelle. Le tiers des modèles arrivent à prévoir les 8 occurrences de l'évènement dont le seuil est le plus élevé. Le nombre de fausses alarmes de ces modèles est très disparate puisqu'il se situe entre une centaine et plus de 20000 fausses alarmes. Ces modèles surestiment globalement la concentration d'ozone. Aucun des modèles n'arrive à avoir un *threat score* aussi élevé que celui donné par le sous-ensemble calibré. Le *threat score* du meilleur modèle reste malgré tout assez proche de celui donné par le sous-ensemble calibré puisqu'il est égal à 0.14 contre 0.17 pour le sous-ensemble calibré. Le nombre de succès est le même dans les deux cas. Seul le nombre de fausses alarmes est plus important pour le meilleur modèle.

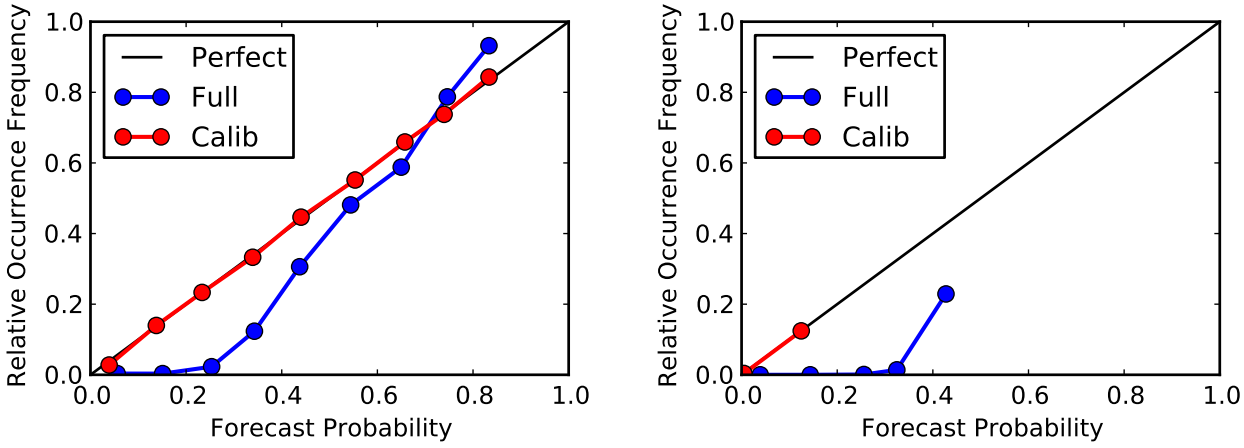
seuil – $\mu g\,m^{-3}$ / ensemble	hit	miss	correct neg.	false alarm
80 / complet	2706	2832	26133	1938
80 / calibré	3697	1841	25297	2774
120 / complet	37	217	33333	22
120 / calibré	129	125	33238	117
140 / complet	0	8	33601	0
140 / calibré	6	2	33574	27

seuil – $\mu g\,m^{-3}$ / ensemble	accuracy	correct. neg. ratio	hit rate	threat score
80 / complet	0.858	0.931	0.489	0.362
80 / calibré	0.863	0.901	0.668	0.445
120 / complet	0.993	0.999	0.146	0.134
120 / calibré	0.993	0.996	0.508	0.348
140 / complet	0.999	1.000	0.000	0.000
140 / calibré	0.999	0.999	0.750	0.171

TABLE 5.10 – Tableaux de contingence et les ratios associés de trois évènements de dépassement de seuil de concentration d'ozone pour le domaine de Cordemais.

Deux diagrammes de fiabilité pour les deux évènements que sont $[O_3] \geq 80\,\mu g\,m^{-3}$ et $[O_3] \geq 120\,\mu g\,m^{-3}$ sont présentés sur la figure 5.21. Le sous-ensemble calibré pour le premier évènement, qui est celui le plus fréquent parmi les trois cités, présente un très bon diagramme de fiabilité. Par contre, pour l'évènement $[O_3] \geq 120\,\mu g\,m^{-3}$ — et a fortiori pour le seuil à $140\,\mu g\,m^{-3}$ beaucoup

moins fréquent — les diagrammes sont de moins bonnes qualité. L'ensemble complet n'est pas fiable pour le peu de probabilités produites, toutes inférieures à 0.45. Au-delà, l'ensemble est incapable d'avoir plus de la moitié de ces membres au-dessus du seuil. En ce qui concerne la calibration, elle semble avoir sélectionné un sous-ensemble qui ne produit que de très faibles probabilités, deux au total et fiables malgré tout. Quand la plupart des modèles n'arrivent pas à atteindre le seuil, la calibration tend à sélectionner un sous-ensemble qui fournit des probabilités peu élevées.



(a) $O_3 \geq 80\,\mu g\,m^{-3}$ (b) $O_3 \geq 120\,\mu g\,m^{-3}$

FIGURE 5.21 – Diagrammes de fiabilité de l'ensemble complet et du sous-ensemble calibré de l'ozone de la région de Cordemais pour les évènements $[O_3] \geq 80\,\mu g\,m^{-3}$ et $\geq 120\,\mu g\,m^{-3}$.

Dioxyde d'azote On peut rappeler que dans le cas du NO_2 dans la région des Pays de la Loire, l'ensemble a été débiaisé, à cause de biais négatif de la moyenne de l'ensemble des simulations. De plus, aux vues des faibles concentrations de NO_2 observées, seules deux valeurs de seuil ont été retenues : $50\,\mu g\,m^{-3}$ et $80\,\mu g\,m^{-3}$. Ces évènements ont une fréquence de 4.2% et 0.63% sur un total de 85 790 données.

Les scores de Brier normalisés de l'ensemble complet et des sous-ensembles calibrés sont décrits dans le tableau 5.11. Ils sont tous positifs même s'ils sont proches de zéro.

Évènement	**Ensemble complet**	**Ensemble calibré**
$50\,\mu g\,m^{-3}$	0.022	0.062
$80\,\mu g\,m^{-3}$	0.002	0.012

TABLE 5.11 – *Brier skill score* de l'ensemble complet de NO_2 et des sous-ensembles calibrés pour le domaine de Cordemais en fonction de deux évènements.

Le tableau de contingence et les scores dédiés à la détection des deux seuils de concentration de NO_2 sont présentés dans le tableau 5.12. Malgré la calibration, le taux de succès n'excède pas les 32% pour le premier seuil avec un nombre de fausses alarmes non négligeable. Le second évènement n'est jamais détecté par l'ensemble complet, tandis que le sous-ensemble calibré ne compte que 4 bonnes prévisions sur les 542 occurrences observées.

Comme précédemment, les scores de détection des modèles seuls ont été calculés. Pour le premier évènement, 5 modèles ont un meilleur taux de réussite que le sous-ensemble calibré — jusqu'à 42% pour le meilleur. Encore une fois, ces modèles surestiment les concentrations de NO_2 et donnent un nombre important de fausses alarmes qui est aux alentours de 7000. Leur *threat score* est donc plus faible que celui donné par le sous-ensemble calibré, mais plus élevé que celui donné par l'ensemble complet. Le meilleur *threat score*, égal à 0.14, ne dépasse cependant pas celui calculé à partir des prévisions produites par le sous-ensemble calibré.

En ce qui concerne le deuxième seuil, ni l'ensemble complet, ni le sous-ensemble calibré n'arrivent à produire des prévisions dont les *hit rates* et *threat scores* soient meilleurs que le meilleur modèle. Le meilleur *hit rate* parmi ces modèles s'élève à 14% pour 2433 fausses alarmes. Le meilleur *threat score* donne quant à lui à 0.06, d'une valeur faible mais supérieure à celui du sous-ensemble calibré.

Comme dans le cas de l'ensemble de NO_2 dans la région Île-de-France pour l'évènement dont le seuil est le plus élevé, la calibration sélectionne un sous-ensemble dont les scores de détection sont plus faibles que ceux du meilleur modèle.

seuil – $\mu g\,m^{-3}$ **/ ensemble**	**hit**	**miss**	**correct neg.**	**false alarm**
50 / complet	175	3439	81931	245
50 / calibré	1153	2461	78227	3949
80 / complet	0	542	85247	1
80 / calibré	4	538	85247	1

seuil – $\mu g\,m^{-3}$ **/ ensemble**	**accuracy**	**correct. neg. ratio**	**hit rate**	**threat score**
50 / complet	0.957	0.997	0.048	0.045
50 / calibré	0.925	0.952	0.319	0.152
80 / complet	0.994	1.000	0.000	0.000
80 / calibré	0.994	1.000	0.007	0.007

TABLE 5.12 – Tableaux de contingence et les ratios associés de deux évènements de dépassement de seuil de concentration de NO_2 pour le domaine de Cordemais.

Le diagramme de fiabilité pour l'évènement $[NO_2] \geq 50\,\mu g\,m^{-3}$ est sur la figure 5.22. Les ensembles produisent des probabilités assez faibles globalement — jusqu'à 0.4–0.5. Même pour un seuil peu élevé, moins de 50% des modèles ne dépassent jamais le seuil en même temps. Le constat est d'autant plus flagrant pour les évènements dont le seuil est plus élevé. Il reste difficile aux ensembles, complet et calibré, de produire un certain nombre de probabilités élevées quand la majeur partie des modèles se situent fréquemment sous la valeur du seuil.

Dioxyde de soufre Tout comme pour la région Île-de-France, la calibration de l'ensemble de simulations de SO_2 est problématique dans la limite où les modèles donnent de faibles performances comparés aux observations — le tableau 5.3 fait mention de corrélations négatives. Le constat est le même, il est impossible d'obtenir des scores d'ensemble suffisamment exploitables, même avec la calibration.

Conclusion

Pour l'ozone, l'ensemble complet présente de meilleures performances que pour les deux autres polluants, et la calibration est efficace. Les sous-ensembles calibrés pour un critère particulier peuvent être performants pour un autre score d'ensemble. Par exemple, la grande part

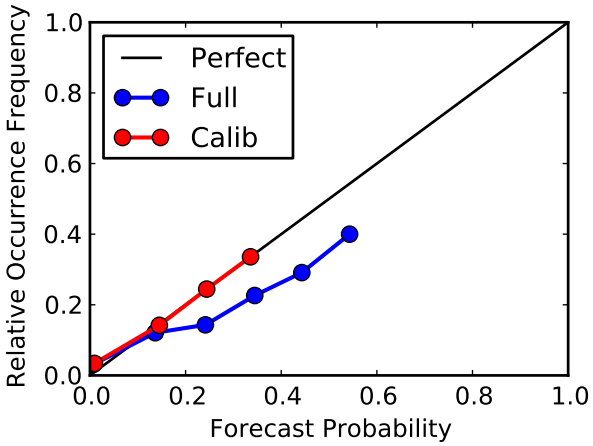


FIGURE 5.22 – Diagramme de fiabilité de l'ensemble complet et du sous-ensemble calibré du NO_2 de la région de Cordemais pour les évènements $[NO_2] \geq 50\ \mu g\,m^{-3}$.

de fiabilité optimisée lors de la calibration du score de Brier permet aux sous-ensembles calibrés avec ce score d'avoir un diagramme de fiabilité correct.

Les évènements dont le seuil de concentration est élevé sont plus difficiles à prévoir de manière efficace. Ces derniers présentent une trop faible occurrence. La calibration de l'ensemble de NO_2 pour le seuil élevé de $200\ \mu g\,m^{-3}$ n'arrive même pas à sélectionner un sous-ensemble meilleur que l'ensemble complet.

Concernant le tableau de contingence, l'optimisation du *threat score*, même s'il est globalement plus élevé que celui de l'ensemble complet, n'empêche pas de compter un nombre parfois important de fausses alarmes. Dans le cas du NO_2, il arrive que le meilleur modèle donne un meilleur *threat score* que le sous-ensemble calibré.

Le cas du SO_2 est assez particulier. Comme dans le cas de l'estimation de l'incertitude, les calibrations dédiées à l'estimation des risques sont incapables de sélectionner des sous-ensembles fiables quel que soit l'évènement. Il est à noter que la calibration ne fait que sélectionner un sous-ensemble parmi l'ensemble complet. Un ensemble dont les modèles ont des performances individuelles très insatisfaisantes restera difficile à calibrer.

5.5 Étude d'impact

Dans cette partie, on tente d'estimer l'impact que peuvent avoir, en moyenne, les émissions des centrales thermiques dans les régions concernées. Dans le cadre de la génération des ensembles photochimiques, toutes les simulations ont été lancées deux fois dans chaque région : avec et sans les émissions de la centrale. La simulation photochimique, dit « de référence », qui sert à l'étude d'impact est, quand cela est possible, la moyenne de l'ensemble calibré dans le but d'avoir un diagramme de rang plat. À défaut, comme nous le verrons pour le cas du dioxyde de soufre où la calibration du diagramme de rang est impossible, c'est la moyenne de l'ensemble complet qui est prise pour « référence ». Malgré l'information que l'on peut tirer de ces scénarios d'émission, il est impossible de vérifier directement les résultats d'une étude d'impact. En effet, on ne peut jamais observer deux situations qui ne différeraient que par la présence ou non des émissions de la centrale. Entre une période où la centrale est active et une période où la centrale est inactive, les conditions météorologiques, les autres émissions de polluants et les polluants transportés à

grande échelle ne sont jamais identiques. On doit donc reposer sur des simulations, incertaines.

On s'intéresse alors à l'ensemble des différences entre les simulations avec ou sans les émissions des centrales. Notons que les ensembles sont constitués des mêmes membres : à l'échelle européenne ou à l'échelle régionale, les mêmes ~ 100 membres (même formulation physique, chimique et numérique ; mêmes données d'entrée) simulent le cas d'intérêt, avec et sans les émissions des centrales. Ainsi, les membres des ensembles avec ou sans les émissions des centrales sont comparables. On note $\mathcal{M}_i(e)$ une concentration calculée par le membre i de l'ensemble sans les émissions de la centrale. On note $\mathcal{M}_i(e+\delta e)$ la même concentration estimée avec les émissions de la centrale. On s'intéresse donc à l'ensemble des différences absolues $(\mathcal{M}_i(e+\delta e)-\mathcal{M}_i(e))_i$ et des différences relatives $([\mathcal{M}_i(e+\delta e)-\mathcal{M}_i(e)]/\mathcal{M}_i(e))_i$.

En plus de fournir les résultats concernant l'étude d'impact de l'ozone, du dioxyde d'azote et du dioxyde de soufre dans les régions autour des centrales thermiques de Porcheville et de Cordemais, une estimation de l'incertitude de ces résultats, via le calcul de l'écart type de l'ensemble des différences, est menée par la suite.

5.5.1 Porcheville

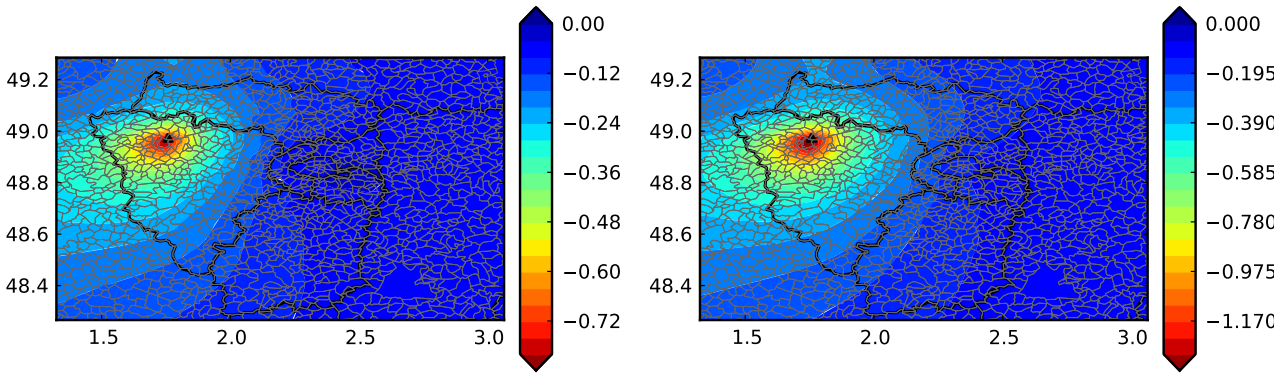
On se propose de comparer les champs de concentration en fonction des deux scénarios d'émission de la centrale thermique de Porcheville. Quand cela est possible, on décide de prendre comme élément de comparaison la moyenne du sous-ensemble sélectionné via la calibration du diagramme de rang.

Concernent les champs d'ozone, on note une légère différence entre les deux champs. En moyenne, on calcule une différence de $-0.14\,\mu g\,m^{-3}$. Le champ d'ozone, dans le cas où les émissions de la centrale de ne sont pas prises en compte, est légèrement supérieur aux champs d'ozone avec les émissions de la centrale. Cet impact est localisé près de la source d'émission comme les deux cartes de la figure 5.23 le montrent. Ces cartes correspondent à la différence, moyennée en temps, entre la moyenne d'ensemble avec et sans les émissions de la centrale ainsi que la différence relative.

Il n'est pas étonnant de constater que les émissions de la centrale diminuent les concentrations d'ozone près de la source d'émission. En effet, les émissions de NOx tendent à diminuer la production d'ozone, par la titration de l'ozone par le monoxyde d'azote (réaction R 1.3). Près de cette zone où les différences sont les plus élevées, la différence maximale n'est que de $0.73\,\mu g\,m^{-3}$. Cette différence n'excède pas les 1.2% du champ de concentration sans les émissions de la centrale (voire la carte des différences relatives).

Pour le dioxyde d'azote, l'impact est bien évidemment positif puisque les émissions de la centrale de Porcheville augmentent les concentrations du NO_2. Néanmoins, l'impact reste assez faible puisque la différence moyenne entre les deux champs est de l'ordre de $+0.12\,\mu g\,m^{-3}$. La figure 5.24 présente les différences absolue et relative en moyenne du champ de NO_2 avec et sans les émissions de la centrale. On note une augmentation de 3.6% de la concentration de NO_2 en moyenne près de la centrale. La différence de concentration entre les deux scénarios est donc assez faible comparée aux champs de NO_2 de fond — c'est-à-dire sans les émissions de la centrale.

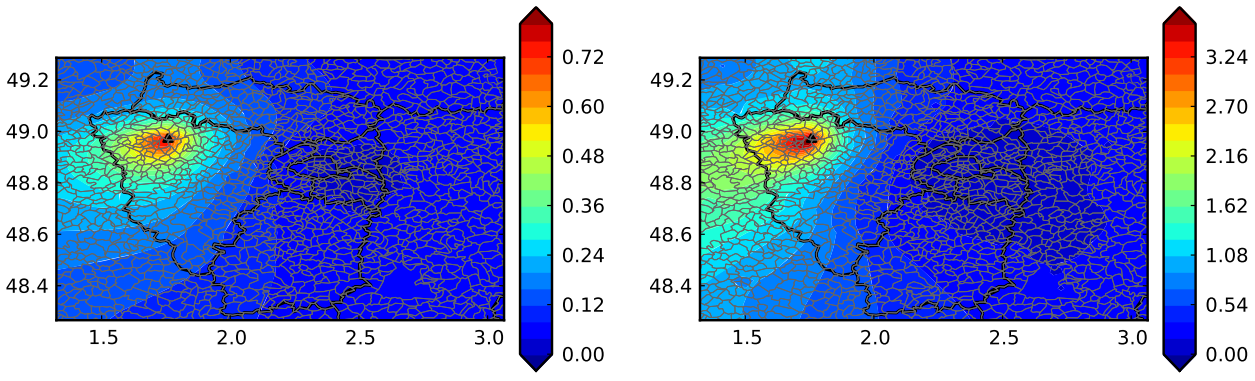
Pour le SO_2, comme indiqué précédemment, la moyenne de l'ensemble complet, et non la moyenne de l'ensemble calibré, est prise en compte dans l'étude d'impact du SO_2. Le SO_2, qui est une espèce primaire et émise par la centrale de Porcheville, voit naturellement ses concentrations augmenter lorsque la centrale est active. La différence, en moyenne, vaut $\sim +0.13\,\mu g\,m^{-3}$. La différence relative donne un écart d'environ 6.5% en moyenne aux alentours de la centrale. La figure 5.25 présente les deux champs de différences absolue et relative des champs moyens de



(a) Différence O_3 ($\mu g m^{-3}$)

(b) Différence relative O_3 (%)

FIGURE 5.23 – Impact moyen absolu ($\mu g m^{-3}$) et relatif (%) du champ d'ozone dans la région Île-de-France.

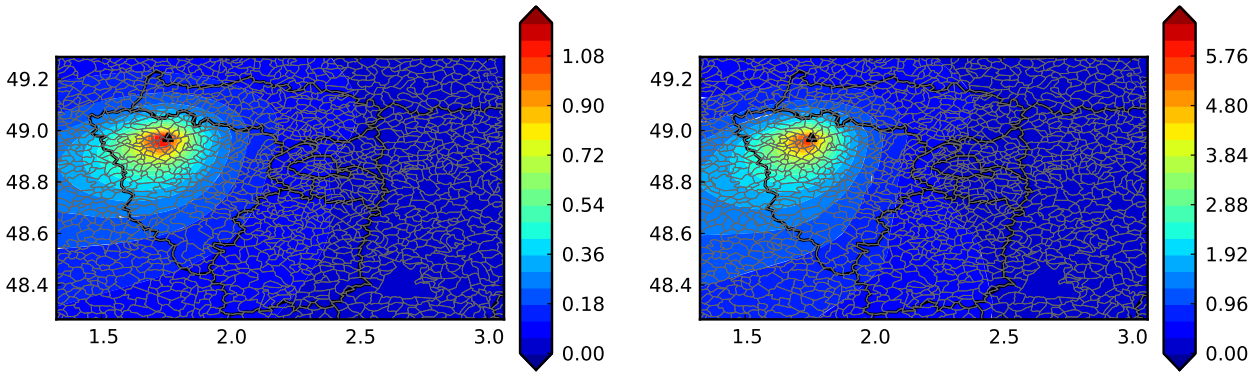


(a) Différence NO_2 ($\mu g m^{-3}$)

(b) Différence relative NO_2 (%)

FIGURE 5.24 – Impact moyen absolu ($\mu g m^{-3}$) et relatif (%) du champ de NO_2 dans la région Île-de-France.

dioxyde de soufre.



(a) Différence SO_2 ($\mu g\,m^{-3}$)

(b) Différence relative SO_2 (%)

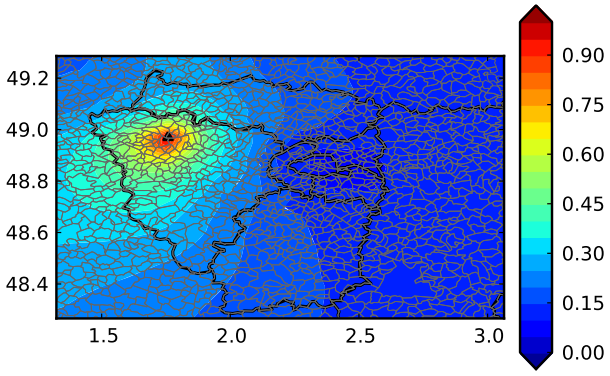
FIGURE 5.25 – Impact moyen absolu ($\mu g\,m^{-3}$) et relatif (%) du champ de SO_2 dans la région Île-de-France.

Quels que soient les polluants étudiés, les différences restent relativement faibles, même autour de la centrale. Globalement, l'impact se situe exclusivement à l'ouest de Paris. Selon ces calculs, les émissions de la centrale thermique de Porcheville ont donc en moyenne un impact faible sur les niveaux de fond en région parisienne.

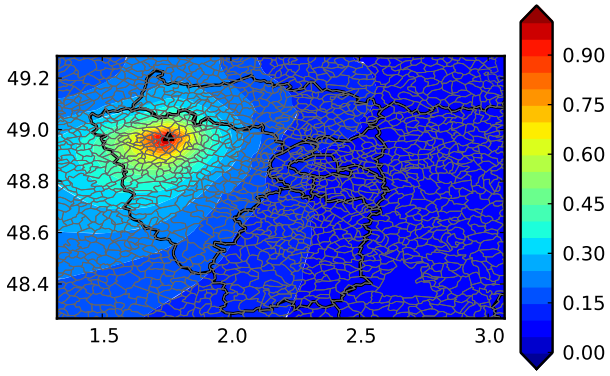
On se propose maintenant de calculer l'écart type des différences des simulations d'ozone et de dioxyde d'azote (sur l'ensemble calibré), puis l'écart type des différences des simulations de dioxyde de soufre (sur l'ensemble complet). Pour l'ozone et le dioxyde d'azote, on utilise des sous-ensembles qui sont calibrés pour être représentatifs des incertitudes sur les concentrations, pas sur l'impact de la centrale (c'est-à-dire, pas sur les différences entre simulations). Il est en effet impossible de calibrer un ensemble précisément sur les différences puisque il n'y aucune donnée d'observation sur ces différences. Comme indiqué précédemment, chaque différence est donc calculée selon $\mathcal{M}_i(e+\delta e)-\mathcal{M}_i(e)$ où i est l'indice d'un membre du sous-ensemble calibré.

La figure 5.26 présente les trois champs d'incertitude sur les différences, moyennés sur la période d'étude, pour les trois polluants. Il s'agit de l'écart type empirique de l'ensemble des différences. L'incertitude maximale, quelle que soit l'espèce étudiée, se situe bien sûr autour de la centrale thermique. L'écart type des différences pour O_3 et NO_2 s'approche des $\sim 1\,\mu g\,m^{-3}$ pour les deux polluants. L'écart type des différences du SO_2 est plus élevé et atteint $\sim 1.3\,\mu g\,m^{-3}$. Cela signifie que, pour les trois polluants, l'incertitude sur l'impact est aussi élevée que l'impact lui-même. On en déduit que l'impact maximal peut être significativement plus élevé que le montre le calcul précédent. Si on suppose que l'incertitude est correctement estimée et que les erreurs sont log-normales, les impacts maximaux pour NO_2 et SO_2 sont respectivement $\sim 3.26\,\mu g\,m^{-3}$ et $\sim 4.69\,\mu g\,m^{-3}$. Ces nombres correspondent aux percentiles 97.5 des distributions log-normales dont les espérances et écarts types ont été estimés avec la centaine de tirages pour chaque polluant.

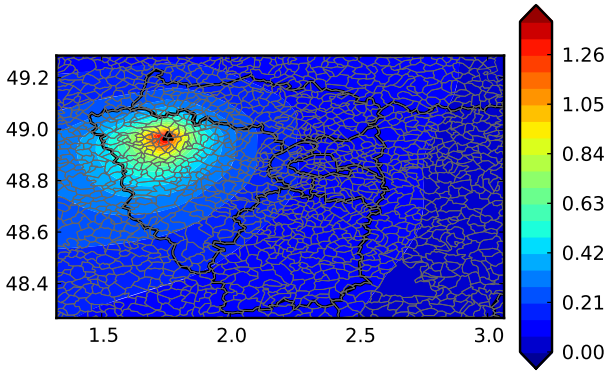
L'impact de l'ozone est très proche en valeur absolue à l'impact du NO_2, comme le montre les cartes 5.23(a) et 5.24(a). Les intervalles de confiance sont donc très proche.



(a) O_3 ($\mu g m^{-3}$)



(b) NO_2 ($\mu g m^{-3}$)



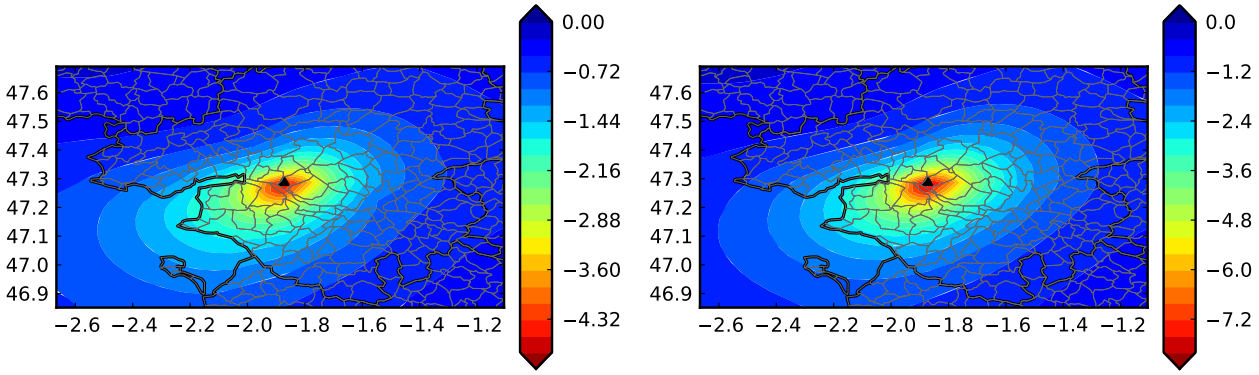
(c) SO_2 ($\mu g m^{-3}$)

FIGURE 5.26 – Estimation de l'incertitude de l'ensemble des différences pour les trois polluants dans la région de Porcheville, moyennée sur toute la période, en $\mu g m^{-3}$.

5.5.2 Cordemais

Comme dans la section précédente, ce sont les moyennes des sous-ensembles calibrés d'ozone et de NO_2 puis la moyenne de l'ensemble complet du SO_2, qui sont utilisées pour étudier l'impact des émissions de la centrale thermique de Cordemais dans la région des Pays de la Loire.

La figure 5.27 présente les différences absolue et relative des simulations d'ozone moyennées sur l'année 2007. Comme dans le cas de Porcheville, l'impact est une diminution des concentrations. La présence d'émissions de NO_2 tend à diminuer la concentration d'ozone autour de la centrale thermique d'environ $\sim 4.5\,\mu g\,m^{-3}$. Cela représente une diminution jusqu'à $\sim 7.5\%$ comparée au scénario sans les émissions de la centrale.



(a) Différence O_3 ($\mu g\,m^{-3}$)

(b) Différence relative O_3 (%)

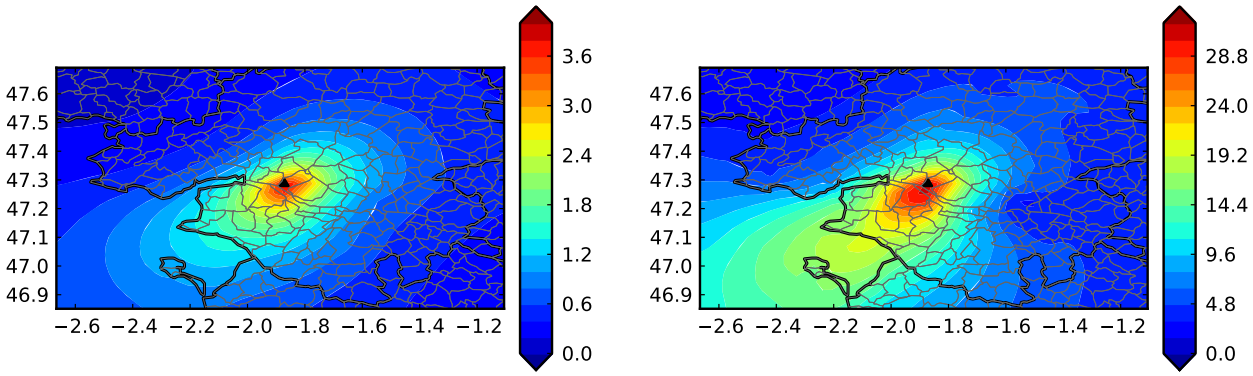
FIGURE 5.27 – Impact moyen absolu ($\mu g\,m^{-3}$) et relatif (%) du champ d'ozone dans la région des Pays de la Loire.

L'impact moyen du NO_2 avec les émissions de la centrale est de $0.73\,\mu g\,m^{-3}$. Le maximum de la différence absolue se situe près de la centrale thermique Cordemais et vaut $3.8\,\mu g\,m^{-3}$ comme le montre la figure 5.28. Cela correspond à une augmentation de $\sim 30\%$ comparée au champ de fond. En moyenne, la différence relative vaut environ 8%.

Les champs de différences absolue et relative du SO_2 ont une structure spatiale très proche de ceux du NO_2, espèces émises par la centrale thermique. Ces champs sont présentés sur la figure 5.29. La différence moyenne ne vaut que $0.2\,\mu g\,m^{-3}$ pour une valeur maximale égale à $1.9\,\mu g\,m^{-3}$ près de la source d'émission. En moyenne, cela correspond à une augmentation de $\sim 10\%$.

Les cartes d'incertitude estimées pour les trois polluants, calculées à partir l'écart type de l'ensemble des différences, sont données sur la figure 5.30. Pour l'ozone et le dioxyde d'azote, ce sont les sous-ensembles calibrés qui ont servi à l'estimation de l'incertitude de la partie 5.4.1 qui sont utilisés dans ce cas. Pour le dioxyde de soufre, c'est à partir de l'écart type de l'ensemble complet des différences que l'incertitude est estimée.

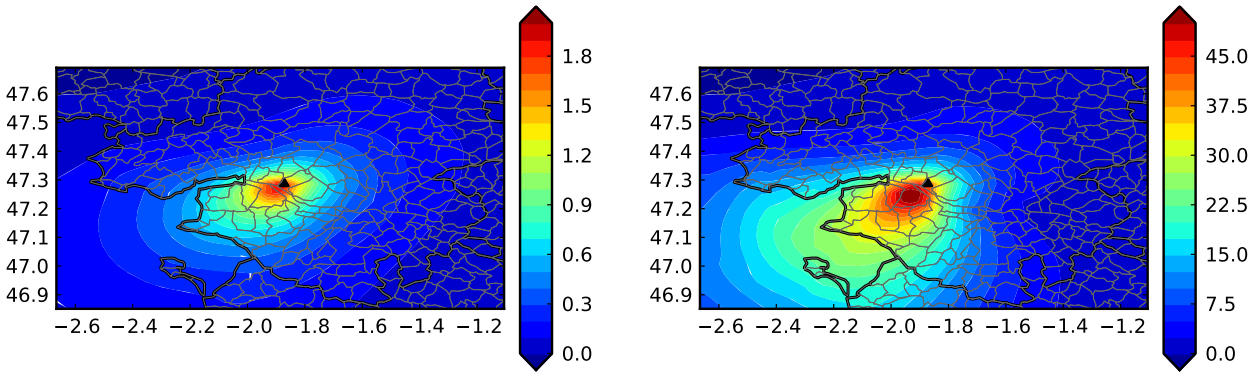
Les valeurs des écarts types sont plus élevées pour l'ozone et le NO_2 que pour le SO_2. Près de la centrale thermique, l'écart type atteint jusqu'à $5\,\mu g\,m^{-3}$ — pour une moyenne égale à



(a) Différence NO_2 ($\mu g m^{-3}$)

(b) Différence relative NO_2 (%)

FIGURE 5.28 – Impact moyen absolu ($\mu g m^{-3}$) et relatif (%) du champ de NO_2 dans la région des Pays de la Loire.



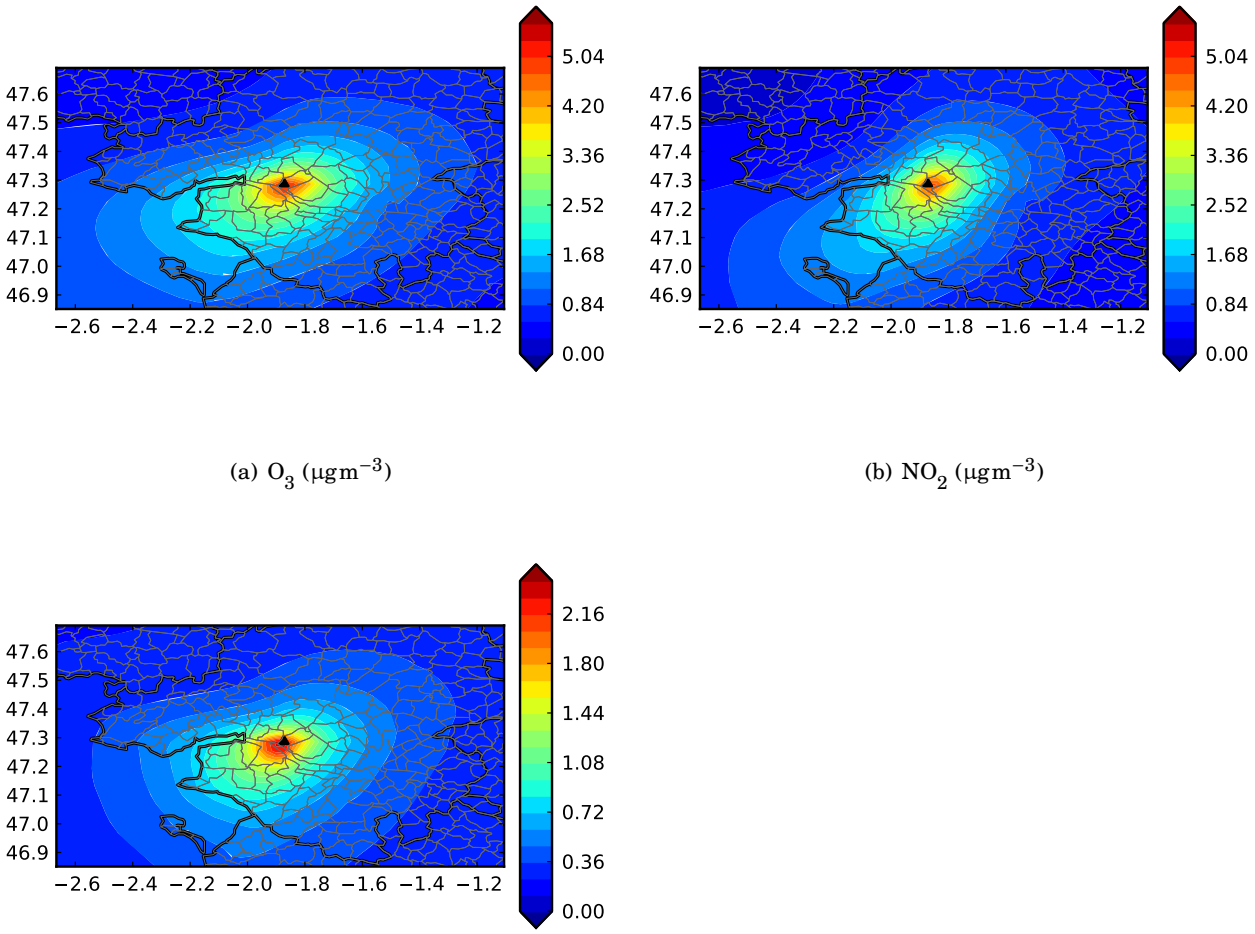
(a) Différence SO_2 ($\mu g m^{-3}$)

(b) Différence relative SO_2 (%)

FIGURE 5.29 – Impact moyen absolu ($\mu g m^{-3}$) et relatif (%) du champ de SO_2 dans la région des Pays de la Loire.

1.1 $\mu g m^{-3}$ pour l'ozone et 0.9 $\mu g m^{-3}$ pour le dioxyde d'azote. Concernant l'incertitude liée aux différences de SO_2, l'écart type atteint 2.3 $\mu g m^{-3}$ près de la centrale, avec une moyenne de 0.46 $\mu g m^{-3}$.

Les valeurs de ces incertitudes sont plus élevées dans la région des Pays de la Loire que dans la région d'Île-de-France. De plus, elles sont fortement supérieures aux valeurs moyennes des différences absolues, quel que soit le polluant étudié. De la même manière que pour le cas de Porcheville, si on suppose que l'incertitude est correctement estimée et que les erreurs sont log-normales, les impacts maximaux (percentile 97.5) pour NO_2 et SO_2 sont respectivement $\sim 16.4\ \mu g m^{-3}$ et $\sim 7.8\ \mu g m^{-3}$.



(a) O_3 ($\mu g m^{-3}$)

(b) NO_2 ($\mu g m^{-3}$)

(c) SO_2 ($\mu g m^{-3}$)

FIGURE 5.30 – Estimation de l'incertitude de l'ensemble des différences pour les trois polluants dans la région des Pays de la Loire, moyennée sur toute la période, en $\mu g m^{-3}$.

Quels que soient le domaine d'étude et la centrale étudiée, l'impact des émissions des deux centrales reste globalement assez faible. Les émissions des centrales thermiques tendent à diminuer la concentration d'ozone. Cette diminution est due aux émissions de NOx et à la titration du O_3 par le monoxyde d'azote. L'impact pour l'ozone par exemple est en moyenne de −0.72

et $-4\,\mu g\,m^{-3}$ autour de Porcheville et Cordemais respectivement. La structure spatiale des différence de champs de concentration d'ozone et de NO_2 sont assez proches — l'impact est évidemment localisé près des centrale thermiques.

L'ensemble des simulations apporte une information sur l'incertitude de l'impact. Cette incertitude est élevée, comparée à l'impact estimé ; l'incertitude est souvent supérieure à l'impact estimé. En prenant en compte une fourchette haute de l'impact (par exemple, le 95e percentile de la distribution de l'impact), on obtient des valeurs parfois non négligeables.

5.6 Robustesse spatiale

Le but de cette partie est de montrer la robustesse spatiale de la calibration pour différents scores d'ensemble. On appelle robustesse spatiale, la capacité d'un sous-ensemble, calibré avec un score d'ensemble donné, à être performant aux endroits non observés.

Dans le cadre d'une étude d'impact, une calibration robuste permet notamment de renseigner sur l'exposition des populations aux endroits où aucune station d'observation n'est disponible. Par exemple, si le diagramme de fiabilité s'avère robuste spatialement, on peut en conclure qu'*en moyenne*, le sous-ensemble calibré estime correctement les fréquences d'occurrence d'un évènement. On peut en particulier estimer la fréquence d'exposition de la population à ces dépassements de seuil.

La section 3.4 du chapitre 3 fait mention de la robustesse spatiale d'un ensemble de simulations photochimiques à l'échelle européenne et de la méthode utilisée. De la même façon, le réseau d'observation disponible est divisé de manière aléatoire en deux sous-réseaux. L'objectif est d'obtenir deux réseaux assez homogènes. Pour chaque score d'ensemble, une calibration est effectuée sur chacun des sous-réseaux et le sous-ensemble calibré est évalué sur le réseau complémentaire.

Seuls les ensembles de simulations d'ozone et de dioxyde d'azote sont étudiés dans cette partie. Le réseau d'observation utilisé est le réseau BDQA dans la région Île-de-France. La robustesse spatiale n'est malheureusement pas étudiée dans la région de Cordemais aux vues du faible nombre de stations disponibles — uniquement quatre stations d'ozone pour toute l'année 2007 par exemple.

5.6.1 Ozone

La robustesse spatiale pour l'ensemble de simulations d'ozone dans la région Île-de-France est étudiée avec un total 31 stations d'observation issues du réseau BDQA. Ce réseau est divisé en deux sous-réseaux, appelés A et B, de 15 et 16 stations respectivement. La figure 5.31 représente ces sous-réseaux A et B dans la région d'Île-de-France.

Ce sont les scores d'ensemble utilisés dans la partie 5.4 qui servent aux différentes calibrations. Dans un premier temps, c'est la robustesse spatiale de la calibration d'un diagramme de fiabilité qui est étudiée. La figure 5.32 présente deux diagrammes de fiabilité pour l'évènement $[O_3] \geq 80\,\mu g\,m^{-3}$. Chaque diagramme correspond au diagramme de fiabilité calculé avec les données d'observation d'un sous-réseau particulier. Le diagramme 5.32(a) a été produit à l'aide du sous-réseau B, tandis que le diagramme 5.32(b) a été produit à l'aide du réseau A. Sur chaque figure apparaît les courbes associées à l'ensemble complet, au sous-ensemble calibré avec le sous-réseau d'où viennent les données d'observations et le sous-ensemble calibré sur l'autre sous-réseau.

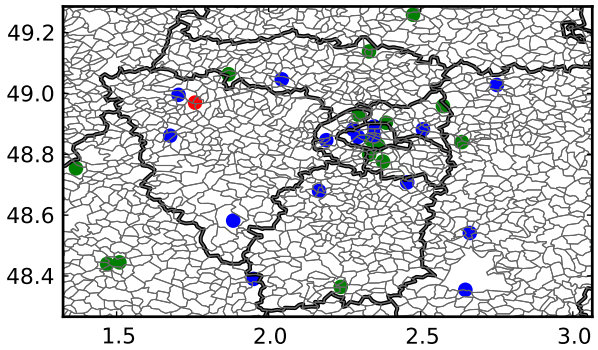
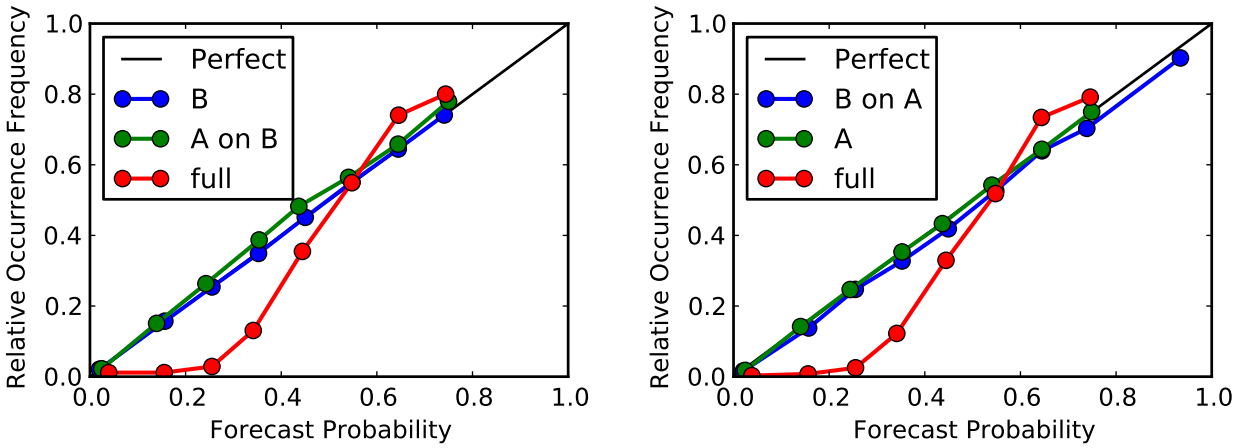


FIGURE 5.31 – Domaine Île-de-France. La centrale de Porcheville est localisée par le point rouge. Deux sous-réseaux d'un total de 31 stations BDQA pour l'ozone ont été sélectionnés de manière aléatoire : le réseau A en vert et le réseau B en bleu.

Quel que soit le sous-réseau d'observation, le diagramme du sous-ensemble appliqué au réseau qui n'a pas servi à la calibration est assez proche du diagramme calibré et reste bien meilleur que le diagramme de l'ensemble complet. La calibration d'un diagramme de fiabilité pour l'évènement considéré semble donc suffisamment robuste pour être appliquée sur des données d'observations indépendantes de ladite calibration.

Aux vues des probabilités produites très faibles de l'ensemble complet et du sous-ensemble calibré pour l'évènement $[O_3] \geq 120\,\mu g\,m^{-3}$ mentionné dans la section 5.4.2, la robustesse spatiale de la calibration du diagramme de fiabilité pour cet évènement n'a pas été réalisée.



(a) A sur B

(b) B sur A

FIGURE 5.32 – Diagrammes de fiabilité pour $[O_3] \geq 80\,\mu g\,m^{-3}$ dans la région Île-de-France. (a) correspond au diagramme de fiabilité du sous-réseau B et (b) correspond au diagramme de fiabilité du sous-réseau A.

Par contre, d'autres évènements peuvent être étudiés dans le cadre de la maximisation du

threat score dédié au tableau de contingence. L'évènement $[O_3] \geq 120\ \mu g\,m^{-3}$ est alors étudié pour la robustesse de la calibration du *threat score*. Le tableau 5.13 montre les scores du tableau de contingence de tous les ensembles, complets et calibrés, sur les deux réseaux d'observation A et B.

L'ensemble complet a un *threat score* de 15 et 14.5 sur le réseau A et B respectivement, tandis que les sous-ensembles calibrés sur ces deux réseaux atteignent 34.47 et 33. Le score de l'ensemble calibré sur la totalité du réseau d'observation présenté dans la partie 5.4.1 était de 34. Les sous-ensembles ont des *threat scores* du même ordre de grandeur sur les sous-réseaux qui n'ont pas servi à leur calibration : 31.83 et 34.9. La calibration de ce score associé au tableau de contingence semble donc aussi robuste.

Ensemble	**Accuracy**	**Correct neg. ratio**	**Hit rate**	**Threat score**
		Comparaison sur le réseau A		
Ensemble complet	98.76	99.84	16.8	15.03
Ensemble calibré avec A	98.7	99.29	54.53	34.47
Ensemble calibré avec B	98.5	98.94	62.7	34.9
		Comparaison sur le réseau B		
Ensemble complet	98.71	99.83	16.3	14.5
Ensemble calibré avec B	98.42	98.97	57.88	33.0
Ensemble calibré avec A	98.6	99.29	48.3	31.83

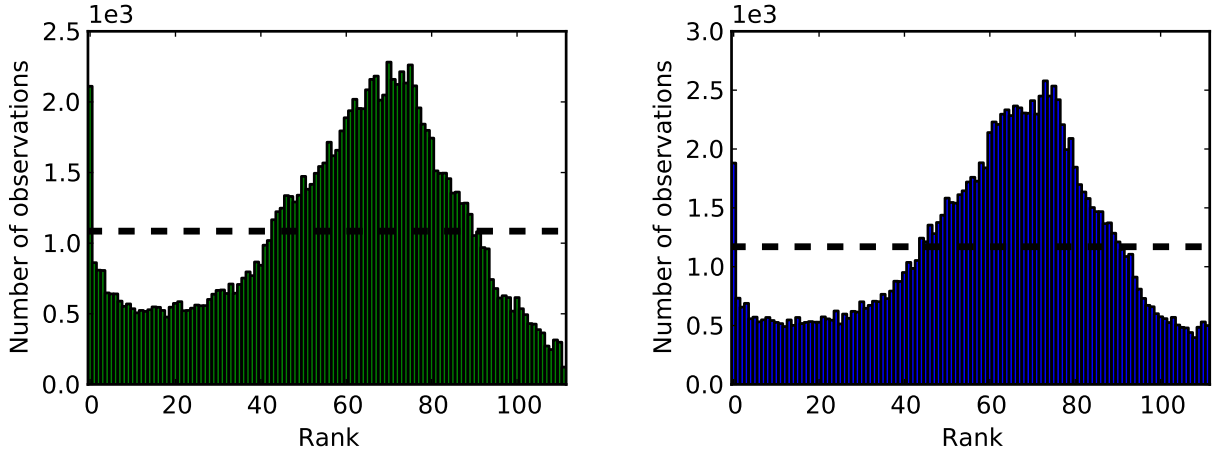
TABLE 5.13 – Scores associés au tableau de contingence pour l'évènement $[O_3] \geq 120\ \mu g\,m^{-3}$ des ensembles complets et calibrés sur chaque sous-réseau A et B. La calibration est effectuée en maximisant le *threat score*.

Enfin, ce sont les diagrammes de rang qui sont soumis à la même expérience. La figure 5.33 présente les deux diagrammes de rang de l'ensemble complet pour les sous-réseaux A et B. Ces derniers ont sensiblement la même forme de cloche et restent semblables au diagramme de rang calculé à partir de la totalité des observations. Les scores normalisés associés à ces diagrammes sont respectivement égaux à 338 et 390.

Une calibration sur chaque sous-réseau est effectuée dans le but d'obtenir un diagramme de rang plat. La figure 5.34 présente deux diagrammes de rang dont les données d'observation proviennent du réseau B. Ces deux diagrammes sont issus des sous-ensembles calibrés à partir des observations du réseau A et du réseau B — en vert et bleu respectivement. Ces sous-ensembles calibrés contiennent tous les deux une trentaine de membres. Un léger biais négatif est à noter sur le diagramme en vert. Ce diagramme, qui n'a pas été calibré sur le réseau en question, a néanmoins une forme convenable. La figure 5.35 présente quant à lui les diagrammes de rang des mêmes sous-ensembles mais calculés à partir des observations du réseau A. Ils ont tous les deux des formes semblables et assez plates, en comparaison des diagrammes de rang de l'ensemble complet.

L'écart type des deux sous-ensembles calibrés a été calculé afin de comparer l'estimation de l'incertitude suivant les réseaux qui ont servis à chacune des calibrations du diagramme de rang. La figure 5.36 présente les champs d'incertitude moyens issus des deux sous-ensembles calibrés. Le sous-ensemble issu du réseau B a un champ légèrement plus élevé que son homologue — $+0.8\ \mu g\,m^{-3}$ en moyenne. Cependant, ces champs sont très semblables, tant au niveau de leur structure spatiale que des valeurs des écarts types.

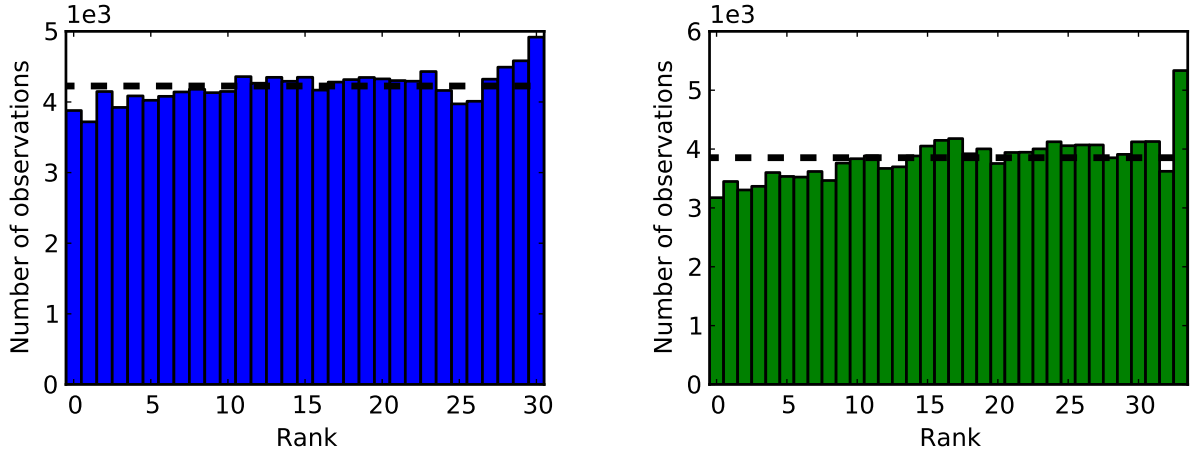
La méthode de calibration d'ensemble utilisée pour la prévision des risques, via le diagramme



(a) Diagramme de rang complet sur A

(b) Diagramme de rang complet sur B

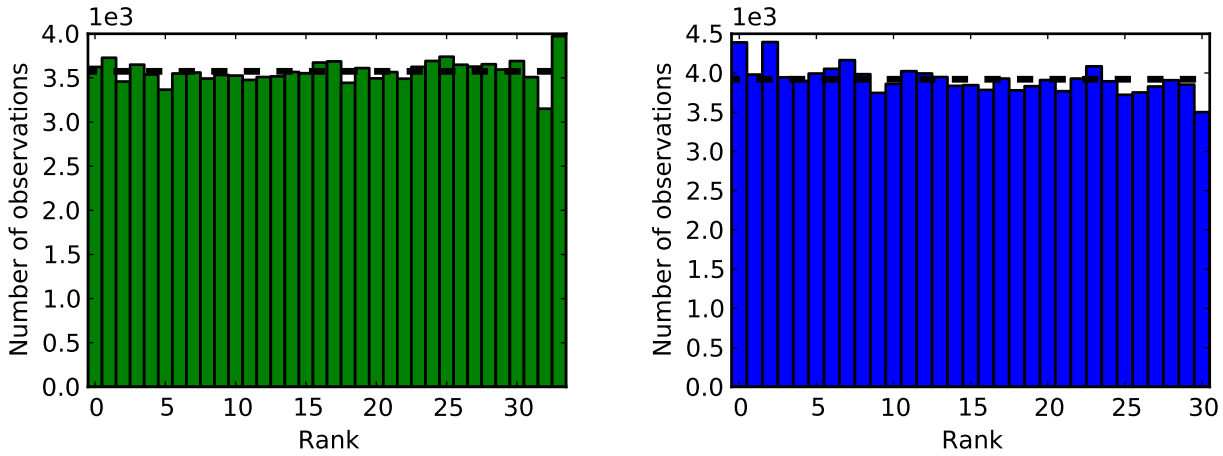
FIGURE 5.33 – Diagrammes de rang de l'ensemble complet d'ozone en concentration horaire dans la région Île-de-France pour les sous-réseaux BDQA A et B.



(a) Diagramme de rang calibré avec B

(b) Diagramme de rang calibré avec A

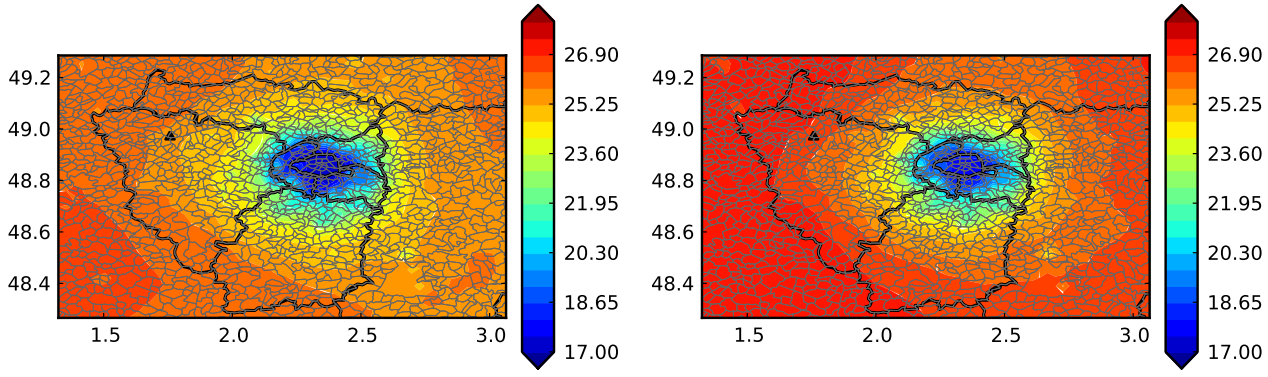
FIGURE 5.34 – Diagrammes de rang O_3 calculés avec les données du sous-réseau B. Le diagramme de rang de gauche a été calibré sur B tandis que celui de droite a été calibré sur A.



(a) Diagramme de rang calibré avec A

(b) Diagramme de rang calibré avec B

FIGURE 5.35 – Diagrammes de rang O_3 calculés avec les données du sous-réseau A. Le diagramme de rang de gauche a été calibré sur A tandis que celui de droite a été calibré sur B.



(a) Incertitude O_3 ($\mu g m^{-3}$) issue de la calibration avec le réseau A

(b) Incertitude O_3 ($\mu g m^{-3}$) issue de la calibration avec le réseau B

FIGURE 5.36 – Champs d'incertitude moyens d'ozone, en $\mu g m^{-3}$, issus des deux sous-ensembles dont les diagrammes de rang ont été calibrés sur le réseau A et B.

de fiabilité et le tableau de contingence, et l'estimation de l'incertitude, via l'optimisation du diagramme de rang, semble donc assez robuste pour l'ensemble d'ozone. La section suivante traite de la robustesse spatiale de la calibration de l'ensemble de simulations NO_2.

5.6.2 Dioxyde d'azote

Le nombre de stations d'observation BDQA pour mesurer le dioxyde de soufre dans la région Île-de-France s'élève à 42. La figure 5.37 représente la localisation de la centrale de Porcheville, en rouge, et l'emplacement des stations des deux sous-réseaux A et B sélectionnés de manière aléatoire, en vert et bleu respectivement.

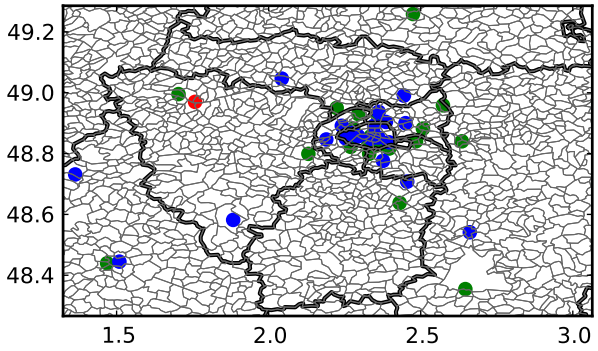
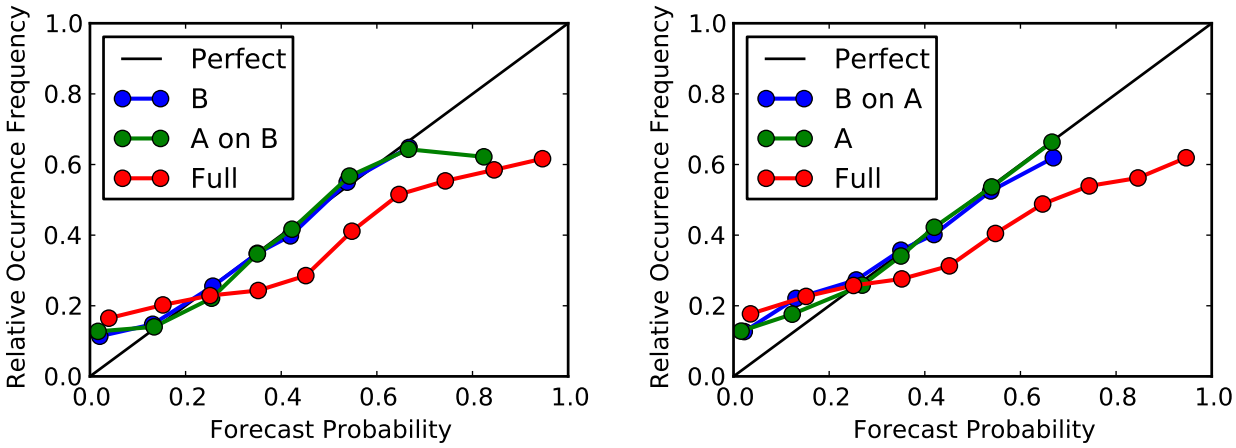


FIGURE 5.37 – Domaine Île-de-France. La centrale de Porcheville est localisée par le point rouge. Deux sous-réseaux d'un total de 42 stations BDQA pour le NO_2 ont été sélectionnés de manière aléatoire : le réseau A en vert et le réseau B en bleu.

Dans un premier temps, c'est la calibration du diagramme de fiabilité de l'évènement $[NO_2] \geq 60\,\mu g\,m^{-3}$ qui est étudiée. La figure 5.38 présente les diagrammes de fiabilité de l'ensemble complet et des sous-ensembles calibrés pour chaque sous-réseau. Les sous-ensembles calibrés, quel que soit le réseau, ne produisent pas assez de probabilités supérieures à 0.7. C'est pourquoi la courbe de ces diagrammes ne s'étend pas au-delà de cette valeur. La courbe de l'ensemble complet reste globalement au-dessous de la diagonale. Les courbes des diagrammes calibrés ont une forme semblable quel que soit le réseau d'observation.

Ensuite, la robustesse de la calibration associée au tableau de contingence via l'optimisation du *threat score* est étudiée. C'est l'évènement $[NO_2] \geq 100\,\mu g\,m^{-3}$ qui est pris en compte. Le tableau 5.14 représente les ratios associés au tableau de contingence pour l'ensemble complet et les sous-ensembles calibrés sur chaque sous-réseau. Quel que soit le réseau d'observation avec lequel les score sont calculés, les *threat scores* issus des sous-ensembles calibrés sont tous autour de 0.16 contre environ 0.045 pour l'ensemble complet. Comme dans le cas de l'ozone, la calibration du *threat score* de l'ensemble de NO_2 reste robuste.

Enfin, nous étudions la robustesse spatiale pour la calibration du diagramme de rang pour l'ensemble de NO_2 puis comparons l'incertitude estimée à partir de l'écart type des deux sous-ensembles calibrés. Comme dans la partie 5.4.1, nous décidons d'enlever les observations de NO_2 qui sont au-dessous de $25\,\mu g\,m^{-3}$ dans le but de maximiser potentiellement le nombre de mem-



(a) A sur B

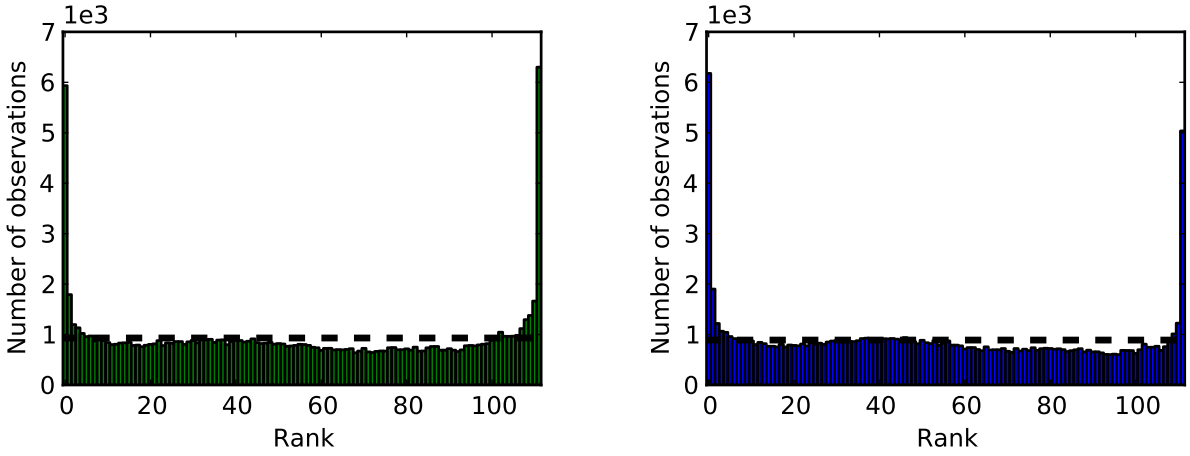
(b) B sur A

FIGURE 5.38 – Diagrammes de fiabilité pour $[NO_2] \geq 60\,\mu g\,m^{-3}$ dans la région Île-de-France. (a) correspond au diagramme de fiabilité du sous-réseau B et (b) correspond au diagramme de fiabilité du sous-réseau A.

Ensemble	**Accuracy**	**Correct neg. ratio**	**hit rate**	**threat score**
	Comparaison sur le réseau A			
Ensemble complet	0.908	0.992	0.046	0.043
Ensemble calibré avec A	0.837	0.884	0.355	0.163
Ensemble calibré avec B	0.860	0.915	0.298	0.159
	Comparaison sur le réseau B			
Ensemble complet	0.910	0.991	0.054	0.049
Ensemble calibré avec B	0.857	0.908	0.326	0.165
Ensemble calibré avec A	0.831	0.874	0.377	0.162

TABLE 5.14 – Scores associés au tableau de contingence pour l'évènement $[NO_2] \geq 100\,\mu g\,m^{-3}$ des ensembles complet et calibrés sur chaque sous-réseau A et B. La calibration est effectuée en maximisant le *threat score*.

bres dans les sous-ensembles calibrés. La figure 5.39 présente les deux diagrammes de rang de l'ensemble complet pour le NO_2 sur chacun des sous-réseaux d'observation A et B. Ils ont une forme en « U » qui indique une sous-estimation de l'incertitude. Aux vues de la forme de ces diagrammes et de la valeur des barres aux extrémités, les diagrammes calibrés peuvent contenir jusqu'à une quinzaine de membres.



(a) Diagramme de rang complet sur A

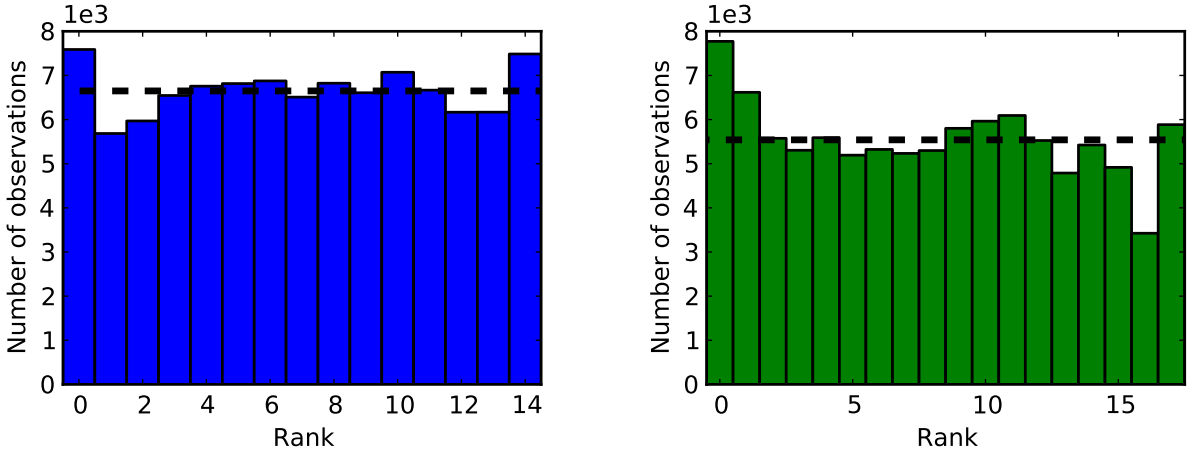
(b) Diagramme de rang complet sur B

FIGURE 5.39 – Diagrammes de rang de l'ensemble complet NO_2 en Île-de-France sur les deux sous-réseaux A en vert et B en bleu.

Deux optimisations combinatoires sont ensuite effectuées afin de fournir un ensemble calibré par sous-réseau. Les figures 5.40 et 5.41 présentent chacune les diagrammes des deux sous-ensembles calibrés sur chaque sous-réseau. Les deux sous-ensembles contiennent tous les deux une quinzaine de membres comme prévu. Même si ces diagrammes calibrés ne sont pas parfaitement plats, ils ont une forme convenable comparée à l'ensemble complet. On peut noter que sur la figure 5.40, le diagramme de rang du sous-ensemble issu du réseau A appliqué sur les réseaux B en vert à un léger biais positif. Ceci peut s'expliquer par le fait que le nombre d'observations sous l'enveloppe inférieure de l'ensemble est plus important dans le réseau B que dans le réseau A. Ceci s'observe en comparant la valeur des barres aux extrémités des deux diagrammes de l'ensemble complet sur la figure 5.39. Un constat équivalent, mais avec un léger biais négatif, est fait sur la figure 5.41. Le diagramme de rang du sous-ensemble calibré issu du réseau B appliqué au réseau A est décalé vers la droite. En effet, le nombre d'observations au-dessus de l'enveloppe supérieure est plus important dans le réseau A que dans le réseau B.

Enfin, l'écart type des deux sous-ensembles précédemment calibrés est calculé afin d'estimer l'incertitude de NO_2 dans la région Île-de-France. La figure 5.42 présente les deux champs moyens d'incertitude issus de l'écart type des sous-ensembles calibrés sur les réseaux A et B. L'incertitude est toujours concentrée autour de Paris et sa banlieue. Les deux champs sont très semblables. La calibration du diagramme de rang de l'ensemble de NO_2 puis l'estimation de l'incertitude semblent donc assez robustes spatialement.

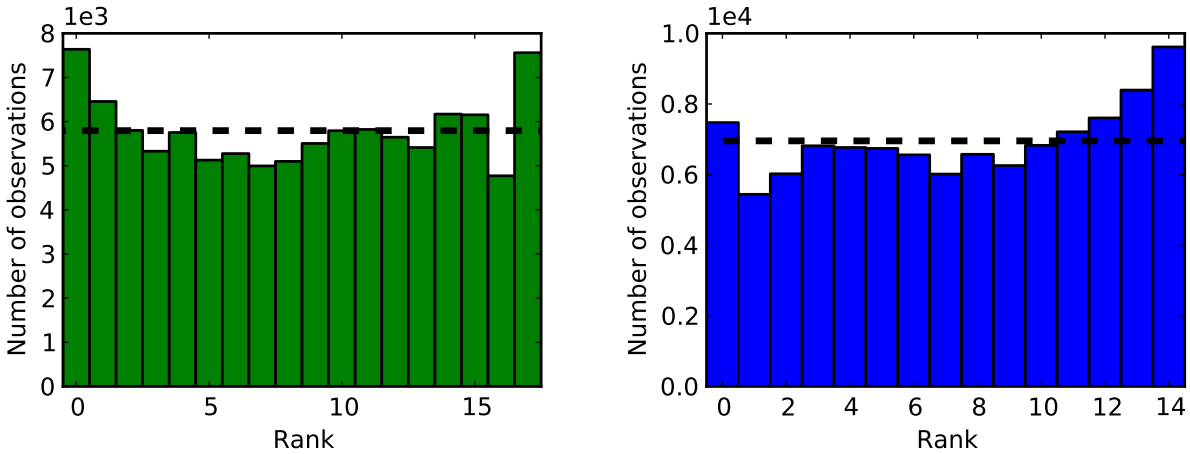
La robustesse spatiale a été étudiée pour la calibration de trois scores d'ensemble : le diagramme de fiabilité, le *threat score* et le diagramme de rang, à la fois pour l'ensemble de simulations d'ozone et de simulations de NO_2. Les résultats sont satisfaisants puisque les performances d'un sous-ensemble calculées à partir d'un réseau indépendant de la calibration sont équivalentes ou très proches des performances du sous-ensemble calibré sur ce même réseaux.



(a) Diagramme de rang calibré sur B

(b) Diagramme de rang calibré sur A

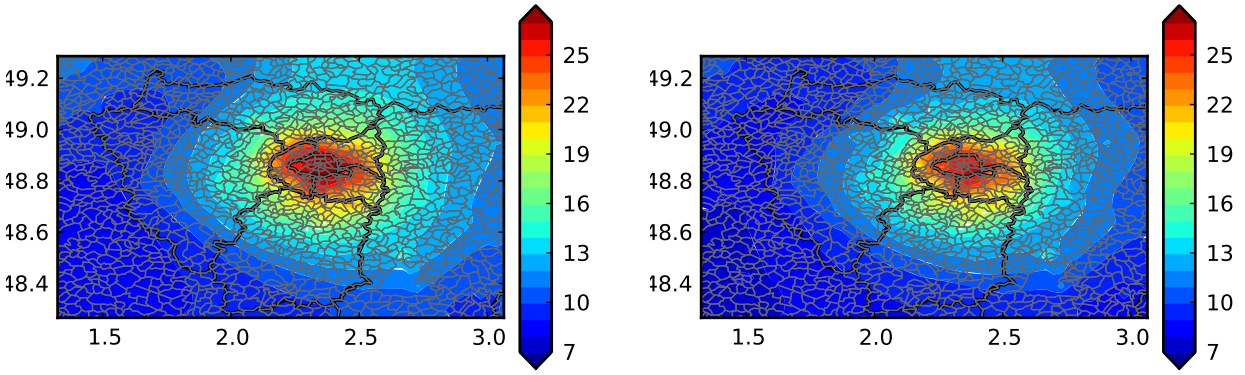
FIGURE 5.40 – Diagrammes de rang NO_2 appliqués sur le sous-réseau B. Le diagramme de rang de gauche a été calibré sur B tandis que celui de droite a été calibré sur A.



(a) Diagramme de rang calibré sur A

(b) Diagramme de rang calibré sur B

FIGURE 5.41 – Diagrammes de rang NO_2 appliqués sur le sous-réseau A. Le diagramme de rang de gauche a été calibré sur A tandis que celui de droite a été calibré sur B.



(a) Incertitude NO_2 ($\mu g m^{-3}$) issue de la calibration avec le réseau A

(b) Incertitude NO_2 ($\mu g m^{-3}$) issue de la calibration avec le réseau B

FIGURE 5.42 – Champs d'incertitude moyen de NO_2, en $\mu g m^{-3}$, issus des deux sous-ensembles dont les diagrammes de rang ont été calibrés sur le réseau A et B.

5.7 Prévision

La calibration pour l'estimation de l'incertitude et pour l'estimation des risques semble robuste spatialement. Qu'en est-il de la robustesse temporelle ? Une calibration pour un score d'ensemble donné est-elle assez robuste pour que le sous-ensemble ainsi sélectionné soit performant en dehors de la période sur laquelle il a été calibré ? Autrement dit, la calibration peut-elle être appliquée à la prévision ?

L'objectif de cette partie est d'appliquer la calibration à (1) la prévision de champs d'incertitude — comme ce qui a été effectué dans la partie 3.4 à l'échelle européenne — et (2) à la prévision des risques de dépassement de seuil. Pour ce faire, il faut comparer une calibration effectuée pendant une période dite d'« apprentissage » — suffisamment longue pour contenir assez d'observations — avec une calibration faite « a posteriori » avec les observations issues de la période sur laquelle on souhaite faire la prévision. Les deux périodes sont de même durée et ne se chevauchent pas. Afin de vérifier la prévision, le sous-ensemble calibré prévu sera comparé — en terme de score d'ensemble et de carte de prévision — au sous-ensemble calibré « a posteriori », ainsi qu'à l'ensemble complet ou à un sous-ensemble sélectionné de manière aléatoire. Plusieurs prévisions à différents moments de l'année 2007 seront effectuées. Le schéma sur la figure 5.43 est un exemple qui illustre la période d'apprentissage et la période de prévision toutes deux d'une durée de deux semaines. On fait glisser sur toute l'année, de semaine en semaine, ce couple période d'apprentissage/période de prévision. Le premier couple s'étend de la première semaine à la quatrième semaine (incluses) de l'année ; le second couple s'étend de la deuxième semaine à la cinquième ; le troisième commence la troisième semaine et se termine la sixième ; etc. Un total de quarante-huit couples sont ainsi considérés.

La période de prévision correspond à une période de vérification : un sous-ensemble calibré (sur la période d'apprentissage) y est évalué avec les observations de cette période. Il faut noter que l'évaluation, comme la calibration, nécessite une quantité de données importante. On peut se

référer à la partie 3.4 où on calcule un diagramme de rang pour différents nombres d'observations (voir figure 3.13). Nous avions conclu qu'environ 8000 à 10000 observations étaient alors requises. La période de prévision doit donc être suffisamment longue, et éventuellement plus longue que la période habituelle des prévisions opérationnelles. On s'autorise à considérer une longue période de prévision, tout en sachant que seuls les résultats des tous premiers jours de la période seraient exploitables en pratique. Notons que les performances sur les premiers jours de prévision sont vraisemblablement meilleures que celles calculées sur toute la période de prévision. Il est en effet probable que la calibration perde de sa pertinence à mesure qu'on s'éloigne de la période d'apprentissage.

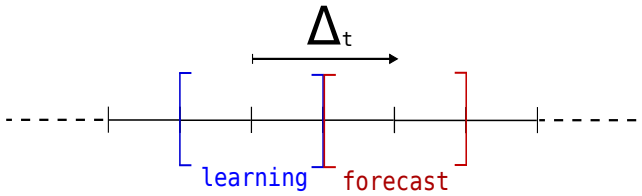


FIGURE 5.43 – Schéma de la fenêtre temporelle constituée de la période d'apprentissage et de la période de prévision — *learning* et *forecast* en anglais. Chaque période a une durée de deux semaines.

Deux ensembles sont étudiés : l'ensemble de simulations d'ozone et l'ensemble de simulations du dioxyde d'azote dans la région autour de la centrale thermique de Porcheville. Ce domaine a été choisi parce qu'il dispose d'un nombre suffisant de stations, contrairement à la région des Pays de la Loire. Il est en effet nécessaire d'avoir un nombre suffisant de stations de mesure, car en prévision, la calibration nécessite un nombre conséquent de mesures sur une période de temps de l'ordre de quelques semaines seulement.

En ce qui concerne les périodes d'apprentissage et de prévision, elles sont de même durée et sont fixées à deux semaines, quel que soit le polluant considéré. Cela permet d'obtenir un nombre de données d'observation autour de 8000 pour chacune des périodes et pour chaque polluant. On fait glisser les fenêtre d'apprentissage et de prévision (qui se suivent toujours) de semaine en semaine, sur toute l'année 2007. Le nombre total d'expériences d'apprentissage/prévision s'élève donc à 48 pour toute l'année 2007. Pour chaque période, nous déployons la même méthode, automatisée grâce à un ensemble de scripts Python : (1) les données d'ensemble et d'observation sont chargées en concentration horaire sur la période « *learning* + *forecast* » — c'est-à-dire un mois, (2) plusieurs calibrations d'ensemble sont effectuées sur chacune des périodes, (3) tous les scores des sous-ensembles calibrés sur la période d'apprentissage sont calculés pour la période de prévision, et (4) ces scores sont comparés à ceux des sous-ensembles calibrés (a posteriori) sur la période de prévision. Ces résultats comprennent aussi les scores de l'ensemble complet sur la période de prévision ainsi que les scores d'un sous-ensemble sélectionné aléatoirement sur cette même période.

Les scores d'ensemble qui ont été pris en compte dans les multiples calibrations et qui sont comparés avec les scores des sous-ensemble calibrés a posteriori sont : le diagramme de fiabilité pour quelques évènements, le *threat score* associé au tableau de contingence et le diagramme de rang. Grâce à la calibration pour le diagramme de fiabilité et le diagramme de rang, on améliore des champs de probabilité et des champs d'incertitude prévus. Ces derniers sont comparés avec les champs calculés à partir des sous-ensembles calibrés a posteriori.

La première partie traite de la calibration du diagramme de rang et de la prévision d'in-

certitude. La deuxième porte sur la prévision de dépassement de seuil, avec la calibration du diagramme de fiabilité et du tableau de contingence, et de la prévision de champs de probabilité.

5.7.1 Incertitudes

Tout comme dans les sections précédentes, l'incertitude est estimée à partir de l'écart type empirique d'un sous-ensemble sélectionné par la calibration du diagramme de rang. Les deux sections qui suivent traitent de la prévision de l'incertitude de l'ozone et du dioxyde d'azote.

Ozone

La totalité des résultats des 48 prévisions de l'ensemble d'ozone n'est pas présentée dans cette partie. Seuls les résultats les plus marquants sont commentés. Nous décidons dans un premier temps de comparer l'incertitude prévue et l'incertitude estimée a posteriori pour une période donnée. Les champs d'incertitude ne sont pas moyennés sur toute la période de prévision — c'est-à-dire deux semaines — mais sur les moyennes journalières des premiers jours de prévision.

Il paraît en effet plus intéressant d'étudier les résultats prévus dont les dates sont encore proches de la période d'apprentissage. Il se peut que l'information utilisée lors de la calibration pendant la période d'apprentissage soit « oubliée » au bout de quelques jours. Les prévisions opérationnelles de la qualité de l'air excèdent rarement quelques jours. Nous décidons donc de nous intéresser aux premiers jours de prévision d'incertitude en moyenne journalière.

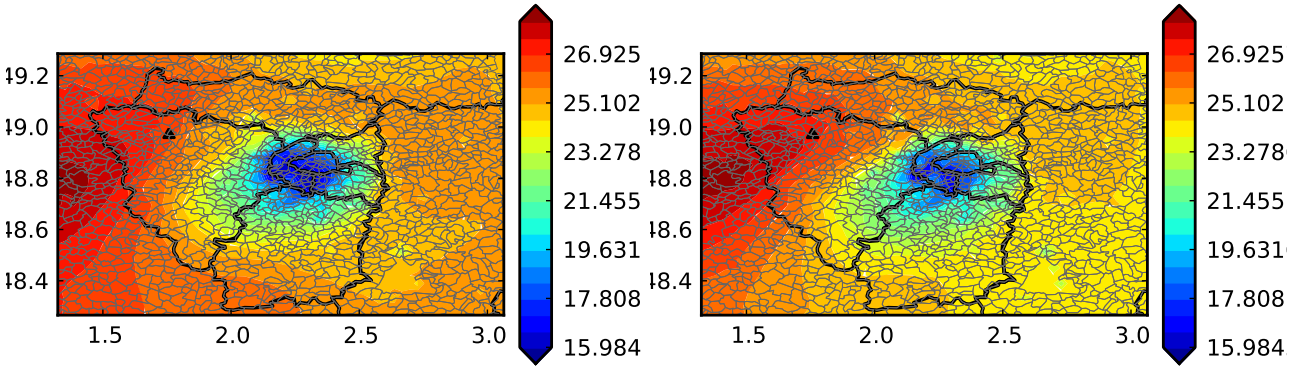
La figure 5.44 montre deux champs d'incertitude moyens, prévu et a posteriori, à la date du 3 février 2007 qui est le premier jour d'une période de prévision. Les champs ont une structure spatiale très similaire et des valeurs d'écart type très proches. Les champs d'incertitude donnés par l'ensemble complet et le sous-ensemble aléatoire sont bien supérieurs aux champs d'incertitude présentés ci-dessus. Une différence de près de $+10\,\mu g\,m^{-3}$ est constatée entre le champ d'écart type de l'ensemble complet et ceux montrés sur la figure 5.44.

Les différents sous-ensembles aléatoires et l'ensemble complet ont tendance, quelle que soit la période de prévision, à surestimer le champ d'écart type, en comparaison des ensembles calibrés. Ils sont en outre trop dispersés.

Les similitudes entre le champ a posteriori et prévu ont tendance à se dégrader en fonction du temps. Plus l'échéance est loin, plus il est difficile de prévoir le champ d'incertitude. La figure 5.45 présente les champs d'incertitude moyennés sur le 4, 5 et 6 février 2007, soit 2, 3 et 4 jours après la fin de la période d'apprentissage. Les champs d'incertitude a posteriori et prévu sont plus dissemblables. Néanmoins, les maxima et les minima de chaque champs se situent à peu près au même endroit. Malgré la différence plus marquante entre les deux champs à échéance plus longue, le champ d'incertitude prévu est toujours plus proche du champ d'incertitude a posteriori que le sont les champs d'incertitude issus de l'ensemble complet et du sous-ensemble aléatoire.

Malheureusement, tous les résultats ne sont pas aussi satisfaisants que les champs du 3 février de la figure 5.44. Les diagrammes de rang prévus ne sont pas toujours parfaitement plats et les champs d'incertitude prévus, comparés aux champs a posteriori sont parfois biaisés. Autrement dit, il arrive que le champ d'incertitude prévu soit surestimé, ou sous-estimé, par rapport aux champs a posteriori. L'ensemble calibré prévu est donc parfois trop ou pas assez dispersé comparé à la distribution des observations de la période de prévision.

Nous décidons de calculer la moyenne des $J+i$ jours de prévision sur toute la période, soit 48 prévisions — i valant 1, 3 ou 5. Le premier jour de prévision correspond à un samedi, le



(a) A posteriori (b) Prévu

FIGURE 5.44 – Champs d'incertitude a posteriori et prévu d'ozone moyennés sur 3 février 2007 (premier jour de prévision après une période d'apprentissage), dans la région Île-de-France (en $\mu g\,m^{-3}$).

troisième à un lundi et le cinquième un mercredi. La figure 5.46 présente les moyennes temporelles des champs d'incertitude a posteriori et prévu du premier, troisième et cinquième jours suivant les différentes périodes d'apprentissage. Les moyennes des champs d'incertitude concernant le premier jour de prévision sont assez proches. Les différences sont ensuite plus notables pour le troisième et cinquième jour. Notons que ces champs d'incertitude sont très semblables au champ d'incertitude moyenné sur toute l'année 2007 montré dans la partie 5.4.1.

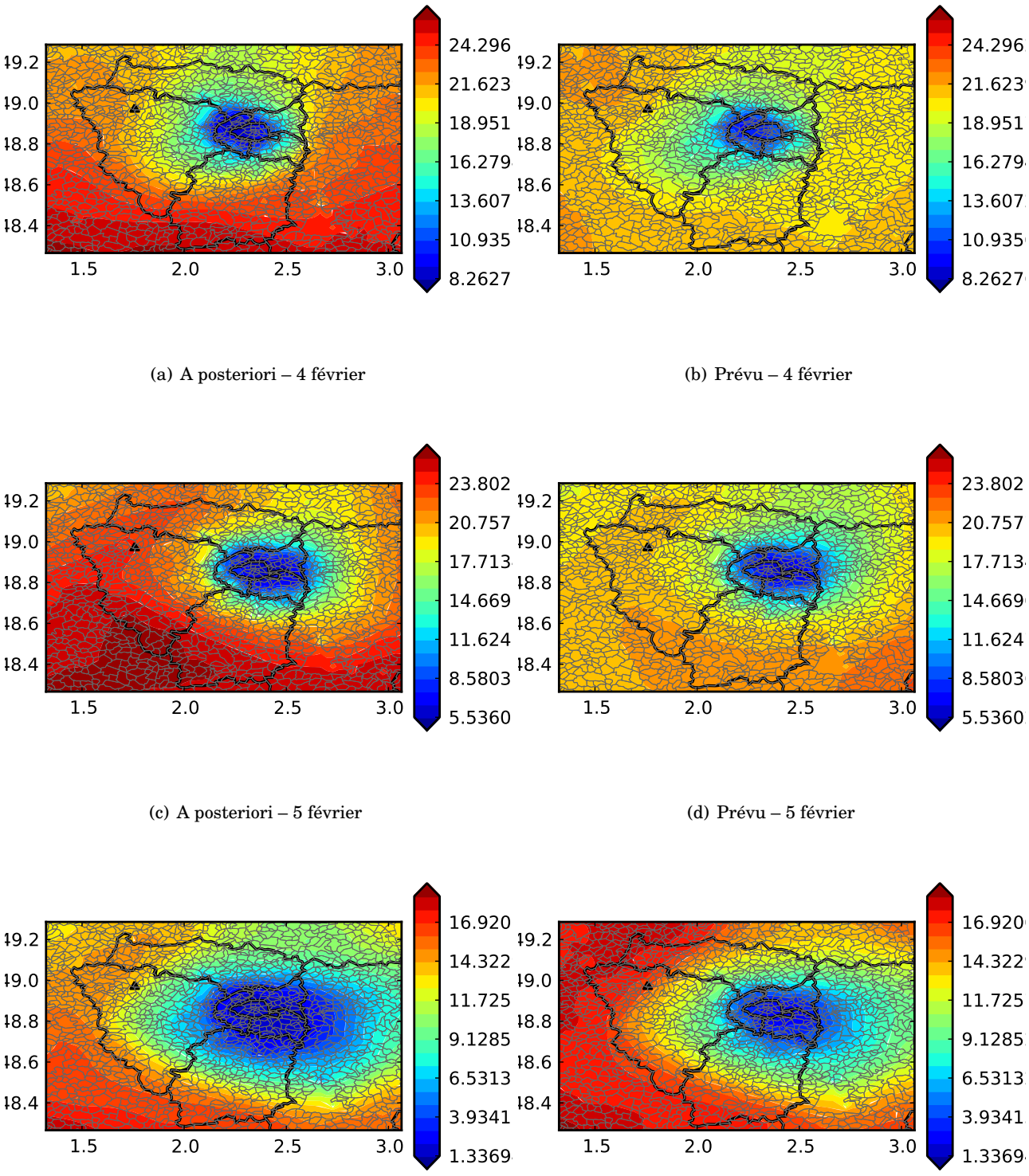
Les résultats sont globalement assez satisfaisants puisque les champs d'incertitude prévus, même s'ils présentent parfois un biais comparé aux champs d'incertitude a posteriori, sont assez proches des champs d'incertitude a posteriori et toujours meilleurs que l'ensemble complet et le sous-ensemble aléatoire.

Dioxyde d'azote

La prévision d'incertitude du dioxyde d'azote est effectuée dans les mêmes conditions que précédemment. Les périodes d'apprentissage et de prévision sont toutes les deux d'une durée de deux semaines. On continues à faire glisser la fenêtre d'apprentissage/prévision de semaine en semaine. Il y a donc 48 prévisions effectuées sur toute l'année 2007.

Un premier exemple de prévision de champs d'incertitude de NO_2 est donné sur la figure 5.47. Les champs d'incertitude a posteriori et prévu sont moyennés sur la journée du 4 août 2007, qui est un jour de prévision suivant immédiatement une période d'apprentissage. L'incertitude du NO_2 est toujours plus élevée au-dessus de Paris et de sa proche banlieue. Les deux champs restent très semblables. L'ensemble complet et le sous-ensemble aléatoire, eux, sous-estiment le champ d'incertitude du NO_2. La différence entre l'écart type de l'ensemble complet et l'écart type de l'ensemble calibré a posteriori est de $-8\,\mu g\,m^{-3}$. L'ensemble complet de NO_2 est donc trop peu dispersé comme l'ont montré les diagrammes de rang de la partie 5.4.1.

Les résultats pour les jours de prévision suivants — le 5, 6 et 7 août 2007 — sont assez



(a) A posteriori – 4 février

(b) Prévu – 4 février

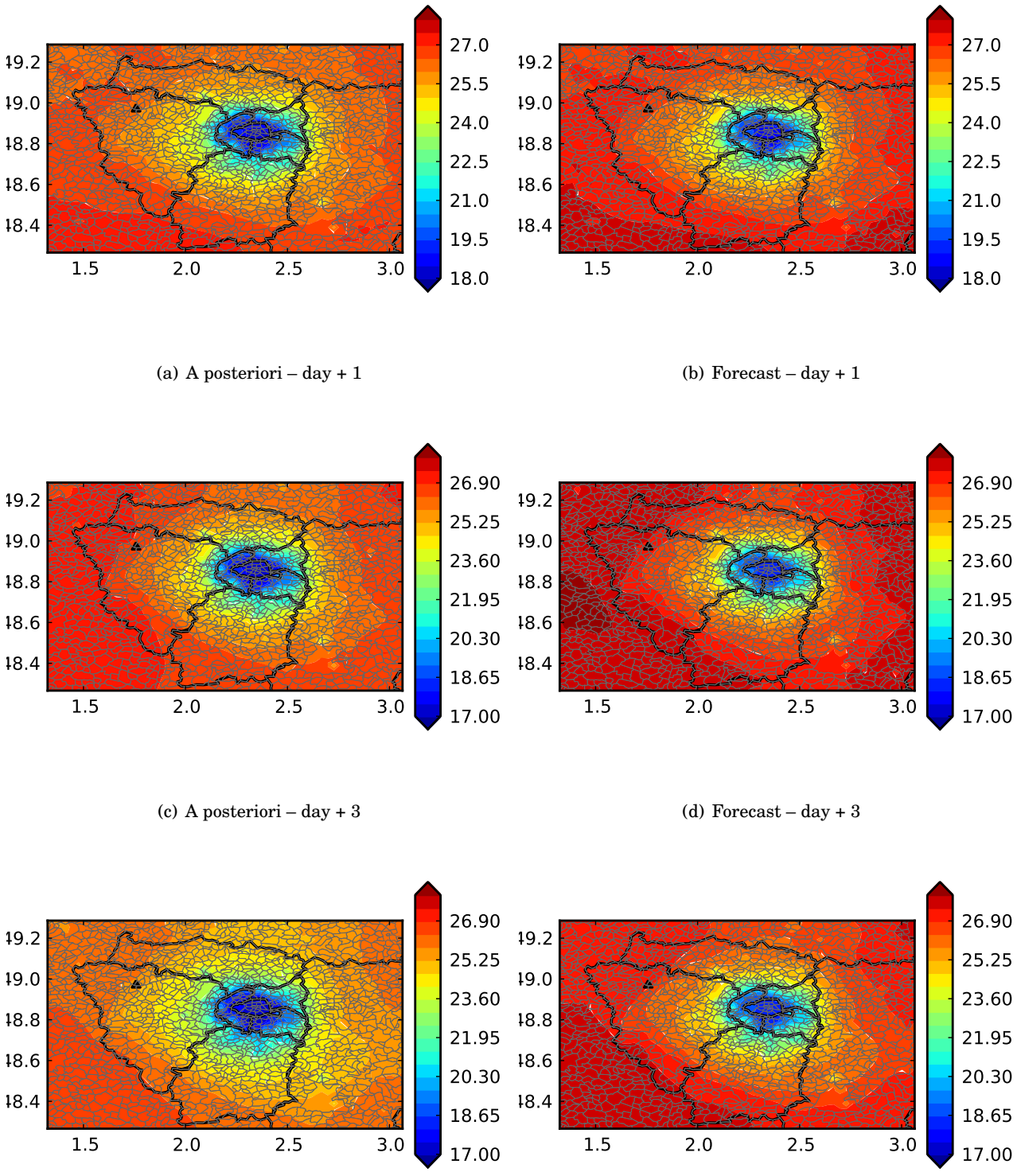
(c) A posteriori – 5 février

(d) Prévu – 5 février

(e) A posteriori – 6 février

(f) Prévu – 6 février

FIGURE 5.45 – Champs d'incertitude a posteriori et prévu d'ozone moyennés pour les 4, 5 et 6 février 2007, dans la région Île-de-France (en $\mu g\,m^{-3}$).



(a) A posteriori – day + 1

(b) Forecast – day + 1

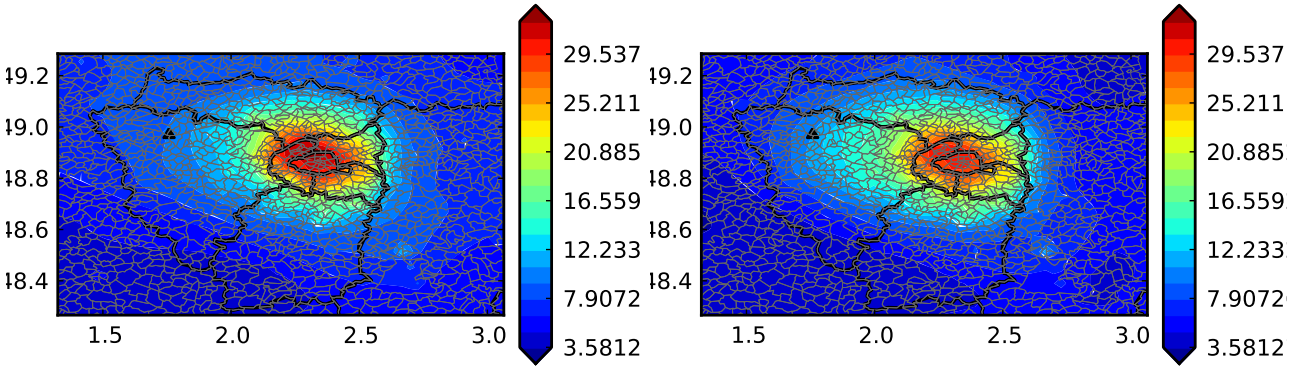
(c) A posteriori – day + 3

(d) Forecast – day + 3

(e) A posteriori – day + 5

(f) Forecast – day + 5

FIGURE 5.46 – Champs d'incertitude d'ozone a posteriori et prévu moyennés sur toute la période, pour les 1er, 3^{e} et 5^{e} jours suivants les périodes d'apprentissage dans la région Île-de-France (en $\mu g m^{-3}$).



(a) A posteriori

(b) Prévu

FIGURE 5.47 – Cartes d'incertitude a posteriori et prévues de dioxyde d'azote moyennées sur un jour de prévision, 4 août 2007, dans la région Île-de-France (en $\mu g\,m^{-3}$).

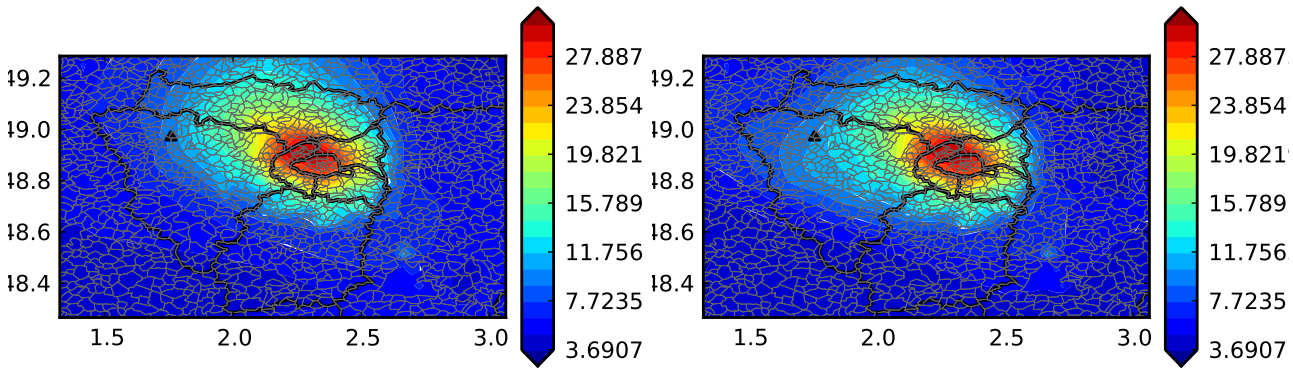
satisfaisants, comme le montrent les cartes de la figure 5.48. Il est intéressant de noter que les valeurs des écarts types fournis par les sous-ensembles restent toujours assez proches. De plus, ces cartes d'écart type, tracées sur trois jours consécutifs, montrent que la répartition spatiale de l'incertitude évolue significativement dans le temps.

Enfin, la moyenne temporelle des champs d'incertitude a posteriori et prévu est calculée pour chaque $J + i$ jour de prévision suivant la période d'apprentissage. On s'intéresse au premier, troisième et cinquième jours suivants la période d'apprentissage. La moyenne est effectuée sur toute la période d'étude qui compte 48 prévisions. La figure 5.49 montre ces champs d'incertitude moyens à la fois pour le sous-ensemble calibré a posteriori et pour le sous-ensemble calibré prévu. Les valeurs restent toujours très élevées au-dessus de Paris et de la petite couronne. Les deux champs, quel que soit le jour de prévision suivant les périodes d'apprentissage, sont assez semblables, tant par leur forme et que par les valeurs de l'incertitude. Concernant l'ensemble complet et le sous-ensemble aléatoire, les champs d'incertitude estimés par ces derniers sont toujours plus faibles de quelques $\mu g\,m^{-3}$.

Les résultats de prévision de champs d'incertitude, à la fois pour l'ensemble d'ozone et de NO_2, sont globalement satisfaisants. Cette partie montre que les champs d'incertitude évolue dans le temps. C'est pourquoi il est important de prévoir « au mieux » les champs d'incertitude pour les jours à venir. Même si les champs prévus présentent parfois des biais comparés aux champs a posteriori, ces derniers sont en moyenne satisfaisants. De plus, on peut noter que les champs prévus restent nettement meilleurs que les champs issus de l'ensemble complet et du sous-ensemble aléatoire.

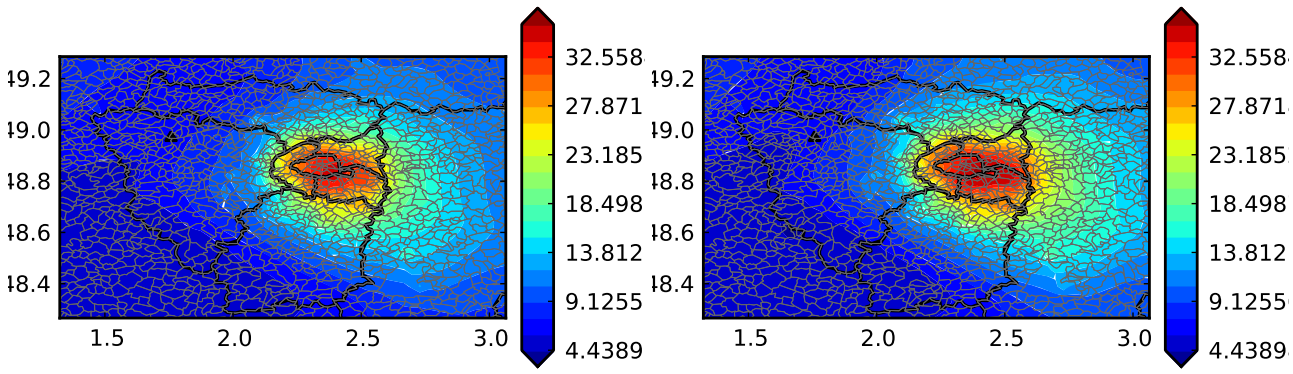
5.7.2 Risques de dépassement de seuil

Les périodes d'apprentissage et de prévision sont exactement les mêmes que pour la prévision d'incertitude : deux fois deux semaines qui glissent de semaine en semaine sur toute l'année. Les scores d'ensemble qui sont utilisés pour la prévision d'évènements relatifs au dépassement de seuil de concentration sont le diagramme de fiabilité et le *threat score*. Après avoir comparé les



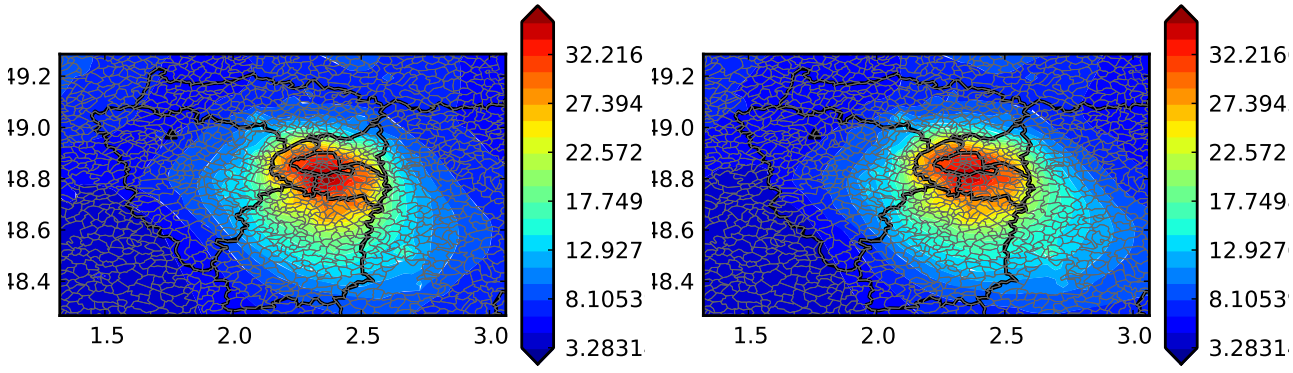
(a) A posteriori – 5 août

(b) Prévu – 5 août



(c) A posteriori – 6 août

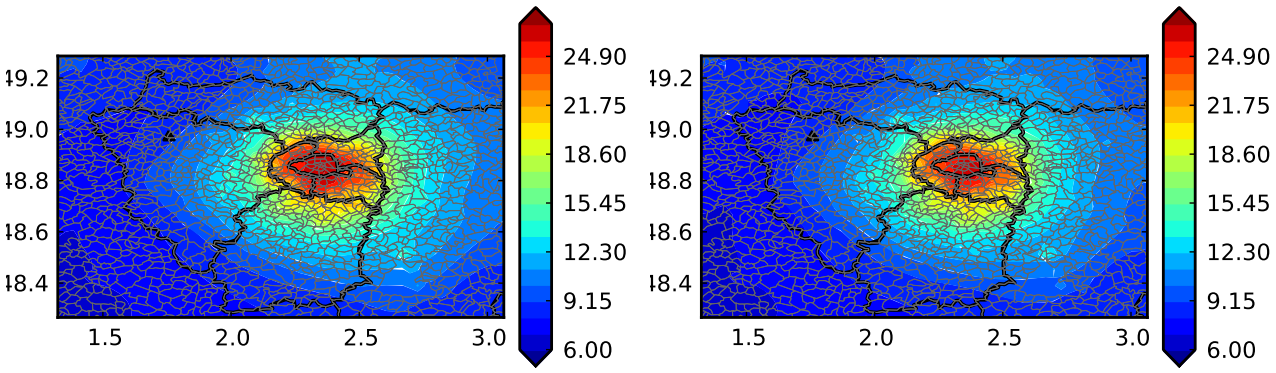
(d) Prévu – 6 août



(e) A posteriori – 7 août

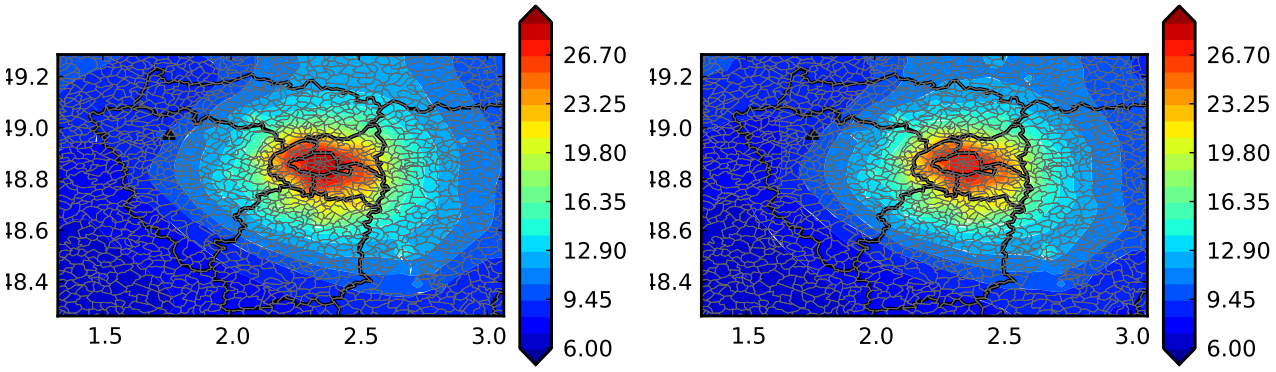
(f) Prévu – 7 août

FIGURE 5.48 – Champs d'incertitude a posteriori et prévu de NO_2 moyennés pour les 5, 6 et 7 août 2007, dans la région Île-de-France (en $\mu g\,m^{-3}$).



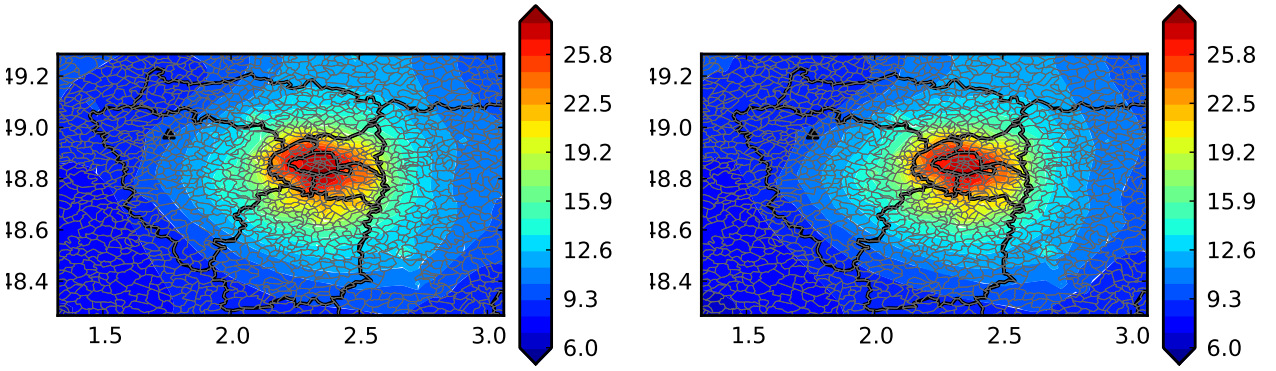
(a) A posteriori – day + 1

(b) Forecast – day + 1



(c) A posteriori – day + 3

(d) Forecast – day + 3



(e) A posteriori – day + 5

(f) Forecast – day + 5

FIGURE 5.49 – Champs d'incertitude de dioxyde d'azote a posteriori et prévu moyennés sur toute la période pour les 1er, 3^{e} et 5^{e} jours suivants les périodes d'apprentissage dans la région Île-de-France (en $\mu g m^{-3}$).

différents scores d'ensemble issus des sous-ensembles a posteriori et des sous-ensembles prévus, des champs de probabilité sont étudiés dans cette partie.

Contrairement aux diagrammes de rang, il est possible de moyenner le diagramme de fiabilité et les valeurs du tableau de contingence, même si le nombre de membres sélectionnés varient au cours du temps. Les résultats des 48 apprentissages/prévisions sont donc cumulés puis moyennés, afin de comparer globalement prévision et estimation a posteriori.

Ozone

Le tableau de contingence de l'évènement $[O_3] \geq 80\ \mu g\,m^{-3}$ et de ses scores associés est présenté dans le tableau 5.15. Les performances de quatre ensembles différents sont mentionnés : l'ensemble complet, le sous-ensemble aléatoire, le sous-ensemble calibré prévu et le sous-ensemble calibré a posteriori. Les calibrations ont été réalisées en maximisant le *threat score*.

Attention cependant, même si ces résultats montrent un tableau de contingence global sur toute la période d'étude, les sous-ensembles calibrés et aléatoires ne sont pas les mêmes à chaque prévision. Ce sont uniquement les résultats — le tableau de contingence dans le cas présent — calculés à chaque période de prévision qui sont cumulés/moyennés.

Ensemble	**hit**	**miss**	**correct neg.**	**false alarm**
Full	27145	23757	444456	11730
Random	26712	24190	439293	16893
Forecast	33487	17415	432280	23906
A posteriori	36752	14150	436297	19889

Ensemble	**accuracy**	**correct. neg. ratio**	**hit rate**	**threat score**
Full	0.930	0.974	0.533	0.433
Random	0.919	0.963	0.525	0.394
Forecast	0.919	0.948	0.658	0.448
A posteriori	0.933	0.956	0.722	0.519

TABLE 5.15 – Tableau de contingence et scores associés pour des prévisions cumulées de l'évènement $[O_3] \geq 80\ \mu g\,m^{-3}$ dans la région de Porcheville. La période d'apprentissage est de deux semaines.

Le *threat score* des sous-ensembles prévus est en moyenne plus élevé qui celui de l'ensemble complet mais reste assez proche de ce dernier — 0.45 et 0.43 respectivement, contre 0.52 pour le *threat score* a posteriori. Le taux de réussite prévu est aussi plus élevé que celui calculé avec l'ensemble complet : 66% contre 53%. Néanmoins, ce sont les sous-ensemble calibrés prévus qui présentent le plus de fausses alarmes : plus de 23000.

Des résultats équivalents sont à noter pour l'évènement $[O_3] \geq 120\ \mu g\,m^{-3}$ présentés sur le tableau 5.16. Le *threat score* prévu est plus de deux fois supérieur au *threat score* de l'ensemble complet mais plus faible que le score a posteriori. Malheureusement, les sous-ensembles calibrés prévus n'arrivent pas à prévoir plus de la moitié de l'occurrence de l'évènement — 48% contre 65% pour le taux de succès issu des sous-ensembles calibrés a posteriori.

On notera qu'il est toujours délicat/difficile d'avoir un bon taux de succès avec un faible nombre de fausses alarmes. Ce même résultat a été constaté dans la partie 5.4.2.

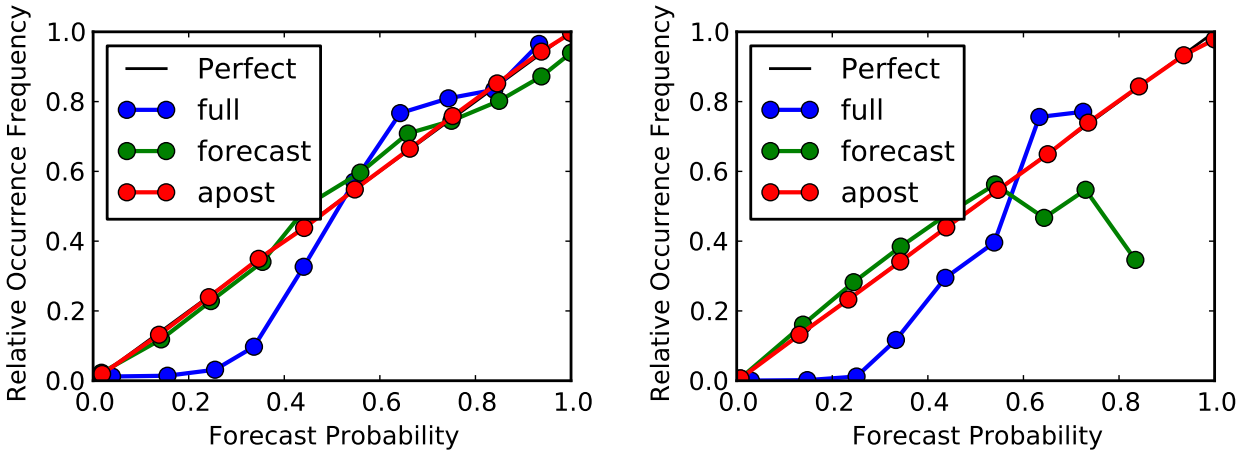
Ensemble	hit	miss	correct neg.	false alarm
Full	1230	6146	498787	925
Random	1228	6148	498104	1608
Forecast	3531	3845	494740	4972
A posteriori	4812	2564	495147	4565

Ensemble	accuracy	correct. neg. ratio	hit rate	threat score
Full	0.986	0.998	0.167	0.148
Random	0.985	0.997	0.166	0.137
Forecast	0.983	0.990	0.479	0.286
A posteriori	0.986	0.991	0.652	0.403

TABLE 5.16 – Tableau de contingence et scores associés pour les prévisions cumulées de l'évènement $[O_3] \geq 120\,\mu g\,m^{-3}$ dans la région de Porcheville. La période d'apprentissage est de deux semaines.

Ensuite, ce sont les diagrammes de fiabilité moyennés sur toutes les périodes de prévision, pour les deux mêmes évènements que précédemment, qui sont calculés. La figure 5.50 montre de tels diagrammes issus de l'ensemble complet, des sous-ensembles calibrés prévus et des sous-ensembles calibrés a posteriori pour $[O_3] \geq 80\,\mu g\,m^{-3}$ et $\geq 120\,\mu g\,m^{-3}$.

Le diagramme de fiabilité prévu pour l'évènement $[O_3] \geq 80\,\mu g\,m^{-3}$ présente de très bons résultats. Il reste légèrement moins bon que le diagramme a posteriori mais est très proche de la diagonale. Le diagramme de fiabilité de l'ensemble complet a une forme équivalente à celui calculé dans la partie 5.4.2 — voir la figure 5.17. Le diagramme issu des sous-ensembles aléatoires, qui n'est pas été présenté ici, a une forme semblable au diagramme de l'ensemble complet.

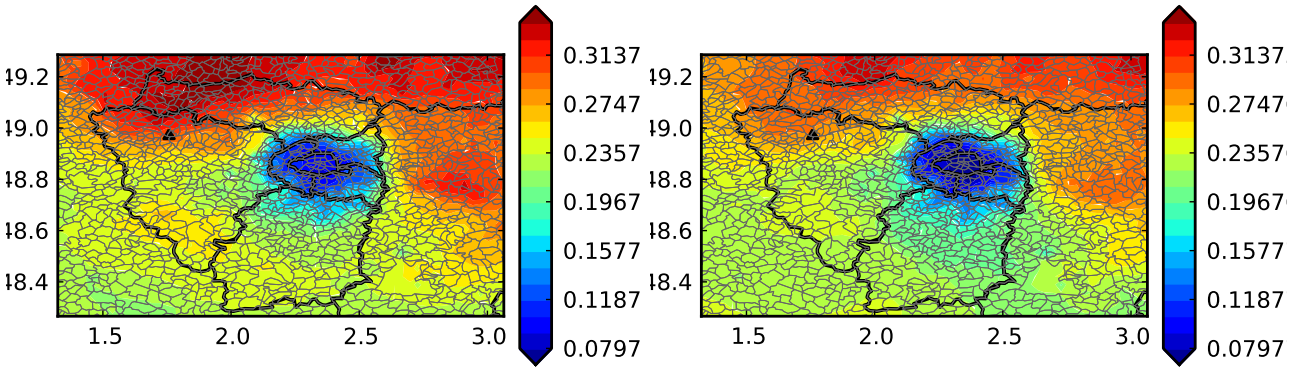


(a) $O_3 \geq 80\,\mu g\,m^{-3}$ (b) $O_3 \geq 120\,\mu g\,m^{-3}$

FIGURE 5.50 – Diagrammes de fiabilité cumulés issus de plusieurs prévisions pour les deux évènements $[O_3] \geq 80\,\mu g\,m^{-3}$ et $\geq 120\,\mu g\,m^{-3}$ dans la région de Porcheville. La courbe verte correspond au diagramme de fiabilité issu des sous-ensembles calibrés prévus, en rouge au diagramme issu des sous-ensembles calibrés a posteriori et en bleu, celui de l'ensemble complet.

En ce qui concerne le second évènement, le diagramme de fiabilité prévu présente de moins bons résultats. Jusqu'à une probabilité produite aux alentours de 0.6, la courbe reste proche de la diagonale. Au-delà de cette valeur, la courbe chute. Ces sous-ensembles prévoient trop souvent des probabilités comprises entre $[0.6, 0.83]$ en comparaison de la fréquence relative de l'occurrence de l'évènement quand ces probabilités sont prévues.

Enfin, des champs de probabilité pour un évènement donné peuvent être calculés à partir des sous-ensembles calibrés a posteriori et prévu à une date donnée. Chaque champ de probabilité a été calculé à partir d'un sous-ensemble calibré sur le diagramme de fiabilité pour l'évènement donné. Dans chaque cellule du domaine et à chaque pas de temps, on compte le nombre de membres du sous-ensemble qui dépassent effectivement le seuil et on le divise par le nombre total de membres. On produit ainsi un champ de probabilité d'occurrence de l'évènement. Comme pour la prévision de l'incertitude, le champ de probabilité est moyenné sur un jour de prévision qui suit une période d'apprentissage. La figure 5.51 montre deux champs de probabilité a posteriori et prévu de l'évènement $[O_3] \geq 120\ \mu g\,m^{-3}$ à la date du 28 avril 2007.



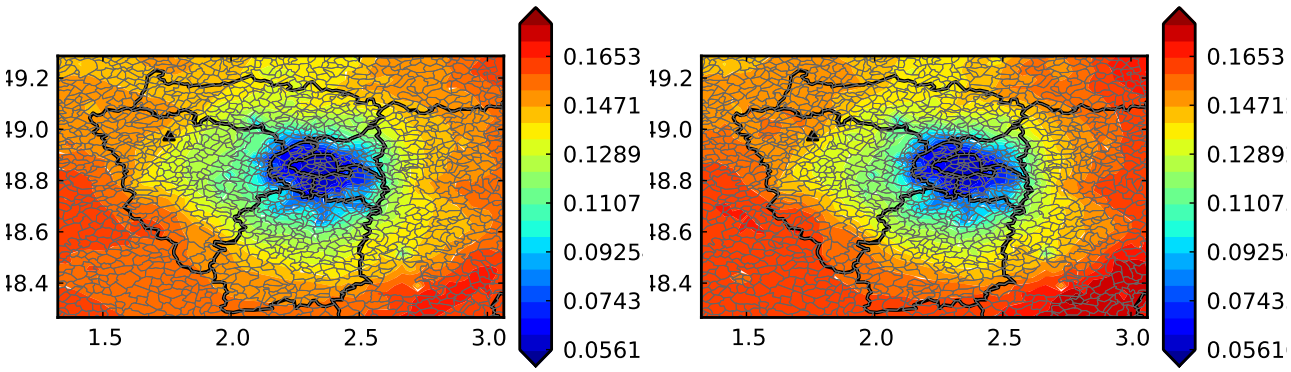
(a) A posteriori (b) Prévu

FIGURE 5.51 – Champs de probabilité a posteriori et prévu pour l'évènement $[O_3] \geq 120\ \mu g\,m^{-3}$ moyennes sur le jour de prévision, le 28 avril 2007, dans la région Île-de-France.

La figure 5.51 montre deux champs de probabilité assez proches. La structure spatiale d'un tel champ apporte une information intéressante. Les probabilités d'occurrence de l'évènement à cette date sont plus importantes dans le nord du domaine. Cependant, celles-ci n'excèdent pas 0.35.

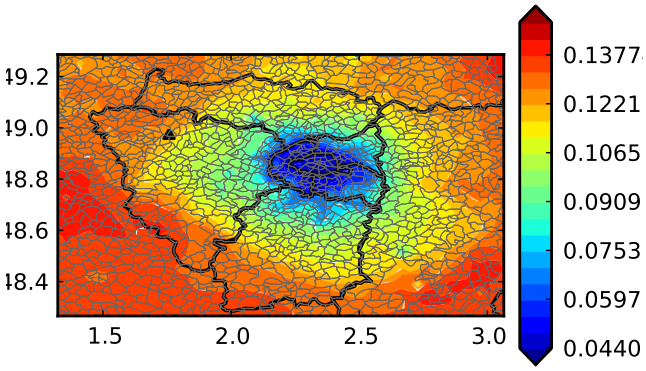
Nous décidons de calculer la moyenne des champs de probabilité sur les $J+i$ jours de prévision sur toute la période, soit 48 prévisions — i valant 1, 3 ou 5. L'évènement considéré ici est $[O_3] \geq 80\ \mu g\,m^{-3}$, car l'évènement $[O_3] \geq 120\ \mu g\,m^{-3}$ est assez peu fréquent, la moyenne des champs de probabilité est très proche de zéro en moyenne. La figure 5.52 présente les champs de probabilité a posteriori et prévu de l'occurrence de l'évènement $[O_3] \geq 80\ \mu g\,m^{-3}$ pour le premier, troisième et cinquième jours suivants les périodes d'apprentissage, moyennés sur toute la période d'étude.

Les champs prévus sont, en moyenne, assez proches des champs a posteriori. Les champs de probabilité issus de l'ensemble complet et du sous-ensemble aléatoire sont plus élevés que les champs issus des sous-ensembles calibrés. À titre d'exemple, les maxima atteints sont deux fois

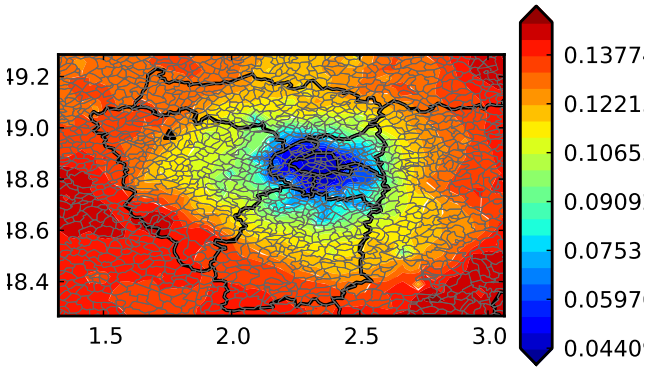


(a) A posteriori – day + 1

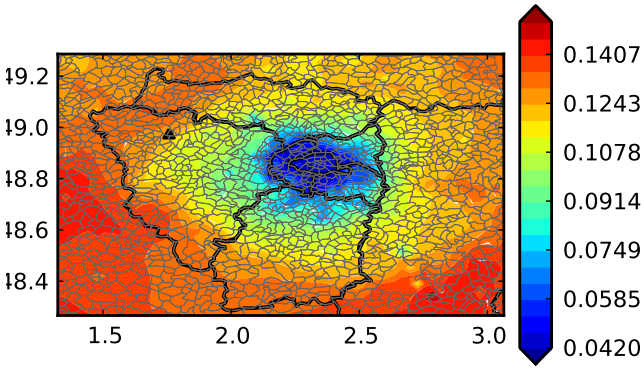
(b) Forecast – day + 1



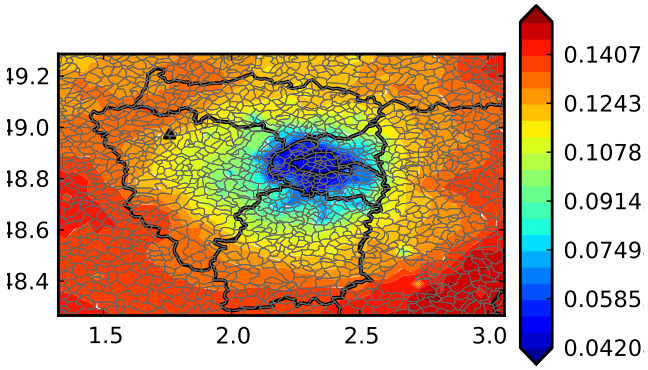
(c) A posteriori – day + 3



(d) Forecast – day + 3



(e) A posteriori – day + 5



(f) Forecast – day + 5

FIGURE 5.52 – Champs de probabilité a posteriori et prévu d'occurrence de l'évènement $[O_3] \geq 80\,\mu g\,m^{-3}$, moyennés sur toute la période, pour les 1er, 3^{e} et 5^{e} jours suivants les périodes d'apprentissage dans la région Île-de-France.

supérieurs pour le champ de probabilité de l'ensemble complet que pour le champ de probabilité prévu.

La section suivante traite de la prévision des risques de dépassement de seuil pour le dioxyde d'azote.

Dioxyde d'azote

Les tableaux 5.17 et 5.18 présentent les tableaux de contingence et les ratios associés pour les deux évènements suivants : $[NO_2] \geq 50\,\mu g\,m^{-3}$ et $\geq 100\,\mu g\,m^{-3}$. Ces résultats sont issus des 48 prévisions effectuées sur toute l'année 2007 dans la région Île-de-France. Chaque tableau contient les résultats pour l'ensemble complet, les sous-ensembles aléatoires, les sous-ensembles calibrés prévus et a posteriori. Le score utilisé pour la calibration est le *threat score*.

Ensemble	**hit**	**miss**	**correct neg.**	**false alarm**
Full	130203	56376	129301	92046
Random	128030	58549	130530	90817
Forecast	152561	34018	96131	125216
A posteriori	158084	28495	98230	123117

Ensemble	**accuracy**	**correct. neg. ratio**	**hit rate**	**threat score**
Full	0.636	0.584	0.698	0.467
Random	0.634	0.590	0.686	0.462
Forecast	0.610	0.434	0.818	0.489
A posteriori	0.628	0.444	0.847	0.510

TABLE 5.17 – Tableau de contingence et scores associés pour les prévisions cumulées de l'évènement $[NO_2] \geq 50\,\mu g\,m^{-3}$ dans la région de Porcheville. La période d'apprentissage est de deux semaines.

Pour le premier évènement, le plus fréquent des deux évènements, les conclusions sont identiques à celles pour le cas de l'ensemble d'ozone : le *threat score* prévu est proche du *threat score* a posteriori et légèrement plus élevé que le *threat score* de l'ensemble complet. Le taux de succès prévu dépasse les 80%. Cependant, les sous-ensembles calibrés prévus obtiennent encore une fois le nombre le plus élevé de fausses alarmes.

Le tableau 5.18 présente les résultats des prévisions de l'évènement $[NO_2] \geq 100\,\mu g\,m^{-3}$. Les sous-ensembles calibrés a posteriori et prévus, dont les taux de réussite sont de 37% et 33% respectivement, améliorent jusqu'à plus de trois fois le *threat score* de l'ensemble complet. Les performances restent cependant insuffisantes pour que ces ensembles soient utilisés pour prédire l'évènement.

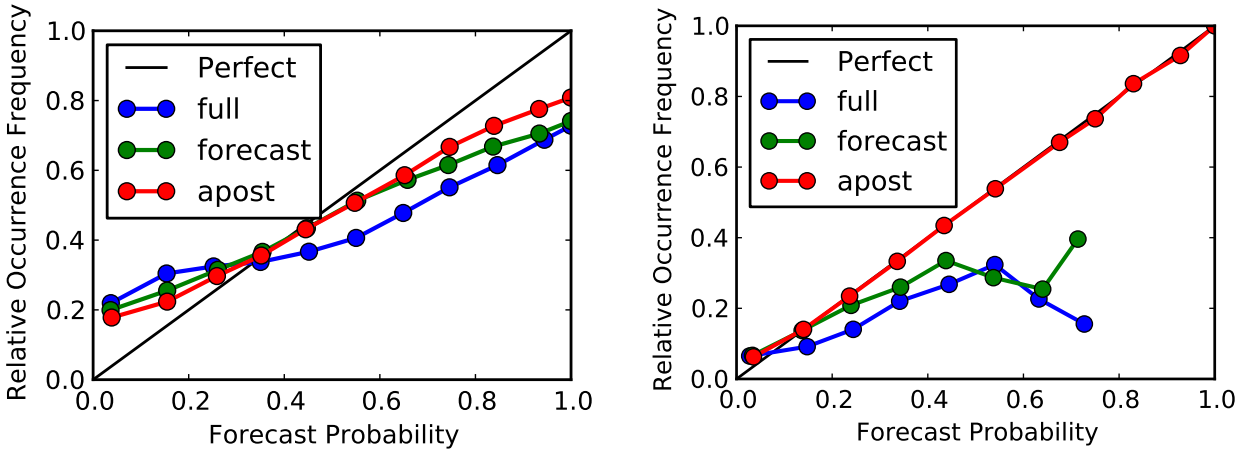
Les diagrammes de fiabilité pour les mêmes évènements sont présentés sur la figure 5.53. Les diagrammes associés à l'évènement $[NO_2] \geq 50\,\mu g\,m^{-3}$ sont pas perfectibles. Les diagrammes de fiabilité issus des ensembles calibrés sont cependant plus proches de la diagonale que celui issu de l'ensemble complet.

En ce qui concerne l'évènement $[NO_2] \geq 100\,\mu g\,m^{-3}$, les sous-ensembles calibrés prévus ne semblent pas être capable de prévoir correctement ledit évènement. Seuls les sous-ensembles calibrés a posteriori permettent d'obtenir un diagramme fiable.

Ensemble	hit	miss	correct neg.	false alarm
Full	1584	33962	368929	3451
Random	1637	33909	367934	4446
Forecast	11975	23571	326128	46252
A posteriori	13268	22278	328687	43693

Ensemble	accuracy	correct. neg. ratio	hit rate	threat score
Full	0.908	0.991	0.045	0.041
Random	0.906	0.988	0.046	0.041
Forecast	0.829	0.876	0.337	0.146
A posteriori	0.838	0.883	0.373	0.167

TABLE 5.18 – Tableau de contingence et scores associés pour les prévisions cumulées de l'évènement $[NO_2] \geq 100\,\mu g\,m^{-3}$ dans la région de Porcheville. La période d'apprentissage est de deux semaines.

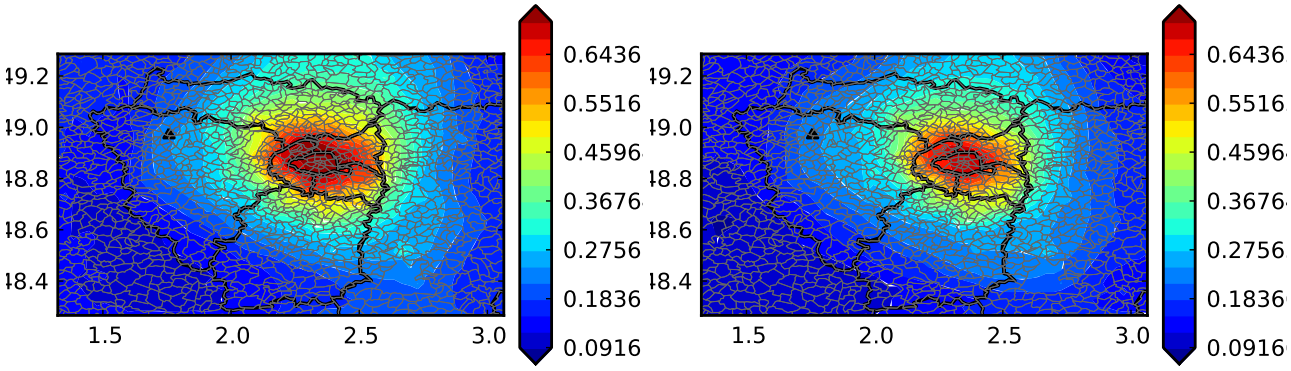


(a) $NO_2 \geq 50\,\mu g\,m^{-3}$ (b) $NO_2 \geq 100\,\mu g\,m^{-3}$

FIGURE 5.53 – Diagrammes de fiabilité cumulés pour les deux évènements $[NO_2] \geq 50\,\mu g\,m^{-3}$ et $\geq 100\,\mu g\,m^{-3}$ dans la région de Porcheville.

De la même manière que pour les diagrammes de fiabilité de l'ensemble d'ozone, le diagramme calculé à partir des sous-ensembles aléatoires, même s'il n'a pas été tracé ici, est très similaires au diagramme de l'ensemble complet.

Malgré les résultats mitigés concernant les diagrammes de fiabilité, des champs de probabilités ont été calculés à partie des sous-ensembles calibrés a posteriori et prévu. La figure 5.54 présente deux champs de probabilité de l'occurrence de l'évènement $[NO_2] \geq 50\,\mu g\,m^{-3}$ à la date du 27 octobre 2007. Les fortes valeurs, jusqu'à plus de 0.65, se situent au-dessus de Paris, où les émissions de NO_2 sont les plus élevées. On notera par ailleurs que la structure spatiale des champs sont assez proches des champs d'incertitude de NO_2. Les valeurs maximales pour le champ prévu sont légèrement plus faibles.



(a) A posteriori (b) Prévue

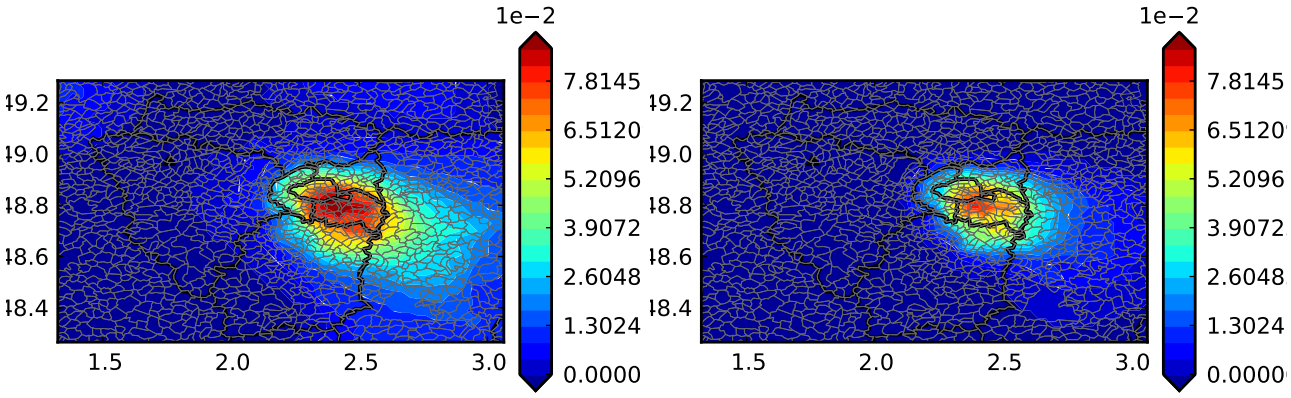
FIGURE 5.54 – Champs de probabilité a posteriori et prévu pour l'évènement $[NO_2] \geq 50\,\mu g\,m^{-3}$ à la date du 27 octobre 2007 dans la région Île-de-France.

La figure 5.55 présente deux champs de probabilité pour $[NO_2] \geq 100\,\mu g\,m^{-3}$ à la date du 27 janvier 2007. Les valeurs maximales des deux champs se situent au même endroit : au sud-est de la banlieue parisienne. Néanmoins, l'évènement étudié reste un évènement improbable à ce jour-ci et même les valeurs les plus élevées n'excèdent pas 0.1.

Pour conclure sur cette partie, quel que soit le polluant — ozone ou dioxyde de soufre — le sous-ensemble prévu donne toujours de meilleurs résultats que l'ensemble complet ou un sous-ensemble aléatoire. Malheureusement, les *threat scores* prévus ne sont pas à la hauteur de ceux calculés a posteriori. Comme dans la section 5.4.2, l'optimisation de ce score augmente à la fois le nombre de succès — tout en faisant diminuer le nombre d'échecs — et le nombre de fausses alarmes.

Les diagrammes de fiabilité d'ozone prévus montrent des résultats assez satisfaisants, mis à part les probabilités prévues de 0.6 ou plus, pour l'évènement $[O_3] \geq 120\,\mu g\,m^{-3}$. Au-delà de cette valeur, le sous-ensemble calibré prévu ne semble pas être capable de fournir de probabilité fiable. La prévision de risque de dépassement de seuil pour le NO_2 est plus difficile. Néanmoins, les résultats prévus restent globalement meilleurs que ceux calculés avec l'ensemble complet.

Un résultat important concerne l'estimation des champs de probabilité pour un évènement donné. L'apport d'un ensemble de simulations est ici non négligeable. Comme pour les champs



(a) A posteriori

(b) Prévue

FIGURE 5.55 – Champs de probabilité a posteriori et prévu pour l'évènement $[NO_2] \geq 100\ \mu g\,m^{-3}$, à la date du 27 janvier 2007 dans la région Île-de-France.

d'incertitude, les champs de probabilité ont une structure spatiale marquée et changent au cours du temps. La calibration du diagramme de fiabilité permet de prévoir « au mieux » les champs de probabilité d'occurrence d'un dépassement de seuil pour les jours à venir et restent assez proches des champs calculés a posteriori. Les faibles probabilités de dépassement de seuil pour l'ozone se situent près de zones d'émission, zones autour desquelles la concentration d'ozone est faible. Au contraire, le champ de probabilité d'occurrence de dépassement de seuil est plus élevé au-dessus de Paris et proche banlieue pour le NO_2. De plus, pour les évènements dont les seuils sont les plus élevés — $120\ \mu g\,m^{-3}$ et $100\ \mu g\,m^{-3}$ pour l'ozone et le NO_2 respectivement — les valeurs du champ de probabilité restent globalement assez faibles.

5.8 Conclusion

Ce cas d'application pour EDF R&D illustre l'utilisation d'ensemble de simulations photochimiques à l'échelle régionale. Des simulations d'ozone, de NO_2 et SO_2 ont été lancées dans les régions d'Île-de-France et des Pays de la Loire dans lesquelles se situent les centrales thermiques de Porcheville et de Cordemais.

Le premier objectif de ce chapitre a été d'étudier les résultats de ces ensembles de simulations comme pour l'ensemble photochimique à l'échelle continentale (chapitres 2 et 3) : méthode de génération d'ensemble afin de construire une centaine de simulations photochimiques, études des performances de l'ensemble en terme de scores d'ensemble — diagramme de rang, score de Brier, fiabilité, ... — et enfin la calibration associée à ces scores afin d'estimer l'incertitude et d'améliorer la fiabilité des prévisions de dépassement de seuil.

L'ensemble présente des simulations photochimiques aux performances statistiques et aux comportements très variés. Les performances pour l'ozone et le NO_2, telles que la corrélation et la RMSE, sont comparables à celles d'autres études effectuées dans la même région. Pour le SO_2, les simulations ont des difficultés à représenter correctement les concentrations, comme le montrent les comparaisons aux observations.

Le calcul des scores d'ensemble et la calibration d'ensemble ont été réalisés pour tous les polluants dans chacune des régions. Sans surprise, la calibration de l'ensemble de SO_2 ne permet pas de compenser les piètres résultats des modèles individuels. Au contraire, il nous a été possi-

ble de calibrer l'ensemble d'ozone et de NO_2 et ainsi de produire des cartes d'incertitude et des diagrammes de rang plus fiables que l'ensemble complet. L'incertitude sur l'ozone reste toujours assez faible au-dessus des sources d'émission, tandis que l'incertitude sur NO_2 y est plus élevée. Il est toujours difficile de produire des probabilités totalement fiables pour des évènements rares, même après calibration. Dans ce cas, l'optimisation semble sélectionner un sous-ensemble qui produit des probabilités fiables mais assez faibles.

Le fait d'avoir lancé les ensembles selon deux scénarios, avec et sans les émissions des centrales thermiques, a donné la possibilité d'effectuer une étude d'impact. Les différences absolues et relatives des champs de concentrations sont évidemment localisées près des centrales et restent souvent faibles, en comparaison des champs sans les émissions de la centrale. On note que l'impact des centrales est toujours plus faible que l'incertitude sur la concentration de fond elle-même. Cela ne signifie pas pour autant que l'impact calculé n'est pas fiable. En effet, l'impact simulé est lié à la réponse du modèle à un changement d'émissions, réponse dont l'incertitude n'est pas celle sur les concentrations. En conséquence, l'ensemble des différences entre les simulations avec et sans les émissions des centrales a été analysé. On constate alors que les incertitudes sur l'impact sont plus élevées que les impacts eux-mêmes. Il semble donc utile de reposer sur un ensemble pour mieux estimer l'impact potentiel des centrales. On peut fournir un impact estimé maximal en sommant l'impact moyen et (par exemple) deux fois l'écart type empirique.

Enfin, la robustesse spatiale et la robustesse temporelle ont pu être étudiées pour l'ensemble d'ozone et de NO_2 dans la région Île-de-France. Cette région a été choisie dans le but d'avoir un maximum de stations d'observation. Les résultats pour l'ozone sont assez satisfaisants. La robustesse spatiale et la prévision des dépassements de seuil pour le NO_2 semblent plus délicats. Cependant, l'estimation de l'incertitude via la calibration du diagramme de rang est assez robuste spatialement et reste globalement valable en prévision. Outre le fait d'estimer les incertitudes via le calcul de l'écart type de l'ensemble, il est parfois possible d'estimer les champs de probabilité d'occurrence d'évènement.

L'objectif à moyen terme est d'obtenir des prévisions d'incertitude et de probabilité de dépassement de seuil de manière opérationnelle. Ces résultats pourraient être couplés aux prévisions de la qualité de l'air et être une bonne source d'information pour l'aide à la décision.

Conclusions

Cette thèse traite de l'estimation des incertitudes et, par extension, de la prévision des risques de dépassement de seuil en qualité de l'air. De nombreuses sources d'erreur conduisent à des simulations ou prévisions de la qualité de l'air fortement entachées d'incertitudes. Ces sources sont (1) les données d'entrée, (2) la formulation physico-chimique du modèle et (3) les approximations numériques. Ces incertitudes peuvent être prises en compte via la construction d'un ensemble de simulations — en perturbant les données d'entrée dans le cadre de simulations Monte Carlo par exemple. Une autre méthode éprouvée durant cette thèse est la génération d'ensembles dits « multi-modèles » qui prend en compte les trois sources d'incertitude précédemment mentionnées.

Outre les incertitudes liées aux modèles de chimie-transport, un ensemble de simulations permet d'améliorer les prévisions de la qualité de l'air. On utilise pour cela des méthodes d'agrégation de prévisions Pagowski *et al.* [2006]; Mallet *et al.* [2009], éventuellement couplées avec l'assimilation de données Mallet [2010]. Par ailleurs, un ensemble de simulations de la qualité de l'air peut, contrairement à une unique simulation déterministe, fournir des prévisions probabilistes. Autrement dit, il a la capacité de produire des probabilités pour l'occurrence d'un évènement tel que le dépassement de seuil de concentration réglementaire.

La première étape importante consiste donc à générer un ensemble qui puisse considérer toutes les sources d'incertitude. Le chapitre 2 décrit une méthode originale de génération automatique d'ensemble de qualité de l'air. Cette méthode permet de prendre en compte toutes les sources d'incertitude : données d'entrée, paramétrisations physiques et approximations numériques. Néanmoins, rien ne garantit que l'ensemble ainsi créé soit représentatif des incertitudes de la cible étudiée (la concentration d'ozone au voisinage du sol par exemple). C'est pourquoi, dans un second temps, il est indispensable de mesurer les performances de l'ensemble, par une comparaison aux observations, et de calibrer ce dernier afin qu'il soit véritablement « représentatif » des incertitudes.

Génération et calibration d'ensemble

Génération automatique d'ensemble Une méthode originale de génération automatique d'ensemble a été mise au point pendant ce travail de thèse et a donné lieu à une publication [Garaud et Mallet, 2010] (voir le chapitre 3). L'objectif est de prendre en considération les trois sources d'incertitude que sont les données d'entrée, les paramétrisations physiques et les approximations numériques dans le cadre de simulations photochimiques. La plate-forme logicielle Polyphemus permet de construire différentes simulations grâce à sa modularité. Au vu du grand nombre d'alternatives possibles dans la construction d'une simulation photochimique, il est indispensable d'automatiser cette tâche. Ainsi, la méthode consiste à sélectionner de manière aléatoire un ensemble de simulations parmi toutes les possibilités, en fixant arbitrairement des poids aux paramétrisations physiques et approximations numériques, tout en perturbant un certain nombre de données d'entrée.

C'est ainsi qu'un ensemble d'une centaine de simulations photochimiques à l'échelle conti-

nentale a été créé puis lancé pour l'Europe pour toute l'année 2001. Le polluant étudié dans cette partie est l'ozone. L'ensemble dénote d'une grande diversité de modèles : performances très différentes — RMSE, corrélation, moyenne et écart type — cartes des meilleurs modèles qui varie fortement en temps et en espace, dispersion importante de l'ensemble, ...

Cette méthode de génération d'ensemble est de nouveau appliquée autour de deux centrales EDF pour une étude d'impact et une estimation des incertitudes sur l'année 2007.

Néanmoins, la génération de tels ensembles ne garantit pas que ces derniers soient à la fois représentatifs des incertitudes ou même fiables pour prévoir l'occurrence d'un évènement. C'est pourquoi il est nécessaire de mesurer les performances de l'ensemble pour ensuite le calibrer.

Scores d'ensemble et calibration Plusieurs scores d'ensemble ont été abordés durant cette thèse. Ces derniers permettent de mesurer la cohérence entre un ensemble de simulations et une cible observée. Deux critères essentiels sont à retenir : fiabilité et résolution. Des indicateurs tels que le diagramme de fiabilité et le score de Brier sont utilisés pour mesurer la performance des probabilités produites par le système, dans le cadre de la prévision d'un dépassement de seuil. Le diagramme de rang, quant à lui, compare les observations à l'estimation de la distribution des concentrations que fournit un ensemble.

Les résultats des ensembles complets sont soit raisonnables mais perfectibles (pour l'ozone), soit insuffisants (dioxyde d'azote ou dioxyde de soufre). La plupart des diagrammes de rang montrent que les ensembles ne sont pas assez dispersés — diagramme de rang en forme de « U » ou de « L » — et montrent parfois un biais systématique. Pour l'ozone, les ensembles surestiment souvent les concentrations observées. De plus, les ensembles ne sont pas parfaitement fiables comme le montrent les différents diagrammes de fiabilité et les scores de Brier des ensembles complets. Même s'ils s'avèrent la plupart du temps meilleurs que la prévision climatologique ou que le meilleur modèle, ils peuvent être considérablement améliorés à l'aide de la calibration d'ensemble.

La calibration d'ensemble est un moyen de sélectionner un sous-ensemble de l'ensemble complet qui minimise ou maximise un critère donné. Ce critère est bien sûr choisi parmi les scores d'ensemble mentionnés précédemment : maximisation du score *skill* de Brier, minimisation de la variance du diagramme de rang ou encore la minimisation de la différence quadratique entre un diagramme parfaitement fiable et le diagramme de fiabilité de l'ensemble. La sélection d'un certain nombre de membres parmi les membres de l'ensemble selon un critère donné est un problème d'optimisation combinatoire. Des algorithmes dits « évolutionnaires » permettent de trouver des minima locaux et ainsi sélectionner un sous-ensemble qui minimise (ou maximise) un critère. Les deux algorithmes utilisés sont un algorithme génétique et le recuit simulé.

Estimation de l'incertitude et prévision des risques

Incertitudes Une mesure de l'incertitude beaucoup employée dans cette thèse est l'écart type empirique d'un ensemble de simulations. L'ensemble est de préférence calibré pour être représentatif de l'incertitude. La variance du diagramme de rang est le critère privilégié dans cette thèse lorsque l'objectif de la calibration est l'estimation d'incertitude.

Les champs d'incertitude évoluent dans le temps et présentent des structures spatiales qui apportent des informations intéressantes. Dans le cas de l'ozone à l'échelle européenne, par exemple, les incertitudes sont élevées au sud de l'Europe et près des côtes — au sud de l'Espagne,

le long de la côte italienne. L'incertitude *relative* de l'ozone montre quant à elle que la part d'incertitude près des sources d'émissions — zone de titration d'ozone par les émissions de monoxyde d'azote — est assez élevée par rapport à la faible concentration moyenne d'ozone dans ces mêmes zones. Au contraire, les cartes d'incertitude du NO_2, dans le domaine Île-de-France, montrent que l'incertitude se situe au niveau des sources d'émission.

Le chapitre 4 montre l'impact des incertitudes liées aux données d'entrée sur les concentrations d'ozone. Une régression linéaire a été effectuée entre les perturbations des champs d'entrée et les simulations d'ozone issues de deux ensembles Monte Carlo et multi-modèles. Dans les deux cas, les champs des conditions aux limites d'ozone et du taux de photolyse du NO_2 montrent un coefficient de régression assez élevé.

Ensuite, on a pu montrer les robustesses spatiale et temporelle de la méthode de calibration d'ensemble dans le cadre de l'estimation de l'incertitude. Les chapitres 3 et 5 vérifient ces deux aspects pour des ensembles de simulations photochimiques à l'échelle continentale et régionale. La robustesse spatiale a été vérifiée grâce à une validation croisée. Un ensemble, calibré avec un score particulier, reste bon sur des points d'observation qui n'ont pas servi à sa calibration. La robustesse temporelle montre que la calibration peut être utilisée dans le cadre de prévision de champs d'incertitude à courte échéance. On peut donc envisager de fournir, avec les prévisions opérationnelles classiques, une bonne estimation de leur incertitude. Le chapitre 3 montre que l'erreur de mesure n'a pas d'impact important sur la calibration et l'estimation de l'incertitude. Concernant l'étude d'impact des centrales thermiques du chapitre 5, l'apport de l'ensemble dans les deux scénarios d'émission a permis d'estimer l'incertitude liée à l'impact présumé des centrales.

Il existe néanmoins des limitations à la méthode de calibration d'ensemble. La calibration du diagramme de rang peut être contrainte par la forme du diagramme de l'ensemble initial. Il est possible qu'un ensemble trop peu dispersé — des observations trop nombreuses en dehors de l'enveloppe de l'ensemble — ne puisse pas être correctement calibré. C'est le cas rencontré avec SO_2 à l'échelle régionale (chapitre 5) pour qui l'ensemble n'est pas assez dispersé et présente un biais fort. La calibration de cet ensemble est incapable de sélectionner plus d'un membre.

Prévision des risques La prévision des risques en qualité de l'air est un enjeu à la fois économique et sanitaire. Des seuils réglementaires ont été fixés pour l'ozone, le dioxyde d'azote et le dioxyde de soufre. Un ensemble de prévisions a la capacité de fournir des prévisions probabilistes d'occurrence de dépassement de seuil. La calibration du diagramme de fiabilité est un bon moyen de fournir un système dit « fiable » afin de produire des probabilités d'occurrence d'un évènement. Une difficulté réside dans la prévision d'évènements rares. La faible occurrence d'évènement tel que $[O_3] \geq 180\,\mu g\,m^{-3}$ rend difficile à la fois la mesure de la fiabilité via le diagramme de fiabilité et par conséquent la calibration de l'ensemble. Globalement, l'algorithme aura tendance à sélectionner un sous-ensemble qui produit de faibles probabilités mais qui reste fiable.

La robustesse spatiale et temporelle d'une telle calibration a aussi été réalisée dans les chapitres 3 et 5. Les diagrammes de fiabilité calculés sur un réseau d'observation qui n'a pas servi à la calibration sont satisfaisants puisqu'ils présentent une forme proche des diagrammes de fiabilité calibrés sur ce même réseaux. Néanmoins, dans un cadre prévisionnel, il semble plus difficile de fournir de bonnes prévisions probabilistes grâce à la calibration du diagramme de fiabilité que d'obtenir de bonnes prévisions d'incertitude.

Il est aussi pertinent de prévoir l'occurrence ou non d'un évènement en particulier — et non plus une probabilité. C'est pourquoi la dernière partie aborde la calibration du tableau de contingence qui mesure les taux de réussite, les fausses alarmes, le *threat score*, ... d'un évènement. Il

m'a paru judicieux, dans le cadre de prévision des risques de la qualité de l'air, de maximiser le *threat score* qui tient compte à la fois du taux de réussite et des fausses alarmes — qui peuvent avoir un coût économique non négligeable. La méthode de calibration améliore globalement le *threat score* pour l'ozone et le dioxyde d'azote — comparé à l'ensemble complet, un sous-ensemble aléatoire ou au meilleur modèle. Les robustesses spatiales et temporelles ont même été étudiées dans ce cadre. Cependant, nous avons pu observer qu'il était difficile de sélectionner un sous-ensemble qui détecte tous les dépassements de seuil sans augmenter de manière parfois significative le nombre de fausses alarmes. De plus, dans le cas d'évènements rares, la méthode de calibration ne permet parfois pas d'améliorer le *threat score*.

Enfin, un ensemble de prévisions peut fournir un champ tridimensionnel de probabilité d'occurrence d'un évènement. Il est possible d'étudier l'évolution temporelle d'un tel champ dans une région donnée. Dans le chapitre 5, des champs de probabilité sont calculés pour l'évènement $[O_3] \geq 80\,\mu g\,m^{-3}$ dans le domaine Île-de-France à différentes dates. À l'inverse du NO_2, les probabilités restent faibles près des zones d'émissions où les concentrations d'ozone sont moins importantes.

Covariance d'erreur et décomposition La calibration d'ensemble pour l'estimation de l'incertitude rend aussi possible l'estimation des covariances d'erreur. Le chapitre 4 fournit deux exemples de champs de covariance pour deux ensembles calibrés : Monte Carlo et multi-modèles. Le champ de covariance estimé par l'ensemble Monte Carlo semble moins réaliste. On peut observer en effet un fort impact des perturbations liées aux conditions aux limites d'ozone, et un problème de localisation de la covariance (trop élevée à grande distance). Concernant l'ensemble multi-modèles, on peut noter que les champs de covariance dépendent fortement du point de grille de référence à partir duquel ils sont calculés. En effet, si le point de grille de référence se trouve en zone d'émission ou en zone dite « de fond » (*background* en anglais), la structure spatiale du champ sera très différente et la distance de décorrélation sera plus importante dans le second cas. Ceci est dû à une échelle de représentativité différente selon la localisation du point de calcul.

En ce qui concerne la représentativité, il a été possible, dans ce même chapitre, d'estimer la part de variance d'erreur en fonction du type d'erreur : erreur de mesure, de représentativité ou de modélisation. Les données d'Airparif concernant les incertitudes liées aux mesures de concentration d'ozone ont permis de retirer la variance de l'erreur de mesure. Ensuite, deux méthodes indépendantes ont été utilisées dans le but d'estimer la variance de l'erreur de représentativité. Ces méthodes donnent des résultats assez similaires et permettent d'estimer une part de la variance de l'erreur de représentativité autour d'un tiers, tandis que la variance de l'erreur de modélisation se trouve à environ 63% dans le cas étudié.

Perspectives

Au regard de ce qui a été réalisé durant ce travail de doctorat, des prolongements et de nouveaux axes de recherche peuvent être identifiés.

Dans un premier temps, il serait utile de prendre en compte les récents changements disponibles dans le système de modélisation de la qualité de l'air Polyphemus, utilisé tout le long de cette thèse. Je pense notamment aux nouveaux mécanismes chimiques RACM 2 et CB05 récemment intégrés à Polyphemus. Cela donnerait la possibilité d'améliorer significativement la variabilité des modèles et ainsi la dispersion de l'ensemble.

Dans le même ordre d'idée, la méthode de génération automatique d'ensemble n'a pas pour vocation à rester confinée aux simulations photochimiques. La génération d'un grand nombre de

simulations avec aérosols peut être réalisée avec la même méthode. Polyphemus propose beaucoup d'options physiques et numériques dédiées à la dispersion et à la dynamique des aérosols. La mise à disposition d'un tel ensemble permettrait aussi d'appliquer la calibration d'ensemble, l'estimation de l'incertitude et la prévision de dépassement de seuil pour les particules fines.

Une amélioration dans la construction de l'ensemble serait, au lieu de perturber les champs météorologiques, de reposer sur un ensemble météorologique. Le centre européen de météorologie [4] fournit par exemple un ensemble de 50 membres, y compris en prévision. Les champs météorologiques de chaque membre conservent une cohérence physique, et les incertitudes sur ces champs sont bien entendu nettement mieux décrites que par une perturbation a posteriori.

On a pu montrer que la méthode de calibration d'ensemble permettait de prévoir l'incertitude à courte échéance et améliorait les prévisions de dépassement de seuil — dans le cas où le seuil n'est pas trop élevé pour qu'il y ait suffisamment d'occurrences de l'évènement dans les observations. De plus, la robustesse spatiale de la méthode de calibration rend l'estimation de l'incertitude et la prévision des risques exploitables aux endroits qui ne sont pas observés. Ainsi, il serait intéressant de fournir de telles données (champs d'incertitude et de probabilités) de manière opérationnelle ou en analyse a posteriori de l'exposition des populations.

Les champs de covariance estimés grâce aux ensembles calibrés peuvent servir en assimilation de données. En effet, les méthodes d'assimilation utilisent des estimations des matrices de covariance d'erreur sur l'état et d'erreur modèle. Concernant l'erreur modèle, un travail spécifique autour des ensembles devrait être mené, puisqu'il s'agit d'estimer l'erreur introduite à chaque pas de temps d'intégration.

Enfin, une des limites de la méthode de calibration d'ensemble concerne la prévision des dépassements de seuils réglementaires pour l'ozone, le NO_2 et SO_2. Il est en effet difficile de produire des probabilités fiables pour ces évènements rares. Des voies d'amélioration se trouveraient éventuellement dans les méthodes d'apprentissage statistique (classification) ou les méthodes statistiques dédiées à la prévision d'évènements rares. Des méthodes bayésiennes, pondérant chaque membre de l'ensemble, peuvent aussi être une voie complémentaire à la méthode de calibration d'ensemble.

4. ECMWF

Bibliographie

AIRPARIF (2007). Guide pratique d'utilisation pour l'estimation de l'incertitude de mesure des concentrations en polluants dans l'air ambiant. Rapport technique Version 9, AIRPARIF.

ANDERSON, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, 9(7):1518–1530.

ANDERSSON, C., LANGNER, J. et BERGSTRÖM, R. (2007). Interannual variation and trends in air pollution over Europe due to climate variability during 1958–2001 simulated with a regional CTM coupled to the ERA40 reanalysis. *Tellus B*, 59(1):77–98.

BEEKMANN, M. et DEROGNAT, C. (2003). Monte Carlo uncertainty analysis of a regional-scale transport chemistry model constrained by measurements from the atmospheric pollution over the Paris area (ESQUIF) campaign. *Journal of Geophysical Research*, 108(D17):8,559.

BOUTAHAR, J., LACOUR, S., MALLET, V., QUÉLO, D., ROUSTAN, Y. et SPORTISSE, B. (2004). Development and validation of a fully modular platform for numerical modelling of air pollution : POLAIR. *International Journal of Environment and Pollution*, 22(1/2):17–28.

BOUTTIER, F. et COURTIER, P. (1999). Data assimilation concepts and methods. Meteorological training course lecture series, ECMWF.

BOYNARD, A., BEEKMANN, M., FORÊT, G., UNG, A., SZOPA, S., SCHMECHTIG, C. et COMAN, A. (2011). An ensemble assessment of regional ozone model uncertainty with an explicit error representation. *Atmospheric Environment*, 45(3):784–793.

BRIER, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

BUILTJES, P. (1992). The LOTOS – LOng Term Ozone Simulation – project, summary report. Rapport technique R92/240, TNO, Delft, the Netherlands.

BYUN, D. W. et CHING, J. K. S., éditeurs (1999). *Science algorithms of the EPA models-3 community multiscale air quality (CMAQ) modeling system*. U.S. Environmental Protection Agency, Washington.

CANDILLE, G. et TALAGRAND, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, 131:2,131–2,150.

CHANG, J., BROST, R., ISAKEN, I., MADRONICH, S., MIDDLETON, P., STOCKWELL, W. et WALCEK, C. (1987). A three-dimensional Eulerian acid deposition model : physical concepts and formulation. *Journal of Geophysical Research*, 92(D12):14,681–14,700.

CROSBY, J. L. (1973). *Computer Simulation in Genetics*. Wiley.

DEBRY, E., FAHEY, K., SARTELET, K., SPORTISSE, B. et TOMBETTE, M. (2007). Technical Note : A new SIze REsolved Aerosol Model (SIREAM). *Atmospheric Chemistry and Physics*, 7:1,537–1,547.

DELLE MONACHE, L., DENG, X., ZHOU, Y. et STULL, R. B. (2006a). Ozone ensemble forecasts : 1. A new ensemble design. *Journal of Geophysical Research*, 111(D05307).

DELLE MONACHE, L., HACKER, J. P., ZHOU, Y., DENG, X. et STULL, R. B. (2006b). Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *Journal of Geophysical Research*, 111(D24307).

DELLE MONACHE, L. et STULL, R. B. (2003). An ensemble air-quality forecast over western Europe during an ozone episode. *Atmospheric Environment*, 37:3,469–3,474.

DRAXLER, R. R. (2003). Evaluation of an ensemble dispersion calculation. *Journal of Applied Meteorology and Climatology*, 42:308–317.

EPSTEIN, E. S. (1969a). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology and Climatology*, 8(6):985–987.

EPSTEIN, E. S. (1969b). Stochastic dynamic prediction. *Tellus*, 21(7):739–759.

ESQUIF (2001). Étude et simulation de la qualité de l'air en île de France – rapport final.

EVENSEN, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99(C5):10,143–10,162.

FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521.

FISHER, R. A. (1921). On the "probable error" of a coefficient of a correlation deduced from a small sample. *Metron*, 1(4):3–32.

FRASER, A. et BURNELL, D. (1970). *Computer Models in Genetics*. McGraw-Hill.

GALMARINI, S., BIANCONI, R., ADDIS, R., ANDRONOPOULOS, S., ASTRUP, P., BARTZIS, J. C., BELLASIO, R., BUCKLEY, R., CHAMPION, H., CHINO, M. *et al.* (2004a). Ensemble dispersion forecasting – part II : application and evaluation. *Atmospheric Environment*, 38(28):4,619–4,632.

GALMARINI, S., BIANCONI, R., KLUG, W., MIKKELSEN, T., ADDIS, R., ANDRONOPOULOS, S., ASTRUP, P., BAKLANOV, A., BARTNIKI, J., BARTZIS, J. C. *et al.* (2004b). Ensemble dispersion forecasting – part I : concept, approach and indicators. *Atmospheric Environment*, 38(28): 4,607–4,617.

GARAUD, D. et MALLET, V. (2010). Automatic generation of large ensembles for air quality forecasting using the Polyphemus system. *Geoscientific Model Development*, 3(1):69–85.

GARAUD, D. et MALLET, V. (2011). Automatic calibration of an ensemble for uncertainty estimation and probabilistic forecast : Application to air quality. *Journal of Geophysical Research*, 116(D19304).

GARAUD, D. et MALLET, V. (2012). Uncertainty estimation and decomposition based on Monte Carlo and multimodel photochemical simulations. Rapport technique RR-7903, INRIA.

GROSS, A. et STOCKWELL, W. R. (2003). Comparison of the EMEP, RADM2 and RACM mechanisms. *Journal of Atmospheric Chemistry*, 44:151–170.

HAMILL, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560.

HAMILL, T. M. et COLUCCI, S. J. (1997). Verification of Eta/RSM short-range ensemble forecasts. *Monthly Weather Review*, 125:1312–1327.

HANNA, S. R., CHANG, J. C. et FERNAU, M. E. (1998). Monte Carlo estimates of uncertainties in predictions by a photochemical grid model (UAM-IV) due to uncertainties in input variables. *Atmospheric Environment*, 32(21):3,619–3,628.

HANNA, S. R., LU, Z., FREY, H. C., WHEELER, N., VUKOVICH, J., ARUNACHALAM, S., FERNAU, M. et HANSEN, D. A. (2001). Uncertainties in predicted ozone concentrations due to input uncertainties for the UAM-V photochemical grid model applied to the July 1995 OTAG domain. *Atmospheric Environment*, 35(5):891–903.

HOGREFE, C., RAO, S. T., KASIBHATLA, P., HAO, W., SISTLA, G., MATHUR, R. et MCHENRY, J. (2001). Evaluating the performance of regional-scale photochemical modeling systems : Part II – ozone predictions. *Atmospheric Environment*, 35:4,159–4,174.

HOLLINGSWORTH, A. et LÖNNBERG, P. (1986). The statistical structure of short-range forecast errors as determined from radiosonde data. Part I : the wind field. *Tellus*, 38A:111–136.

HOPSON, T. M. et WEBSTER, P. J. (2010). A 1–10-day ensemble forecasting scheme for the major river basins of bangladesh : Forecasting severe floods of 2003–07*. *Journal of Hydrometeorology*, 11(3):618–641.

HOROWITZ, L. W., WALTERS, S., MAUZERALL, D. L., EMMONS, L. K., RASCH, P. J., GRANIER, C., TIE, X., LAMARQUE, J.-F., SCHULTZ, M. G., TYNDALL, G. S., ORLANDO, J. J. et BRASSEUR, G. P. (2003). A global simulation of tropospheric ozone and related tracers : description and evaluation of MOZART, version 2. *Journal of Geophysical Research*, 108(D24).

KIRKPATRICK, S., GELATT, C. D. et VECCHI, M. P. (1983). Optimization by simulating annealing. *Science*, 220(4598):671–680.

KROL, M., HOUWELING, S., BREGMAN, B., van den BROEK, M., SEGERS, A., van VELTHOVEN, P., PETERS, W., DENTENER, F. et BERGAMASCHI, P. (2005). The two-way nested global chemistry-transport zoom model tm5 : algorithm and applications. *Atmospheric Chemistry and Physics*, 5(2):417–432.

LANSER, D. et VERWER, J. G. (1999). Analysis of operator splitting for advection-diffusion-reaction problems from air pollution modelling. *Journal of Computational and Applied Mathematics*, 111:201–216.

LATTUATI, M. (1997). *Contribution à l'étude du bilan de l'ozone troposphérique à l'interface de l'Europe et de l'Atlantique Nord : Modélisation lagrangienne et mesures en altidude*. Thèse de doctorat, Université Paris 6.

LORENZ, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20:131–140.

LOUIS, J.-F. (1979). A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteorology*, 17:187–202.

MADRONICH, S. (1987). Photodissociation in the atmosphere : 1. actinic flux and the effects of ground reflections and clouds. *Journal of Geophysical Research*, 92(D8):9,740–9,752.

MALLET, V. (2005). *Estimation de l'incertitude et prévision d'ensemble avec un modèle de chimie-transport – Application à la simulation numérique de la qualité de l'air*. Thèse de doctorat, École nationale des ponts et chaussées.

MALLET, V. (2010). Ensemble forecast of analyses : Coupling data assimilation and sequential aggregation. *Journal of Geophysical Research*, 115(D24303).

MALLET, V., POURCHET, A., QUÉLO, D. et SPORTISSE, B. (2007a). Investigation of some numerical issues in a chemistry-transport model : Gas-phase simulations. *Journal of Geophysical Research*, 112(D15).

MALLET, V., QUÉLO, D., SPORTISSE, B., Ahmed de BIASI, M., DEBRY, É., KORSAKISSOK, I., WU, L., ROUSTAN, Y., SARTELET, K., TOMBETTE, M. et FOUDHIL, H. (2007b). Technical Note : The air quality modeling system Polyphemus. *Atmospheric Chemistry and Physics*, 7(20):5,479–5,487.

MALLET, V. et SPORTISSE, B. (2006a). Ensemble-based air quality forecasts : A multimodel approach applied to ozone. *Journal of Geophysical Research*, 111(D18).

MALLET, V. et SPORTISSE, B. (2006b). Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations : An ensemble approach applied to ozone modeling. *Journal of Geophysical Research*, 111(D1).

MALLET, V., STOLTZ, G. et MAURICETTE, B. (2009). Ozone ensemble forecast with machine learning algorithms. *Journal of Geophysical Research*, 114(D05307).

MARTIEN, P. T. et HARLEY, R. A. (2006). Adjoint sensitivity analysis for a three-dimensional photochemical model : application to Southern California. *Environmental Science & Technology*, 40(13):4,200–4,210.

MCKEEN, S., CHUNG, S. H., WILCZAK, J., GRELL, G., DJALALOVA, I., PECKHAM, S., GONG, W., BOUCHET, V., MOFFET, R., TANG, Y., CARMICHAEL, G. R., MATHUR, R. et YU, S. (2007). Evaluation of several $PM_{2.5}$ forecast models using data collected during the ICARTT/NEAQS 2004 field study. *Journal of Geophysical Research*, 112(D10S20).

MCKEEN, S., WILCZAK, J., GRELL, G., DJALALOVA, I., PECKHAM, S., HSIE, E.-Y., GONG, W., BOUCHET, V., MENARD, S., MOFFET, R., MCHENRY, J., MCQUEEN, J., TANG, Y., CARMICHAEL, G. R., PAGOWSKI, M., CHAN, A., DYE, T., FROST, G., LEE, P. et MATHUR, R. (2005). Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. *Journal of Geophysical Research*, 110(D21).

MCKEEN, S. A., HSIE, E.-Y., TRAINER, M., TALLAMRAJU, R. et LIU, S. C. (1991). A regional model study of the ozone budget in the eastern United States. *Journal of Geophysical Research*, 96(D6):10,809–10,845.

MONAHAN, E. C., SPIEL, D. E. et DAVIDSON, K. L. (1986). *Oceanic Whitecaps – and Their Role in Air-Sea Exchange Processes*, chapitre A model of marine aerosol generation via whitecaps and wave disruption, pages 167–174. Kluwer Academic.

MURPHY, A. H. (1971). A note on the ranked probability score. *Journal of Applied Meteorology and Climatology*, 10(2):155–156.

MURPHY, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600.

PAGOWSKI, M., GRELL, G., DEVENYI, D., PECKHAM, S., MCKEEN, S., GONG, W., DELLE MONACHE, L., MCHENRY, J., MCQUEEN, J. et LEE, P. (2006). Application of dynamic linear regression to improve the skill of ensemble-based deterministic ozone forecasts. *Atmospheric Environment*, 40:3,240–3,250.

PAGOWSKI, M., GRELL, G. A., MCKEEN, S. A., DÉVÉNYI, D., WILCZAK, J. M., BOUCHET, V., GONG, W., MCHENRY, J., PECKHAM, S., MCQUEEN, J., MOFFET, R. et TANG, Y. (2005). A simple method to improve ensemble-based ozone forecasts. *Geophysical Research Letters*, 32.

PINDER, R. W., GILLIAM, R. C., APPEL, K. W., NAPELENOK, S. L., FOLEY, K. M. et GILLILAND, A. B. (2009). Efficient probabilistic estimates of surface ozone concentration using an ensemble of model configurations and direct sensitivity calculations. *Environmental Science & Technology*, 43(7):2,388–2,393.

QUÉLO, D., MALLET, V. et SPORTISSE, B. (2005). Inverse modeling of NO_x emissions at regional scale over northern France : Preliminary investigation of the second-order sensitivity. *Journal of Geophysical Research*, 110(D24).

RUSSELL, A. et DENNIS, R. (2000). NARSTO critical review of photochemical models and modeling. *Atmospheric Environment*, 34:2,283–2,234.

SANDU, A., VERWER, J. G., BLOM, J. G., SPEE, E. J., CARMICHAEL, G. R. et POTRA, F. A. (1997a). Benchmarking stiff ODE solvers for atmospheric chemistry problems II : Rosenbrock solvers. *Atmospheric Environment*, 31(20):3,459–3,472.

SANDU, A., VERWER, J. G., van LOON, M., CARMICHAEL, G. R., POTRA, F. A., DABDUB, D. et SEINFELD, J. H. (1997b). Benchmarking stiff ODE solvers for atmospheric chemistry problems-I. Implicit vs explicit. *Atmospheric Environment*, 31(19):3,151–3,166.

SARTELET, K., HAYAMI, H. et SPORTISSE, B. (2008). MICS Asia Phase II–Sensitivity to the aerosol module. *Atmospheric Environment*, 42(15):3,562–3,570.

SCHMIDT, H., DEROGNAT, C., VAUTARD, R. et BEEKMANN, M. (2001). A comparison of simulated and observed ozone mixing ratios for the summer of 1998 in Western Europe. *Atmospheric Environment*, 35:6,277–6,297.

SEGERS, A. (2002). *Data assimilation in atmospheric chemistry models using Kalman filtering*. Thèse de doctorat, Delft University.

SIMPSON, D. (1992). Long-period modelling of photochemical oxidants in Europe. Model calculations for July 1985. *Atmospheric Environment*, 26(9):1,609–1,634.

SIMPSON, D., WINIWARTER, W., BÖRJESSON, G., CINDERBY, S., FERREIRO, A., GUENTHER, A., HEWITT, C. N., JANSON, R., KHALIL, M. A. K., OWEN, S., PIERCE, T. E., PUXBAUM, H., SHEARER, M., SKIBA, U., STEINBRECHER, R., TARRASÓN, L. et ÖQUIST, M. G. (1999). Inventorying emissions from nature in Europe. *Journal of Geophysical Research*, 104(D7):8,113–8,152.

SMITH, M. et HARRISON, N. (1998). The sea spray generation function. *Journal of Aerosol Science*, 29:189–190.

SPORTISSE, B. (2000). An analysis of operator splitting techniques in the stiff case. *Journal of Computational Physics*, 161(1):140–168.

SPORTISSE, B. (2008). *Pollution atmosphérique. Des processus à la modélisation*. Springer.

SPORTISSE, B. et MALLET, V. (2005 & 2006). Calcul scientifique pour l'environnement. Cours de deuxième année à l'ENSTA.

STERN, R. (1994). *Entwicklung und Anwendung eines dreidmensionalen photochemischen Ausbreitungsmodells mit verschiedenen chemischen Mechanismen*. Thèse de doctorat, Freie Universität Berlin.

STERN, R., YAMARTINO, R. et GRAFF, A. (2003). Dispersion modelling within the European Community's air quality directives : Long term modelling of O3, PM10 and NO2. *In 26th ITM on Air Pollution Modelling and its Application*, Istanbul, Turkey.

STOCKWELL, W. R., KIRCHNER, F., KUHN, M. et SEEFELD, S. (1997). A new mechanism for regional atmospheric chemistry modeling. *Journal of Geophysical Research*, 102(D22):25,847–25,879.

STOCKWELL, W. R., MIDDLETON, P., CHANG, J. S. et TANG, X. (1990). The second generation regional acid deposition model chemical mechanism for regional air quality modeling. *Journal of Geophysical Research*, 95(D10):16,343–16,367.

STRAUME, A. G. (2001). A more extensive investigation of the use of ensemble forecasts for dispersion model evaluation. *Journal of Applied Meteorology and Climatology*, 40:425–445.

STRAUME, A. G., KOFFI, E. N. et NODOP, K. (1998). Dispersion modeling using ensemble forecasts compared to ETEX measurements. *Journal of Applied Meteorology and Climatology*, 37:1,444–1,456.

STULL, R. B. (1988). *An introduction to boundary layer meteorology*. Kluwer Academic Publishers.

TALAGRAND, O., VAUTARD, R. et STRAUSS, B. (1999). Evaluation of probabilistic prediction system. Proceedings of the ECMWF Workshop on Predictability.

TAYLOR, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, 106(D7):7,183–7,192.

TOMBETTE, M. et SPORTISSE, B. (2007). Aerosol modeling at a regional scale : Model-to-data comparison and sensitivity analysis over Greater Paris. *Atmospheric Environment*, 41(33): 6,941–6,950.

TROEN, I. et MAHRT, L. (1986). A simple model of the atmospheric boundary layer ; sensitivity to surface evaporation. *Boundary-Layer Meteorology*, 37:129–148.

US EPA (1991). Guideline for regulatory application of the urban airshed model. Rapport technique EPA-450/4-91-013, US EPA.

van LOON, M., VAUTARD, R., SCHAAP, M., BERGSTRÖM, R., BESSAGNET, B., BRANDT, J., BUILTJES, P., CHRISTENSEN, J., CUVELIER, C., GRAFF, A., JONSON, J., KROL, M., LANGNER, J., ROBERTS, P., ROUÏL, L., STERN, R., TARRASÓN, L., THUNIS, P., VIGNATI, E., WHITE, L. et WIND, P. (2007). Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble. *Atmospheric Environment*, 41:2,083–2,097.

VAUTARD, R., BEEKMANN, M., ROUX, J. et GOMBERT, D. (2001). Validation of a hybrid forecasting system for the ozone concentrations over the Paris area. *Atmospheric Environment*, 35(14):2,449–2,461.

VAUTARD, R., SCHAAP, M., BERGSTRÖM, R., BESSAGNET, B., BRANDT, J., BUILTJES, P., CHRISTENSEN, J., CUVELIER, C., FOLTESCU, V., GRAFF, A., KERSCHBAUMER, A., KROL, M., ROBERTS, P., ROUÏL, L., STERN, R., TARRASON, L., THUNIS, P., VIGNATI, E. et WIND, P. (2009). Skill and uncertainty of a regional air quality model ensemble. *Atmospheric Environment*, 43:4,822–4,832.

VERWER, J. G., HUNDSDORFER, W. et BLOM, J. G. (1998). Numerical time integration for air pollution models. Rapport technique, CWI.

VERWER, J. G., HUNDSDORFER, W. et BLOM, J. G. (2002). Numerical time integration for air pollution models. *Surveys on Mathematics for Industry*, 10:107–174.

VERWER, J. G., SPEE, E. J., BLOM, J. G. et HUNDSDORFER, W. (1999). A second-order Rosenbrock method applied to photochemical dispersion problems. *SIAM Journal on Scientific Computing*, 20(4):1,456–1,480.

WARNER, T. T., SHEU, R.-S., BOWERS, J. F., SYKES, R. I., DODD, G. C. et HENN, D. S. (2002). Ensemble simulations with coupled atmospheric dynamic and dispersion models : illustrating uncertainties in dosage simulations. *Journal of Applied Meteorology and Climatology*, 41:488–504.

WESELY, M. L. (1989). Parameterization of surface resistances to gaseous dry deposition in regional-scale numerical models. *Atmospheric Environment*, 23:1,293–1,304.

WILKS, D. S. (2005). *Statistical Methods in the Atmospheric Sciences*, volume 100 de *International Geophysics Series*. Academic Press, second édition.

WU, L., MALLET, V., BOCQUET, M. et SPORTISSE, B. (2008). A comparison study of data assimilation algorithms for ozone forecasts. *Journal of Geophysical Research*, 113(D20310).

ZHANG, L., BROOK, J. R. et VET, R. (2003). A revised parameterization for gaseous dry deposition in air-quality models. *Atmospheric Chemistry and Physics*, 3:2,067–2,082.