Thèse de doctorat de l'École nationale des ponts et chaussées

Présentée et soutenue publiquement le 6 décembre 2005 par

Vivien MALLET

pour l'obtention du diplôme de docteur
de l'École nationale des ponts et chaussées

Spécialité : mathématiques et informatique

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

Jury composé de

P[r] François-Xavier Le Dimet	Université J. Fourier et INRIA	président
P[r] Steven Hanna	Harvard University	rapporteur*
P[r] Robert Rosset	Université P. Sabatier	rapporteur
P[r] Georges Oppenheim	Université Paris XI	directeur de thèse
D[r] Bruno Sportisse	ENPC	co-directeur de thèse
D[r] Gilles Bergametti	CNRS et Université Paris XII	examinateur

*Absent de la soutenance

Remerciements

Je remercie d'abord la région Île-de-France et l'École nationale des ponts et chaussées pour avoir financé ma thèse.

Je remercie principalement Bruno Sportisse pour avoir accepté de diriger ma thèse. J'ai eu la chance d'apprécier ses qualités scientifiques. Je lui suis très reconnaissant de la confiance qu'il m'a toujours accordée. Je le suis aussi reconnaissant d'avoir su se rendre disponible malgré son activité débordante. J'ai également apprécié son dynamisme.

Je remercie Georges Oppenheim pour avoir assumé la direction de ma thèse et pour m'avoir guidé vers les méthodes d'apprentissage statistique.

Je remercie les rapporteurs, Steven Hanna et Robert Rosset, ainsi que les autres membres du jury, dont François-Xavier Le Dimet et Gilles Bergametti, pour avoir évalué mon travail.

Je remercie particulièrement Denis Quélo pour l'aide qu'il m'a souvent apportée. Je remercie aussi toutes les personnes qui ont significativement contribué à l'avancement de la thèse : Luc Musson-Genon, Vincent Picavet, Hervé Njomgang, Mohamed Aissaoui et Cécile Honoré. J'ai enfin apprécié les interventions de Laurence Rouïl, Gilles Stoltz, Adélaïde Pourchet, Jaouad Boutahar, Chi-Sian Soh, Bernard Aumont, Kathleen Fahey, Germàn Torres, Karine Sartelet, Yelva Roustan et Frédérik Meleux.

Je remercie l'ECMWF, l'INERIS, EMEP, la BDQA (et les 40 associations agréées de surveillance de la qualité de l'air), Robert Vautard (Pioneer), l'USGS, GLCF, et l'équipe de Mozart 2 pour les données mises à disposition.

Table des matières

Introduction

Modèles de chimie-transport

Les modèles de chimie-transport sont des implémentations numériques de modèles physiques qui décrivent l'évolution de polluants atmosphériques. En estimant les concentrations de divers polluants, ils permettent notamment de simuler la qualité de l'air.

La qualité de l'air renvoie généralement à la pollution chronique de l'air au voisinage du sol, c'est-à-dire à la pollution atmosphérique potentiellement néfaste aux humains ou aux cultures agricoles. La définition étant fluctuante, on peut éventuellement l'étendre à la pollution accidentelle (par exemple, au rejet d'éléments radioactifs suite à un accident nucléaire). La qualité de l'air est décrite objectivement par les concentrations de polluants.

Concernant la qualité de l'air, les modèles de chimie-transport interviennent dans quatre grandes classes d'activité :

1. la *prévision* de la qualité de l'air. Les concentrations de polluants sont prévues quotidiennement sur quelques jours suivant le jour courant. Une prévision est généralement initialisée par la prévision du jour précédent. Des prévisions météorologiques sont fournies par un modèle météorologique, ce qui suffit au modèle de chimie-transport pour générer des prévisions de concentrations de polluants. Un bon exemple d'une telle application est la plate-forme de prévision Prév'air (`http://www.prevair.org/`), opérée par l'INERIS[1], qui propose quotidiennement des simulations sur quatre jours (depuis la veille jusqu'au surlendemain). Par abus de langage, on peut également parler de prévision dans le cas de simulations effectuées sur une période passée. Les simulations en question ne doivent cependant pas faire appel à de l'assimilation de données (voir ci-dessous) puisque cette dernière repose sur des observations inconnues en prévision.
2. la *modélisation inverse*. Il est possible de tirer parti des données d'observation dans le but d'améliorer les champs d'entrée du modèle de chimie-transport. Des méthodes dédiées permettent de corriger les champs d'entrée de sorte à diminuer l'écart entre les concentrations simulées et les observations. Ces méthodes sont des procédures d'assimilation de données. Lorsque les champs d'entrée optimisés par la procédure d'assimilation peuvent être utilisés pour d'autres simulations (reproductibilité), on parle de modélisation inverse. La modélisation inverse est une procédure d'amélioration des simulations d'un modèle de chimie-transport, mais elle peut être une finalité puisqu'elle affine des champs d'entrée (par exemple, des émissions). Si une procédure de modélisation inverse est fiable, elle constitue une voie d'amélioration de données mal connues.
3. le développement de la *modélisation physique*. La modélisation physique, c'est-à-dire la description mathématique des phénomènes physiques, peut bénéficier des simulations des modèles de chimie-transport. Il est parfois impossible de rendre compte en laboratoire de l'ensemble des phénomènes qui interviennent dans l'évolution d'un polluant. Les modèles

[1] Institut national de l'environnement industriel et des risques

de chimie-transport permettent de tester et donc de développer des modèles physiques via des expériences numériques comparées aux observations. Ces expériences numériques ont l'avantage de simuler une grande partie de la complexité de l'atmosphère.

4. les *études d'impact*. Lorsque les simulations sont effectuées sur plusieurs années futures, le terme de prévision, qui se réfère à des simulations détaillées et de courte durée, est abandonné au profit du terme impact. Ces études sont en effet centrées sur l'impact de la pollution sur de longues périodes futures et généralement sur l'impact de changements en amont. En particulier, les études d'impact servent souvent à évaluer les conséquences de réductions d'émissions. Elles nécessitent des approximations, du fait de leur coût calcul, et à cause du manque de connaissance des situations futures (conditions météorologiques, niveaux d'émissions). Il faut noter deux points très importants : les conditions de ces simulations sont généralement inédites (par exemple, avec des émissions divisées par deux), et il n'y a pas d'observations pour contraindre les modèles.

Limitations des modèles

Les quatre activités identifiées trouvent bien sûr leurs limitations dans la qualité des modèles de chimie-transport. Il y a d'abord des *limitations numériques* dues aux coûts des calculs. Les schémas numériques induisent des coûts importants du fait du couplage entre toutes les concentrations, des inhomogénéités fortes (qui requièrent des schémas précis) et de la raideur des équations (grande dispersion des temps caractéristiques en chimie). Dans ce contexte, la discrétisation (spatiale, temporelle et le nombre d'espèces chimiques) est nécessairement fortement limitée. Actuellement, les modèles de chimie-transport intègrent en temps environ un à dix millions de variables. Leur résolution est souvent volontairement dégradée pour les études les plus coûteuses. Les performances numériques des modèles dépendent aussi de leur architecture informatique, principalement via leurs capacités en calcul parallèle.

Une deuxième limitation réside dans les *données d'entrée* des modèles. Parmi ces données, on compte les inventaires d'émission, les données météorologiques, les données de sol, etc. L'évaluation par des experts des données d'entrée [Hanna *et al.*, 1998, 2001] révèle de fortes incertitudes, souvent de l'ordre de 50% ou 100%. Ces incertitudes touchent l'ensemble des masses de polluants injectées (conditions initiales, conditions aux limites, émissions), les processus de pertes (dépôts), les réactions chimiques (constantes de réaction) et les caractéristiques du transport (champs météorologiques). Dans ces conditions, les concentrations simulées par les modèles de chimie-transport ne peuvent être qu'incertaines – dans des limites à quantifier.

La dernière grande limitation concerne la *formulation des modèles*. La modélisation physique n'est pas en mesure de décrire avec une extrême précision les phénomènes atmosphériques complexes. Il faut alors recourir à des modèles simplifiés, appelés paramétrisations. En particulier, toutes les échelles n'étant pas résolues par un modèle (dont l'échelle caractéristique est souvent de plusieurs kilomètres), les phénomènes correspondants sont approchés par des paramétrisations sous-maille. Ainsi plusieurs paramétrisations physiques « concurrentes » alimentent les modèles. Une même quantité peut être raisonnablement estimée par plusieurs paramétrisations physiques qu'il est difficile de départager. Un aperçu des fluctuations dans les formulations est donné par la disparité des modèles de chimie-transport existants.

Ces limitations ont pour conséquence le choix d'approximations numériques (maîtrise des coûts calcul), le choix voire la correction de données et la composition d'un modèle sur la base des paramétrisations existantes. Une grande liberté est ainsi laissée au modélisateur. L'expérience du modélisateur lui permet d'effectuer les bons choix dans le but d'approcher au mieux les observations. On parle généralement d'*ajustement* des modèles (ou de « tuning » en anglais).

La pratique de l'ajustement de modèle soulève plusieurs difficultés. Il est possible que les optimisations du modélisateur ne soient pas conformes à la physique. Par exemple, un choix de paramétrisation peut permettre de compenser une erreur faite sur un champ d'entrée. La validité du modèle pose alors question. La formulation du modèle n'est pas assurée d'être la meilleure d'un point du vue physique et les données choisies ne sont pas nécessairement les plus proches de la réalité. En sortie de modèle, les concentrations qui ne sont pas contrôlées par des observations – la plupart des espèces sont dans ce cas – sont peu fiables. Ces difficultés se déclinent sur les quatre applications introduites précédemment :

- En prévision, les ajustements du modélisateur sont centraux puisqu'ils vont déterminer les performances (scores de prévision). Les meilleurs ajustements sur une période courte peuvent donner de très bons résultats mais seront peu robustes. À l'inverse, des ajustements pour de longues périodes seront plus robustes mais moins efficaces. Ceci n'est cependant pas trop problématique : les ajustements font partie intégrante de la démarche de prévision.
- En modélisation inverse, les ajustements effectués ont un impact sur les champs retrouvés. Les champs sont eux-mêmes optimisés pour compenser des erreurs dont ils ne sont pas la cause. C'est la raison pour laquelle la sensibilité des champs optimisés par rapport aux limitations du modèle est une donnée importante. On parle parfois de sensibilité du second ordre.
- De même, en modélisation physique, les erreurs sont compensées par des ajustements via les leviers dont dispose le modélisateur. Il est alors hasardeux de discriminer les modèles physiques sur la base des choix de la configuration optimale (au titre des comparaisons aux observations). Quant aux phénomènes observés dans les sorties de modèle, il convient de les considérer à la lumière des incertitudes.
- Les études d'impact sont encore plus sujettes aux incertitudes. Elles sont réalisées dans des conditions inédites : émissions fortement diminuées, ou conditions météorologiques différentes. Les ajustements effectués dans des conditions plus communes (et observées) n'ont aucune raison d'être adaptés aux nouvelles conditions. L'absence d'observation est un point-clé. Elle impose une plus grande prudence vis-à-vis des résultats, c'est-à-dire qu'elle nécessite une estimation de l'incertitude.

Incertitude

Sur la base des constatations précédentes, on voit bien l'importance d'estimer l'incertitude des sorties des modèles de chimie-transport. Avant d'aller plus loin, il faut définir plus précisément ce qu'est l'incertitude. Pour cela, le modèle doit être considéré comme un modèle statistique. On note F le modèle (aléatoire) et f une réalisation de F, c'est-à-dire un modèle particulier. Le modèle f correspond à un jeu de paramétrisations et d'approximations numériques. On définit les variables aléatoires X et Y représentant les entrées du modèle et ses sorties respectivement. On leur associe les réalisations x et y (y étant les sorties de f avec les entrées x).

Ainsi, sous forme statistique :

$$Y = F(X) \tag{1}$$

qui se décline en une réalisation

$$y = f(x) \tag{2}$$

La variable aléatoire Y décrit les sorties que l'ensemble des modèles et données possibles peuvent générer. On peut alors associer aux sorties du modèle F une densité de probabilité.

Pour simplifier la présentation, on suppose que Y suit une loi normale de moyenne $\overline{Y}$ et d'écart-type σ :

$$Y \sim \mathcal{N}\left(\overline{Y}, \sigma^2\right) \tag{3}$$

On suppose maintenant que la vraie valeur de Y est observée et on la note o, pour « observation ». L'observation est elle-même une variable aléatoire O dont o est une réalisation, mais cet aspect est écarté car ce n'est pas essentiel ici. L'observation o est supposée parfaite.

Aujourd'hui, les modèles subissent essentiellement une évaluation de l'*erreur*, ce qui consiste à calculer la distance entre l'observation o et la réalisation y :

$$\text{erreur} = \mathcal{D}(y, o) \tag{4}$$

où $\mathcal{D}$ est une mesure de la distance, par exemple une erreur quadratique moyenne.

L'autre mesure de la qualité des modèles est l'*incertitude a priori* :

$$\text{incertitude} = \sigma_Y = \sigma \tag{5}$$

On peut relever un abus de langage dans les équations 4 et 5 puisqu'elles introduisent des *mesures* de l'erreur et de l'incertitude. L'incertitude est plus précisément décrite par la densité de probabilité des concentrations. L'écart-type en est une mesure commode et parlante.

On parle d'incertitude *a priori* car les observations n'interviennent pas dans l'estimation. Seule la distribution de probabilité des concentrations de sortie est nécessaire à la mesure de l'incertitude. En revanche, l'incertitude *a posteriori* est l'incertitude estimée avec les observations, ou, dans la terminologie probabiliste, « sachant les observations ». Sur la base des observations, les concentrations (simulées) deviennent plus ou moins probables, selon leur distance aux observations. La densité de probabilité qu'on peut associer aux concentrations est alors modifiée : on introduit une probabilité conditionnelle (aux observations). L'écart-type qui découle de la nouvelle densité de probabilité est une mesure de l'incertitude *a posteriori* :

$$\text{incertitude } \textit{a posteriori} = \sigma_{Y|o} \tag{6}$$

Seule l'incertitude *a priori* est étudiée dans cette thèse.

Estimation de l'incertitude

Les trois sources d'incertitude précédemment listées sont les approximations numériques, la formulation du modèle et les données d'entrée. Des méthodes différentes permettent de les estimer :

- l'incertitude liée aux approximations numériques est étudiée grâce à des *comparaisons entre schémas numériques*. Les changements ayant un impact important sur les concentrations désignent les schémas associés à une forte sensibilité et donc sources d'incertitude.
- l'incertitude due à la formulation du modèle est estimée par des *simulations d'ensemble*, c'est-à-dire par des ensembles de simulations avec des modèles différents. On parle aussi de simulation multi-modèles.
- l'incertitude due aux données d'entrée s'estime naturellement via des *simulations Monte Carlo*. Avec un nombre suffisant de simulations, il est possible d'obtenir une distribution de probabilité des sorties.

Ces trois aspects sont abordés dans cette thèse.

Traitement de l'incertitude

Après avoir estimé l'incertitude, on peut chercher à l'atténuer ou à la « contourner ». Pour cela, les pistes les plus évidentes consistent en la réduction des trois limitations précédemment exposées :

- les approximations numériques peuvent être moindres si une puissance de calcul supérieure est disponible, ou si de nouveaux schémas numériques sont proposés ;
- les paramétrisations physiques des modèles peuvent être améliorées, grâce à des études de processus ;
- les données d'entrée peuvent être affinées, soit par observation ou modélisation directe, soit par modélisation inverse.

Toutes ces études n'ont bien sûr pas été menées dans cette thèse. Une étude de sensibilité aux émissions prépare à la modélisation inverse des émissions à l'échelle continentale. En outre, une participation à une expérience de modélisation inverse des émissions à l'échelle régionale a abouti à la publication[2] suivante :
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. *J. Geophys. Res.*, 110(D24)

Ce travail analyse aussi l'incertitude dans les émissions optimisées. La sensibilité (du second ordre) des émissions à la procédure d'assimilation et aux autres paramètres est suffisamment forte pour que les résultats soient modérément robustes. Cette conclusion montre l'importance de l'évaluation de l'incertitude.

Une stratégie moins conventionnelle consiste à dépasser les limites de l'incertitude en combinant des sorties de modèles. Sur la base d'un ensemble de modèles (donc d'un ensemble de prévisions), une prévision unique est construite pour minimiser la distance aux observations. On parle parfois de *prévision d'ensemble*.

Plan de la thèse

Après une introduction à la physique et à la chimie atmosphériques (orientée vers l'ozone qui est le principal polluant auquel les méthodes sont appliquées), le chapitre 1 introduit la modélisation (paramétrisations physiques et données) sur laquelle reposent les simulations de la thèse.

Avant de mener les études annoncées ci-dessus, il faut disposer d'un système de simulation complet qui permet la prévision (éventuellement opérationnelle), qui joue le rôle de plate-forme multi-modèles, et qui accueille des procédures d'assimilation de données. L'architecture de ce système est présentée, avec sa réalisation partielle qu'est le système Polyphemus, au chapitre 2 – il s'agit d'une partie technique. Une simulation de référence est ensuite introduite pour évaluer le système et montrer qu'il est raisonnable de l'utiliser pour les études qui suivent.

Le chapitre 3 estime l'incertitude liée aux approximations numériques, à la formulation du modèle et aux données d'entrée.

Le chapitre 4 est une étude de sensibilité aux émissions. Il s'agit de travaux préparant la modélisation inverse d'émissions à l'échelle européenne.

Enfin, le chapitre 5 traite de la prévision d'ensemble. Des résultats encourageants montrent le potentiel des méthodes d'ensemble pour la prévision.

[2] L'auteur principal étant Denis Quélo et l'étude ayant déjà été publiée dans sa thèse [Quélo, 2004], elle n'est pas présentée ici.

Le chapitre 2 et la première partie du chapitre 3 sont l'objet d'articles en préparation. Une version anglaise du chapitre 5 est un article soumis. Le chapitre 4 et la partie centrale du chapitre 3 sont des articles publiés, et insérés tel quel.

Avertissement et conventions

Avertissement

La plupart des courbes sont issues de travaux publiés, soumis ou présentés. Elles sont donc en anglais.

Conventions

Sauf mention contraire, l'heure est toujours reportée en temps universel, parfois désigné par UT (« universal time »).

La virgule des nombres décimaux est remplacée par un point, comme il est d'usage en anglais. Cette convention est adoptée car elle permet de conserver une cohérence avec les contraintes informatiques (axes des courbes) et avec les articles ou annexes en anglais.

Chapitre 1

Physique et modélisation

Ce chapitre introduit succinctement les phénomènes physiques de la prévision photochimique en phase gazeuse. La première section consiste en une présentation du transport pertinente pour tous les polluants atmosphériques. Une deuxième section traite de la photochimie, principalement pour la production et la consommation d'ozone. La modélisation de ces phénomènes conduit à la formulation de l'équation de transport-réaction – section 1.4. Cette équation repose sur diverses variables (température, vitesses de dépôt, émissions, etc.) dont un certain nombre sont estimées par des paramétrisations physiques – section 1.5.

Sommaire

La description des phénomènes physiques et chimiques s'inspire de

– Stull, R. B. (1988). *An introduction to boundary layer meteorology*. Kluwer Academic Publishers, et
– Honoré, C. (2000). *La photochimie de l'ozone à l'échelle urbaine, un système dynamique non-linéaire*. Thèse de doctorat, Université Paris 6.

Afin d'approfondir cette partie, on pourra consulter

– Jacob, D. J. (1999). *Introduction to atmospheric chemistry*. Princeton University Press,
– Seinfeld, J. H. et Pandis, S. N. (1998). *Atmospheric chemistry and physics : from air pollution to climate change*. Wiley-Interscience,
– Garrat, J. R. (1992). *The atmospheric boundary layer*. Cambridge University Press,
– Holton, J. R. (2004). *An introduction to dynamic meteorology*. Academic Press, quatrième édition.

1.1 Transport de polluants

1.1.1 Structure de l'atmosphère

L'atmosphère est divisée en plusieurs couches définies à partir du profil vertical de température. La figure 1.1 illustre les quatre premières couches de l'atmosphère :

1. la troposphère entre 0 km et 8–18 km caractérisée par une décroissance de la température avec l'altitude ;
2. la stratosphère, au-dessus de la troposphère, s'étendant jusqu'à environ 50 km et caractérisée par une température croissante avec l'altitude (réchauffement dû l'absorption du rayonnement solaire par la couche d'ozone) ;
3. la mésosphère qui est située entre la stratosphère et 80 km ;
4. la thermosphère qui s'étend jusqu'à environ 500 km.

On appelle tropopause, stratopause et mésopause les limites supérieures respectives de la troposphère, de la stratosphère et de la mésosphère.

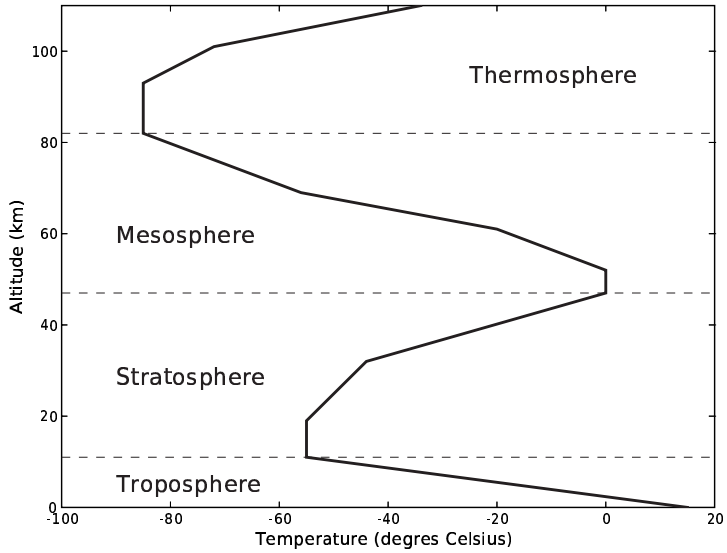


FIG. 1.1 – Profil vertical typique de température et principales couches de l'atmosphère.

En qualité de l'air, les cibles sont les concentrations de certains polluants uniquement dans les premiers mètres, c'est-à-dire les quantités de polluants ayant un impact potentiel (sur les populations ou les cultures agricoles). Or les transferts entre la stratosphère et la troposphère sont suffisamment lents pour que les polluants (concernés dans ce travail) ne soient pas directement sensibles aux variations de ces échanges. On se restreint donc à la troposphère et, en fait, à une partie de celle-ci appelée couche limite atmosphérique.

Ainsi que proposé dans Stull [1988], on peut définir la couche limite (atmosphérique – « couche limite » se réfère désormais à « couche limite atmosphérique ») comme la partie de la troposphère qui est influencée en une heure ou moins par des changements au voisinage de la surface terrestre (réchauffement, évaporation, émissions de polluant, etc.). Comme représentée par la figure 1.2, la couche limite atmosphérique s'étend sur environ un kilomètre. Plus précisément, sa hauteur peut varier entre quelques dizaines de mètres (la nuit) et quelques kilomètres.

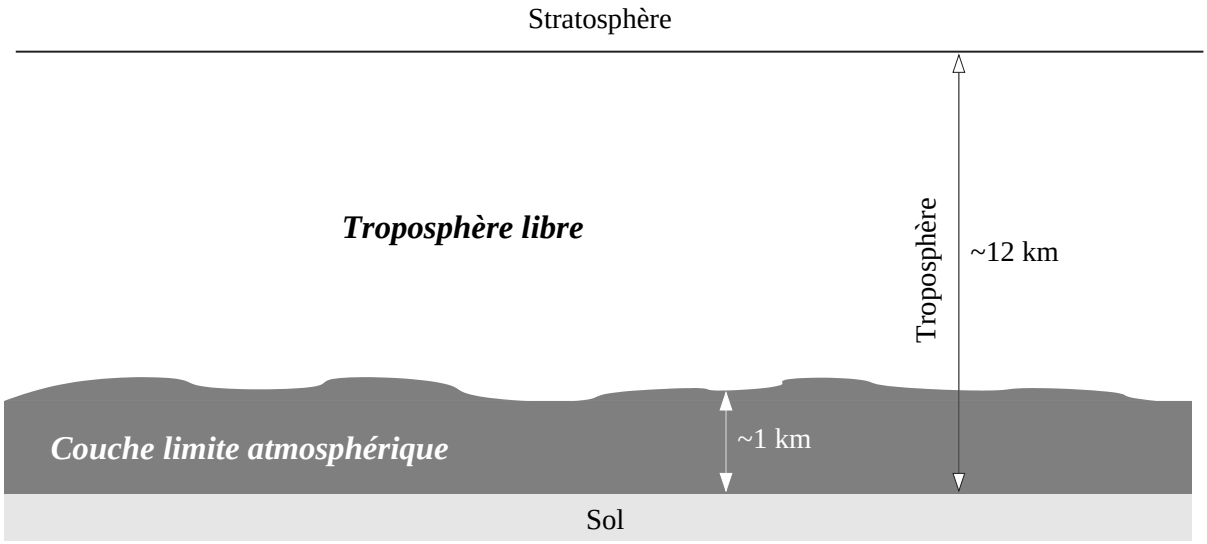


FIG. 1.2 – Structure de la troposphère.

Au-dessus de la couche limite se situe la troposphère libre. Les polluants présents dans cette dernière contribuent assez peu aux concentrations observées au sol. La description de cette partie de l'atmosphère a cependant une importance du fait des nuages qui s'y développent. Ces derniers ont un impact sur les concentrations de polluant au sol via leur atténuation du rayonnement solaire (température au sol, réactions photolytiques présentées dans la section 1.2), la convection qu'ils peuvent engendrer (voir section 1.1.3) et éventuellement les précipitations dont ils sont à l'origine.

1.1.2 Transport par le vent

Lorsqu'il s'agit du transport de polluants, on distingue généralement le transport dû à la convection (section 1.1.3) et celui dû aux vents (rigoureusement appelés vents moyens – les déplacements d'air convectifs correspondant aux fluctuations du vent).

Les principales composantes du vent (moyen) sont horizontales. Le vent est d'ailleurs, et de loin, la source essentielle du transport horizontal. Il s'annule au voisinage du sol sous l'effet de la rugosité et augmente avec l'altitude. Dans la couche limite atmosphérique, les vents horizontaux sont typiquement de l'ordre de quelques mètres par seconde (2–10 $\mathrm{m} \cdot \mathrm{s}^{-1}$).

Les vents verticaux sont très faibles : ils sont de l'ordre du millimètre ou du centimètre par seconde. Ils sont négligeables par rapport à la turbulence.

1.1.3 Turbulence

On peut représenter les mouvements turbulents par des tourbillons dont l'amplitude peut aller jusqu'à la hauteur de la couche limite. La turbulence naît principalement (par échelle croissante)

1. d'obstacles au sol qui laissent une traînée turbulente,
2. du cisaillement du vent,
3. du réchauffement au sol (dû au rayonnement solaire) qui élève les masses d'air réchauffées au contact du sol,
4. de la convection nuageuse.

À l'échelle de la couche limite et pour de grandes échelles horizontales, la présence d'obstacles n'est pas une source importante de turbulence. Le cisaillement du vent joue un rôle important puisqu'il détermine la hauteur de couche limite la nuit. Le réchauffement au sol est le phénomène principal en journée et il conditionne la hauteur de la couche limite. La convection nuageuse n'apparaît pas systématiquement en présence de nuages. Lorsqu'elle se produit, elle domine la turbulence et génère un transport vertical important sur de longues distances (jusqu'à quelques kilomètres).

Puisque le vent vertical est très faible, le transport vertical est principalement turbulent. En conséquence, la turbulence définit la hauteur de la couche limite. La figure 1.3 représente l'évolution temporelle de la couche limite. Le jour, un peu après le lever du soleil, la couche limite dite instable croît pour atteindre son maximum un peu après midi. Cette couche est mélangée sous l'action de la turbulence générée par un réchauffement du sol. Elle peut atteindre 2 à 3 kilomètres. Les polluants y sont dispersés rapidement. Pendant la nuit, le mélange est beaucoup moins efficace, dans une couche limite de quelques centaines de mètres.

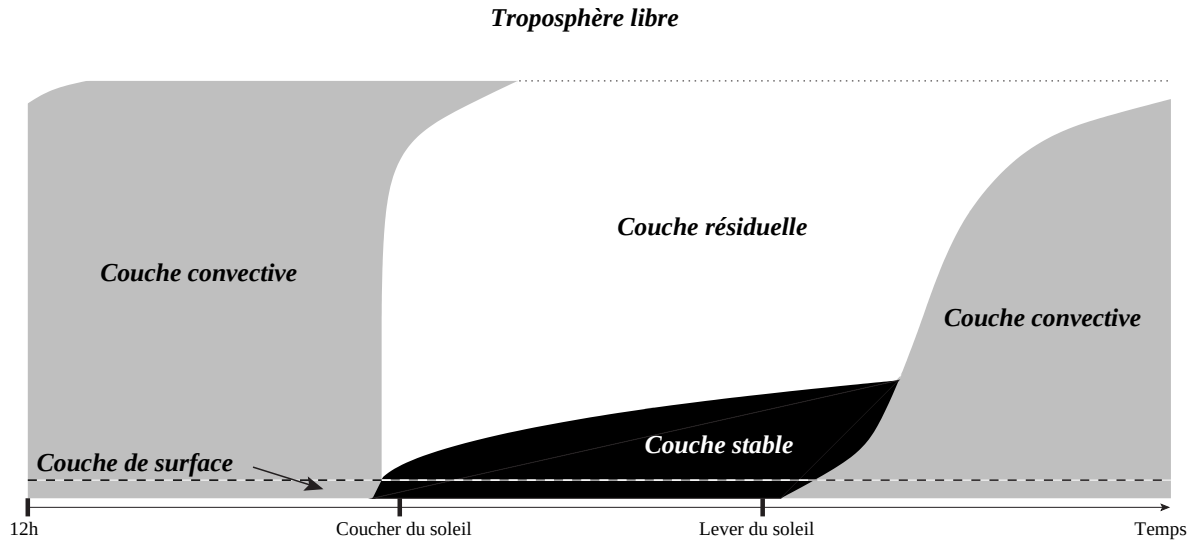


FIG. 1.3 – Évolution temporelle de la couche limite atmosphérique. Inspiré par Stull [1988].

Au-dessus de la couche (nocturne) stable se situe la couche résiduelle. Rigoureusement parlant, elle fait partie de la troposphère libre puisqu'elle n'a pratiquement pas d'échanges avec la couche stable. Sa hauteur est celle de la couche limite du jour précédent. On y trouve des polluants mélangés pendant le jour précédent. Environ trente minutes avant le coucher du soleil, la turbulence due au réchauffement par le sol s'éteint ; les polluants, au-dessus de la couche stable qui se forme, demeurent dans la couche résiduelle.

On définit enfin une couche de surface dans laquelle les flux turbulents varient de moins de 10%. On parle parfois de couche à flux constants. On considère que la hauteur de la couche de surface est environ 10% de celle de la couche limite atmosphérique.

1.2 Photochimie

1.2.1 Composition chimique de l'atmosphère

Outre les molécules présentes en quantité importante (N_2, O_2 et Ar), de nombreuses molécules, appelées espèces chimiques, se trouvent à l'état de trace dans l'atmosphère. Leur concentration varie généralement entre quelques picogrammes et quelques milligrammes par mètre cube.

Les espèces peuvent être en phase gazeuse, liquide ou solide. Dans les deux derniers cas, on parle d'aérosols ou de particules. Les aérosols possèdent des tailles entre quelques nanomètres et plus de 100 μm. Ils peuvent avoir un impact sur le rayonnement solaire en l'absorbant. Dans la suite, on ne s'intéresse qu'à la phase gazeuse, sauf mention contraire.

À chaque espèce dans l'atmosphère, on peut associer un temps de vie qui dépend de ses réactions chimiques (section 1.2.2) et de ses pertes par dépôt au sol (section 1.3). Certaines espèces instables se combinent très rapidement avec d'autres molécules et ont donc des temps de vie très courts (quelques secondes). D'autres réagissent peu et déposent peu au regard de leur concentrations moyennes, elles résident alors jusqu'à plusieurs années dans l'atmosphère. De telles espèces sont transportées sur de très longues distances (elles parcourent tout l'hémisphère). Elles disposent du temps suffisant pour diffuser jusque dans la stratosphère.

On distingue les espèces primaires qui sont émises dans l'atmosphère des espèces secondaires qui sont uniquement le produit de réactions chimiques (section suivante).

1.2.2 Chimie de l'atmosphère

Si certains polluants sont inertes, de nombreux polluants se transforment ou réagissent entre eux selon des réactions allant de la photolyse à des réactions impliquant plusieurs molécules.

Les réactions photolytiques conduisent à la décomposition d'une molécule sous l'effet du rayonnement solaire. Leurs constantes de réaction (constantes photolytiques) sont directement dépendantes du flux solaire. La photolyse est d'autant moins efficace que le rayonnement solaire parcourt de longues distances dans l'atmosphère. Ainsi il y a une forte dépendance envers l'angle zénithal[1] et l'altitude dans l'atmosphère. De plus les constantes photolytiques varient fortement avec la période de l'année et la couverture nuageuse. Dans le cas d'une diminution des constantes due à la présence de nuages, on parle d'atténuation nuageuse. Il est à noter que les nuages augmentent les constantes photolytiques au-dessus d'eux puisqu'ils réfléchissent les rayons. Enfin, d'une molécule à l'autre, la sensibilité au rayonnement solaire diffère. En particulier, chaque liaison moléculaire peut uniquement être cassée par certaines longueurs d'onde.

Les autres réactions en phase gazeuse sont plus communes. Elles impliquent plusieurs espèces et leur efficacité dépend des conditions (concentrations, température, pression, humidité).

Les espèces en phase gazeuse peuvent aussi réagir avec les aérosols. Il peut s'agir de réactions de surface lorsqu'elles ont lieu à la surface des particules solides. Certaines réactions se déroulent aussi dans les gouttes d'eau (nuages) avec des espèces qui ont au préalable diffusé dans ces gouttes.

1.2.3 Cycle de l'ozone troposphérique

L'ozone, O_3, est un polluant secondaire dont le cycle comprend des dizaines de réactions importantes. La réaction de formation de l'ozone est la suivante :

$$O_2 + O + M \longrightarrow O_3 + M \qquad \text{(R 1.1)}$$

où M est un tiers corps qui est en général N_2 ou O_2. L'oxygène O est le résultat de la photolyse de NO_2 :

$$NO_2 + h\nu \xrightarrow{J_{NO_2}} NO + O \qquad \text{(R 1.2)}$$

[1] Angle formé entre un rayon incident et la verticale qui se confond avec les rayons au zénith.

$h\nu$ représente la lumière qui vient dissocier la molécule NO_2 (ν étant un nombre d'onde correspondant à des fréquences inférieures à 424 nm). Dans le même temps, une troisième réaction vient consommer l'ozone :

$$O_3 + NO \xrightarrow{k} NO_2 + O_2 \qquad \text{(R 1.3)}$$

On parle souvent de titration de l'ozone par le monoxyde d'azote. L'ozone est aussi détruit par photolyse :

$$O_3 + h\nu \longrightarrow O_2 + O^{1D} \qquad \text{(R 1.4)}$$

où O^{1D} est un atome d'oxygène excité.

Une production d'ozone efficace repose sur l'oxydation des composés organiques volatils (désignés par COV), c'est-à-dire des molécules organiques telles que les alkanes, les alcènes ou les aldéhydes. Après leur oxydation, ils oxydent à leur tour NO et produisent ainsi des molécules de dioxyde d'azote NO_2. Ces dernières contribuent à la formation d'ozone par les réactions R 1.2 et R 1.1.

L'oxydation des COV puis de NO a lieu dans la journée et a pour bilan

$$COV + 2\,NO + 2\,O_2 \longrightarrow 2\,NO_2 + H_2O + RCHO \qquad \text{(R 1.5)}$$

où RCHO est un aldéhyde qui peut être photolysé et ensuite contribuer aussi à l'oxydation du monoxyde d'azote.

Un point essentiel à la réalisation de la réaction R 1.5 est la présence du radical hydroxyle OH. Ce dernier peut être formé à la suite de la photolyse de l'ozone (réaction R 1.4) par

$$O^{1D} + H_2O \longrightarrow 2\,OH \qquad \text{(R 1.6)}$$

La photolyse de l'acide nitreux HONO est une autre source de OH :

$$HONO + h\nu \longrightarrow OH + NO \qquad \text{(R 1.7)}$$

La photolyse des aldéhydes produit aussi des radicaux hydroxyles.

Des réactions concurrentes peuvent diminuer l'efficacité de la réaction R 1.5. La perte du radical OH modère aussi le processus de formation d'ozone. La réaction principale de ce point de vue est

$$OH + NO_2 + M \longrightarrow HNO_3 + M \qquad \text{(R 1.8)}$$

Dans le bilan, il convient de souligner le rôle prépondérant des oxydes d'azote NO et NO_2 (réactions R 1.2 et R 1.3). Les COV sont des polluants contribuant à la formation d'ozone. Il faut noter que la formation d'ozone à partir des COV est un processus plus long que la titration R 1.3. Par exemple, près des sources d'oxydes d'azote (principalement de monoxyde d'azote) et de COV, on observe souvent une diminution des concentrations d'ozone (titration par NO) puis, dans un second temps, une augmentation des concentrations d'ozone (cycle des COV). Ce phénomène s'observe dans les villes et leur panache de polluants.

Un phénomène notable est l'existence, pour l'ozone, de deux régimes chimiques dits « COV limité » et « NO_x limité ». En régime « NO_x limité », les concentrations d'ozone sont sensibles

à l'augmentation des concentrations des NO_x, mais ne le sont pas à celles des COV. Les concentrations d'ozone augmentent avec celles des NO_x. En régime « COV limité », l'ozone est sensible aux COV qui, en augmentant, accroissent la production d'ozone. Dans ce régime, l'ozone reste sensible aux NO_x, mais une augmentation des concentrations des NO_x diminue celles d'ozone [Seinfeld et Pandis, 1998].

1.3 Autres processus et remarques

1.3.1 Émissions

Les polluants primaires sont rejetés dans l'atmosphère par l'industrie, le trafic et les sources naturelles (principalement la biomasse, mais aussi les volcans, les feux de forêt et, pour NOx, les éclairs lors d'orages). On appelle émissions anthropogéniques les émissions issues de l'activité humaine, et émissions biogéniques celles issues de la biomasse.

Les émissions anthropogéniques constituent une partie importante des émissions et sont principalement localisées dans et autour des grandes villes. Elles peuvent être rejetées au sol (trafic) ou en hauteur (cheminées d'usine). Elles ont généralement une température supérieure à celle de l'air environnant, ce qui les élève un peu dans la couche limite (« surhauteur »). Elles sont plus importantes en journée du fait de l'activité humaine, mais leur niveau n'est pas négligeable la nuit du fait de l'industrie.

Les émissions biogéniques sont présentes partout (même en mer) ; on parle d'émissions diffuses. On considère que ces émissions sont principalement du monoxyde d'azote et quelques COV (isoprène et terpènes). Leur niveau est plus faible que celui des émissions anthropogéniques, mais les COV impliqués sont plus réactifs. Leur contribution à l'ozone n'est pas négligeable [par exemple, Derognat *et al.*, 2003].

1.3.2 Dépôt sec

On dit que les polluants se déposent au sol lorsqu'ils sont « absorbés » par l'eau, le sol ou la végétation. Ce phénomène constitue un terme de perte élevé. Son intensité dépend des polluants, des conditions météorologiques, de l'éclairement, du lieu (type et densité de végétation), de la saison (état de la végétation). Le dépôt est plus fort en journée et il est accru par l'éclairement. Au-dessus des masses d'eau, il croît avec la solubilité de l'espèce considérée.

1.3.3 Dépôt humide

On qualifie de dépôt humide ou de lessivage la perte due aux transferts de masse avec la phase aqueuse (nuages ou pluies). Les polluants solubles peuvent pénétrer les gouttes de pluie lors de leur chute et sont ainsi précipités au sol. Une autre forme de lessivage se déroule dans les nuages où les polluants solubles ont des échanges (transferts de masse) avec les gouttes d'eau. Il faut noter que l'ozone, peu soluble, n'est pas affecté directement par le dépôt humide, mais indirectement puisque plusieurs molécules intervenant dans son cycle de formation sont solubles.

1.3.4 Remarques à propos des concentrations d'ozone

L'ozone est directement sujet à tous les phénomènes précédemment décrits, excepté le dépôt humide et les émissions. Les émissions ont cependant un impact indirect fort sur les concentrations d'ozone.

Environ 8% de l'ozone se trouve dans la troposphère. La majeure partie de l'ozone se trouve dans la stratosphère et constitue une source pour l'ozone troposphérique. Les concentrations

augmentent généralement dans les premières centaines de mètres de la troposphère. Sur l'Europe, l'ozone est présent partout avec des concentrations moyennes typiques entre 40 $\mu g \cdot m^{-3}$ et 120 $\mu g \cdot m^{-3}$ au voisinage du sol.

La figure 1.4 présente le profil journalier moyen des concentrations d'ozone observées sur quatre-vingt-six stations européennes pendant les mois de juin à août 2001. Le profil typique d'ozone est donc une courbe « en cloche » dont le minimum est atteint dans la nuit et le maximum s'observe vers 15 h UT. Le minimum est conditionné par le dépôt sec, par la chimie de nuit et la turbulence. Le maximum est principalement dépendant du mélange vertical. En effet, les concentrations d'ozone croissent avec l'altitude sur plusieurs centaines de mètres. Or le mélange vertical tend à homogénéiser les concentrations ; il contribue donc à augmenter les concentrations au voisinage du sol. La hauteur de couche limite étant maximale vers 15 h UT, le maximum d'ozone s'observe à ce même moment.

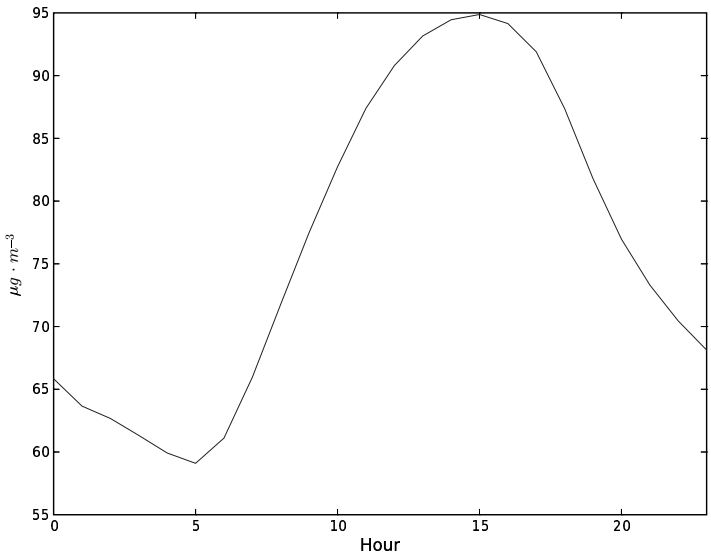


FIG. 1.4 – Profil journalier moyen des concentrations d'ozone observées sur quatre-vingt-six stations européennes pendant les mois de juin à août 2001.

Sur l'année, l'ozone atteint ses maxima pendant l'été. La figure 1.5 montre l'évolution sur l'année 2001 des maxima d'ozone. Les maxima peuvent varier fortement d'un jour à l'autre, mais la présence de concentrations plus élevées pendant l'été est claire. Ceci s'explique en partie par un mélange vertical plus important. De plus, le flux solaire étant plus important en été, le cycle de formation d'ozone est plus efficace. Au voisinage des points d'émission, des conditions météorologiques stationnaires (plus communes en été) permettent aux polluants de s'accumuler et à l'ozone d'atteindre des niveaux élevés.

La distribution spatiale d'ozone varie selon les conditions météorologiques. On repère souvent les points d'émission majeurs (généralement les grandes villes) car l'ozone y est titré par le monoxyde d'azote. Une carte d'ozone typique (simulé) est proposée par la figure 1.6.

1.4 Bilan : équation de chimie-transport

Dans cette section, on expose les modèles représentant les phénomènes précédemment exposés.

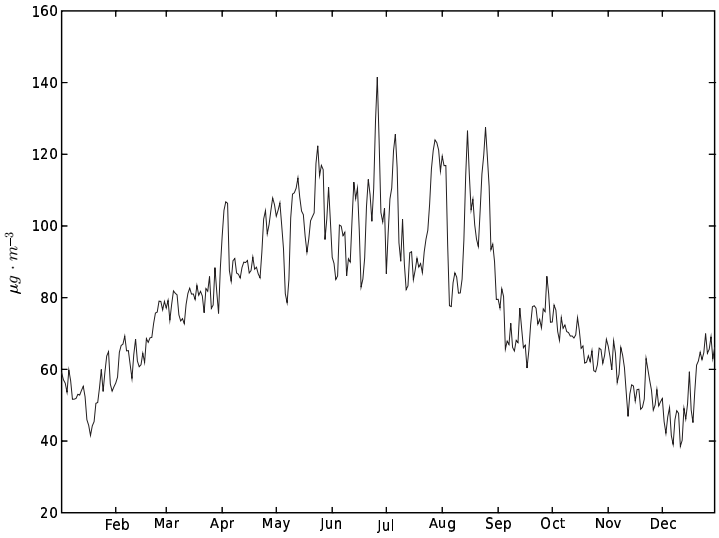


FIG. 1.5 – Les maxima journaliers d'ozone sur quatre-vingt-six stations européennes sont moyennés (sur toutes les stations). La figure représente l'évolution au cours de l'année 2001 de cette moyenne de maxima.

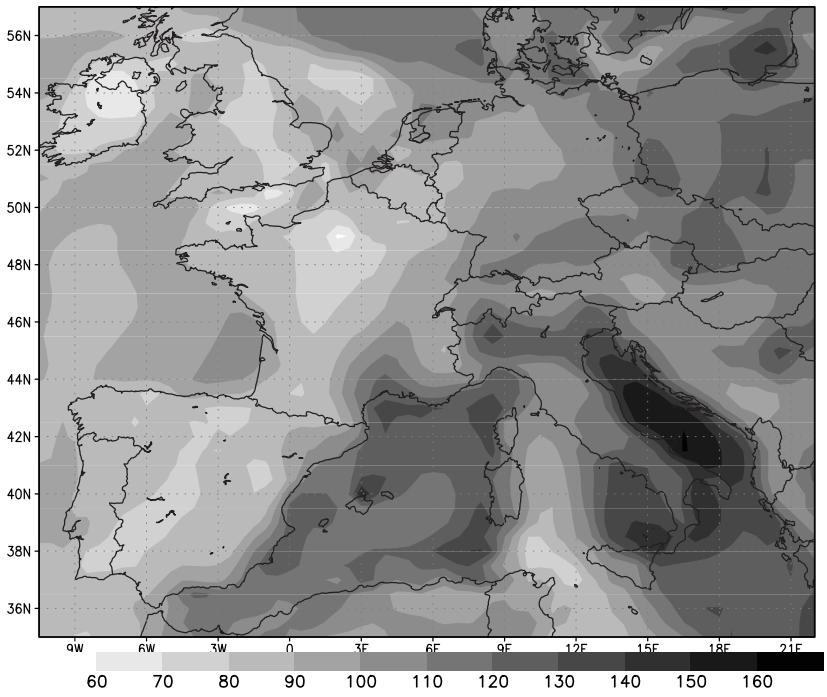


FIG. 1.6 – Carte d'ozone simulé (concentrations en $\mu g \cdot m^{-3}$) à 15 h UT en août 2004.

1.4.1 Transport

Le transport peut être modélisé par un terme d'advection :

$$\frac{\partial \widetilde{c}}{\partial t} = -\mathrm{div}\left(\widetilde{V}\widetilde{c}\right) \tag{1.1}$$

où $\widetilde{c}$ est un champ tridimensionnel de concentrations (pour un polluant) et $\widetilde{V}$ est le vent.

Puisqu'il est impossible de décrire numériquement toutes les échelles, il s'agit maintenant d'écrire l'équation vérifiée par des quantités moyennes. Le vent peut être décomposé en un vent moyen V et des perturbations V' (de moyenne nulle) : $\widetilde{V} = V + V'$. Dans la suite, on appelle abusivement V le vent, abus déjà mentionné dans la section 1.1.2. Le transport dû à la convection (section 1.1.3) est porté par V'.

On décompose $\widetilde{c}$ de la même manière : $\widetilde{c} = c + c'$. Alors, en moyennant l'équation 1.1, on obtient

$$\frac{\partial c}{\partial t} = -\mathrm{div}(Vc) - \mathrm{div}\left(\overline{V'c'}\right) \tag{1.2}$$

où $\overline{V'c'}$ est la moyenne de $V'c'$.

Pour fermer cette équation, on peut approcher $\overline{V'c'}$ ainsi :

$$\overline{V'c'} \simeq -K\nabla c \tag{1.3}$$

où K est une matrice de diffusion turbulente. Cette fermeture est souvent appelée théorie K.

Cependant, l'équation de continuité moyenne s'écrit

$$\frac{\partial \rho}{\partial t} = -\mathrm{div}(\rho V) \tag{1.4}$$

où ρ est la densité de l'air. On souhaite naturellement que, si $c = \rho$, la concentration c suive la même évolution que la densité ρ du fluide porteur. Or, si on reporte l'équation 1.3 dans l'équation 1.2, $c = \rho$ ne redonne pas l'équation de continuité 1.4. On modifie en conséquence l'approximation de l'équation 1.3 par

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

L'équation de transport devient finalement

$$\frac{\partial c}{\partial t} = -\mathrm{div}(Vc) + \mathrm{div}\left(\rho K\nabla\frac{c}{\rho}\right) \tag{1.6}$$

1.4.2 Transport réactif

On suppose maintenant que c est un vecteur de champs tridimensionnels de concentrations. À chaque élément c_i de ce vecteur correspond une espèce chimique. Le champ c_i est sujet au transport (équation 1.6) mais aussi à des réactions chimiques (voir section 1.2). Le champ c_i vérifie une équation de la forme

$$\frac{\partial c_i}{\partial t} = \underbrace{-\mathrm{div}(Vc_i)}_{\text{advection}} + \underbrace{\mathrm{div}\left(\rho K\nabla\frac{c_i}{\rho}\right)}_{\text{diffusion}} + \underbrace{\chi_i(c)}_{\text{chimie}} + S_i - P_i \tag{1.7}$$

où χ_i est le bilan des productions et pertes par réaction chimique de l'espèce i, S_i représente les sources (émissions) de l'espèce i et P_i représente les pertes par lessivage. χ est une fonction des concentrations de toutes les espèces (seulement à la position courante). Il faut bien noter que les dépendances ont été omises. Le vent V, la matrice de diffusion K, les sources S_i et les pertes P_i dépendent de la position et du temps. La fonction χ_i dépend également du temps et de l'espace (via les constantes photolytiques), mais aussi des conditions météorologiques : éclairement (pour la photolyse), température, pression et humidité.

Mécanisme chimique

Le terme χ_i intervenant dans l'équation 1.7 s'estime par l'utilisation d'un mécanisme chimique adapté. On s'intéresse ici aux mécanismes adaptés à la chimie de l'ozone.

Le mécanisme comprend un nombre d'espèces fixé. Il peut être plus ou moins détaillé, selon la précision souhaitée et la complexité de la chimie des espèces impliquées. Un mécanisme pour l'ozone comprend quelques dizaines à quelques centaines d'espèces. Ces espèces ne sont pas nécessairement représentatives d'espèces réelles : elles peuvent regrouper les concentrations de plusieurs espèces réelles. Elles sont qualifiées « d'espèces modèles ». Ces espèces interviennent dans les centaines de réactions qui composent généralement un mécanisme chimique.

Dans cette thèse, on utilise les mécanismes RADM 2 [Stockwell *et al.*, 1990] et RACM [Stockwell *et al.*, 1997]. RADM 2 inclut 61 espèces et 157 réactions chimiques. RACM est composé de 237 réactions pour 72 espèces. Les résultats issus des deux mécanismes sont notamment comparés dans Gross et Stockwell [2003].

RACM inclut 23 réactions photolytiques. Les autres réactions ont des constantes suivant diverses lois, par exemple la loi d'Arrhenius (réactions thermiques).

1.4.3 Conditions aux limites

Pour compléter la description du modèle, il faut traiter des conditions aux limites. En terme de flux, les conditions aux limites latérales (faces verticales du domaine, en supposant que le domaine est un parallélépipède rectangle) et au sommet du domaine sont effectives pour des vents entrant et sont de la forme Vc_i (conditions associées à l'advection). Au sol, les seuls flux pris en compte sont ceux liés au dépôt, notés D_i, et ceux liés aux émissions de surface, notés E_i :

$$K\nabla c_i \cdot n = E_i - D_i \tag{1.8}$$

où n est la normale au sol orientée dans le sens des altitudes croissantes.

1.5 Paramétrisations

Les paramétrisations principales sont reportées dans cette section. Elles peuvent être considérées comme l'état de l'art ou sont au moins largement utilisées dans les modèles de la qualité de l'air.

Elles sont exposées afin de compléter la présentation de la modélisation, et de décrire le système de simulation utilisé (dont l'architecture est présentée au chapitre 2). Une description plus complète des paramétrisations utilisées se trouve dans Njomgang *et al.* [2005], documentation scientifique de la bibliothèque AtmoData présentée plus loin (section 2.2).

1.5.1 Notations

Une liste des notations est disponible page 183.

Les formules sont reportées sous forme discrétisée, avec les approximations numériques utilisées par les simulations présentées dans cette thèse. On suppose que l'espace est discrétisé horizontalement et verticalement par une grille (parallélépipèdes rectangles – appelés mailles ou cellules – dans les coordonnées latitude, longitude et altitude en mètres). La lettre k désigne un indice vertical. Le niveau (centre de maille) k a pour coordonnée verticale z_k et la hauteur de son interface supérieure est notée $\tilde{z}_k$. L'interface k est l'interface supérieure du niveau (ou de la cellule) k.

L'opérateur Δ renvoie la différence entre deux niveaux ou deux interfaces, par exemple :

$$\Delta z_k = z_k - z_{k-1} \tag{1.9}$$

Nombre de Richardson

Le nombre de Richardson, Ri, est calculé ainsi au niveau k :

$$Ri_k = \frac{g\Delta\theta_k}{\max(D\mathcal{W}_k^2, WT^2)\theta_{k-1}\Delta z_k} \tag{1.10}$$

où θ est la température potentielle, WT un seuil minimal sur le module du vent fixé à 0.001 $\mathrm{m}\cdot\mathrm{s}^{-1}$ et $D\mathcal{W}_k$ est le cisaillement du vent donné par

$$D\mathcal{W}_k = \sqrt{\left(\frac{\Delta U_k}{\Delta z_k}\right)^2 + \left(\frac{\Delta V_k}{\Delta z_k}\right)^2} \tag{1.11}$$

avec U_k et V_k les composantes zonales et méridionales du vent respectivement.

1.5.2 Diffusion verticale

Le terme prépondérant de la diffusion est le terme correspondant au transport vertical, sur la diagonale de la matrice de diffusion K (équation 1.7). Ce terme est noté K_{zz} ou simplement K_z (les termes extra-diagonaux étant négligeables). Deux paramétrisations sont décrites : celle de Louis [1979] et celle de Troen et Mahrt [1986].

Paramétrisation de Louis

La paramétrisation proposée dans Louis [1979] introduit la fonction de stabilité

$$F(Ri) = \begin{cases} \left(1 + 3BRi\sqrt{1+DRi}\right)^{-1} & \text{si } Ri \geq 0 \\ 1 - 3BRi\left(1 + \frac{3BCL^2}{(z+z_0)^2}\sqrt{\frac{|Ri|}{27}}\right)^{-1} & \text{si } Ri < 0 \end{cases} \tag{1.12}$$

où z est l'altitude, z_0 est la hauteur de rugosité, généralement fixée à 1 m pour le calcul de $F(Ri)$, et la longueur de mélange L s'écrit

$$L = \frac{\kappa(z+z_0)}{1 + \frac{\kappa(z+z_0)}{\lambda}} \tag{1.13}$$

B, C et D sont des coefficients à ajuster et sont pris égaux à 5. La longueur de mélange asymptotique λ est fixée à 100 m. La constante de Von Kármán κ est prise égale à 0.40.

Comme proposé dans Nordeng [1986]; Pielke [2002], on introduit un seuil minimal pour le nombre de Richardson associé à l'interface k :

$$Ri_k^{\min} = a\left(\frac{\Delta z_k}{\Delta z_0}\right)^b \tag{1.14}$$

avec $a = 0.115$, $b = 0.175$ et $\Delta z_0 = 0.01$ m.

À l'interface k, le coefficient de diffusion verticale vaut

$$\boxed{K_{z,k} = L_k^2 F(Ri_k)\left[\left(\frac{\Delta U_k}{\Delta z_k}\right)^2 + \left(\frac{\Delta V_k}{\Delta z_k}\right)^2\right]} \tag{1.15}$$

Cette paramétrisation repose sur le gradient vertical des vents au point où K_z est estimé.

Paramétrisation de Troen et Mahrt

Troen et Mahrt [1986] propose d'approcher le coefficient de diffusion verticale par

$$\boxed{K_{z,k} = u_* \kappa z_k \Phi_{m,k}^{-1} \left(1 - \frac{z_k}{PBLH}\right)^p} \tag{1.16}$$

où u_* est la vitesse de friction du vent, $PBLH$ est la hauteur de la couche limite atmosphérique et $\Phi_{m,k}$ est une valeur définie ci-dessous.

En couche limite stable, $\Phi_{m,k}$ est donné par

$$\Phi_{m,k} = 1 + 4.7 \frac{z_k}{LMO} \tag{1.17}$$

LMO étant la longueur de Monin-Obukhov, qu'on ne détaille pas ici (voir [Seinfeld et Pandis, 1998], par exemple).

En couche limite convective et dans la couche limite de surface, $\Phi_{m,k}$ devient

$$\Phi_{m,k} = \left(1 - 7 \frac{z_k}{LMO}\right)^{-\frac{1}{3}} \tag{1.18}$$

Enfin, en couche limite convective et au-dessus de la couche limite de surface, $\Phi_{m,k}$ vaut

$$\Phi_{m,k} = \frac{u_*}{w_s} \tag{1.19}$$

où w_s s'exprime en fonction de u_* et des flux de chaleur à la surface.

L'exposant p de l'équation 1.16 est fixé à 2 mais peut raisonnablement varier entre 1.5 et 3. Il faut noter que K_z est fonction décroissante de p. On remarque aussi que la formulation de Troen et Mahrt [1986] est plus paramétrique que celle de Louis [1979]. Elle est en conséquence plus robuste pour de faibles discrétisations verticales.

1.5.3 Dépôt

La perte due au dépôt s'exprime par le flux

$$F_d = -v_d c_1 \tag{1.20}$$

où v_d est la vitesse de dépôt et c_1 la concentration du polluant considéré au premier niveau.

L'approximation classique consiste en une analogie électrique [Wesely, 1989] où apparaît une résistance au transfert, composée de trois résistances en série :

$$\boxed{v_d = \frac{1}{R_a + R_b + R_c}} \tag{1.21}$$

où R_a est la résistance aérodynamique (transport dans l'air vers le sol), R_b la résistance de couche quasi-laminaire (résistance au transport à travers une fine couche au « contact » du sol) et R_c la résistance de canopée (végétation, sol).

La majeure partie de l'incertitude réside dans R_c, sauf éventuellement en conditions stables [Massman *et al.*, 1994; Zhang *et al.*, 2003a].

Résistance aérodynamique

Le calcul de la résistance aérodynamique est classiquement effectué à partir des flux de chaleur ou de moment au sol.

On l'écrit ici pour le flux de chaleur pour lequel on introduit la fonction de stabilité

$$F_m = \begin{cases} 1 - 3B\frac{R_i}{1+3\widetilde{C}\sqrt{-Ri}} & \text{si } Ri < 0 \\ \left(1 + 2B\frac{R_i}{\sqrt{1+DRi}}\right)^{-1} (1 + 3B(1 + DRi)R_i)^{-1} & \text{si } Ri > 0 \end{cases} \tag{1.22}$$

avec $B = D = 5$,

$$\widetilde{C} = BCA_uA_t\sqrt{1 - \frac{z_t}{z_1 + z_t}}\left(\left(\frac{z_1 + z_t}{z_t}\right)^{\frac{1}{3}} - 1\right)^{\frac{3}{2}} \tag{1.23}$$

avec $C = 5$ et où z_t est la hauteur de rugosité associée à la température, supposée égale à $\frac{z_0}{10}$, et

$$A_u = \frac{\kappa}{\ln\left(\frac{z_1+z_0}{z_0}\right)} \tag{1.24}$$

$$A_t = \frac{\kappa}{\ln\left(\frac{z_1+z_t}{z_t}\right)} \tag{1.25}$$

La résistance aérodynamique s'écrit

$$\boxed{R_a = \frac{1}{A_uA_tF_h\mathcal{W}}} \tag{1.26}$$

où $\mathcal{W}$ est le module du vent.

La résistance aérodynamique peut aussi être diagnostiquée sur la base du coefficient de traînée.

Résistance de couche quasi-laminaire

La résistance de couche quasi-laminaire est estimée par

$$\boxed{R_b = \frac{2}{\kappa u_*}\left(\frac{Sc}{Pr}\right)^{\frac{2}{3}}} \tag{1.27}$$

où Sc est le nombre de Schmidt pour l'espèce considérée et Pr le nombre de Prandtl pris égal à 0.74.

Résistance de canopée

La résistance de canopée R_c est plus complexe. Plusieurs paramétrisations existent pour l'approcher. Dans cette thèse, les paramétrisations issues de Wesely [1989] et Zhang *et al.* [2003b] sont utilisées.

Pour ces deux paramétrisations, la résistance R_c s'écrit aussi par analogie électrique et introduit plusieurs résistances, en série ou en parallèle. Ainsi, la résistance R_c proposée dans Wesely [1989] est

$$\boxed{\frac{1}{R_c} = \frac{1}{R_{stom} + R_m} + \frac{1}{R_{lu}} + \frac{1}{R_{dc} + R_{cl}} + \frac{1}{R_{ac} + R_{gs}}} \tag{1.28}$$

où, pour l'espèce s de diffusivité moléculaire D_s, de constante de Henry H_s et de réactivité f_s,

– R_{stom} est la résistance stomatique, qui vaut, pour une température de surface T_{srf} en °C telle que $0\,°\mathrm{C} < T_{srf} < 40\,°\mathrm{C}$,

$$R_{stom} = r_i \left(1 + \left[\frac{200}{SR + 0.1}\right]^2\right) \frac{400}{T_{srf}(40 - T_{srf})} \frac{D_s}{D_{\mathrm{H_2O}}} \tag{1.29}$$

où r_i est la résistance minimale de canopée pour la vapeur d'eau, SR est le rayonnement solaire et la diffusivité moléculaire de H_2O est prise égale à $D_{\mathrm{H_2O}} = 0.25\ \mathrm{cm}^2 \cdot \mathrm{s}^{-1}$. Si T_{srf} sort de l'intervalle $]0\,°\mathrm{C}, 40\,°\mathrm{C}[$, $R_{stom} = +\infty$.

– R_m est la résistance de mésophylle

$$R_m = \frac{1}{\frac{H_s}{3000} + 100 f_s} \tag{1.30}$$

– R_{lu} est la résistance cuticulaire

$$R_{lu} = \frac{R_{LUC}}{10^{-5} H_s + f_s} \tag{1.31}$$

où R_{LUC} est une résistance cuticulaire de référence pour le type de terrain courant (LUC signifie « land use coverage », soit occupation des sols).

– R_{dc} est la résistance de convection thermique

$$R_{dc} = \frac{100}{1 + \frac{1000}{SR+10}} \tag{1.32}$$

– R_{cl} (résistance des surfaces exposées dans la basse canopée) et R_{gs} (résistance de la surface du sol) sont estimées par pondération des résistances associées à SO_2 et O_3 :

$$\frac{1}{R^s_{cl,gs}} = \frac{10^{-5} H_s}{R^{\mathrm{SO_2}}_{cl,gs}} + \frac{f_s}{R^{\mathrm{O_3}}_{cl,gs}} \tag{1.33}$$

– R_{ac} dépend de la hauteur de la canopée et de sa densité. Cette résistance prend une valeur constante par type de terrain et par saison.

La résistance de canopée R_c dépend donc de paramètres météorologiques (température et rayonnement solaire), du type de terrain, de la saison et de l'espèce chimique. Les données manquantes (r_i, par exemple) sont, par exemple, fournies dans Wesely [1989]. Quelques cas particuliers (neige, seuils sur l'humidité) ne sont pas reportés ici.

Dans Zhang *et al.* [2003b], la résistance R_c est estimée de manière plus fine. D'autres types de terrain sont considérés, des données (telle que R_{ac}) diffèrent et la décomposition en résistances est modifiée.

1.5.4 Émissions

Émissions biogéniques

Les émissions biogéniques sont issues de la biomasse. On considère que les polluants sont émis à la surface. Les polluants concernés sont l'isoprène, des terpènes et le monoxyde d'azote. L'approximation proposée dans Simpson *et al.* [1999] est utilisée pour calculer l'émission d'une espèce s, dans la maille x :

$$\boxed{E_{s,x} = \sum_{i=1}^{N_c} \gamma_{s,x}(i) d(i) EF_s(i) LUC_x(i)} \tag{1.34}$$

où $LUC_x(i)$ est la proportion (entre 0 et 1) d'occupation du type de terrain i dans la maille x, N_c est le nombre de classes de terrain, $d(i)$ est la densité foliaire de la biomasse dans la classe i, $EF_s(i)$ est le facteur d'émission pour l'espèce s, et $\gamma_{s,x}$ est un facteur de correction qui tient compte de l'ensoleillement et de la température.

Les facteurs d et EF_s sont fournis pour différents types de végétation notamment dans Simpson *et al.* [1995]; Makar *et al.* [1999]; Simpson *et al.* [1999]. Pour l'isoprène, on a

$$\gamma = \frac{\alpha c_l PAR \exp\left(\frac{c_{t_1}(T_{srf}-T_r)}{RT_rT_{srf}}\right)}{\sqrt{1+\alpha^2 PAR^2}\left[1+\exp\left(\frac{c_{t_2}(T_{srf}-T_m)}{RT_rT_{srf}}\right)\right]} \tag{1.35}$$

où $c_l = 1.066$, $\alpha = 0.0027$, $T_r = 303$ K, $T_m = 314$ K, $c_{t_1} = 95\,000$, $c_{t_2} = 230\,000$, R est la constante des gaz parfaits, T_{srf} est la température de surface et PAR est la part du rayonnement active pour la photosynthèse.

Le facteur pour les émissions de terpènes s'écrit

$$\gamma = e^{0.09\left(T_{srf}-T_r\right)} \tag{1.36}$$

Les émissions de monoxyde d'azote reposent sur

$$\gamma = e^{0.071\left(T_r^{\mathrm{NO}}\right)} \tag{1.37}$$

où

$$T_r^{\mathrm{NO}} = \begin{cases} 0.67\,(T_{srf}-273.15)+8.8 & \text{si } EF_{\mathrm{NO}}(i)=0.9 \\ 0.84\,(T_{srf}-273.15)+3.6 & \text{sinon} \end{cases} \tag{1.38}$$

Les émissions biogéniques sont donc déterminées le type de terrain (en fait, par la végétation), le rayonnement solaire et la température.

Émissions anthropogéniques

Le traitement des émissions anthropogéniques diffère selon les données disponibles. À l'échelle européenne, ces émissions sont fournies par l'organisme européen EMEP sur un maillage de 50 km par 50 km sous forme de taux annuels pour les catégories de polluant NO_x, COV, SO_2 et CO, et pour dix types d'émetteurs (dont le trafic et diverses industries) appelés classes SNAP. Les données sont disponibles à l'adresse `http://webdab.emep.int/`.

Les émissions sont réparties temporellement successivement par mois, par type de jour (jour de semaine, samedi et dimanche), et puis par heure. Les facteurs appliqués dépendent du pays et de la classe SNAP. Spatialement, les émissions d'une maille sont réparties de manière privilégiée sur les forêts et surtout villes selon des facteurs fixes. Une répartition verticale peut aussi être appliquée, surtout dans le cas de polluants principalement d'origine industrielle (cheminées) tel que SO_2.

Les émissions de la catégorie NO_x sont réparties en masse à 90% sur NO, 9.2% sur NO_2 et à 0.8% sur HONO. La répartition des COV est réalisée selon Middleton *et al.* [1990]. Il s'agit d'abord de désagréger les émissions en émissions d'espèces chimiques réelles ; c'est l'étape de « spéciation ». Cette spéciation est fournie par Passant [2002]. Elle s'applique normalement au Royaume-Uni mais est utilisée sur l'ensemble de l'Europe, faute de spéciation adaptée sur les autres pays.

Afin d'obtenir les émissions pour les espèces modèles, les émissions des espèces réelles sont agrégées ; c'est l'étape d'agrégation. Elle s'effectue généralement en deux temps. Les espèces sont d'abord regroupées dans trente-deux classes (proposées dans Middleton *et al.* [1990]) dont les espèces ont des réactivités par rapport à OH similaires. Ensuite, les émissions de ces classes

sont réparties sur la vingtaine (pour les mécanismes chimiques courants – voir section 1.4.2) d'espèces émises.

Les émissions d'espèces réelles, en molécules, doivent être ajoutées avec un facteur correctif tenant compte de leur réactivité par rapport à OH. Par exemple, si une espèce réelle possède une réactivité (par rapport à OH) inférieure à celle de l'espèce modèle (aux émissions de laquelle elle contribue), l'émission de cette espèce est artificiellement augmentée. L'augmentation doit compenser pour la perte de réactivité des émissions attribuées à l'espèce modèle. Le facteur correctif est

$$corr = \frac{1 - \exp\left(-k_{r,\mathrm{OH}} \int [\mathrm{OH}]\,\mathrm{d}t\right)}{1 - \exp\left(-k_{m,\mathrm{OH}} \int [\mathrm{OH}]\,\mathrm{d}t\right)} \tag{1.39}$$

où $k_{r,\mathrm{OH}}$ et $k_{m,\mathrm{OH}}$ sont les réactivités par rapport à OH de l'espèce réelle et de l'espèce modèle respectivement. Les concentrations de OH sont généralement intégrées sur un jour, sur la base de concentrations typiques, ce qui conduit souvent à prendre $\int [\mathrm{OH}]\,\mathrm{d}t = 110$ ppt $\cdot$ min [Middleton *et al.*, 1990; Stockwell *et al.*, 1990].

1.5.5 Atténuation nuageuse

Afin d'estimer l'atténuation des taux de photolyse due aux nuages, on dispose de deux paramétrisations. La première, issue de Madronich [1987]; Chang *et al.* [1987], repose sur un diagnostic de l'extension verticale des nuages. La seconde est plus simple, mais aussi plus robuste, puisqu'elle ne fait intervenir que les couvertures nuageuses (nébulosité) moyenne et haute totales ; elle est proposée dans ESQUIF [2001].

Ces deux paramétrisations nécessitent des informations plus ou moins détaillées sur les nuages. La seconde paramétrisation ne requiert que les nébulosités moyenne et haute, mais elles ne sont pas toujours fournies par le modèle météorologique. Un diagnostic de la distribution des nuages est donc nécessaire.

Diagnostic des nuages

Il est possible d'effectuer un diagnostic des nuages sur la base du contenu en eau liquide (variable souvent difficilement calculée par les modèles météorologiques), en estimant qu'un nuage est présent au-delà d'un certain seuil. Du fait d'une forte sensibilité au seuil, ce diagnostic n'est pas utilisé.

Un seuil sur l'humidité relative est plus robuste. Il est appelé humidité relative critique et vaut

$$\boxed{CRH_k = 1 - \lambda_0 \sigma_k^{a_0} (1 - \sigma_k)^{a_1} (1 + \lambda_1 (\sigma_k - 0.5))} \tag{1.40}$$

où CRH est l'humidité relative critique, σ_k est le rapport entre la pression au niveau k et la pression de surface. Les constantes de la paramétrisation sont mal connues et peuvent varier beaucoup entre les auteurs. Dans cette thèse, on fixe $\lambda_0 = 1.1$, $\lambda_1 = \sqrt{1.3}$, $a_0 = 0$ et $a_1 = 1.1$.

Une paramétrisation plus simple est aussi possible :

$$CRH_k = \begin{cases} CRH_{l_0} & \text{si } P_{l_0} < P_k \\ CRH_{l_1} & \text{si } P_{l_1} < P_k \leq P_{l_0} \\ CRH_{l_2} & \text{si } P_k \leq P_{l_1} \end{cases} \tag{1.41}$$

avec $P_{l_0} = 70\,000$ Pa, $P_{l_1} = 40\,000$ Pa, $CRH_{l_0} = 0.75$, $CRH_{l_1} = 0.95$, $CRH_{l_2} = 0.95$.

Ainsi que Byun et Ching [1999], on impose $CRH = 0.98$ dans la couche limite atmosphérique, ceci pour les deux paramétrisations précédentes.

Les calculs de la fraction nuageuse CF et de la nébulosité sont effectués comme dans Byun et Ching [1999]. La fraction nuageuse estimée par

$$\boxed{CF_k = \begin{cases} 0.34\dfrac{h_k - CRH}{1 - CRH} & \text{si } z_k \leq PBLH \\ \left(\dfrac{h_k - CRH}{1 - CRH}\right)^2 & \text{si } z_k > PBLH \end{cases}} \tag{1.42}$$

où h_k est l'humidité relative au niveau k.

La nébulosité est calculée pour trois couches (en-dessous de 80 000 Pa, jusqu'à 45 000 Pa et au-dessus). Dans chaque couche, on considère que le nuage est situé dans la plage où la fraction nuageuse est supérieure 50% de son maximum dans la couche en question. La nébulosité est alors estimée par intégration de la fraction nuageuse sur la couche :

$$\mathcal{N} = \frac{\sum_{k_{\text{base}}}^{k_{\text{sommet}}} CF_k \Delta\tilde{z}_k}{\sum_{k_{\text{base}}}^{k_{\text{sommet}}} \Delta\tilde{z}_k} \tag{1.43}$$

où k_{base} et k_{sommet} sont les indices correspondant à la base et au sommet du nuage.

Atténuation des taux de photolyse

La première méthode, issue de Madronich [1987]; Chang *et al.* [1987], corrige un taux de photolyse par temps clair J_{clair} par un coefficient $\mathcal{A}_u$ au-dessus des nuages et par un coefficient $\mathcal{A}_d$ en-dessous. Dans les nuages, le coefficient correctif est obtenu par interpolation linéaire. Les coefficients sont calculés ainsi :

$$\boxed{\begin{array}{c} \mathcal{A}_d = 1 - \min(1, \mathcal{N}_m + \mathcal{N}_h)(1 - 1.6Tr\cos Z) \\ \mathcal{A}_u = 1 + \min(1, \mathcal{N}_m + \mathcal{N}_h)(1 + (1 - Tr)\cos Z) \end{array}} \tag{1.44}$$

où Z est l'angle zénithal, $\mathcal{N}_m$ et $\mathcal{N}_h$ sont les nébulosités moyennes et hautes respectivement, et Tr est la transmissivité estimée par

$$Tr = \frac{5 - e^{-\tau}}{4 + 0.42\tau} \tag{1.45}$$

où τ est l'épaisseur optique calculée sur la base du contenu en eau liquide à l'intérieur des nuages diagnostiqués.

La seconde paramétrisation a été proposée dans ESQUIF [2001]. Le coefficient d'atténuation est constant sur la verticale et vaut

$$\mathcal{A} = (1 - b\mathcal{N}_m)(1 - a\mathcal{N}_h)e^{-cB} \tag{1.46}$$

avec

$$B = \frac{\int_{z=0}^{z<1500\text{ m}} \min(0, h(z) - 0.7)\,\mathrm{d}z}{\int_{z=0}^{z<1500\text{ m}} (1 - 0.7)\,\mathrm{d}z} \tag{1.47}$$

où h est l'humidité relative, $a = 0.1$, $b = 0.3$ et $c = 1.5$.

1.6 Données

Dans les paramétrisations précédemment exposées interviennent de nombreuses données, depuis les constantes de réaction jusqu'aux champs météorologiques. Les résultats de simulation sont très sensibles à ces données. Or la plupart d'entre elles sont fortement entachées d'incertitudes. Elle jouent donc un rôle important dans la qualité des résultats. Les principales données sont listées dans cette section.

1.6.1 Occupation des sols

L'occupation des sols, déjà mentionnée précédemment sous l'acronyme LUC (pour « land use cover »), décrit les types de terrain sur un domaine considéré, souvent avec une résolution de l'ordre du kilomètre carré. Cette donnée permet d'abord de distinguer les terres des étendues d'eau. Sur terre, plusieurs types de terrain sont aussi identifiés : type de culture agricole, type de forêt, étendue urbaine.

Dans cette thèse, deux descriptifs d'occupation des sols sont utilisés :

- les données de l'USGS[2] (U.S. Geological Survey) couvrent la surface de la Terre avec une résolution de 1 km^2 Lambert. Vingt-quatre types de terrain sont identifiés. Des observations satellitaires faites en 1992 et 1993 ont servi à produire ces données.
- les données GLCF[3] (Global Land Cover Facility) couvrent la surface de la Terre avec une résolution de 1 km^2 et la décrivent par quatorze classes. Elles sont fondées sur des observations satellitaires récoltées entre 1981 et 1994.

Une résolution de 1 km^2 signifie qu'un type de terrain est associé à chaque kilomètre carré. Les données de l'USGS contiennent plus de classes, mais celles de GLCF sont distribuées en proportions plus égales sur l'Europe.

1.6.2 Émissions

Les émissions ont été décrites dans la section 1.5.4.

1.6.3 Données météorologiques

Les données météorologiques utiles aux simulations ont été introduites dans la section 1.5, au fur et à mesure de la description des paramétrisations utilisées. Pour mener à bien une simulation, il faut donc disposer, pour mémoire, d'une vingtaine ou d'une trentaine de variables météorologiques. Parmi les plus importantes, on trouve bien sûr le vent, la température, les flux de surface, la hauteur de couche limite ou encore l'intensité du rayonnement solaire.

Les modèles météorologiques ne proposent pas toujours toutes les variables nécessaires aux paramétrisations. En conséquence, quelques paramétrisations sont proposées pour diagnostiquer les variables indisponibles (par exemple, la nébulosité – section 1.5.5).

Il faut aussi noter que les modèles météorologiques ont leurs propres systèmes de coordonnées, horizontalement et verticalement. Ils génèrent des fichiers dans des formats qui peuvent leur être propres. Dans cette thèse, les champs météorologiques proviennent de l'ECMWF (European Centre for Medium-Range Weather Forecasts) ou du modèle MM5 (Fifth-Generation NCAR / Penn State Mesoscale Model).

1.6.4 Constantes de réaction

Les constantes de réaction des mécanismes chimiques sont de deux natures. Il y a d'une part les réactions dont les constantes sont estimées à partir de quelques coefficients et quelques variables météorologiques ; un bon exemple concerne les réactions suivant la loi d'Arrhenius. D'autre part, il faut estimer les constantes photolytiques. Les constantes photolytiques par temps clair sont corrigées comme indiqué à la section 1.5.5. Ces constantes par temps clair sont des données générées par JPROC, le module dédié du système CMAQ (Community Multiscale Air Quality Model). Pour chaque espèce, elles dépendent du jour de l'année, de l'angle horaire, de l'altitude et de la latitude.

[2]Données disponibles à l'adresse `http://edcsns17.cr.usgs.gov/glcc/`.

[3]Données disponibles à l'adresse `http://glcf.umiacs.umd.edu/data/landcover/data.shtml`.

1.6.5 Concentrations de polluants

Des concentrations de polluant sont nécessaires pour initialiser les simulations et surtout pour fournir des conditions aux limites. Les simulations présentées dans cette thèse intègrent des concentrations issues du modèle de chimie-transport global Mozart 2 [Horowitz *et al.*, 2003]. Les données correspondantes sont disponibles sur le portail de données du NCAR (National Center for Atmospheric Research, États-Unis), à l'adresse `https://cdp.ucar.edu/`. Il s'agit de concentrations disponibles sur tout le globe pour une année typique. Une année météorologique typique a été construite sur la base de dix années météorologiques.

1.6.6 Remarques

La liste précédente ne détaille pas l'ensemble des données et n'indique pas les formats et tailles des données manipulées. Un point à noter cependant est la multitude des sources de données. Divers formats, diverses coordonnées, diverses natures de données, ... doivent être gérés. Des données multidimensionnelles et de grande taille sont traitées dans le processus de simulation.

Chapitre 2

Système de simulation

L'architecture d'un système de simulation complet pour la qualité de l'air est proposée. Elle répond aux contraintes identifiées : la gestion de données multidimensionnelles, l'existence de nombreuses paramétrisations physiques, et l'intégration de méthodes de haut niveau telles que l'assimilation de données ou la prévision d'ensemble. Les deux premières contraintes sont gérées par une bibliothèque dédiée. La dernière contrainte suggère la conception de pilotes en charge du déroulement des calculs. Polyphemus est une réalisation partielle du système préconisé. Les schémas numériques choisis et les fonctionnalités pour le traitement des résultats de calcul sont aussi présentés. Sur cette base, une évaluation de Polyphemus, sur quatre mois de l'année 2001 et à l'échelle européenne, est exposée.

Sommaire

Les notes techniques suivantes servent de base à ce chapitre :

– Mallet, V., Quélo, D. et Sportisse, B. (2005). Software architecture of an ideal modeling platform in air quality – A first step : Polyphemus. Rapport technique 11, CEREA,
– Mallet, V. et Sportisse, B. (2005b). Data processing and parameterizations in atmospheric chemistry and physics : the AtmoData library. Rapport technique 12, CEREA.

La seconde note étant à forte tonalité informatique, elle ne se retrouve que très partiellement dans ce chapitre et est placée en annexe (chapitre A).

2.1 Architecture du système Polyphemus

2.1.1 Introduction

Cette thèse repose sur le système de simulation Polyphemus [Mallet *et al.*, 2005] qui a été développé en grande partie lors de la thèse. Il vise à répondre aux besoins exigeants des études et applications évoluées qui seront des standards dans les années à venir. Son élaboration a donc été l'objet d'une réflexion sur l'architecture qu'aurait un système de simulation « idéal ». La réflexion a été guidée par les seuls besoins et les seules contraintes de la simulation de la qualité de l'air. Elle ne cherche pas à reposer sur un système ou sur des composants établis. La démarche est par conséquent ambitieuse, mais il convient de noter qu'elle n'est pas une fin. Elle vise à faciliter grandement une recherche qui s'appuie sur des modèles et des méthodes de plus en plus complexes. Ce chapitre est lui-même un préalable, volontairement technique, aux études qui suivent.

Polyphemus est une réalisation partielle du système préconisé. Il faut noter que le système proposé doit permettre de traiter des problèmes autres que la photochimie ; il s'agit d'un système complet de simulation de la qualité de l'air, voire de ses impacts.

L'architecture logicielle des systèmes de simulation de la qualité de l'air doit être soigneusement étudiée du fait de la diversité des applications visées, des grandes quantités de données manipulées et de la complexité des modèles et méthodes utilisés (par exemple en assimilation de données).

En qualité de l'air, la modélisation concerne à la fois des cas relativement simples comme la dispersion de traceurs passifs à petite échelle (quelques kilomètres) et des cas complexes tels l'estimation de distributions d'aérosols sur un continent ou sur le globe. Ces applications reposent sur des modèles communs ou similaires (au moins pour le transport), mais il existe de grandes différences entre elles dues aux ajouts spécifiques tels que les mécanismes chimiques. De plus, pour une même application, une certaine diversité demeure dans les données utilisées, pour la chimie ou dans les champs météorologiques par exemple. Les objectifs peuvent aussi fortement varier, depuis de simples simulations jusqu'à l'assimilation de données ou la prévision d'ensemble.

La maturité des modèles actuels permet justement la mise en œuvre des méthodes de haut niveau que sont l'assimilation de données et la prévision d'ensemble. Un système de modélisation bien conçu doit permettre l'application de ces méthodes sans nécessiter de longs ou fastidieux développements.

Un point important, déjà mentionné dans la section 1.6, concerne la grande quantité de données manipulées. Ceci impose une architecture logicielle sûre et robuste. Les performances (coût calcul) doivent aussi être bien maîtrisées puisque plusieurs applications conduisent à des coûts très élevés (aérosols, assimilation de données, prévision d'ensemble).

À partir des remarques précédentes, l'importance d'une conception soignée apparaît clairement pour qui souhaite mener des travaux avancés dans un cadre sûr, efficace et pérenne. On peut noter que les modèles de chimie-transport actuels comprennent les éléments suivants :

1. des bases de données et des capacités de traitement de données ;
2. des paramétrisations physiques (section 1.5) ;
3. un coeur numérique (intégration en temps) ;
4. des méthodes de haut niveau (c'est-à-dire s'appliquant à l'intégration numérique de l'équation de dispersion-réaction 1.7, intégration traitée comme une « boîte noire »), comme l'assimilation de données et la prévision d'ensemble.

Ces quatre composants sont généralement imbriqués dans un seul système, fait d'un bloc, appelé modèle de chimie-transport (en anglais, « chemistry-transport model »). Cette conception, héritée des premiers modèles encore simplifiés (et qui étaient donc légitimement d'un seul tenant), associe des composants de natures très différentes. Il s'agit désormais de proposer une architecture plus saine pour porter les nouvelles applications de haut niveau. Un effort doit aussi être consenti pour permettre l'intégration d'autres modèles de chimie-transport, ou de modèles en amont (météorologique) ou en aval (impact sanitaire).

Dans la section suivante, on décrit plus précisément les contraintes d'un système de modélisation et ses applications. L'architecture proposée est ensuite présentée et le contenu de ses composantes est détaillé. Enfin, le système Polyphemus est décrit.

2.1.2 Contraintes

Modèle numérique

Un modèle de chimie-transport calcule des concentrations de polluants dans un domaine discrétisé tridimensionnel. Il résout l'équation de dispersion-réaction 1.7 discrétisée. Un système de modélisation est un système de simulation numérique gérant les flux de données et les calculs (« workflow », en anglais), en particulier ceux du modèle de chimie-transport, dans une série d'applications pour la modélisation.

Des schémas numériques sont dédiés au transport (advection et diffusion). Ils reposent généralement sur des schémas aux différences finies puisque la plupart des modèles de chimie-transport sont discrétisés ainsi. Parfois, des paramétrisations sous-maille sont utilisées pour affiner les calculs près des sources. On peut néanmoins prévoir que des discrétisations à base d'éléments finis apparaissent pour traiter les problèmes fortement dépendant de quelques sources ponctuelles. La chimie pose des problèmes numériques plus difficiles car elle introduit un système raide lié à la dispersion des temps caractéristiques (plusieurs décades – voir section 2.3) [Sportisse, 1999].

Pour les applications multiphasiques, des problèmes numériques encore délicats persistent [Debry, 2004]. Cependant, pour nombre d'applications, les schémas numériques sont satisfaisants. S'il convient de garder une certaine flexibilité numérique, les contraintes de ce point de vue sont faibles. Sachant de plus que les modèles de chimie-transport ont des discrétisations similaires, il est possible de concevoir un système accueillant plusieurs modèles aux intégrations numériques très différentes.

Entrées du modèle numérique

Contrairement aux schémas numériques, les données d'entrée (celles qui interviennent directement ou indirectement dans l'équation 1.7, à l'exception des concentrations) demeurent un point-clé. On peut distinguer les données fournies (voir section 1.6), telles que les données météorologiques, et les données calculées par des paramétrisations (dont celles introduites dans la section 1.5). Le système de simulation doit pouvoir gérer les nombreuses données d'entrée et paramétrisations.

Les données proviennent de diverses sources, et sont dans des formats différents et sur des grilles qui leur sont propres. Elles représentent souvent plusieurs giga-octets. Enfin, elles interviennent à plusieurs niveaux, soit très en amont (occupation des sols, par exemple) soit très en aval (constantes de réaction chimique, utilisées au moment de l'intégration numérique). Un système performant doit donc être muni de bonnes capacités de gestion des données.

Les nombreuses paramétrisations physiques (section 1.5) vont de paire avec les données puisqu'elles les transforment dans le but de générer des données utiles à l'intégration numérique

des concentrations de polluants. Il faut bien noter que la plupart des paramétrisations ont des paramétrisations alternatives. On peut citer l'exemple des paramétrisations de Louis [1979] et de Troen et Mahrt [1986] pour le calcul du coefficient de diffusion verticale (voir section 1.5.2). Le système de modélisation doit être suffisamment flexible pour opérer des changements entre paramétrisations estimant les mêmes quantités.

On conclut donc que le système doit disposer de bons outils pour la gestion des données et doit garantir une grande flexibilité dans l'utilisation des paramétrisations physiques.

Applications visées

Un modèle de chimie-transport peut être utilisé pour des applications assez diverses. Les polluants concernés par les simulations peuvent être des traceurs passifs, c'est-à-dire des espèces ne participant pas à des réactions chimiques. Il s'agit des simulations parmi les plus simples. On peut par exemple citer la dispersion de radionucléides où seul un terme de décroissance radioactive est ajouté à l'équation 1.7 (où il prend la place du terme de chimie χ). Les métaux lourds peuvent faire partie de ces applications encore assez simples puisqu'ils sont raisonnablement simulés avec peu d'espèces chimiques et une chimie linéaire. À un niveau de complexité supérieur, on trouve les mécanismes chimiques comme ceux du cycle de l'ozone ou des oxydes d'azote. Les cas les plus complexes concernent les aérosols qui interagissent avec la phase gazeuse, et dont la distribution (en masse) et la composition évoluent selon des processus physiques difficiles à simuler (nucléation, évaporation/condensation, coagulation).

Les applications peuvent concerner des échelles spatiales très différentes, depuis l'échelle dite locale (quelques dizaines de mètres) jusqu'à l'échelle globale. À très petite échelle, l'équation 1.7 et plusieurs paramétrisations ne s'appliquent plus (la description eulérienne de la diffusion n'étant plus valable). Des échelles régionales à l'échelle globale, l'équation 1.7 et les paramétrisations présentées au chapitre 1 sont couramment utilisées.

Les objectifs du système peuvent aussi varier fortement. Il y a tout d'abord la prévision qui consiste simplement en une simulation délivrant des concentrations sur quelques jours. Des simulations sur plusieurs années sont réalisées, souvent pour tester l'impact de réductions d'émissions. Des simulations intermédiaires, de quelques jours à quelques mois, sont des études d'épisodes particuliers où l'objectif est d'obtenir les meilleures performances dans les comparaisons aux observations. On peut aussi relever les simulations grossières qui constituent des cas réalistes pour des travaux académiques.

L'intégration de l'équation 1.7 permet d'obtenir une évolution temporelle de concentrations de polluants. De nouvelles applications délivrent un ensemble de telles évolutions. Dans un contexte opérationnel, on parle de prévision d'ensemble : le système génère plusieurs prévisions, souvent dans le but de décrire l'incertitude associée aux prévisions. Il s'agit alors d'approcher la densité de probabilité des concentrations. Un ensemble de prévisions peut être généré sur la base de plusieurs modèles [Delle Monache et Stull, 2003], ce qui requiert une plate-forme de simulation avec plusieurs modèles. Une autre possibilité réside dans l'utilisation d'un modèle suffisamment flexible pour proposer plusieurs jeux de données, plusieurs paramétrisations pour un même champ et plusieurs schémas numériques pour une même équation [Mallet et Sportisse, 2006]. Un même système dispose alors de plusieurs configurations qui peuvent *de facto* être vues comme autant de modèles.

La dernière application majeure est l'assimilation de données. Elle introduit des méthodes pour tenir compte des mesures passées dans le but d'améliorer les prévisions [par exemple, Elbern et Schmidt, 2001; Segers, 2002] ou, s'il s'agit de modélisation inverse, dans le but d'affiner certaines données d'entrées (les émissions en sont un exemple classique, Quélo *et al.* [2005]). Une exigence importante réside dans la nécessité (pour certaines méthodes) de disposer d'un modèle linéaire tangent (assimilation séquentielle avec le filtre de Kalman) ou d'un modèle adjoint

(assimilation variationnelle). Un système de modélisation doit viser l'inclusion des deux types d'assimilation (séquentiel et variationnel), le choix dépendant du type d'application visée.

Conclusion

Sur la base des sections précédentes, les principales exigences d'un système de modélisation complet sont listées. Il apparaît d'abord que le système doit intégrer plusieurs applications spécifiques, notamment du point de vue des polluants traités. Il convient donc d'avoir un système flexible. Un deuxième point-clé concerne les paramétrisations utilisées dans la génération des champs d'entrée, et la grande quantité de données gérées par le système. Le dernier point à relever réside dans les natures différentes des applications : simulations académiques, prévisions, prévisions d'ensemble, assimilation de données. Les systèmes de simulations modernes doivent porter toutes ces applications, et il s'agit d'une contrainte forte dans leur conception.

2.1.3 Architecture proposée

Modèles existants

Avant de décrire l'architecture proposée, il est utile de considérer la structure des modèles existants. De nombreux modèles ont été développés, parmi lesquels on trouve : Chimere [Schmidt *et al.*, 2001], CMAQ [Byun et Ching, 1999], DEHM [Christensen, 1997], EMEP [Simpson *et al.*, 2003], Eurad [Hass, 1991], Lotos [Builtjes, 1992], Polair3D [Boutahar *et al.*, 2004], etc. Une revue assez complète est proposée dans Russell et Dennis [2000]. Les modèles actuels ne satisfont pas aux contraintes précédemment exposées. Quelques-uns ne disposent pas d'outils performants pour la gestion des données ou ne proposent que peu d'options (concernant les paramétrisations ou jeux de données) ; ils sont donc restreints à un petit nombre d'applications. D'autres n'ont pas de capacités avancées et bien intégrées en assimilation de données. Rares sont ceux qui proposent à la fois l'assimilation de données séquentielle et variationnelle. Enfin, la prévision d'ensemble n'est pas une possibilité offerte par tous ces modèles.

Le nombre important de modèles pose la question de l'interopérabilité et de la mise en commun de développements. Il serait intéressant de pouvoir comparer ces modèles de manière équitable et de pouvoir les utiliser en prévision d'ensemble, ce qui imposerait l'élaboration d'une plate-forme commune. Il faut noter que les mêmes paramétrisations et des capacités similaires de gestion des données ont été développées dans ces modèles. Un développement communautaire aurait permis à la communauté de réduire ses efforts de développement. Il paraît dans ce sens pertinent de viser des développements communs pour les prochaines applications.

La conception historique des systèmes en qualité de l'air veut que les modèles soient plus ou moins faits d'un seul tenant. Ils ont donc une modularité très limitée et le partage de développement ne peut se faire que difficilement en dehors d'un modèle spécifique (modèle communautaire, rarement accepté, même si le modèle CMAQ joue raisonnablement ce rôle aux États-Unis).

Structure d'ensemble

On peut distinguer trois grands pôles dans un système de modélisation :

1. les *données* : les données « brutes » intervenant dans l'équation de dispersion-réaction 1.7, telles que le vent V fourni par un modèle météorologique, et les données brutes utilisées par les paramétrisations, comme les données de sol. Sur la base du descriptif de la section 1.6, les données brutes sont principalement les champs météorologiques, les inventaires d'émission, l'occupation des sols (et les données associées aux différents types de sols) et les constantes intervenant en chimie.

2. les *paramétrisations* : les paramétrisations physiques ont été détaillées à la section 1.5. Un ensemble de paramétrisations à jour est un élément essentiel d'un bon système de simulation. Les paramétrisations sont déterminantes sur les processus prédominants, par exemple dans l'estimation de la diffusion turbulente.
3. l'*intégration numérique* : il s'agit de la dernière étape du processus de simulation puisqu'elle rend les concentrations en intégrant numériquement l'équation 1.7.

La stratégie proposée pour porter ces pôles consiste en :

1. l'établissement de *bases de données*. Le système doit pouvoir s'appuyer sur des bases de données suffisamment fournies afin d'alimenter plusieurs modèles de chimie-transport, et adaptées de sorte à satisfaire à toutes les applications possibles.
2. la constitution d'une *bibliothèque* (ou de plusieurs bibliothèques) pour la gestion des données et pour rassembler les paramétrisations. L'idée est d'extraire des modèles de chimie-transport, qui sont faits d'un bloc, les fonctionnalités de gestion des données et les paramétrisations physiques qui peuvent alors être utilisées dans différents programmes voire différente équipes de recherche. Une bibliothèque est la structure la plus adaptée pour remplir cet objectif, comme cela est expliqué à la section 2.1.3.
3. des *programmes pour la génération des données d'entrée* (au modèle de chimie-transport) reposant sur des appels aux fonctions des bibliothèques. Les bibliothèques pourraient être simplement appelées par le modèle de chimie-transport. Néanmoins, disposer de plusieurs programmes externes, exécutés préalablement au modèle de chimie-transport, semble une meilleure solution, détaillée dans la suite.
4. un *modèle de chimie-transport* limité à l'intégration en temps de l'équation 1.7. Il s'agit donc de ne conserver que le coeur numérique de ce que sont actuellement les modèles de chimie-transport. Quelques calculs physiques, appels à des paramétrisations, restent cependant acceptables dans certains cas.

La figure 2.1 schématise la structure d'ensemble. On remarque qu'un point essentiel de l'architecture est une séparation nette entre les composants qui sont par nature différents. La conséquence est un accroissement de la flexibilité du système.

On détaille désormais le contenu des quatre niveaux précédemment identifiés (bases de données, bibliothèques, programmes de génération des données d'entrée et modèle de chimie-transport).

Bases de données

Ainsi qu'il est mentionné précédemment, de grandes quantités de données sont manipulées lors d'une simulation. Des bases de données peuvent faciliter la gestion de ces données. Il est au moins nécessaire de rassembler l'ensemble des données provenant de diverses sources : modèles météorologiques, modèles de chimie-transport, bases de données pour la chimie, etc.

Une avancée consisterait à définir des formats de fichiers standards. Il paraît difficile de se restreindre à un seul type de fichiers à cause de la diversité des données manipulées. Néanmoins, un type de fichier standard pour les données météorologiques et pour les concentrations de polluants semblerait raisonnable. Par exemple, un format tel que NetCDF pourrait convenir. Les données d'observation pourraient être stockées dans un format spécifique. Enfin, les données chimiques, généralement classées par espèce ou par réaction chimique, pourraient aussi bénéficier d'un format particulier.

Quoi qu'il en soit, définir des fichiers standards nécessite l'accord d'une grande partie des acteurs du domaine.

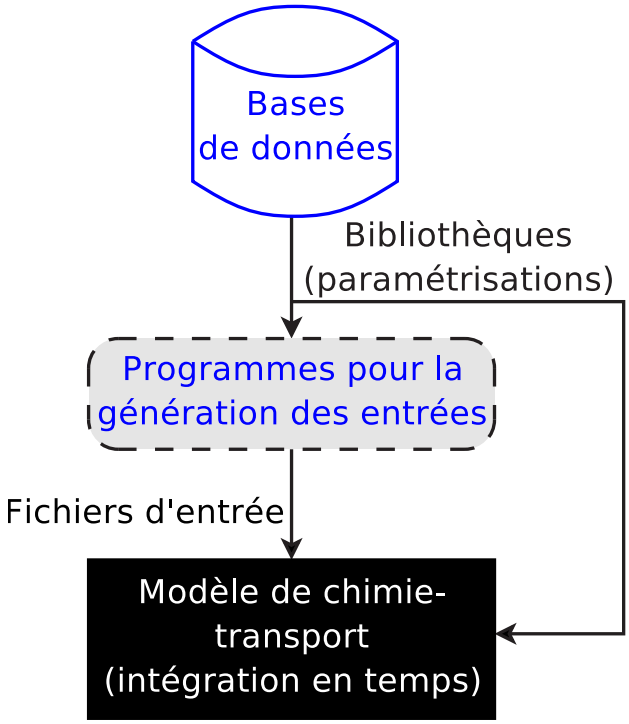


FIG. 2.1 – Architecture proposée pour un système de modélisation flexible et complet.

Bibliothèques pour la physique et la chimie atmosphériques

De la nécessité de bibliothèques Les langages évolués, qui ne sont pas dévolus au calcul scientifique, nécessitent l'introduction de bibliothèques implémentant des tableaux multidimensionnels, des listes chaînées, etc. Ces bibliothèques proposent donc des structures, appelées *objets* dans les langages orientés objet. D'autres bibliothèques sont une collection de *fonctions*; Blas [Lawson *et al.*, 1979] et Lapack [Anderson *et al.*, 1999] sont des exemples connus pour l'algèbre linéaire. Dans le contexte de la chimie atmosphérique, les deux types de bibliothèques sont utiles puisqu'il faut proposer des structures pour une gestion efficace des données et des fonctions pour les paramétrisations.

En effet, étant donné l'usage intensif de données multidimensionnelles lors des simulations, des structures de données adaptées peuvent être d'une grande aide. Une bibliothèque est le cadre logiciel le plus pertinent pour fournir de telles structures à de nombreux programmes.

Il en est de même pour les paramétrisations physiques. Une fois implémentée, une paramétrisation est simplement une fonction que le système de modélisation appelle. Collecter l'ensemble de ces paramétrisations (fonctions) dans une bibliothèque est une démarche naturelle. Ceci est d'autant plus vrai pour les paramétrisations les plus simples qui sont appelées maintes fois, comme par exemple la conversion de l'humidité spécifique en humidité relative. Une bibliothèque évite les duplications.

On peut citer d'autres avantages. D'abord, une bibliothèque peut facilement être étoffée : il suffit de lui ajouter des fonctions. La plupart des paramétrisations qui peuvent être implémentées sont indépendantes les unes des autres. Ensuite, les améliorations introduites dans une bibliothèque deviennent du même coup disponibles dans tous les programmes qui l'appellent. Par exemple, la correction d'une erreur dans une paramétrisation est immédiatement propagée à tous les programmes qui utilisent cette paramétrisation. Enfin, il convient d'ajouter qu'une bibliothèque est aisément partagée dans une même équipe ou même dans une communauté scientifique. Une forte coordination entre les intervenants n'est pas nécessaire.

Contenu des bibliothèques Comme indiqué précédemment et comme illustré sur la figure 2.1, les bibliothèques sont en charge

1. de la gestion des données par le biais de structures complètes et associées à des fonctions pour les manipuler. Une bibliothèque orientée objet est fortement recommandée puisqu'elle permet de définir des structures avancées. Parmi les fonctionnalités requises, on peut relever la possibilité de lire et écrire des fichiers de différents formats, la disponibilité de fonctions d'interpolation et d'analyse statistique, et la manipulation de données associées à des coordonnées.
2. des paramétrisations et, plus généralement, des fonctions physiques. Ces fonctions sont supposées manipuler les structures de données précédemment introduites. Elles couvrent la plupart des besoins depuis les fonctions physiques basiques (température potentielle, nombre de Richardson, etc.) jusqu'aux paramétrisations complexes pour évaluer des flux de surface ou pour détecter des nuages.

Un exemple de bibliothèque pour le traitement des données et des paramétrisations en chimie atmosphérique est AtmoData [Mallet et Sportisse, 2005b, note technique placée en annexe – chapitre A – dans laquelle on trouve aussi des arguments supplémentaires en faveur d'une telle bibliothèque]. Cette bibliothèque est utilisée par Polyphemus.

Les bibliothèques peuvent aussi faciliter le traitement des mécanismes chimiques, des dates, des fichiers de configurations, des interfaces graphiques, etc., qui interviennent éventuellement dans le processus de simulation.

Génération des données d'entrée

Le troisième niveau de la figure 2.1 est un ensemble de programmes destiné à générer les données d'entrée du modèle de chimie-transport. Ils font les appels nécessaires aux bibliothèques précédemment décrites. Par exemple, l'un de ces programmes peut estimer les coefficients de diffusion verticale selon une des paramétrisations de la bibliothèque. Un tel programme extrait les champs météorologiques d'une base de données, fait les appels nécessaires aux bibliothèques afin de calculer les coefficients de diffusion verticale, et écrit les résultats dans un fichier lu plus tard par le modèle de chimie-transport.

Une question est de savoir si ces programmes ne doivent pas être intégrés au modèle de chimie-transport, comme il est d'usage. Il faut noter que, dans la figure 2.1, des appels aux bibliothèques sont autorisés dans le modèle de chimie-transport. Cependant, il convient d'éviter ces appels. L'inclusion ou non d'un calcul de paramétrisation dans le modèle de chimie-transport dépend :

1. du nombre de paramétrisations disponibles pour le calcul du champ en question et de la qualité de ces paramétrisations. S'il existe plusieurs paramétrisations pour l'estimation d'un même champ (comme c'est le cas pour la diffusion verticale, par exemple – voir 1.5.2), un travail de modélisation et d'intercomparaison doit être effectué sur ces paramétrisations. Il est alors préférable d'effectuer les calculs en dehors du modèle de chimie-transport. Cela permet de s'abstraire des contraintes du modèle (qui effectue d'autres calculs sans lien avec la paramétrisation travaillée) et donc de gagner en flexibilité lors du travail sur les paramétrisations. À l'inverse, si une paramétrisation n'est pas destinée à changer (si elle est très bien connue), elle peut être intégrée dans le modèle de chimie-transport.
2. le temps de calcul associé à la paramétrisation. Si le temps de calcul est élevé, la paramétrisation ne doit être appelée qu'une fois, et donc à l'extérieur du modèle de chimie-transport. En effet, une simulation donnée est souvent exécutée plusieurs fois (notamment

en prévision d'ensemble ou en assimilation de données). Conserver des champs calculés une seule fois par des programmes dédiés est plus efficace.

3. la taille des données. Si les données (champs calculés par la paramétrisation) ne peuvent être sauvegardés à cause de leur grande dimension, elles sont nécessairement calculées pendant l'intégration en temps, donc dans le modèle de chimie-transport. Un exemple concerne les coefficients de lessivage humide qui sont tridimensionnels, dépendant du temps et calculés pour toutes les espèces.

En pratique, il est nettement plus commode de travailler sur des programmes externes dans le but d'améliorer des paramétrisations. Il est aussi plus sûr de générer les champs pas à pas. Ainsi il est recommandé de calculer la plupart des champs à l'extérieur du modèle de chimie-transport, lequel est uniquement en charge de l'intégration numérique de l'équation 1.7 discrétisée.

Modèle de chimie-transport

Comme mentionné précédemment, un modèle de chimie-transport doit permettre de mener à bien plusieurs types d'application. En particulier, l'équation 1.7 introduit un terme de chimie χ. Puisque les mécanismes chimiques diffèrent fortement entre eux, le modèle de chimie-transport doit être flexible de ce point de vue. La majeure partie du modèle doit donc être indépendante du mécanisme chimique. La chimie est alors un module qui peut être modifié sans affecter l'intégration numérique des autres processus. Le système Polyphemus utilisé dans cette thèse intègre le modèle de chimie-transport Polair3D [Boutahar *et al.*, 2004] qui répond à cette exigence.

Une contrainte supplémentaire provient de l'assimilation de données variationnelle qui requiert un modèle adjoint. On rappelle qu'un modèle adjoint est une version différenciée du modèle. Un appel à ce modèle adjoint permet de calculer les dérivées d'une sortie par rapport à toutes les entrées. Pour l'obtenir, il est conseillé d'utiliser la différenciation automatique [par exemple avec le logiciel O∂yssée, Faure et Papegay, 1998], technique qui permet de générer le code informatique du modèle adjoint sur la base du code informatique du modèle (dit direct). La différenciation automatique permet aussi de générer un modèle linéaire tangent, utile en assimilation de données séquentielle. Le modèle linéaire tangent est aussi une version différenciée du code. Une intégration (en temps) de ce modèle permet d'obtenir les dérivées de toutes les sorties du modèle par rapport à une entrée. Un code informatique ne peut être différencié automatiquement que s'il est construit d'une manière adéquate. Par exemple, tous les langages informatiques ne peuvent pas être différenciés automatiquement à cause de l'absence de logiciels dédiés. La structure du code a aussi son importance car toutes les opérations d'un code ne peuvent pas être différenciées (les opérations de lecture et écriture sur fichier, principalement).

Indépendamment des contraintes de la différenciation automatique, il est tout de même recommandé de bien séparer les schémas numériques des opérations de lecture et écriture. Un modèle de chimie-transport peut en fait être structuré autour d'un *pilote* gérant les opérations de lecture et d'écriture ainsi que les appels aux schémas numériques. Les schémas numériques doivent être implémentés sous forme de fonctions indépendantes intégrant une équation sur un pas de temps. Le pilote se charge d'initialiser la simulation, et puis d'effectuer la boucle en temps, boucle dans laquelle il fait les appels nécessaires aux schémas numériques. De plus, le pilote gère les lectures et écritures classiques : lecture des données d'entrées (générées par les programmes présentés précédemment) et sauvegarde des concentrations des polluants visés. Il est aussi possible que le pilote fasse quelques appels aux bibliothèques afin de calculer les champs absents des fichiers d'entrée (champs nécessairement calculés par le modèle de chimie-transport, pour une des raisons déjà indiquées). La figure 2.2 schématise l'architecture proposée.

Une perspective est l'inclusion d'autres modèles de chimie-transport dans le même cadre.

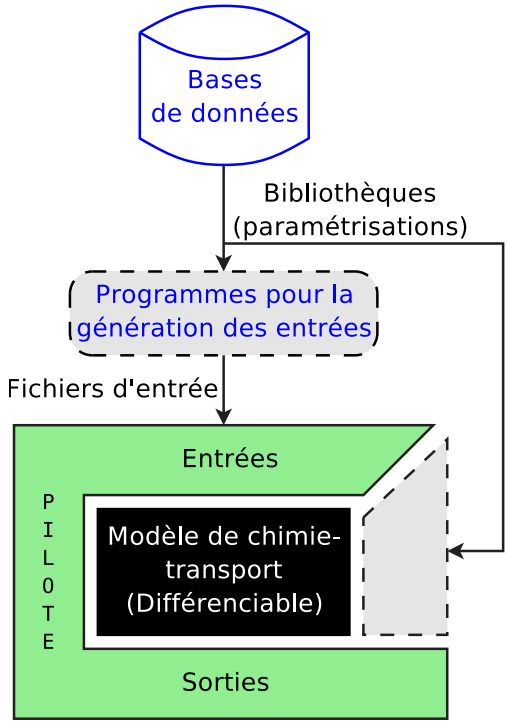


FIG. 2.2 – Architecture proposée pour un système de modélisation flexible et complet.

L'étape finale serait la conception d'un pilote générique capable de gérer n'importe quel modèle. Ceci est envisageable puisque tous les modèles de chimie-transport possèdent des structures similaires : une initialisation suivie par une boucle en temps avec des appels aux schémas numériques et à des paramétrisations. La stratégie informatique proposée est l'utilisation d'un langage orienté *objet* (voir ci-dessous) et la définition d'une interface régissant des communications standards avec le modèle de chimie-transport. Il s'agirait d'implémenter cette interface sous forme d'un objet. On rappelle qu'un objet est une structure composée :

1. d'*attributs* qui sont les données contenues par l'objet. Par exemple, un objet `Vector` aurait pour attributs un tableau de valeurs ainsi que la longueur du vecteur.
2. de *méthodes* pour la manipulation et la communication avec l'objet. Une méthode est une fonction associée à l'objet. Les méthodes permettent généralement d'accéder, de modifier ou d'effectuer des opérations sur les attributs. L'objet `Vector` posséderait notamment les méthodes `GetLength` et `Resize`. L'appel à une méthode est désigné par le point : `Vector.Resize(5)` redimensionne le vecteur pour qu'il ait une taille 5.

Dans le contexte qui nous intéresse, l'objet central gère le modèle de chimie-transport et est utilisé par le pilote. Une interface définit les méthodes que cet objet doit présenter pour que le pilote puisse effectivement piloter le modèle de chimie-transport. Supposons qu'on associe au modèle Polair3D l'objet `Polair3D`. L'interface peut imposer, pour l'advection, la présence des méthodes suivantes : `Init` et `Advection`. `Init` initialise Polair3D et `Advection` intègre les concentrations (attributs) sur un pas de temps. Le pilote effectue d'abord un appel à `Polair3D.Init` (c'est-à-dire la méthode `Init` de l'objet `Polair3D`) et effectue une boucle temporelle appelant `Polair3D.Advection`. Si un autre modèle de chimie-transport dispose d'une telle interface, le même pilote peut gérer cet autre modèle, ce qui est particulièrement commode pour la constitution d'une plate-forme de simulation. Il semble que la plupart des modèles de chimie-transport puissent être interfacés de la sorte sans changements importants dans leur code.

L'interface inclurait aussi quelques méthodes d'accès à l'état du modèle (c'est-à-dire aux concentrations). Ceci permettrait au pilote de procéder aux opérations d'écriture des résultats. De plus, plusieurs pilotes peuvent être écrits afin de proposer d'autres types de calculs. Il y aurait un pilote pour les simulations directes (simulations classiques), pour un filtre de Kalman (assimilation de données séquentielle), pour l'assimilation variationnelle, etc. Par exemple, le pilote en charge d'une assimilation de données variationnelle effectuerait une simulation directe, puis un calcul d'adjoint (boucle temporelle renversée). Une piste à explorer est un pilote pour le couplage à un modèle météorologique ou avec d'autres modèles (radiatif, économique ou de santé).

Tout modèle de chimie-transport, pour lequel une interface serait effective, pourrait être utilisé comme modèle sous-jacent à l'un des pilotes disponibles. En conséquence, tout modèle pourrait bénéficier de l'ensemble de l'architecture et en particulier des pilotes. Ceci offre une possibilité de partage de développement. Même pour un modèle unique, cette architecture est valable puisqu'elle clarifie la structure, en rend la gestion plus facile, notamment pour des applications de haut niveau telles que la prévision d'ensemble ou l'assimilation de données.

2.1.4 Réalisation partielle avec Polyphemus

Polyphemus[1] est un système réalisé sur la base des considérations précédentes. Toute l'architecture proposée n'est pas respectée ; l'état actuel de Polyphemus en est une réalisation partielle, sans les pilotes, le modèle de chimie-transport utilisé par Polyphemus, Polair3D, ne disposant pas des pilotes suggérés. Ainsi, Polyphemus possède l'architecture schématisée par la figure 2.1.

La bibliothèque AtmoData

La bibliothèque AtmoData[2] [Mallet et Sportisse, 2005b, note technique placée en annexe – chapitre A] est en charge de la gestion des données et des paramétrisations physiques. Elle regroupe l'ensemble des fonctionnalités (identifiées précédemment) qu'il convient d'apporter au système sous forme de bibliothèques. Une description de cette bibliothèque est fournie à la section 2.2.

Génération des données d'entrée

Plusieurs programmes sont disponibles dans Polyphemus pour la génération des données d'entrée (du modèle de chimie-transport). Ils permettent de générer la plupart des données nécessaires à la réalisation de simulations complexes. Pour mémoire, les principales données transformées ou estimées par ces programmes sont :

1. l'occupation des sols (à partir des données de l'USGS et de GLCF),
2. les conditions aux limites et des conditions initiales générées sur la base des simulations du modèle global Mozart 2 [Horowitz *et al.*, 2003],
3. les émissions anthropogéniques à partir des données brutes d'EMEP, suivant Middleton *et al.* [1990],
4. les émissions biogéniques [Simpson *et al.*, 1999],
5. les données météorologiques issues soit de l'ECMWF soit du modèle MM5,
6. les coefficients de diffusion verticale [Louis, 1979; Troen et Mahrt, 1986],
7. les diagnostics de la couverture nuageuse et de l'atténuation des coefficients photolytiques [Chang *et al.*, 1987; Madronich, 1987],

[1] `http://www.enpc.fr/cerea/polyphemus/`
[2] `http://www.enpc.fr/cerea/atmodata/`

8. les vitesses de dépôt [Wesely, 1989; Zhang *et al.*, 2003b].

Tous ces traitement font bien sûr écho aux paramétrisations de la section 1.5 et aux données mentionnées à la section 1.6.

Les programmes associés sont contrôlés par des configurations qui permettent facilement d'utiliser des paramétrisations, des données et des réglages différents d'une simulation à l'autre. Ceci constitue l'une des fonctionnalités permettant de constituer des prévisions d'ensemble. La diversité des configurations autorisées permet de considérer que chaque configuration est un nouveau modèle. Polyphemus est donc une *plate-forme multi-modèles* construite comme un système à multiples configurations.

Le modèle de chimie-transport

Le modèle de chimie-transport intégré à Polyphemus est Polair3D [Boutahar *et al.*, 2004]. Il est écrit en Fortran 77 avec une structure lui permettant d'être différencié automatiquement [Quélo, 2004; Mallet et Sportisse, 2004]. De plus, ce modèle est suffisamment flexible pour mener plusieurs applications-cibles en qualité de l'air : les simulations de la phase gazeuse (prévisions d'ozone, par exemple), simulations avec aérosols, assimilation de données et prévision d'ensemble.

Plus de détails sont fournis dans Boutahar *et al.* [2004]. Polair3D est le seul élément de Polyphemus dont il existait une version avant la constitution de l'ensemble du système. Il n'inclut pas, pour le moment, les pilotes recommandés, même s'il en existe des versions expérimentales (l'une d'elle a été utilisée en assimilation de données dans Quélo *et al.* [2005]).

Applications gérées par Polyphemus

Le système actuel permet de simuler les concentrations de polluants gazeux (passifs ou réactifs, radionucléides) et solides (aérosols). Il peut gérer des études de cas, des études d'impact (sur plusieurs années) comme des prévisions opérationnelles. Grâce aux multiples configurations disponibles, Polyphemus permet de générer des prévisions d'ensemble pour ces applications.

Autour de Polyphemus (c'est-à-dire via des programmes extérieurs – destinés à être adaptés avant inclusion définitive), des programmes permettent d'effectuer des simulations Monte Carlo par perturbation des entrées du modèle de chimie-transport [Aissaoui, 2004; Mallet et Sportisse, 2005a] ; cette fonctionnalité est aussi utilisée aux chapitres 3 et 4. De plus, des expériences d'assimilation de données variationnelle ont été menées avec Polyphemus [Quélo *et al.*, 2005]. Des expériences d'assimilation de données séquentielle ont aussi été réalisées avec Polair3D.

2.2 Gestion des paramétrisations physiques et des données

Ainsi qu'il est indiqué à la section 2.1.4, Polyphemus intègre la bibliothèque AtmoData. Elle remplit le rôle des bibliothèques introduites dans l'architecture proposée – voir section 2.1.3. Elle gère donc les données et les paramétrisations physiques. Ces deux fonctionnalités sont en réalité bien distinctes.

Concernant la gestion des données, AtmoData repose sur des structures de données évoluées. Pour fournir de telles structures, il faut disposer d'un langage orienté *objet*. Une introduction aux objets a été proposée à la section 2.1.3.

Dans le cas d'AtmoData, les objets doivent stocker des données multidimensionnelles et les coordonnées qui vont avec. Les langages les plus couramment utilisés en calcul scientifique, pour les calculs lourds, sont certainement les trois langages de la série Fortran (77, 90 et 95), le C et le C++. Le Fortran 77 et le C ne sont pas des langages objet et ne conviennent donc pas.

D'autre part, il manque au Fortran 90/95 des capacités qui plaident clairement en faveur du C++ [Cary *et al.*, 1997], notamment la généricité et les exceptions. La généricité permet d'écrire des fonctions indépendantes du type des données manipulées, ce qui est utile à Polyphemus qui manipule des données parfois en simple précision, parfois en double précision et au nombre de dimensions variable. Les exceptions sont un mécanisme de gestion des erreurs dont les langages modernes disposent pour d'améliorer la lisibilité et la sécurité des codes. Afin d'écarter les écueils parfois rencontrés, il convient de préciser que le C++ est au moins aussi rapide que le Fortran 90/95, qu'il ne consomme pas plus de mémoire, qu'il est standardisé [ISO/IEC, 1998] et que tous les compilateurs sérieux respectent presque entièrement ce standard. Le C++ s'impose donc techniquement. Du point de vue de l'utilisateur, la différence avec le Fortran 90/95 est faible puisqu'il s'agit, dans les deux cas, de manipuler des objets de haut niveau.

Sans entrer dans les détails techniques, on peut indiquer qu'AtmoData repose sur les bibliothèques Blitz++ [Veldhuizen, 1998], Talos et SeldonData. Elle propose une classe `Data` gérant les données, deux classes pour la gestion des coordonnées et autant de classes que de formats de fichier. Les paramétrisations sont une série de fonctions génériques (indépendantes du type des données) agissant sur les données multidimensionnelles (`Data`).

2.3 Intégration numérique

L'intégration numérique est la fonction première du modèle de chimie-transport tel qu'il est présenté dans l'architecture proposée. Les schémas numériques utilisés pour l'advection et la chimie dans Polair3D sont tirés de Verwer *et al.* [1998] ; ils sont aussi décrits dans [Sportisse et Mallet, 2005].

2.3.1 Notations

On se place en dimension 1 avec un pas de discrétisation spatial de Δx. Le pas de temps temporel est noté Δt. La concentration dans la cellule i et au pas de temps n est notée c_i^n. Les indices spatiaux demi-entiers se réfèrent aux interfaces (entre les cellules). Le vent u est connu au interfaces ; il est donc noté $u_{i+\frac{1}{2}}$. On note le nombre de Courant $\nu_{i+\frac{1}{2}} = \left|u_{i+\frac{1}{2}}\right| \frac{\Delta t}{\Delta x}$.

2.3.2 Advection

Le schéma numérique pour l'advection est un schéma d'ordre trois (en espace) auquel un limiteur de flux (de type Koren) vient s'ajouter. Ce limiteur de flux pondère le schéma d'ordre trois avec un schéma décentré d'ordre un (« upwind » en anglais). Il permet de conserver la positivité du schéma au niveau des fronts qui, intégrés par le schéma d'ordre trois, induiraient des oscillations dans la solution.

On écrit le schéma sous forme conservative :

$$c_i^{n+1} = c_i^n + F_{i-\frac{1}{2}}^n - F_{i+\frac{1}{2}}^n \tag{2.1}$$

où $F_{i+\frac{1}{2}}^n$ est le flux numérique entre les cellules i et $i+1$.

Le schéma décentré d'ordre a pour flux numérique

$$F_{i+\frac{1}{2}} = \begin{cases} \nu_{i+\frac{1}{2}} c_i & \text{si } u_{i+\frac{1}{2}} \geq 0 \\ -\nu_{i+\frac{1}{2}} c_{i+1} & \text{si } u_{i+\frac{1}{2}} < 0 \end{cases} \tag{2.2}$$

Le schéma d'ordre trois a pour flux numérique

$$F_{i+\frac{1}{2}} = \begin{cases} \nu_{i+\frac{1}{2}} \left(c_i + d_0(\nu_{i+\frac{1}{2}})(c_{i+1} - c_i) + d_1(\nu_{i+\frac{1}{2}})(c_i - c_{i-1}) \right) & \text{si } u_{i+\frac{1}{2}} \geq 0 \\ -\nu_{i+\frac{1}{2}}(c_{i+1} + d_0(\nu_{i+\frac{1}{2}})(c_i - c_{i+1}) + d_1(\nu_{i+\frac{1}{2}})(c_{i+1} - c_{i+2})) & \text{si } u_{i+\frac{1}{2}} < 0. \end{cases} \tag{2.3}$$

avec

$$d_0(\nu) = \frac{1}{6}(2 - \nu)(1 - \nu) \tag{2.4}$$

$$d_1(\nu) = \frac{1}{6}(1 - \nu^2) \tag{2.5}$$

Avec limiteur de flux, le flux numérique vaut

$$F_{i+\frac{1}{2}} = \begin{cases} \nu_{i+\frac{1}{2}} \left(c_i + \psi \left(\nu_{i+\frac{1}{2}}, \theta_i \right) (c_{i+1} - c_i) \right) & \text{si } u_{i+\frac{1}{2}} \geq 0 \\ -\nu_{i+\frac{1}{2}} \left(c_{i+1} + \psi \left(\nu_{i+\frac{1}{2}}, \frac{1}{\theta_{i+1}} \right) (c_i - c_{i+1}) \right) & \text{si } u_{i+\frac{1}{2}} < 0 \end{cases} \tag{2.6}$$

où

$$\psi(\nu, \theta) = \max \left(0, \min \left(1, d_0(\nu) + d_1(\nu)\theta, \frac{1 - \nu}{\nu}\theta \right) \right). \tag{2.7}$$

et

$$\theta_i = \frac{c_i - c_{i-1}}{c_{i+1} - c_i} \tag{2.8}$$

2.3.3 Chimie

Un point important dans l'intégration de la chimie concerne la grande disparité des temps caractéristiques des réactions chimiques. On parle alors de problème raide. Plusieurs temps caractéristiques sont beaucoup plus petits que le pas de temps d'intégration. L'utilisation d'un schéma implicite (ou semi-implicite) est adapté à un tel cas. Le schéma de Rosenbrock d'ordre deux est proposé dans Verwer *et al.* [1998].

Si on cherche à résoudre $c' = \chi(t, c)$, on introduit d'abord la matrice jacobienne $A^n = \partial_c F(t^n, c^n)$. La raideur du système est illustrée par la figure 2.3 où sont reportées les valeurs propres d'une matrice A^n (pour le mécanisme chimique RADM 2).

Le schéma de Rosenbrock d'ordre 2 s'écrit

$$c^{n+1} = c^n + \Delta t(b_1 k_1 + b_2 k_2) \tag{2.9}$$

avec k_1 et k_2 solutions des systèmes linéaires suivants :

$$\begin{cases} k_1 = \chi(t^n, c^n) + \Delta t \gamma A^n k_1 \\ k_2 = \chi(t^n + \alpha_{12}\Delta t, c^n + \alpha_{21}\Delta t k_1) + \Delta t \gamma_{21} A^n k_1 + \Delta t \gamma A^n k_2 \end{cases} \tag{2.10}$$

où

$$b_1 = 1 - b_2\,; \qquad \gamma_{21} = -\frac{\gamma}{b_2}\,; \qquad \alpha_{21} = \frac{1}{2b_2} \tag{2.11}$$

La méthode est A-stable si et seulement si $\gamma \geq \frac{1}{4}$. Elle est L-stable pour $\gamma = 1 \pm \frac{\sqrt{2}}{2}$. En pratique, on prend $b_2 = \frac{1}{2}$ et $\gamma = 1 + \frac{\sqrt{2}}{2}$. De plus, en remplaçant k_2 par $k_2 - 2k_1$, on obtient le schéma

$$c^{n+1} = c^n + \frac{3}{2}\Delta t k_1 + \frac{1}{2}\Delta t k_2 \tag{2.12}$$

avec

$$\begin{cases} (I - \gamma \Delta t A^n) \cdot k_1 = \chi(t^n, c^n) \\ (I - \gamma \Delta t A^n) \cdot k_2 = \chi(t^{n+1}, c^n + \Delta t k_1) - 2k_1 \end{cases} \tag{2.13}$$

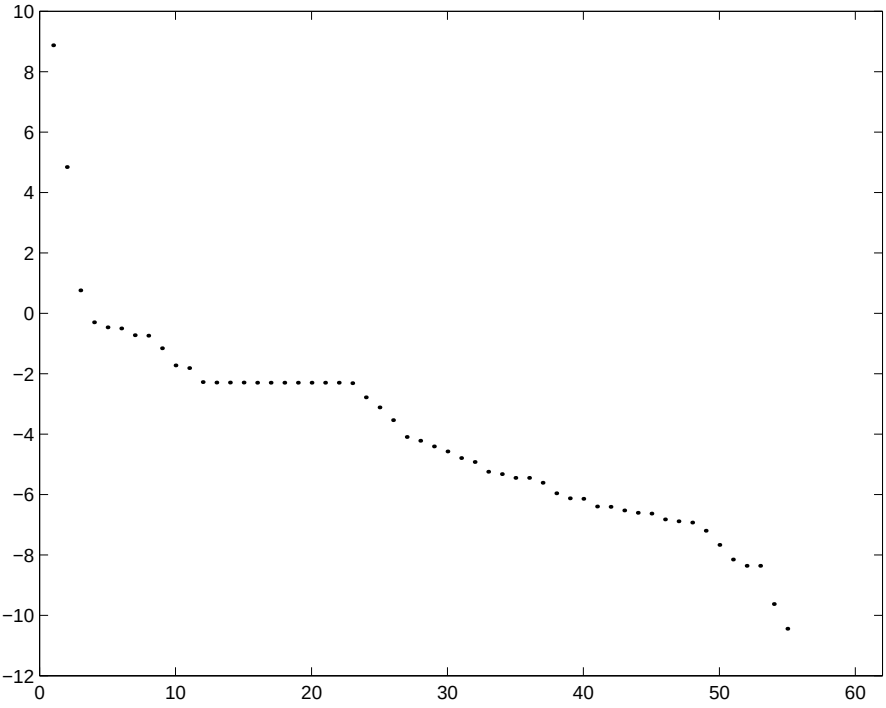


FIG. 2.3 – Valeurs propre (valeurs absolues) d'une matrice jacobienne pour le mécanisme chimique RADM 2 (échelle semi-logarithmique). Plus de vingt décades séparent les valeurs propres reportées. Les plus petites valeurs propres sont omises à cause d'imprécisions numériques.

Le coût du schéma implicite tient principalement à la résolution des systèmes linéaires pour le calcul de k_1 et k_2. La même matrice $(I - \gamma \Delta t A^n)$ est « inversée » deux fois ; une décomposition LU est efficace. Or, la matrice jacobienne A^n est creuse. Pour le mécanisme RADM 2, elle contient environ 85% de zéros. De plus, la position des zéros de la matrice sont connus avant son calcul. Suivant Sandu *et al.* [1996], la décomposition LU de la matrice est écrite sans boucle et sans indirection (alors que c'est l'usage dans le stockage des matrices creuse). La remontée (résolution) est aussi écrite de la sorte. Enfin, les espèces sont permutées de sorte à ce que la décomposition LU de la matrice demeure creuse. Ceci permet d'économiser le calcul d'un grand nombre de coefficients. La technique utilisée pour trouver la permutation est la méthode de Markowitz.

Comme l'indique Mallet et Sportisse [2004], cette procédure tenant compte du caractère creux de la matrice jacobienne peut conduire à de forts gains en temps calcul. Pour le schéma RADM 2, le temps calcul est réduit d'un facteur 3.6 pour l'intégration de la chimie.

2.3.4 Diffusion

La diffusion est intégrée par le schéma de Rosenbrock utilisé pour la chimie. La matrice jacobienne est alors tridiagonale. La résolution du système est effectuée grâce à l'algorithme de

Thomas. On suppose qu'on résout $Tc = b$, avec

$$T = \begin{pmatrix} \alpha_1 & \gamma_1 & 0 & & & \\ \beta_2 & \alpha_2 & \gamma_2 & 0 & & (0) \\ 0 & \ddots & \ddots & \ddots & 0 & \\ & 0 & \ddots & \ddots & \ddots & 0 \\ (0) & & 0 & \beta_{n-2} & \alpha_{n-2} & \gamma_{n-2} \\ & & & 0 & \beta_{n-1} & \alpha_{n-1} \end{pmatrix} \tag{2.14}$$

On peut éliminer le terme en c_1 de la deuxième équation du système ($\beta_2 c_1 + \alpha_2 c_2 + \gamma_2 c_3 = b_2$) en utilisant la première équation ($\alpha_1 c_1 + \gamma_1 c_2 = b_1$). Par combinaison linéaire, la deuxième équation devient : $\left(\alpha_2 - \frac{\beta_2}{\alpha_1}\gamma_1\right) c_2 + \gamma_2 c_3 = b_2 - \frac{\beta_2}{\alpha_1} b_1$. On réitère le processus sur toutes les équations, ce qui aboutit au nouveau système $T'c = b'$:

$$T' = \begin{pmatrix} \alpha'_1 & \gamma_1 & 0 & & & \\ 0 & \alpha'_2 & \gamma_2 & 0 & & (0) \\ & \ddots & \ddots & \ddots & 0 & \\ & & \ddots & \ddots & \ddots & 0 \\ (0) & & & 0 & \alpha'_{n-2} & \gamma_{n-2} \\ & & & & 0 & \alpha'_{n-1} \end{pmatrix} \tag{2.15}$$

où $\alpha'_1 = \alpha_1$ et, si $i > 1$, $\alpha'_i = \alpha_i - \frac{\beta_i}{\alpha'_{i-1}}\gamma_{i-1}$. Pour le second membre, $b'_1 = b_1$ et, si $i > 1$, $b'_i = b_i - \frac{\beta_i}{\alpha'_{i-1}} b_{i-1}$.

Il ne reste qu'à effectuer une remontée (i.e. résoudre $T'c = b'$).

2.3.5 Intégration de l'ensemble

Les schémas numériques précédents permettent d'intégrer les processus importants en temps. Pour intégrer l'équation couplée advection-diffusion-chimie, on sépare les opérateurs (en anglais, on parle de « splitting »), c'est-à-dire qu'on les intègre les uns après les autres sur un pas de temps. L'ordre adopté pour la séquence est advection, diffusion puis chimie [Sportisse, 2000].

2.4 Traitement des sorties

Pour être complet quant à la présentation du système Polyphemus, il faut présenter les outils d'analyse des concentrations calculées par le système (sorties de Polair3D).

Les deux éléments les plus courants dans le traitement des résultats sont l'affichage (graphique) et la comparaison aux observations. Plus encore que dans le calcul des concentrations, une grande flexibilité est requise car il ne s'agit pas d'effectuer des traitements systématiques. Pour répondre à ces objectifs, Polyphemus est doté d'une bibliothèque en Python appelée AtmoPy. Le choix du Python est motivé par

1. le caractère objet et complet du langage,
2. les modules (terme désignant les bibliothèques Python) dédiés à l'affichage scientifique,
3. l'interactivité du Python (compilation inutile, mode « shell »),
4. la disponibilité de modules de calcul scientifique,

5. la complémentarité avec le C++ qui permet de restreindre les dépendances de Polyphemus au nombre de trois (C++, Python et Fortran) – tous les besoins sont ainsi couverts, Python étant aussi le langage dévolu aux scripts.

En un mot, Python permet de remplacer Matlab (interactivité) avec en plus un langage moderne (pérennité). La bibliothèque AtmoPy gère

1. l'affichage graphique[3] (1D et 2D),
2. l'analyse des concentrations calculées (statistiques, comparaison entre simulations, comparaison aux observations),
3. certaines fonctionnalités orientées « ensemble » (celles présentées au chapitre 5 notamment).

AtmoPy utilise les modules numarray [Greenfield *et al.*, 2003], Matplotlib (`http://matplotlib.sourceforge.net/`) et SciPy [Jones *et al.*, 2001].

Dans la suite, les résultats présentés reposent en grande partie sur la bibliothèque AtmoPy. L'objectif de la bibliothèque est de fournir tous les outils nécessaires à des analyses représentatives, c'est-à-dire appuyées sur des indicateurs statistiques adapatés. Un exemple est fourni dans ce chapitre avec l'évaluation de simulations (section 2.5). Pour cette évaluation, la bibliothèque permet d'interpoler les concentrations simulées aux stations et d'effectuer des comparaisons via plusieurs indicateurs statistiques pertinents. Ces indicateurs sont présentés dans le tableau 2.1.

D'autres modes d'analyse sont possibles pour des comparaisons entre simulations, avec par exemple des distributions de différences entre simulations (section 3.1). La distribution des concentrations elles-mêmes peut être une information pertinente (section 3.2). Des cartes de variabilité sont aussi introduites pour localiser les incertitudes (toujours section 3.2).

Outre les outils d'analyse précédents, AtmoPy fournit aussi un support à la prévision d'ensemble, support qui intervient dans les traitements effectués au chapitre 5.

2.5 Évaluation de Polyphemus

2.5.1 Introduction

L'évaluation d'un modèle permet de déterminer s'il est raisonnable de fonder des études sur ses résultats. Comme Russell et Dennis [2000] l'ont souligné, il ne s'agit pas de validation ou de vérification puisque ces termes renvoient à un accord strict avec les quantités calculées. Un modèle numérique associé à un système environnemental ne peut faire l'objet que d'une évaluation.

Une évaluation n'a de sens que si elle est réalisée sur une période assez longue et avec un nombre important d'observations. Des indicateurs statistiques caractérisent l'*erreur*, c'est-à-dire la distance entre les concentrations calculées et celles mesurées (observations). Il faut relever deux limitations de cette procédure :

1. Les mesures sont elles-mêmes entachées d'erreur. Par exemple, ESQUIF [2001] estime les incertitudes associées aux mesures – voir tableau 2.2.
2. Les concentrations calculées ne correspondent pas aux mesures auxquelles elles sont comparées. Elles sont des moyennes dans une cellule du modèle (ou une combinaison de telles moyennes) alors que les mesures ne représentent qu'une moyenne selon une trajectoire des vents. On parle classiquement d'erreur de représentativité. Elle est plus élevée au voisinage des sources qui sont artificiellement diluées dans les cellules du modèle.

[3] La bibliothèque AtmoPy étant récente, toutes les figures présentes dans cette thèse n'en sont pas issues.

Tab. 2.1 – Indicateurs statistiques évaluant les performances d'un modèle. $(y_i)_i$ est la série temporelle simulée. $(o_i)_i$ est la série correspondante observée. n est le nombre d'éléments dans chaque série. Dans cette thèse, pour le calcul de UPA, le maximum est pris par jour puis moyenné sur l'ensemble des jours.

(Indicateur) Nom anglais	Notation(s)	Formule
Écart quadratique moyen Root mean square error	RMS ou RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - o_i)^2}$
Biais Bias	Bias	$\frac{1}{n}\sum_{i=1}^{n}(y_i - o_i)$
Bias factor	BF	$\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{o_i}$
Corrélation Correlation	Corr	$\frac{\sum_{i=1}^{n}(y_i - \bar{y})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(o_i - \bar{o})^2}}$
Bias normalisé moyen Mean normalized bias error	MNBE	$\frac{1}{n}\sum_{i=1}^{n}\frac{y_i - o_i}{o_i}$
Mean normalized gross error	MNGE	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i - o_i\rvert}{o_i}$
Unpaired peak prediction accuracy	UPA	$\frac{y_{\max} - o_{\max}}{o_{\max}}$

Tab. 2.2 – Incertitudes (ppb) associées aux mesures.

Concentrations (ppb)	O_3	NO	NO_2
10	±3.1	±3.4	±6.8
30	±3.1	±3.4	±6.8
50	±3.6	±3.6	±7.2
70	±4.2	±4.1	±8.2
90	±4.9	±4.7	±9.4
100	±5.3	±5.0	±6.8
150	±7.6	±6.9	±13.8
200	±10.0	±9.0	±18.0
250	±12.5	±11.2	±22.4

2.5.2 Les observations

Les comparaisons aux mesures sont effectuées avec des observations d'ozone. Les travaux qui suivent s'intéressent principalement à l'ozone, même s'ils pourraient être appliqués à d'autres polluants – la démarche serait similaire. De plus, le grand nombre d'observations d'ozone permet des comparaisons plus complètes.

Trois réseaux d'observation fournissent des mesures horaires :

1. le réseau EMEP constitué de stations régionales (c'est-à-dire loin des sources d'émission) distribuées sur l'Europe. 85 stations délivrent des observations horaires d'ozone. Voir figure 2.4.
2. le réseau utilisé par l'expérience Pioneer[4]. Il contient 241 stations urbaines, péri-urbaines et rurales dans la terminologie ADEME (Agence de l'environnement et de la maîtrise de l'énergie). Selon cette terminologie, une station urbaine possède un rayon de représentativité de 100 m à 2 km, une station péri-urbaine un rayon de 1 km à 5 km et une station rurale un rayon de plus de 5 km. Par la suite, la terminologie adoptée se réfère principalement aux stations urbaines (incluant les stations péri-urbaines) et aux stations régionales (stations rurales). 116 stations sont situées en France, 81 stations en Allemagne. Voir figure 2.5.
3. le réseau de la BDQA (Banque de données sur la qualité de l'air), géré par l'ADEME et 40 agences agréées pour la surveillance de la qualité de l'air[5]. 356 stations urbaines et régionales couvrent le territoire français. Voir figure 2.6.

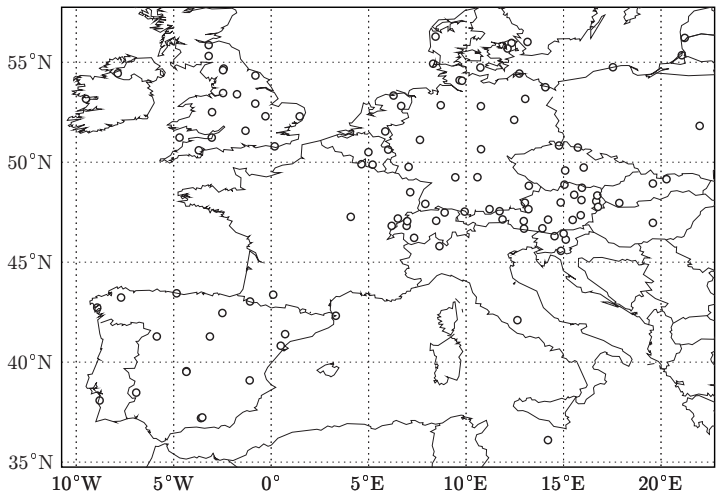


FIG. 2.4 – Réseau d'observation EMEP.

[4] `http://euler.lmd.polytechnique.fr/pioneer/`

[5] Agences agréées pour la surveillance de la qualité de l'air en France : AERFOM, AIRFOBEP, AIRMARAIX, AIRPARIF, ALPA, AREMASSE, ATMO AUVERGNE, AIR LANGUEDOC ROUSSILLON, ATMO POITOU CHARENTES, OPALAIR, AREMA LILLE METROPOLE, ORAMIP, ARPAM, ATMO CHAMPAGNE ARDENNE, ASCOPARG, ASPA, ASQAB, ATMO PICARDIE, AIR BREIZH, COPARLY, AIRCOM, ESPOL, AIR PAYS DE LA LOIRE, QUALITAIR, REMAPPA, ATMOSFAIR BOURGOGNE CENTRE NORD, SUPAIRE, AREMARTOIS, AMPASEL, AIRLOR, AIRAQ, ATMOSFAIR BOURGOGNE SUD, L'AIR de l'Ain et des pays de SAVOIE, LIGAIR, LIMAIR, ASQUADRA, GWADAIR, ORA, MADININAIR, ORA de GUYANE.

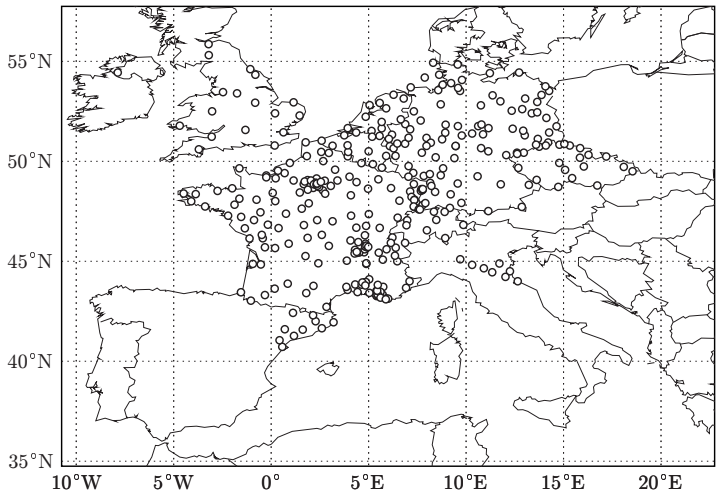


FIG. 2.5 – Réseau d'observation Pioneer.

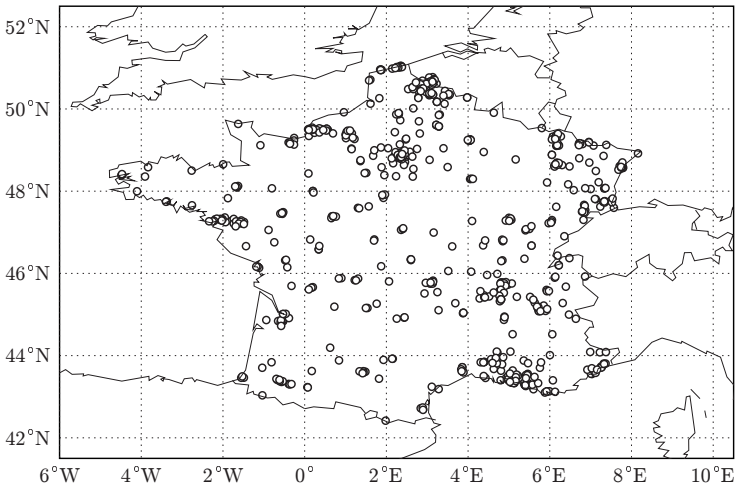


FIG. 2.6 – Réseau d'observation de la BDQA.

2.5.3 Procédure de comparaison aux observations

L'évaluation repose sur les indicateurs statistiques du tableau 2.1. Les recommandations de l'EPA [EPA, 1991; Hogrefe *et al.*, 2001; Russell et Dennis, 2000] sont des limites sur les valeurs de MNBE, MNGE et UPA. Ces limites s'appliquent à des statistiques calculées uniquement avec les concentrations dont les mesures dépassent un seuil de 40 ppb à 60 ppb, soit environ 80 $\mu g \cdot m^{-3}$ et 120 $\mu g \cdot m^{-3}$ respectivement. Plus les concentrations sont faibles plus les performances se dégradent (il convient de noter que les indicateurs sont relatifs). Dans la suite, on utilise une valeur limite de 80 $\mu g \cdot m^{-3}$, qui est donc le seuil le plus exigeant.

Les performances recommandées vont de $\pm 5\%$ à $\pm 15\%$ pour MNBE, de $\pm 30\%$ à $\pm 35\%$ pour MNGE et de $\pm 15\%$ à $\pm 20\%$ pour UPA. On choisit les critères suivants :

$$|\text{MNBE}| < 15\% \tag{2.16}$$

$$|\text{MNGE}| < 30\% \tag{2.17}$$

$$|\text{UPA}| < 15\% \tag{2.18}$$

soit les critères les plus restrictifs pour MNGE et UPA, et un critère plus lâche sur MNBE car la limite de 5% ne paraît pas cohérente avec celle imposée sur MNGE – même si elle est là pour éviter les biais systématiques.

Dans la suite, on estime aussi MNGE sur les pics journaliers, bien que cela ne rentre pas dans les évaluations recommandées par l'EPA. On applique les mêmes critères pour les pics, même s'ils sont généralement mieux prévus. Il faut noter que MNBE sur les pics correspond à UPA.

2.5.4 Évaluation sur l'année 2001

Configuration

La configuration d'une simulation désigne le choix des données, des paramétrisations, des approximations numériques et des schémas numériques. Elle peut varier selon les applications et la version du système de simulation. Pour des raisons chronologiques, la configuration utilisée pour valider le modèle n'est donc pas celle sur laquelle les travaux de cette thèse reposent. Elle est cependant proche de toutes les configurations utilisées dans cette thèse. Il ne s'agit pas nécessairement de la meilleure configuration possible avec la dernière version de Polyphemus, mais elle constitue une configuration représentative de celles introduites par la suite.

Le domaine est [40.25°N, 10.25°W]×[56.75°N, 22.25°E] – voir figure 2.7. Le domaine est discrétisé horizontalement par des mailles régulières en latitude/longitude de dimension $0.5^\circ \times 0.5^\circ$. Verticalement, les cinq niveaux sont compris entre des interfaces à 0 m, 50 m, 600 m, 1200 m, 2000 m et 3000 m. L'interface supérieure du domaine est généralement au-dessus de la couche limite atmosphérique, ce qui autorise à ne pas dépasser 3000 m.

Notons que la faible discrétisation verticale dégrade nécessairement les résultats, mais ceux-ci restent satisfaisants. Même dans le cadre de prévisions, la discrétisation est souvent plus fine car les puissances de calcul disponibles le permettent. Dans cette thèse, le choix de la discrétisation verticale se justifie par le coût calcul très élevé des études qui y sont menées. Les études de sensibilité du chapitre 4 sont réalisées via de multiples simulations avec le modèle direct (simulations Monte Carlo), avec le modèle linéaire tangent et avec le modèle adjoint (dont le coût est environ sept fois celui du modèle direct). Les chapitres 3 et 5 reposent sur des calculs d'ensemble, c'est-à-dire sur des dizaines à des milliers de simulations. En conséquence, les coûts calcul sont très élevés et justifient une discrétisation verticale faible.

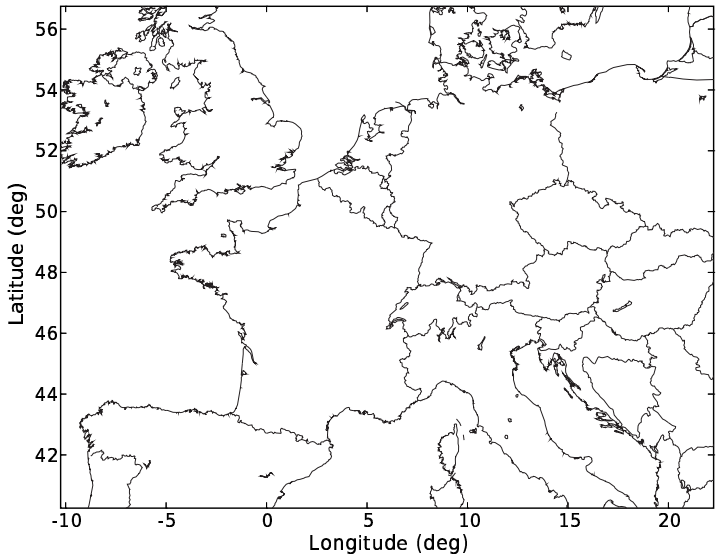


FIG. 2.7 – Domaine de la simulation – [40.25°N, 10.25°W]×[56.75°N, 22.25°E].

Les concentrations considérées couvrent la période du 27 avril 2001 au 31 août 2001 (soit environ quatre mois). Le pas de temps d'intégration est 600 s. On prévoit des concentrations horaires d'ozone puisque les observations sont elles-mêmes horaires.

La configuration de la simulation est la suivante (se reporter au chapitre 1 pour les détails) :

1. occupation des sols : données de GLCF pour les émissions biogéniques, données USGS pour le reste.
2. données météorologiques : champs ECMWF, résolution de 0.36°×0.36°, résolution spectrale horizontale TL511, 60 niveaux, pas de temps de 3 heures, cycles de prévision de 12 heures démarrant à partir de champs analysés.
3. mécanisme chimique : RACM [Stockwell *et al.*, 1997], avec 72 espèces et 237 réactions.
4. émissions : inventaire EMEP pour l'année 2001 à l'exclusion de la classe SNAP 11, contenue dans les émissions biogéniques (ci-dessous). L'inventaire est converti selon Middleton *et al.* [1990].
5. émissions biogéniques : estimées selon Simpson *et al.* [1999].
6. vitesses de dépôt : issues de la paramétrisation révisée et proposée dans Zhang *et al.* [2003b]. Le flux de surface utilisé est le flux de chaleur.
7. diffusion verticale : dans la couche limite et en conditions instables, on utilise la paramétrisation de Troen et Mahrt introduite dans Troen et Mahrt [1986], avec la hauteur de couche limite fournie dans les données ECMWF. Dans les autres cas, on repose sur la paramétrisation de Louis [Louis, 1979]. Les paramètres à fixer, selon Troen et Mahrt [1986], sont $p = 3$, $C = 6.5$, $\varepsilon = 0.1$ (rapport entre la hauteur de la couche de surface et la couche limite atmosphérique), et le nombre de Richardson à 0.21.
8. diffusion horizontale : elle est fixée à $10\,000\ \mathrm{m}^2 \cdot \mathrm{s}^{-1}$.
9. conditions aux limites : elles sont générées sur la base des sorties, pour une année météorologique typique, du modèle de chimie-transport global Mozart 2 [Horowitz *et al.*, 2003].

Notes techniques

Le temps d'intégration peut varier avec les versions (dans un sens ou dans l'autre). Généralement, l'intégration numérique prend environ 5 min par jour (réel), sur un Pentium IV cadencé à 3 Ghz, et compilé avec IFC 7.0 (Intel Fortran Compiler). Environ 60% du temps est passé dans l'intégration de la chimie, 22% dans l'intégration de la diffusion et 15% dans l'intégration de l'advection.

Pour la simulation présentée ici, on utilise la version de Polyphemus du 7 février 2005 (ultérieure à la version 0.2). Cette version repose sur AtmoData 1.0. La version de Polair3D du 20 août 2005 effectue l'intégration de l'équation de réaction-diffusion.

Résultats

Lors la comparaison aux observations, on ne sélectionne que les stations dont le nombre de mesures correspond à plus de 30% du nombre total de mesures possibles sur la période.

Les statistiques d'erreur sur les concentrations horaires d'ozone sont reportées dans le tableau 2.3. Celles pour les pics journaliers sont rassemblées dans le tableau 2.4. Il est difficile de comparer ces résultats à ceux d'autres modèles car il existe peu d'études où les statistiques d'erreur sont calculées sur de longues périodes, avec un nombre d'observations important et dans des conditions comparables. Sur la base de Russell et Dennis [2000]; Schmidt *et al.* [2001]; van Loon *et al.* [2004], on peut estimer les résultats de la simulation satisfaisants. Ils sont en tout cas comparables à l'état de l'art, et ce malgré la faible résolution.

Tab. 2.3 – Statistiques d'erreur de la simulation pour les concentrations horaires d'ozone.

Indicateur	Réseau EMEP	Réseau Pioneer	Réseau BDQA
Moyenne des observations ($\mu g \cdot m^{-3}$)	78.0	70.4	67.0
Moyenne simulée ($\mu g \cdot m^{-3}$)	83.3	80.2	83.9
RMSE ($\mu g \cdot m^{-3}$)	27.4	30.0	33.7
Corrélation (%)	59.1	66.0	63.6
BF	0.99	1.02	1.05
MNBE[1] (%)	-1.4	1.6	5.1
Stations réalisant le critère EPA sur MNBE (%)	86.0	88.0	87.6
MNGE[1] (%)	17.6	17.7	17.5
Stations réalisant le critère EPA sur MNGE (%)	98.8	98.8	97.5

[1] Moyenne sur l'ensemble des stations.

Les distributions spatiales de l'erreur sur les pics sont représentées sur les figures 2.8, 2.9 et 2.10. Il est difficile d'analyser ces distributions, et certainement hasardeux de leur trouver des explications.

Quatre chroniques temporelles sont rassemblées dans la figure 2.11. Deux stations représentatives y sont incluses. La station (Lyon–Garibaldi) est associée à la plus mauvaise RMSE sur le réseau Pioneer. Il s'agit d'une station urbaine proche de fortes sources que le modèle ne peut reproduire correctement (les émissions étant diluées dans des mailles de $0.5^\circ \times 0.5^\circ$). La forte surévaluation

TAB. 2.4 – Statistiques d'erreur de la simulation pour les pics journaliers d'ozone.

Indicateur	Réseau EMEP	Réseau Pioneer	Réseau BDQA
Moyenne des observations ($\mu g \cdot m^{-3}$)	101.9	102.3	103.3
Moyenne simulée ($\mu g \cdot m^{-3}$)	106.2	107.2	109.0
RMSE ($\mu g \cdot m^{-3}$)	21.4	22.9	25.1
Corrélation (%)	64.6	72.1	69.9
BF	1.06	1.02	1.03
MNGE[1] (%)	16.5	14.0	15.0
Stations réalisant le critère EPA sur MNGE (%)	94.2	98.8	98.0
UPA[1] (%)	6.0	2.3	3.0
Stations réalisant le critère EPA sur UPA (%)	84.9	94.6	90.8

[1] Moyenne sur l'ensemble des stations.

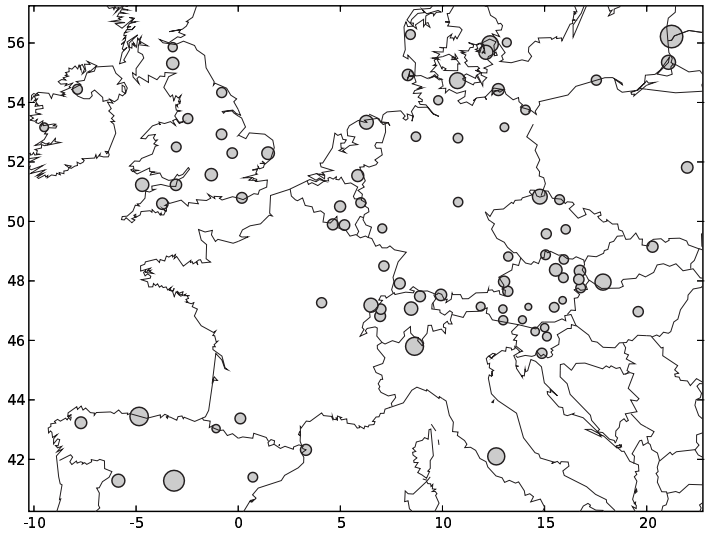


FIG. 2.8 – Distribution spatiale de la RMSE des pics journaliers d'ozone sur le réseau EMEP. Les diamètres sont proportionnels à la RMSE.

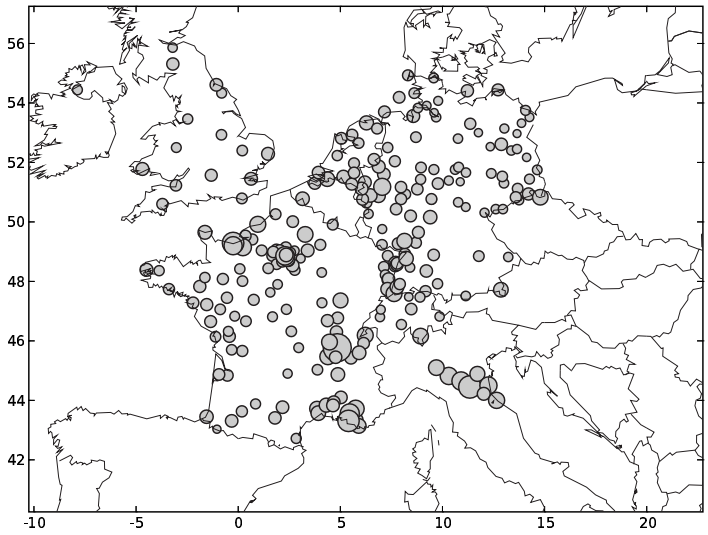


FIG. 2.9 – Distribution spatiale de la RMSE des pics journaliers d'ozone sur le réseau Pioneer. Les diamètres sont proportionnels à la RMSE.

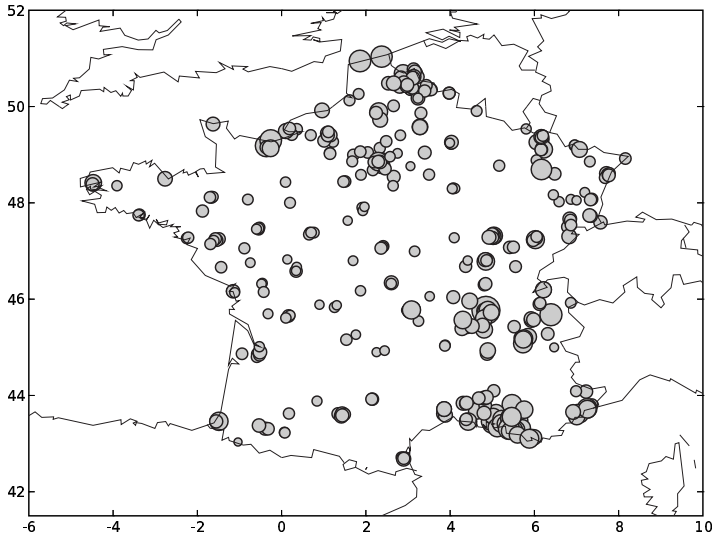
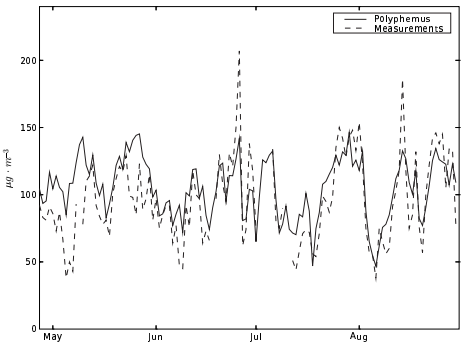
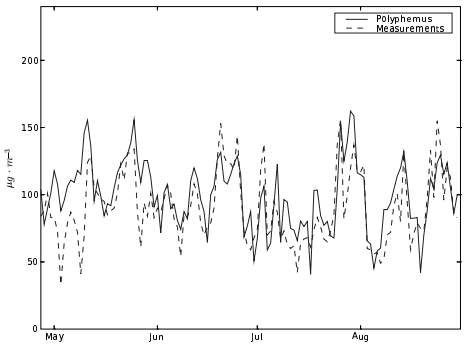


FIG. 2.10 – Distribution spatiale de la RMSE des pics journaliers d'ozone sur le réseau BDQA. Les diamètres sont proportionnels à la RMSE.

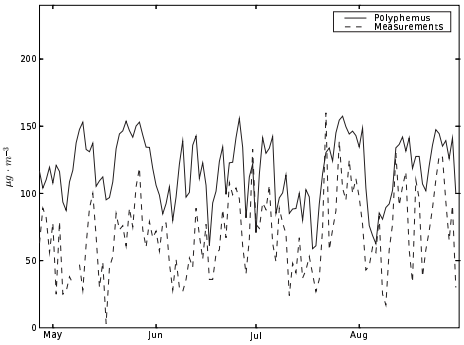
des concentrations provient vraisemblablement d'un titrage trop faible de l'ozone par le monoxyde d'azote (ce dernier étant justement dilué).



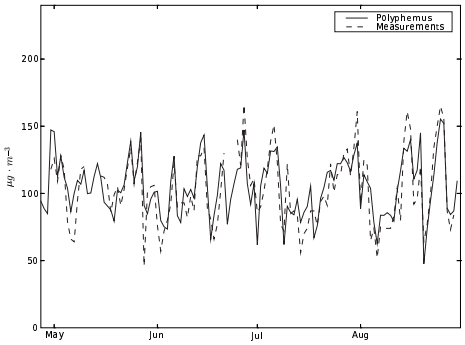
(a) Montceau-les-Mines – RMSE : 23.4 $\mu g \cdot m^{-3}$, corrélation : 74.2%



(b) Saint Nazaire – RMSE : 22.2 $\mu g \cdot m^{-3}$, corrélation : 69.4%



(c) Lyon–Garibaldi – RMSE : 53.8 $\mu g \cdot m^{-3}$, corrélation : 57.5%



(d) Possen – RMSE : 15.5 $\mu g \cdot m^{-3}$, corrélation : 81.8%

FIG. 2.11 – Chroniques temporelles des pics journaliers d'ozone en quatre stations du réseau Pioneer. Les deux premières stations (a, b) ont des statistiques représentatives de la simulation. La station Lyon–Garibaldi (c) est associée à la plus mauvaise RMSE (sur le réseau Pioneer) et celle de Possen (d) à la meilleure RMSE (idem).

Chapitre 3

Estimation de l'incertitude

Les modèles de chimie-transport sont principalement évalués sur la base de comparaisons aux concentrations (voire aux quantités déposées) mesurées. Cette évaluation n'est pas suffisante pour juger des performances d'un modèle. Un autre indicateur est l'incertitude a priori (en l'absence d'observations) qui quantifie la confiance à accorder à des simulations. Les sources d'incertitude sont l'intégration numérique, la formulation du modèle et les données d'entrée. Chaque source d'incertitude est évaluée grâce à une stratégie adaptée : sensibilité aux schémas numériques (section 3.1), approche multi-modèles (section 3.2) et simulations Monte Carlo (section 3.3).

Sommaire

La section 3.1 est une variation de
Pourchet, A., Mallet, V., Quélo, D. et Sportisse, B. (2005). Some numerical issues in Chemistry-Transport Models – a comprehensive study with the Polyphemus/-Polair3D platform. Rapport technique 26, CEREA. *En préparation pour soumission à J. Comp. Phys.*

La section 3.2 est constituée de
Mallet, V. et Sportisse, B. (2006). Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations : an ensemble approach applied to ozone modeling. *J. Geophys. Res.*, 111(D1)

3.1 Sensibilité aux schémas numériques

3.1.1 Introduction

Ainsi que le présente le chapitre 2, la dernière étape identifiée du processus de simulation est l'intégration numérique de l'équation de dispersion-réaction 1.7. Cette intégration numérique dépend

1. d'approximations numériques : type de maillage, raffinement du maillage, étendue du domaine de simulation, imbrication de domaines (« nesting », en anglais) et discrétisation temporelle ;
2. de schémas numériques, présentés à la section 2.3.

Une partie de l'erreur et de l'incertitude provient des approximations et des schémas numériques. Il est difficile d'estimer l'incertitude due aux seules approximations numériques. Par exemple, l'incertitude liée au raffinement du maillage dépend de l'incertitude des données d'entrée qui sont connues ou estimées à une certaine résolution. Une partie de l'incertitude due aux approximations numériques est traitée à la section 3.2.

L'incertitude due aux schémas numériques peut être évaluée avec une démarche plus systématique, ce qui est réalisé dans cette section. Les variations dans les concentrations calculées, lors de changements de schémas numériques, sont analysées. La procédure consiste à tester un grand nombre de schémas numériques et à conclure sur leur impact. Il s'agit donc d'une étude de sensibilité. L'étude ne peut prétendre évaluer précisément l'incertitude due aux schémas numériques car les changements de schéma ne sont pas cumulés et tous les schémas ne sont pas objectivement de même valeur *a priori*. Néanmoins le coût des calculs peut conduire à simplifier les schémas numériques et à introduire une erreur, quantifiée ici, qui contribue à l'incertitude.

Des schémas numériques adaptés sont nécessaires à la simulation de la qualité de l'air du fait de la dimension des systèmes à résoudre (environ un million de variables suivies en temps), du coût calcul important de certaines applications (assimilation de données, prévision d'ensemble), de difficultés numériques propres (raideur de la chimie – voir section 2.3.3, figure 2.3) et de la combinaison du transport et de la chimie. De nombreux travaux ont été menés, par exemple McRae *et al.* [1982]; Zlatev [1995].

Des développements et des évaluations de schémas numériques ont été réalisés dans le but de construire des modèles de chimie-transport à la fois efficaces (coûts calculs faibles) et précis [Verwer *et al.*, 1998]. On peut citer les études spécifiques aux problèmes raides [Verwer *et al.*, 1999; Sandu *et al.*, 1997b,a] ou à la séparation d'opérateurs [Lanser et Verwer, 1999; Browning et Kreiss, 1994; Sportisse, 2000]. La plupart de ces études ont été réalisées en 0D ou 1D. Elles ne rendent pas compte de la diversité des situations rencontrées dans les modèles, en 3D. Seule une étude sur un cas réaliste permet d'estimer le véritable impact des choix numériques.

Le cadre de l'étude est brièvement présenté à la section 3.1.2. Dans cette même section sont introduits les outils permettant l'analyse systématique des résultats. Les sections suivantes présentent les résultats concernant la séparation d'opérateurs (section 3.1.3), l'intégration de la chimie (section 3.1.4), l'advection (section 3.1.5), la diffusion horizontale (section 3.1.6) et le pas de temps d'intégration (section 3.1.7).

3.1.2 Étude et procédure d'analyse

Une simulation réaliste d'une semaine de l'été 2001 (voir chapitre 2) sert de simulation de test. La description de cette simulation du point de vue physique se trouve en section 3.3. La simulation étudiée dans cette section se déroule du 1er juillet au 7 juillet inclus. Le domaine est [40.25°N, 10.25°W]×[56.75°N, 22.25°E], discrétisé comme au chapitre 2 avec 65 cellules d'ouest

en est, avec 33 cellules du sud au nord, et avec 5 couches verticales. Il y a donc 10 725 cellules. Soixante-douze espèces sont intégrées en temps, ce qui porte le nombre de variables intégrées en temps (sur tout le domaine) à 772 200. Le pas de temps de référence est 600 s, afin de respecter la condition de Courant-Friedrichs-Lewy (CFL dans la suite).

On ne dispose pas de solution exacte et on s'attache donc principalement à constater les effets des changements opérés. La plupart des changements peuvent être identifiés *a priori* comme des dégradations. On en vient donc à juger de l'importance de ces dégradations. Pour cela, on compare les simulations entre elles ; par exemple, une simulation avec un schéma numérique précis et une autre avec un schéma plus simple. Les comparaisons se font au premier niveau vertical puisqu'il s'agit du niveau pour lequel la connaissance de l'incertitude est la plus utile et parce que le traitement numérique des flux au sol (émission, dépôt) est délicat.

Soient $A = (A)_{h,i,j}$ et $B = (B)_{h,i,j}$ les résultats de deux simulations A et B. h est un indice temporel, et i et j sont des indices spatiaux. La moyenne spatio-temporelle de A est notée $\overline{A}$.

Une partie de l'analyse est faite sur la base de courbes qui représentent l'évolution au cours du temps de moyennes ou d'écarts-types (spatiaux) de concentrations. En comparant les évolutions de deux simulations, une première appréciation de l'impact des schémas numériques est fournie.

Une seconde partie de l'analyse se fonde sur une mesure de la différence. Il s'agit d'un nombre (résultat agrégé) qui quantifie la distance entre deux simulations, de la même manière que des mesures de la distance ont été introduites au chapitre 2 pour la comparaison entre une simulation et des observations. En considérant qu'une erreur numérique supérieure à 5% n'est plus négligeable, on utilise l'indicateur suivant :

$$c(A,B) = \frac{\text{card}\{(h,i,j)/|\Delta_{h,i,j}| < 5\%\}}{\text{card}\{(h,i,j)\}}, \quad \text{avec} \quad \Delta_{h,i,j} = \frac{A_{h,i,j} - B_{h,i,j}}{\frac{1}{2}(\overline{A} + \overline{B})} \tag{3.1}$$

où card est le cardinal. $c(A,B)$ représente donc la proportion d'erreur relative inférieure à 5%. Dans la suite, on se réfère à $c(A,B)$ comme étant le *coefficient d'accord*. Si le coefficient d'accord vaut 100%, on jugera qu'il y a un accord très fort entre les simulations comparées.

Les polluants étudiés sont

1. l'ozone O_3, polluant servant aux applications de cette thèse, souvent impliqué dans l'étude la qualité de l'air et sujet à une grande partie des phénomènes physico-chimiques de l'atmosphère ;
2. le monoxyde d'azote NO, très localisé près de ses sources d'émissions (grandes villes) ;
3. dioxyde d'azote NO_2, polluant essentiel dans le cycle de l'ozone (section 1.2.3) ;
4. le dioxyde de soufre SO_2, polluant d'importance pour les aérosols, souvent simulé, peu impliqué dans la chimie de l'ozone ;
5. le radical hydroxyle HO, espèce à temps de vie très court, très réactif.

On ne peut pas exiger la même précision numérique pour toutes les espèces du fait de leurs répartitions spatiales (forts gradients pour certaines espèces) et de leurs temps de réaction. L'ozone O_3 et le dioxyde de soufre SO_2 ne doivent être altérés que modérément, notamment du fait de leur relative homogénéité sur le domaine. Le monoxyde d'azote NO est lui très localisé en ses lieux d'émission. Numériquement comme physiquement, il est difficile de le simuler à l'échelle continentale. Le radical HO est aussi difficile à simuler du fait de ses temps de réaction très courts. Le dioxyde d'azote NO_2 est intermédiaire. En conclusion, si un changement de schéma numérique n'altère pas HO et NO, on peut considérer qu'il n'a pas d'impact. S'il modifie les concentrations de NO_2, le changement n'est pas négligeable. Si les concentrations de O_3 et SO_2 sont significativement modifiées (ce qui est mesuré avec le coefficient d'accord), l'intégration est sensible au schéma.

3.1.3 Séparation d'opérateurs

Ordre de la séquence

Ainsi que le présente la section 2.3, Polair3D utilise la séparation d'opérateurs. Il s'agit d'une approximation nécessaire pour conserver des temps de calcul raisonnables. S'il fallait coupler tous les processus, seul un schéma explicite permettrait d'éviter l'inversion d'un système de taille 772 200×772 200 (pour cette étude), ce qui poserait alors un problème pour l'intégration de la chimie (raideur).

La première question concerne la sensibilité à l'ordre de la séquence d'intégration. L'ordre choisi dans Polair3D est advection–diffusion–chimie de sorte à placer l'opérateur le plus raide (chimie) en fin de séquence [Sportisse, 2000] ; cet ordre est noté ADC. Il faut noter que les conditions aux limites de surface (émissions et vitesses de dépôt) sont incluses dans la diffusion. Les tests montrent que l'ordre essentiel est celui entre la diffusion et la chimie. La figure 3.1 l'illustre : les simulations avec les ordres ADC, DCA et DAC sont confondues et la simulation avec l'ordre ACD présente un comportement différent. Les simulations ACD, CDA et CAD (chimie puis diffusion) apparaîtraient confondues si elles étaient tracées.

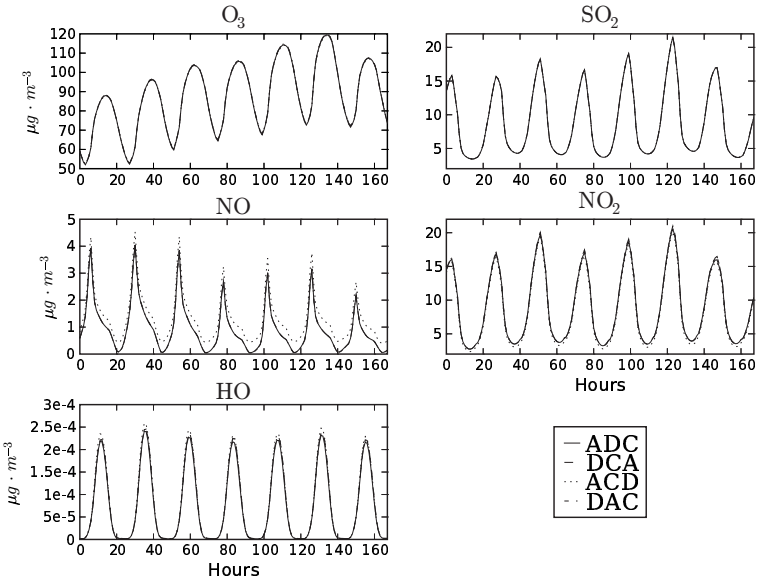


FIG. 3.1 – Évolutions temporelles des concentrations moyennes selon l'ordre de la séquence d'intégration. Les trois simulations pour lesquelles la chimie est intégrée après la diffusion sont confondues. Seule la simulation avec l'ordre ACD (chimie et diffusion permutées par rapport à la simulation de référence) se distingue.

En analysant plus finement, on constate tout de même de faibles différences lorsque la chimie reste intégrée après la diffusion. La figure 3.2 l'illustre. Dans le cas où la chimie et la diffusion sont permutées (ACD), les différences quantifiées sont nettement plus importantes – voir figure 3.3.

Conditions aux limites

La diffusion porte les conditions aux limites (au sol). L'ordre de la séquence entre la chimie et la diffusion est peut-être sensible de ce fait. La figure 3.4 indique que le placement des conditions aux limites n'explique pas les différences observées précédemment. Les différences entre les

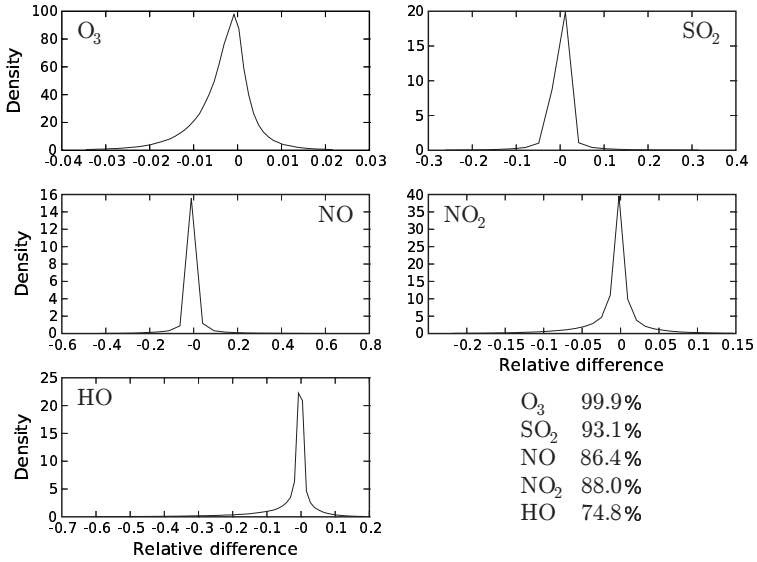


FIG. 3.2 – Distribution de la différence relative Δ (équation 3.1) entre les simulations avec ordres ADC (référence) et DCA, et coefficients d'accord pour les cinq espèces. On constate des différences non négligeables entre ces deux simulations. L'ozone est cependant peu affecté.

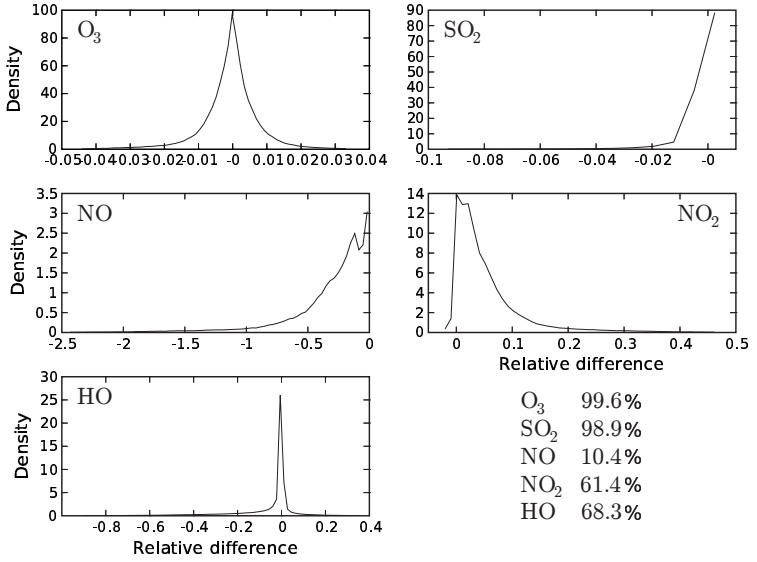


FIG. 3.3 – Distribution de la différence relative Δ (équation 3.1) entre les simulations avec ordres ADC (référence) et ACD, et coefficients d'accord pour les cinq espèces. L'ozone et le dioxyde de soufre sont peu affectés, mais les autres espèces sont très sensibles au changement d'ordre.

simulations sont moins importantes que précédemment pour toutes les espèces, sauf pour l'ozone. Les perturbations des concentrations d'ozone restent modérées. Sur les cinq polluants analysés, on peut conclure que l'impact d'un changement d'ordre dans la séquence n'a pas sa source dans le placement des conditions aux limites. Cependant, l'impact de ce placement n'est pas négligeable.

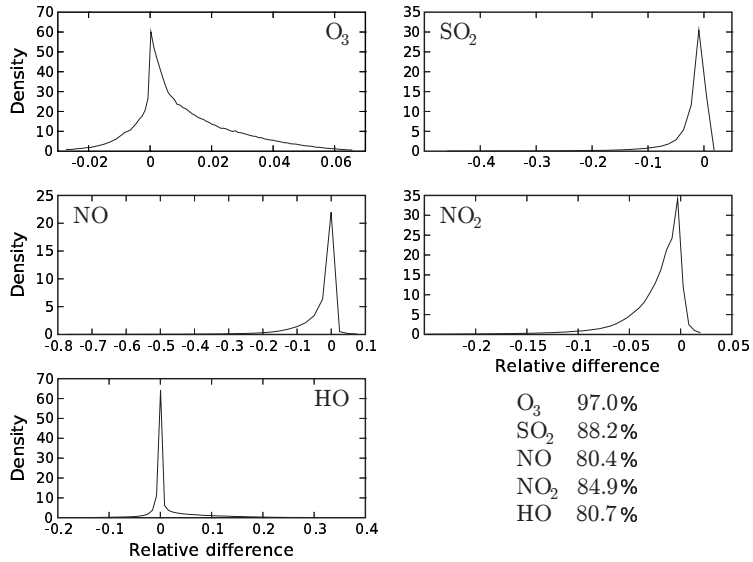


FIG. 3.4 – Distribution de la différence relative Δ (équation 3.1) entre les simulations dont les conditions aux limites (au sol) sont soit intégrées avec la diffusion (référence) soit avec la chimie, et coefficients d'accord pour les cinq espèces. L'ozone est assez peu affecté, mais les autres espèces sont sensibles au changement d'ordre.

Méthode de séparation d'opérateurs

La séparation d'opérateurs conduit à simplement intégrer les processus les uns après les autres, chaque processus intégrant les concentrations issues de l'intégration du processus précédent. Il existe plusieurs variantes.

La première [Strang, 1968] consiste à intégrer les processus les uns après les autres sur un demi pas de temps, et puis à intégrer les processus dans l'ordre inverse sur un demi pas de temps. Ceci permet d'augmenter l'ordre du schéma. Un léger impact est observé, figure 3.5.

Une autre méthode (« internal splitting ») consiste à séparer les opérateurs au niveau de l'algèbre linéaire [Verwer *et al.*, 1996]. Par exemple, la première équation du système 2.13 demande l'inversion d'une matrice $I - \gamma\Delta t A^n$. Sans séparation d'opérateurs, A^n est une approximation de la matrice jacobienne pour l'ensemble des processus. L'idée est de la décomposer : $A^n = A_1^n + A_2^n$; et d'approcher $I - \gamma\Delta t A^n$ par $(I - \gamma\Delta t A_1^n)(I - \gamma\Delta t A_2^n)$. L'inversion de $I - \gamma\Delta t A^n$ peut alors se faire en deux temps, avec les inversions successives de $(I - \gamma\Delta t A_1^n)$ et puis $(I - \gamma\Delta t A_2^n)$. Ces deux inversions sont moins coûteuses du fait de la structure avantageuse (par bloc, par exemple, si les matrices correspondent à des processus physiques) des matrices A_1^n et A_2^n. On teste cette méthode sur le couple diffusion verticale–chimie. L'impact reste faible sur l'ozone et sur le dioxyde d'azote. Les autres espèces sont perturbées, mais moins que dans les tests précédents. Les résultats sont présentés figure 3.6.

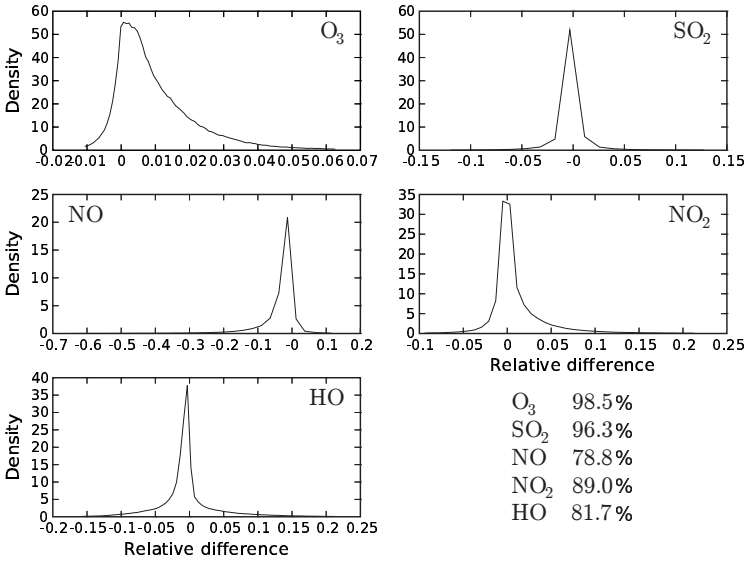


FIG. 3.5 – Distribution de la différence relative Δ (équation 3.1) entre la simulation de référence et la simulation avec séparation du « second ordre » [Strang, 1968], et coefficients d'accord pour les cinq espèces. L'ozone et le dioxyde de soufre sont assez peu affectés, mais les autres espèces sont sensibles au changement d'ordre.

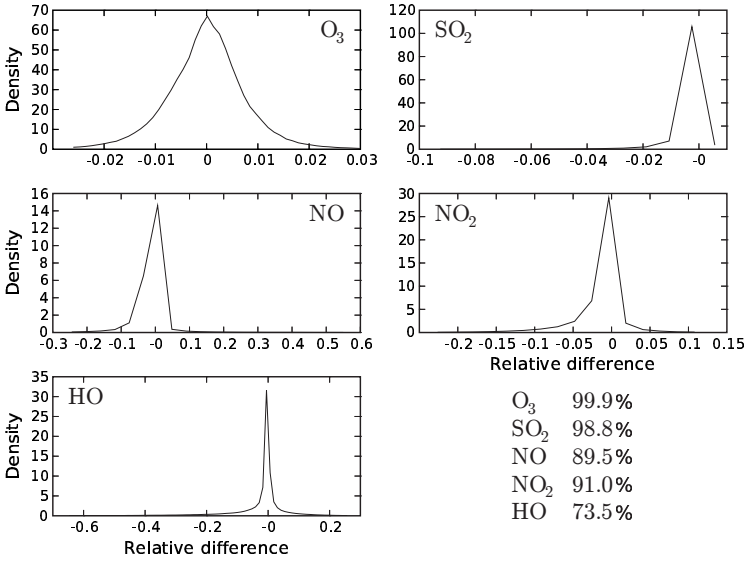


FIG. 3.6 – Distribution de la différence relative Δ (équation 3.1) entre la simulation de référence et la simulation avec séparation au niveau de l'algèbre linéaire (« internal splitting »), et coefficients d'accord pour les cinq espèces. L'impact est faible.

Un dernier test concerne la séparation d'opérateurs dite « source splitting » ou « no time splitting » [Sun, 1996]. Cette méthode permet d'intégrer tous les processus à partir de la même condition initiale, ce qui évite les ruptures dans le processus d'intégration. Les tests montrent que les concentrations sont très peu affectées par cette méthode (figure 3.7). Par contre, la méthode contribue à stabiliser les calculs. Des pas de temps de 1800 s, par exemple, sont envisageables grâce à elle. C'est la raison pour laquelle Polair3D utilise le « source splitting ».

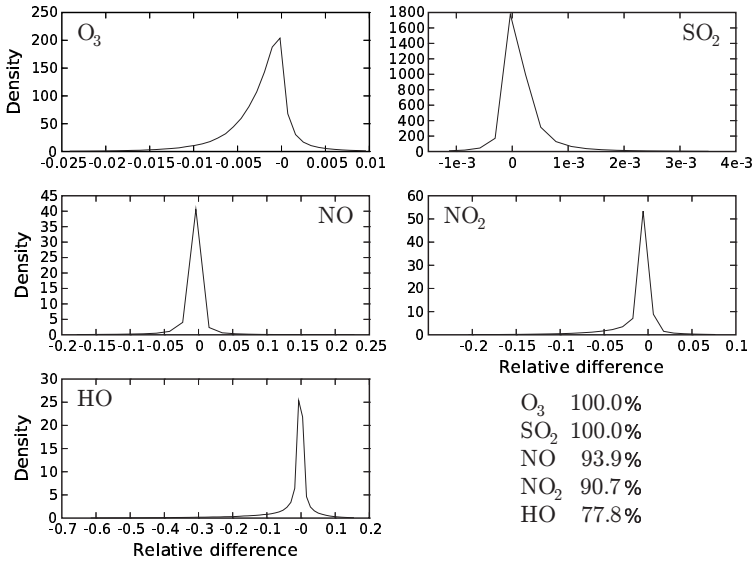


FIG. 3.7 – Distribution de la différence relative Δ (équation 3.1) entre la simulation de référence et la simulation avec « source splitting », et coefficients d'accord pour les cinq espèces. L'impact est très faible.

3.1.4 Intégration de la chimie

Outre le temps calcul et la précision des calculs, l'intégration de la chimie soulève le problème de la positivité et de la stabilité. Le schéma de Rosenbrock (équations 2.12–2.13), appliqué à l'équation scalaire $\frac{\mathrm{d}c}{\mathrm{d}t} = f(c,t)$ d'un pas de temps n à un pas de temps $n+1$, avec $f(c,t^n) = -k^n c$ et $f(c,t^{n+1}) = -\beta k^n c$, a pour fonction de stabilité[1] :

$$R(\lambda) = \frac{(2\gamma^2 - (3+\beta)\gamma + \beta)\lambda^2 + (4\gamma - 1 - \beta)\lambda + 2}{2(1+\gamma\lambda)^2} \tag{3.2}$$

où $\lambda = k^n \Delta t$. Dans le cas autonome ($\beta = 1$), le schéma est positif et L-stable (c'est-à-dire $\lim_{\lambda\to+\infty} R(\lambda) = 0$) avec $\gamma = 1 + \frac{\sqrt{2}}{2}$. En pratique, le système n'est pas autonome puisque les constantes photolytiques (principalement) varient en temps. La valeur de β associée aux constantes photolytiques diffère fortement de 1 au lever et au coucher du soleil.

Dans le cas non-autonome, on peut prendre :

$$\gamma(\beta) = \frac{3 + \beta + \sqrt{\beta^2 - 2\beta + 9}}{4} \tag{3.3}$$

[1] La fonction de stabilité R est telle que $c^{n+1} = R(\lambda)c^n$.

qui garantit la positivité et la L-stabilité. Dans le test qui suit, β est estimé avec les matrices jacobiennes calculées, à concentrations constantes, aux temps n et $n+1$. Les deux diagonales en sont extraites et β est pris égal au rapport maximal entre éléments diagonaux correspondants.

Le système étant non-autonome, le schéma n'est plus positif. Les concentrations négatives en sortie d'intégration sont mises à zéro, on parle de « clipping ». Il s'agit d'un ajout de masse artificiel qu'il convient d'évaluer. Un premier diagnostic de positivité consiste à comparer les concentrations moyennes aux concentrations mises à zéro. Le tableau 3.1 indique le taux de « clipping », c'est-à-dire le rapport entre les concentrations moyennes mises à zéro et les concentrations moyennes. Peu d'espèces (sur les 72 espèces du mécanisme chimique) ont des rapports élevés. Cependant de fortes corrections peuvent être opérées au lever ou au coucher du soleil. Le tableau 3.2 reporte le rapport entre les concentrations *maximales* mises à zéro et les concentrations moyennes. Les rapports peuvent alors être très élevés.

TAB. 3.1 – Taux de « clipping » : rapport entre les concentrations (en valeur absolue) moyennes mises à zéro et les concentrations moyennes. Seules les espèces dont le rapport est supérieur à 0.005% sont incluses.

Espèce	Rapport (%)	Espèce	Rapport (%)	Espèce	Rapport (%)
NO_3	1.01	N_2O_5	0.53	ETHP	0.52
OLND	0.52	OLNN	0.48	APIP	0.26
KETP	0.22	ETEP	0.06	HNO_4	0.03
HC_3P	0.02	XO_2	0.02	XYLP	0.02
HC_5P	0.02	HC_8P	0.02	TOLP	0.02
OLTP	0.01	OLIP	0.01	ISOP	0.01

TAB. 3.2 – Taux de « clipping » : rapport entre les concentrations (en valeur absolue) *maximales* mises à zéro et les concentrations moyennes. Seules les espèces dont le rapport est supérieur à 0.1 sont incluses. Contrairement au tableau 3.1, les rapports ne sont pas en pourcentage.

Espèce	Rapport	Espèce	Rapport	Espèce	Rapport
ETHP	121.39	APIP	95.49	KETP	60.00
NO_3	15.58	OLND	13.53	ETEP	11.76
N_2O_5	7.00	HC_3P	6.04	XO_2	4.54
OLNN	4.42	XYLP	4.09	TOLP	3.48
HC_5P	3.36	HC_8P	3.29	OLTP	2.12
OLIP	2.06	HNO_4	1.39	CSLP	1.29
ISOP	0.70	PHO	0.70	MO_2	0.22

C'est au lever du soleil que les valeurs remises à zéro sont les plus élevées (en valeur absolue). Il s'agit du cas où les constantes photolytiques croissent, soit $\beta > 1$ et donc $\gamma(\beta) > 1 + \frac{\sqrt{2}}{2}$. On constate effectivement que de plus fortes valeurs de γ diminuent les concentrations ajoutées (figure 3.8). Dans le même temps, l'impact sur les concentrations des autres espèces est faible (figure 3.9). On ne retient pourtant pas cette valeur de γ car elle peut conduire à des instabilités (à $\gamma = 5$, la simulation devient instable), et son apport est faible pour les polluants généralement étudiés à cette échelle (comme O_3, SO_2 voire NO_2).

La valeur de γ proposée dans l'équation 3.3 n'est pas efficace pour préserver la positivité, ainsi que le montre la figure 3.10. L'approximation de β (rapport des éléments diagonaux des matrices jacobiennes) est probablement insuffisante.

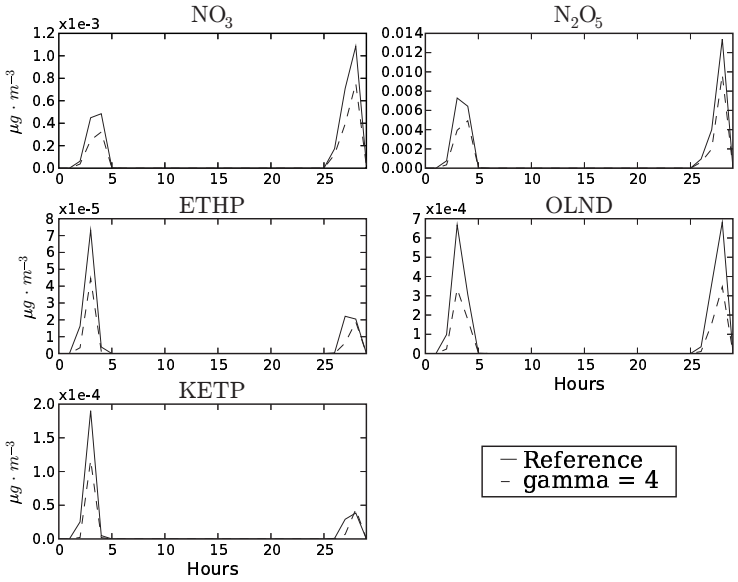


FIG. 3.8 – Évolution temporelle des concentrations moyennes ajoutées (« clipping ») pour cinq espèces des tableaux 3.1 et 3.2. La simulation de référence ($\gamma = 1 + \frac{\sqrt{2}}{2}$) est comparée à la simulation avec $\gamma = 4$. La première heure (UT) est indicée par 0. Seules les 30 premières heures sont affichées par souci de lisibilité. On constate que $\gamma = 4$ est plus favorable au lever du soleil.

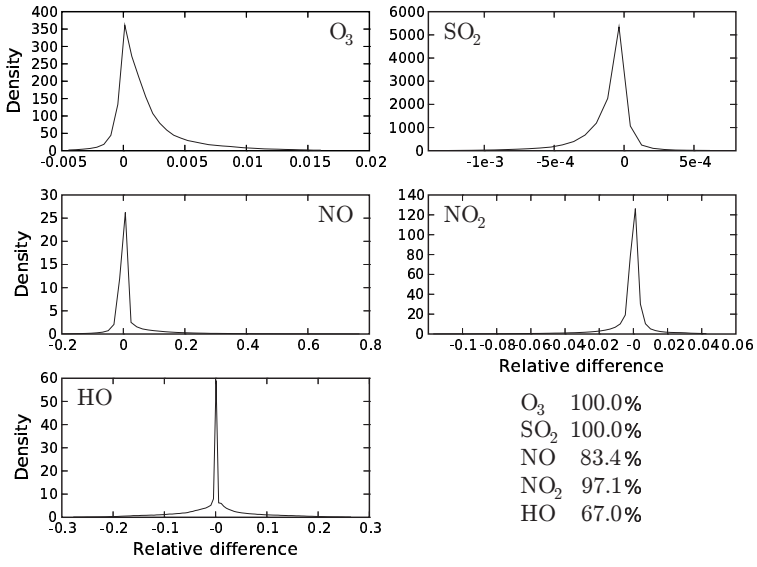


FIG. 3.9 – Distribution de la différence relative Δ (équation 3.1) entre la simulation de référence et la simulation avec $\gamma = 4$, et coefficients d'accord pour les cinq espèces. L'impact du changement est faible.

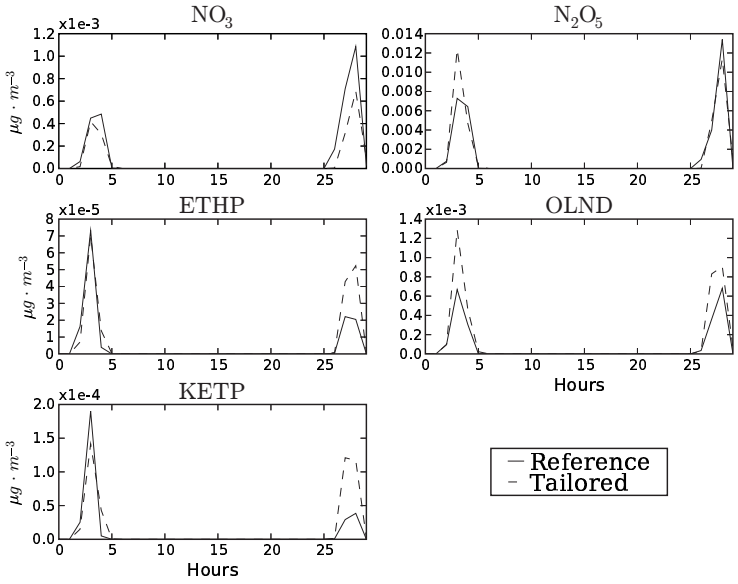


FIG. 3.10 – Évolution temporelle des concentrations moyennes ajoutées (« clipping ») pour cinq espèces des tableaux 3.1 et 3.2. La simulation de référence ($\gamma = 1 + \frac{\sqrt{2}}{2}$) est comparée à la simulation avec $\gamma(\beta)$ issu de l'équation 3.3 où β est estimé par les rapports des éléments diagonaux des matrices jacobiennes (détails dans le corps du texte). La première heure (UT) est indicée par 0. Seules les 30 premières heures sont affichées par souci de lisibilité.

3.1.5 Schéma d'advection

On compare le schéma d'advection présenté à la section 2.3 au schéma d'ordre 1 décentré (« upwind », en anglais). On rappelle que le schéma de référence, avec limiteur de flux, est une pondération entre un schéma d'ordre 3 et le schéma d'ordre 1 décentré. Les résultats sont résumés par la figure 3.11. On note une différence importante entre les deux simulations. La diffusion numérique du schéma d'ordre 1 a un fort impact.

3.1.6 Diffusion horizontale

Le coefficient de diffusion horizontale est un paramètre physique mal connu. La diffusion horizontale est parfois retirée des modèles de chimie-transport car le schéma d'advection, par sa diffusion numérique, joue un rôle *a priori* similaire. Trois simulations sont effectuées, avec des coefficients de diffusion horizontale de $K_h = 0$ m$^2 \cdot$ s^{-1}, $K_h = 10\,000$ m$^2 \cdot$ s^{-1} et $K_h = 50\,000$ m$^2 \cdot$ s^{-1}. On constate que les moyennes spatiales sont peu affectées. En revanche, les maxima ainsi que les écarts-types (spatiaux) sont modifiés nettement pour $K_h = 50\,000$ m$^2 \cdot$ s^{-1} et légèrement pour $K_h = 10\,000$ m$^2 \cdot$ s^{-1}. La figure 3.12 l'illustre pour les écarts-types.

3.1.7 Pas de temps

Les modèles de chimie-transport permettent d'intégrer l'équation de réaction-diffusion avec des pas de temps largement supérieurs aux temps caractéristiques de plusieurs réactions. Certains schémas numériques permettent même de s'affranchir de la condition de CFL [Hundsdorfer et Spee, 1995]. Le pas de temps peut donc varier sur un intervalle très large. Compte tenu de la dynamique, un pas de temps d'une heure pourrait être raisonnable.

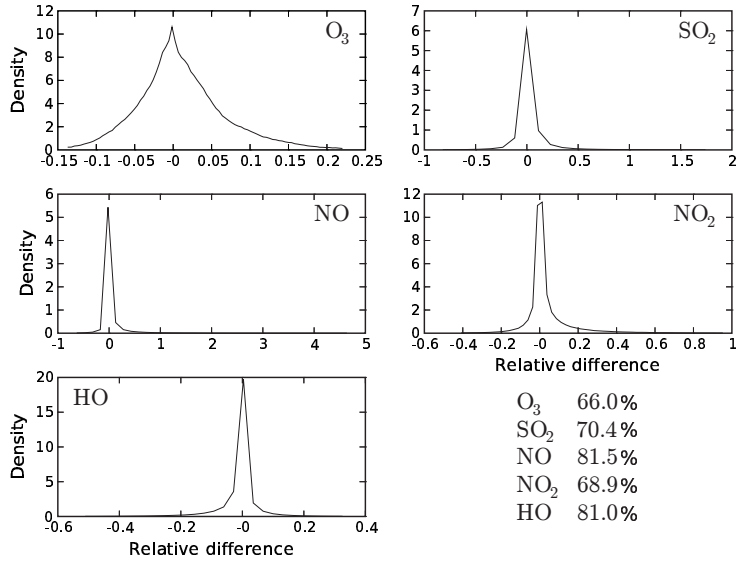


FIG. 3.11 – Distribution de la différence relative Δ (équation 3.1) entre la simulation de référence et la simulation avec le schéma d'advection décentré d'ordre 1, et coefficients d'accord pour les cinq espèces.

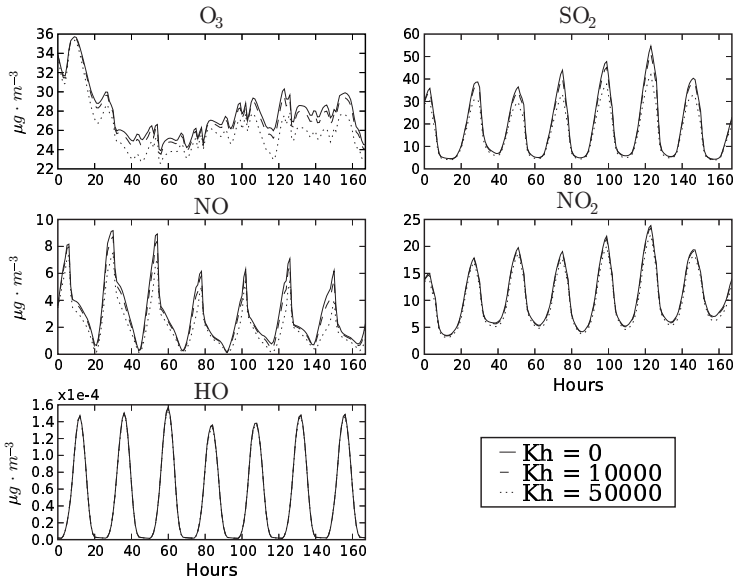


FIG. 3.12 – Évolutions temporelles des écarts-types spatiaux selon le coefficient de diffusion horizontale ($K_h = 0\ \mathrm{m^2 \cdot s^{-1}}$, $K_h = 10\,000\ \mathrm{m^2 \cdot s^{-1}}$ et $K_h = 50\,000\ \mathrm{m^2 \cdot s^{-1}}$).

Des pas de temps d'intégration (parfois dits de « splitting ») de 30 s, 60 s, 100 s, 200 s et 300 s, 600 s, 900 s, 1200 s et 1800 s. Les résultats sont tous très proches jusqu'à 600 s, et puis des différences notables apparaissent. La figure 3.13 l'illustre par les simulations avec les pas de temps 30 s, 600 s et 1800 s.

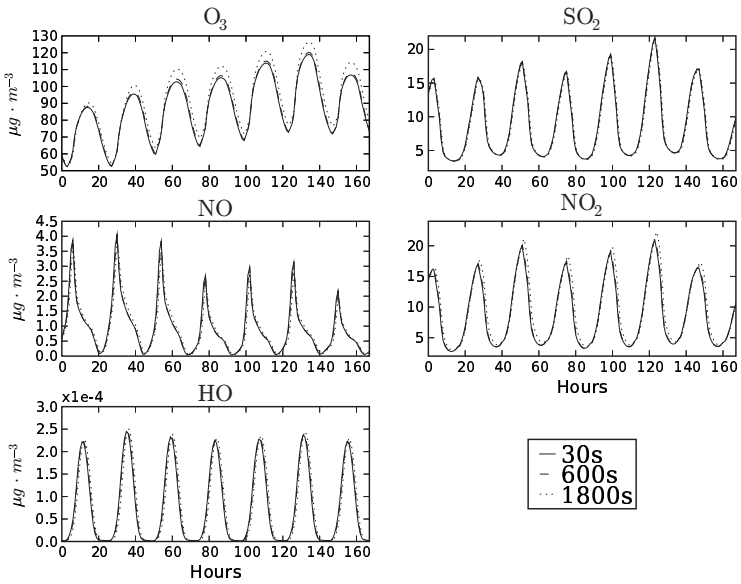


FIG. 3.13 – Évolutions temporelles de la moyenne spatiale des concentrations pour trois pas de temps d'intégration. Seule l'intégration avec un pas de temps de 1800 s se détache.

3.1.8 Bilan

Les schémas testés ont des impacts divers qu'il convient de hiérarchiser. Le tableau 3.3 condense les principaux résultats qui mettent en évidence l'influence du pas de temps (qu'il est préférable de ne pas trop augmenter), l'importance du schéma d'advection et du coefficient de diffusion horizontale. Les problèmes liés à la séparation d'opérateurs sont moins prépondérants.

Néanmoins, l'objectif des simulations doit être pris en compte avant de formuler des conclusions. Les espèces ne sont pas toutes sensibles aux mêmes schémas. Et pour une même espèce, un schéma peut par exemple avoir un fort impact sur les maxima mais aucun sur la moyenne. Le bilan précédent est essentiellement valable pour les concentrations horaires d'ozone.

D'après ce bilan, les schémas numériques sont source d'erreur et donc potentiellement d'incertitude. Il faut cependant comparer leur impact avec les autres incertitudes. La section 3.2 étudie l'incertitude due aux paramétrisations et permet une comparaison avec l'incertitude numérique (notamment via une simulation avec un pas de temps de 1800 s). Les conclusions de la section 3.2 modèrent l'impact des schémas numériques tant l'incertitude globale est élevée.

TAB. 3.3 – Bilan des coefficients d'accord, pour les principales comparaisons de l'étude, classés par ordre croissant d'accord sur l'ozone.

Comparaison	O_3	SO_2	NO	NO_2	HO
$\Delta t = 600$ s / $\Delta t = 1800$ s	54.7	80.1	28.7	59.4	43.3
Référence / ordre 1 en advection	66.0	70.4	81.5	68.9	81.0
$K_h = 10\,000$ m$^2 \cdot$ s^{-1} / $K_h = 50\,000$ m$^2 \cdot$ s^{-1}	80.0	81.9	18.4	65.7	83.9
$\Delta t = 600$ s / $\Delta t = 30$ s	96.4	89.4	83.9	79.7	57.2
Cond. lim. diffusion / cond. lim. chimie	97.0	88.2	80.4	84.9	80.7
$K_h = 10\,000$ m$^2 \cdot$ s^{-1} / $K_h = 0$ m$^2 \cdot$ s^{-1}	97.9	84.2	90.7	85.4	94.5
Référence / « Strang splitting »	98.5	96.3	78.8	89.0	81.7
ADC / ACD	99.6	98.9	10.4	61.4	68.3
Référence / $\gamma(\beta)$	99.7	100.0	94.6	97.7	97.9
ADC / DCA	99.9	93.1	86.4	88.0	74.8
Référence / « internal splitting »	99.9	98.8	89.5	91.0	73.5
Référence / $\gamma = 4$	100.0	100.0	83.4	97.1	67.0
Référence / « source splitting »	100.0	100.0	93.9	90.7	77.8

3.2 Incertitude liée aux paramétrisations physiques et aux approximations numériques

Cette section est constituée de
MALLET, V. et SPORTISSE, B. (2006). Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations : an ensemble approach applied to ozone modeling. *J. Geophys. Res.*, 111(D1)

3.2.1 Introduction

Chemistry-transport models are now widely used in air-quality applications ranging from impact studies to daily forecasts. To date, they perform satisfactory simulations, in both basic cases such as passive tracer tracking [e.g., Nodop, 1997] and in complex cases involving photochemical mechanisms. The reliability of the models is partially assessed through comparisons with measurements and numerous statistical-measures [as those defined in EPA, 1991]. These comparisons are performed with intensive observation periods from specific campaigns or with daily measurements from regular monitoring sites. A large set of comprehensive and reliable 3D Eulerian chemistry-transport models has been "validated" this way, such as Chimere [Schmidt *et al.*, 2001], CMAQ [Community Multiscale Air Quality, Byun et Ching, 1999], DEHM [Danish Eulerian Hemispheric Model, Christensen, 1997], EMEP [European Monitoring and Evaluation Programme, Simpson *et al.*, 2003], Eurad [European Air Pollution Dispersion, Hass, 1991], Lotos [Long Term Ozone Simulation, Builtjes, 1992], Polair3D [Boutahar *et al.*, 2004].

These models have usually been "tuned" in order to deliver satisfactory model-to-observation statistics. Also while the "validations" give the error of the simulations, they do not give information on the uncertainty associated with these simulations. The origin of the uncertainty is threefold: the underlying physical parameterizations (biogenic emissions, deposition velocities, turbulent closure, chemical mechanism, etc.), the input data (land use data, emission inventories, raw meteorological fields, chemical data, etc.) and the numerical approximations (mesh sizes, time step and number of chemical species). The best characterization of the uncertainty would be the probability density functions of the simulation errors. Computing a probability density function (PDF) for given model outputs (such as forecast error statistics) is in practice

a difficult task primarily because of the computational costs.

There are specific techniques to assess uncertainties. The first-order derivatives of model outputs with respect to model inputs can give "local" estimates of uncertainties [e.g., Schmidt, 2002]. Monte Carlo simulations based on different values for given input parameters or fields can provide an approximation to the probability density functions if the number of simulations is large enough [Hanna *et al.*, 2001]. An alternative approach, which is now widely used in meteorology [Toth et Kalnay, 1993; Houtemaker *et al.*, 1996; Buizza *et al.*, 1999] and which is a promising method in air quality modeling (e.g., Delle Monache et Stull [2003] for photochemistry or Galmarini *et al.* [2004] for radionuclides), is the so-called ensemble approach based on a set of models supposed to account for the range of uncertainties.

This paper uses an ensemble approach to provide estimates of the uncertainty in photochemical forecasts due to the parameterizations and some data associated with them. It also deals with numerical issues such as mesh size.

The study is performed with a four-month European-scale simulation, from May to August 2001. A comparison between the reference simulation and a similar simulation but for one change in a parameterization enables us to estimate the impact of this parameterization. For each modified parameterization, the reliability of the simulation is checked with comparisons to measurements, which allows us to assess the robustness of the whole modeling system. The same experiment is finally performed with a set of simulations in which several parameterizations may be changed (at the same time, in the same simulation). It allows us to study the robustness of the system with respect to cumulated uncertainties.

This paper is organized as follows. Section 3.2.2 briefly summarizes the relevant methods to estimate uncertainties, details the specific aims of this paper and describes the methodology. Section 3.2.3 details the model, the reference simulation and the involved parameterizations. In the last section, the results are analyzed with intercomparisons of the simulations and comparisons to observations.

3.2.2 Methodology

Definitions

We define:

- The error. It is the discrepancy between model outputs and field observations.
- The uncertainty. It is the range of values in which the model outputs may lie with a high degree of confidence. In this paper, we only deal with a priori uncertainties, i.e. uncertainties estimated without taking into account observations.
- The spread. Hereafter we refer to the variability of an ensemble as its spread. The spread is a measure of the uncertainty and it can be quantified by a standard deviation.
- The variability. Herein the variability solely refers to the spatial or/and temporal variabilities of a concentration field. For the sake of clarity, the variability of an ensemble is called a spread.

Motivation

Assessing the uncertainties in model outputs is a field of growing interest in environmental forecasting, especially in meteorology. In meteorology, the dynamics of models have a "chaotic" behavior. The uncertainties in initial conditions have therefore a strong impact and the issue is to propagate these uncertainties through "ensemble forecasts" [Toth et Kalnay, 1993;

Houtemaker *et al.*, 1996; Buizza *et al.*, 1999]. In air quality applications, there is not such a strong dependence on initial conditions. The impact of uncertainties in the input data (e.g., emissions, meteorological fields), in the parameterizations (e.g., deposition velocities, turbulence closure) and in the numerical algorithms is much stronger.

The actual errors of a model, given by comparisons to observational data, may be low with high uncertainties in the results. A model may be tuned to fit the observations (and all models are improved this way), which leads to low errors. Nevertheless, if this model is used with different parameterizations (assumed to be valid physical parameterizations), other data or alternative numerical schemes, then it could lead to very different results, including those far from the measurements, with the magnitude of spread depending on the actual uncertainty. This is, of course, a strong limitation of the models, and the uncertainty has to be estimated in order to assess the "robustness" of the models. One may refer to Russell et Dennis [2000] for an overview of the strengths and limitations of photochemical models.

It is impossible to compute the error in all meteorological conditions, at every point in a given simulation domain (even at ground level), for all chemical species, and at every time. In the absence of observations, an estimation of the uncertainty is essentially the only means to assess the quality of the results. In an operational context, the models may be used for risk assessment. The reliability of the results is then a crucial issue and, if available, the full PDFs associated with these results would be highly valuable. For instance, in prospective or screening studies (e.g., impact studies related to different emission scenarios), the models may be used with uncommon input data (e.g., strongly corrected emissions) and without any available observations with which to tune the models. From the research point of view, an estimate of the uncertainty is necessary for other communities to assess the feasibility and the relevance of given applications. For instance, the effect of pollution on health may or may not be effectively estimated, depending on the accuracy of the underlying air-quality models. For each model, the development is also oriented to improve the description in the parameterizations responsible for the main uncertainty.

A review of existing methods

There are several methods to estimate the uncertainty and to identify its sources. As for the uncertainty due to the input data, one can compute first-order derivatives of the model outputs with respect to the model inputs [e.g., Schmidt, 2002]. This provides "local" sensitivities from which the uncertainty in the outputs can be derived, taking into account the uncertainty in the input data.

Ideally one would want to compute the full PDF associated with the results. It would mean solving the Fokker-Planck equation [the equation satisfied by the output PDF, Gardiner, 1996] which is unfeasible. Instead, the Monte Carlo methods can generate approximations of the PDF. The idea is to generate a set of N input fields that roughly describe the PDF associated with the input data. The model is then run N times, which provides an approximation of the output PDF. These methods may be well suited but they are restricted to the uncertainty due to input data or parameters in parameterizations, that is, due to continuous variables. A related method, which could be viewed as a Monte Carlo method too, is the use of a set of N input fields generated by another model. In practice, the ensemble forecasts from the meteorological centers may be used as input to the air quality models. It leads to promising applications but it is restricted to the meteorological fields [Warner *et al.*, 2002].

Another method is the use of different air quality models. This technique has already been used but with a fairly low number of models [e.g., four models in Delle Monache et Stull, 2003]. It is hard to assemble enough models to claim a reliable estimate of the uncertainty. Moreover intercomparisons are difficult because the models may not be operated under the

same conditions (e.g. with the same meteorological fields). Note that this technique involves the uncertainties of several models and is not suited to assess the uncertainty of a given model. Moreover the models have usually been tuned in comparisons to measured data; hence they do not embrace the whole uncertainty in the physics and the chemistry.

The method applied in this paper mainly takes advantage of the multiple parameterizations that should be available in a well designed modeling system [Mallet *et al.*, 2005]. The model is run in many configurations with respect to the available state-of-the-art parameterizations, but also with respect to changes in the parameters and the base input-data needed for these parameterizations. The impact of the numerical approximations is studied as well. This method allows fair comparisons since the framework is exactly the same for all simulations. It gives an accurate view of the uncertainty due to the parameterizations of a given model. Notice that the method introduces discrete changes, which is the only means to assess the impact of the parameterizations. There is no continuous transition between two parameterizations or between their base input-data sets. Details about the method are provided below.

The multi-configurations approach

The air quality system with which the experiments have been performed relies on many parameterizations (see Section 3.2.3). There are often several valid parameterizations to compute the same field. Furthermore most parameterizations depend on input-data sets (including scalar parameters). For instance, the deposition velocities depend on the land use coverage which may be given by USGS[2] data or by GLCF[3] data (see below). The alternatives between the parameterizations themselves and their input-data sets introduce a finite number of choices. Hence the method deals with discrete dependencies.

The impact of numerical options are also assessed through discrete changes, e.g. by changing a numerical scheme. Nonetheless a few values that belong to a continuous interval are studied as well. They are modified as if they were discrete variables, i.e. only a few values are allowed for them. For example, the time step is a continuous variable but it can be restricted to a set of three values (a reference time step, a small one and a large one).

For the sake of clarity, the changes in the input data to the parameterizations will be viewed as changes in the parameterizations themselves. Since the numerical issues are treated in the same way as the parameterizations (they are associated with a finite number of choices), they are also viewed as parameterizations hereafter.

Assume that the model is written in the form:

$$y = f(p_1, p_2, ..., p_N) = f(p) \tag{3.4}$$

Every input parameter $p_i \in \{0, ..., n_i - 1\}$ is associated with a given parameterization that has n_i possible values. f is the model itself. The output y may be the pollutant concentrations, deposition fields, evaluation statistics, etc. Notice that f is already a discretized model.

The reference simulation is associated with a reference vector assumed to be zero: $p_{\text{ref}} = 0$. The idea is to estimate the uncertainty and the impact of every parameterization by changing one parameterization at a time, i.e. computing all $f(p)$ where $p_i = 0$ for all i except for one component. There are $\sum_{i=1}^{N}(n_i - 1)$ such simulations. This is only a small subset of the $\Pi_{i=1}^{N} n_i$ possible combinations, but the computational cost makes it impossible to run all simulations.

This method allows us to estimate the impact of each parameterization. The impact is estimated with the resulting changes in the output concentrations. It is analyzed with the concentration distributions and their spatial and temporal variabilities. In addition, for each

[2]U.S. Geological survey.
[3]Global Land Cover Facility.

change, an evaluation of the output can be performed. It shows whether the modified parameterization leads to an improved agreement with the measurements and, therefore, maybe to a better description of the physics. The fact that not all combinations $(p_i)_i$ are available restricts the study: it is hard to decide whether a parameterization should be discarded because its drawbacks may be canceled by changes in other parameterizations. There are still useful conclusions to draw: for instance, it may be shown that a given parameterization limits the variability in the results.

Furthermore, the results are enhanced by combined changes, but only with a few selected parameterizations to reduce the computational cost of the study. Four parameterizations are selected mainly due to their significant impact (even if other parameterizations have a similar importance). The model is then put in the form $y = f(\tilde{p})$ where the vector $\tilde{p}$ has four components. Each component can take two values (0 or 1); therefore there are 16 possible combinations. It provides a rough estimate of the overall uncertainty.

3.2.3 The Experiments Setup

The Modeling System

This study is based on the modeling system Polyphemus (available under the GNU General Public License at `http://www.enpc.fr/cerea/polyphemus/`). This system is divided into four parts:

1. the databases: they incorporate the data needed in the parameterizations (one may also include the meteorological fields here).

2. the libraries: they provide (1) facilities to manage the multidimensional data involved in atmospheric chemistry, (2) useful functions associated with the physical and chemical fields (e.g. coordinate transformations) and (3) the parameterizations.

3. a set of programs: these programs make the calls to the libraries to generate the input data needed by the chemistry-transport model. Their flexibility is made possible by the input configuration files that they read.

4. the chemistry-transport model: it is responsible for the time integration of the chemistry-transport equation. It therefore computes the output concentrations.

The databases contain the raw data: the land use coverage, the anthropogenic emission inventories, chemical constants, etc. The meteorological fields may also be included even if they strongly depend on the application.

The libraries play a major role in this study since they provide the basis of the flexibility of the parameterizations. They first provide the data structures and functions needed for data processing. They then provide a set of parameterizations. Most of the changes to the simulations are made with a different call to the libraries, specifically to the library dedicated to physical parameterizations, the C++ library AtmoData [Mallet et Sportisse, 2005b].

The programs of Polyphemus make calls to the libraries in order to process the raw data. They format the raw data for the chemistry-transport model, but the primary function of the programs is to use the parameterizations from the library AtmoData to compute the needed fields. These programs read configuration files in which many options are specified, including which parameterizations are to be used and with which input data and parameters. Roughly speaking, there exists a set of configuration files for every vector p (vector defined in Section 3.2.2). This study therefore relies heavily on the flexibility characteristic of the programs.

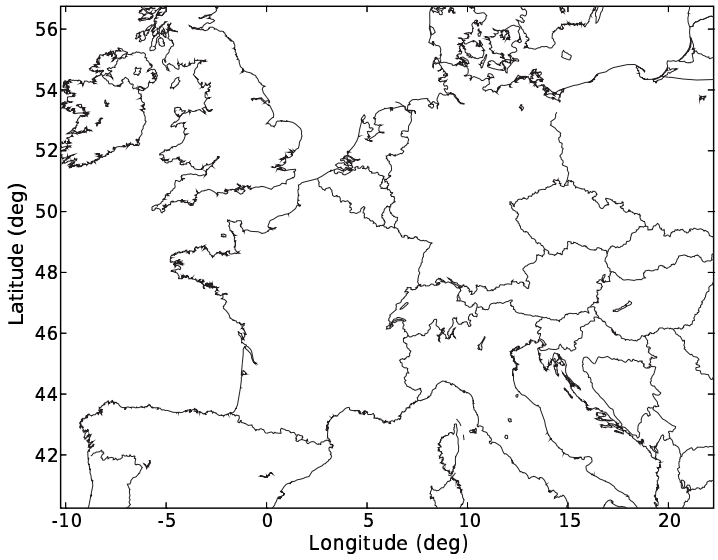


Figure 3.14: Domain $[40.25°N, 10.25°W] \times [56.75°N, 22.25°E]$ of the reference simulation.

Finally, the Eulerian chemistry-transport model Polair3D computes the output concentrations through the numerical integration of the transport-chemistry equation. With respect to this study, a strong advantage of Polair3D is its ability to deal with multiple chemical mechanisms. Details about Polair3D may be found in Boutahar *et al.* [2004].

Further details about the architecture of the whole system may be found in Mallet *et al.* [2005]. A complete description is not relevant here because not all features of the system are used in this study. The system is able to handle many applications (many chemical mechanisms, data assimilation, Monte Carlo simulations, etc.) and its flexibility enables the multiple experiments presented in this paper. The next subsection describes the base application.

The Reference Simulation

The impact of the parameterizations is evaluated by the changes they introduce with respect to the reference simulation. The reference simulation takes place at European scale during summer 2001 (22 April 2001 to 31 August 2001). A validation, over the same domain and the same period, similar to the reference simulation, may be found in Mallet et Sportisse [2004].

The domain is $[40.25°N, 10.25°W] \times [56.75°N, 22.25°E]$ and is shown in Figure 3.14. The first layer is located between 0 m and 50 m; the concentrations are thus computed at 25 m. The thickness of the other layers is about 600 m with the top of the last layer at 3000 m. RACM is the photochemical mechanism used in this simulation [Stockwell *et al.*, 1997]. Since the best results are obtained for ozone and the number of ozone measurements is significantly higher than for other species, this study focuses on ozone. We are notably concerned with the ozone peaks since they are often of high interest in forecasts (due to the regulations that mostly limit the peaks).

Here is a review of the main components of the reference simulation:

1. meteorological data: the best ECMWF data available for the period (i.e. $0.36° \times 0.36°$, the TL511 spectral resolution in the horizontal, 60 levels, time step of 3 hours, 12 hours forecast-cycles starting from analyzed fields);

2. land use coverage: USGS finest land cover map (24 categories, 1 km Lambert);

3. emissions: the EMEP[4] inventory, converted according to Middleton *et al.* [1990];

4. biogenic emissions: computed as advocated in Simpson *et al.* [1999];

5. deposition velocities: the revised parameterization proposed in Zhang *et al.* [2003b];

6. vertical diffusion: within the boundary layer, the Troen and Mahrt parameterization as described in Troen et Mahrt [1986], with the boundary-layer height provided by the ECMWF; above the boundary layer, the Louis parameterization [Louis, 1979];

7. boundary conditions: output of the global chemistry-transport model Mozart 2 [Horowitz *et al.*, 2003] run over a typical year;

8. numerical schemes: a first-order operator splitting, the sequence being advection–diffusion–chemistry; a direct space-time third-order advection scheme with a Koren flux-limiter [Verwer *et al.*, 1998]; a second-order order Rosenbrock method for diffusion and chemistry.

The performance of the reference simulation has been evaluated through a comparison of the forecasted ozone peaks with the observations from 242 stations distributed over Europe (in a network with mixed stations: urban, peri-urban and rural stations). With the first five days excluded (because of the rough initial conditions), the root mean square (with all observations put together) is $23.5\mu g \cdot m^{-3}$, the correlation is 71.4% and the bias $-4.5\mu g \cdot m^{-3}$ (the mean of observed values being $94.7\mu g \cdot m^{-3}$) – the statistical measures are defined in the appendix. The results therefore show a reasonable agreement with observations [Hass *et al.*, 1997; Schmidt *et al.*, 2001].

The Parameterizations

The modified parameterizations were chosen according to the relevance and the availability of alternative parameterizations. Only state-of-the-art parameterizations or, at least, widely used parameterizations are involved. The list of the parameterizations (and the data associated with them) used in this study is shown in Table 3.4.

[4]Co-operative Programme for Monitoring and Evaluation of the Long-range Transmission of Air Pollutants in Europe.

Table 3.4: Parameterizations, raw input data and numerical choices for the reference simulation and their alternatives. The impact of the parameterizations are assessed in this study through the use of the alternatives shown in this table.

#	Parameterization	Reference	Alternative(s)	Comment
		Physical parameterizations		
1.[a]	Chemistry	RACM	RADM 2 [Stockwell *et al.*, 1990]	
2.	Vertical diffusion	Troen & Mahrt	Louis [Louis, 1979]	
3.			Louis in stable conditions	Troen & Mahrt kept in unstable conditions
4.	Deposition velocities	Zhang [Zhang *et al.*, 2003b]	Wesely [Wesely, 1989]	
5.	Surface flux	Heat flux[b]	Momentum flux[b]	For the aerodynamic resistance (in deposition velocities)
6.	Cloud attenuation	RADM method [Chang *et al.*, 1987; Madronich, 1987]	Esquif (ESQUIF [2001])	
7.	Critical relative humidity	Depends on σ	Two layers	Used in the RADM method to compute cloud attenuation
		Input data		
8.	Emissions vertical distribution	All in the first cell	All in the two first cells	
9.	Land use coverage	USGS	GLCF	For deposition velocities[c]
10.	Land use coverage	USGS	GLCF	For biogenic emissions[c]
11.	Exponent p in Troen & Mahrt	2	3	
12.	Photolytic constants	JPROC	Depends on the zenith angle (only)	
		Numerical issues		
13.	Time Step	600 s	100 s	
14.			1800 s[d,e]	
15.	Splitting method	First order	Strang splitting	
16.	Horizontal resolution	0.5°	0.1° [e]	
17.			1.0°	
18.	Vertical resolution	5 layers	9 layers	The first layer height remains 50 m
19.	First layer height	50 m	40 m	The top of every other layer does not change

[a] The reference simulation will be referred to as simulation #0.

[b] Computed using the Louis formulae.

[c] The consistency between the land use coverage used for the deposition velocities and for the biogenic emissions is not required. Indeed a large part of the uncertainty lies in the data associated with the land use categories (e.g. resistances for deposition and emission factors for the biogenic emissions). Moreover a given description may be more suited only for the emissions (vegetation) or only for the deposition (roughness, etc.).

[d] The advection is integrated over submultiples of 1800 s so as to satisfy the CFL (Courant-Friedrichs-Lewy) condition.

[e] The numerical scheme is also slightly modified in this simulation: it uses source splitting. It is used to enforce the stability but has only slight consequences in the results.

The changes first include prominent processes such as the chemistry (RADM 2). Several chemical mechanisms are available in Polair3D but reliable emission inventories were available only for RACM and for RADM 2. The same speciation [for volatile organic compounds, Passant, 2002] was used for the two mechanisms. A drawback is that both mechanisms are too close to embrace the diversity of the chemical mechanisms available in air quality modeling. Nevertheless, as shown hereafter, there is a substantial difference between the two mechanisms.

The sensitivity to the turbulence closure is assessed with the comparison between the Troen & Mahrt parameterization (well suited for models with a low vertical resolution – which is the case with only five layers) and the Louis parameterization. The Louis closure is used above the boundary layer (for all simulations), in the boundary layer in stable conditions (simulation 3) and in any condition (simulation 2). One should note that the leading contribution to ground concentrations comes from the vertical diffusion coefficient at the top of the first layer. It determines the transfer between this and the above layer (up to 600m) which roughly corresponds to the residual layer in the night. The Troen & Mahrt parameterization and the Louis parameterization are designed in two different ways: the first one is independent of the vertical discretization while the second one relies on finite differences. There is a clear difference in the coefficients computed by the two parameterizations: the averages at the top of the first layer are $7.6m^2 \cdot s^{-1}$ (Troen & Mahrt) and $5.7m^2 \cdot s^{-1}$ (Louis). The correlation of 60% also shows the gap between the two parameterizations for coarse vertical discretizations (the differences decrease as the vertical mesh is refined).

Another known important process with multiple parameterizations is dry deposition. An alternative to the reference velocities computed as proposed in Zhang *et al.* [2003b] is based on the method by Wesely [1989] (simulation 4) which includes a reasonable parameterization and is widely used. The two parameterizations rely on the same fundamentals and the differences in the computed deposition velocities come as much from the input data (resistances, land use descriptions) as from the parameterization itself. As for ozone deposition velocities, the relative bias between the two parameterizations is only 3%, the correlation is 96% but the ratio of the standard deviation of the difference and the mean velocity is high: 0.31. In addition, the surface flux used to compute the aerodynamic resistance can be the heat flux or the momentum flux, although the heat flux is usually assumed to be more suitable for a scalar variable such as the concentration of a pollutant.

The two last physical-parameterizations (simulations 6 and 7) deal with the attenuation coefficients. The reference option is based on the optical depth, as described in Chang *et al.* [1987] and Madronich [1987], estimated with the cloud liquid water content. The liquid water content is integrated within the clouds and the cloud fraction is calculated based on the relative humidity q and its critical value q_c:

$$\text{cloud fraction} = \frac{1-q}{1-q_c} \tag{3.5}$$

$$q_c = 1 - \alpha\sigma^a(1-\sigma)^b\left(1+\beta\left(\sigma-\frac{1}{2}\right)\right) \tag{3.6}$$

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 an alternative simulation (#7), the critical relative humidity is simply constant over two distinct layers: $q_c = 0.75$ below $700hPa$ and $q_c = 0.95$ above.

Another set of simulations is derived from changes in the input data. The land use coverage is described by the USGS data (24 categories, 1 km Lambert) or by the GLCF data (14 categories, 0.0083°). The GLCF data contain less categories worldwide but they involve more categories over Europe than the USGS data do. The impact of the land use description is assessed

through the deposition velocities (simulation 9) and the biogenic emissions (simulation 10) independently.

In the Troen & Mahrt parameterization, the vertical diffusion coefficients depend on several parameters, particularly an exponent p [see Troen et Mahrt, 1986] which determines the shape of the vertical profile. Since it is a free parameter (with $p = 2$ or $p = 3$ recommended), the exponent is set to 2 in the reference simulation and also set to 3 as an alternative (which increases the diffusion coefficients – simulation 11).

The emission inventories are a concern of most modelers, especially the time and spatial distributions associated with them. The time distribution is known to have a slight impact at continental scale [Tao *et al.*, 2004]. The horizontal distribution is given with the EMEP inventory. Meanwhile the vertical distribution is not well known and is chosen by the modeler. In the reference simulation, all emissions are released in the first layer (therefore below 50 m). In an alternative simulation (#8), the emissions from industrial combustion (sectors 1 and 3 in the EMEP inventory) and from the waste treatment (sector 9 in the EMEP inventory) are released in the second layer. The emissions due to the combustion of non-industrial plants (sector 2 in the EMEP inventory) are released in both layers (one half in each).

Finally the impact of the database for the photolytic constants is estimated. The reference simulation takes advantage of the photolytic constants computed by JPROC [part of CMAQ, Byun et Ching, 1999] that are a function of the latitude, the altitude, the day in the year and the hour angle. The alternative simulation uses a coarser description with a single dependence on the zenith angle derived from the values given in Stockwell *et al.* [1997].

The last set of simulations involves changes in the numerical approximations. The time step is set to 100 s (simulation 13) and 1800 s (simulation 14) instead of 600 s (reference simulation). In the reference simulation, the splitting method is a first-order method (advection–diffusion–chemistry). An alternative simulation (#15) takes advantage of the Strang splitting method (advection–chemistry–diffusion over $\frac{\Delta t}{2}$ and then diffusion–chemistry–advection still over $\frac{\Delta t}{2}$), Sportisse [2000]. Finally the spatial discretization is changed horizontally (simulations 16 and 17) and vertically (simulation 18 and 19). When the spatial discretization changes, the raw meteorological fields (ECMWF fields) are interpolated on the new grid.

The nineteen alternative simulations address a reasonable range of the choices that can be made in a forecasting system, in the parameterizations, their input data and the numerical options.

3.2.4 Results and Discussion

Due to the coarse initial conditions, the five first days of the simulations are excluded in the following comparisons. The comparison is limited to hourly ozone concentrations in the first layer. Moreover the comparisons between the computed fields (not with the observations) are not performed in all cells to avoid the influence of the boundary conditions. A three-cell band at the domain borders is excluded from the comparisons in Subsections 3.2.4, 3.2.4 and 3.2.4.

The first subsection compares all simulations in order to estimate the spread due to the parameterizations (and numerical choices). The second subsection focuses on the impact of each parameterization. The comparisons are relative to the reference simulation. In the third subsection, a comparison with the observations evaluates the impact of the parameterizations on the forecasts. In the last section, the impact of combined changes in the parameterizations is performed to give an estimate of the overall uncertainty.

Intercomparison between the Simulations

The distribution of the spatio-temporal means and standard deviations of the fields is shown in Table 3.5. The means and the standard deviations are well spread considering that the simulations differ only at most in two parameterizations. The mean is particularly affected by the splitting method, the number of layers, the time step (1800 s), the chemical mechanism and the vertical diffusion. The standard deviation increases due to the turbulent closure and decreases with RADM 2, the vertically distributed emissions and the land use coverage used to compute the biogenic emissions. From these comparisons, it appears that the turbulent closure and the chemical mechanism have a strong impact on the output ozone concentrations. Even the use of the Louis closure only in stable conditions modifies both the mean and the standard deviation. The numerical issues also have a clear impact on the ozone mean. Finally the emissions can modify the standard deviation of the output concentrations. These conclusions may already be known issues, but this study shows the prominent impacts.

Sorted by mean			*Sorted by standard deviation*		
#	Mean	Standard deviation	#	Mean	Standard deviation
15	90.30	25.73	2	68.94	34.23
18	89.15	24.27	3	79.15	28.83
14	87.49	26.14	11	84.14	26.22
7	85.92	25.16	16	83.27	26.20
6	85.92	24.75	4	85.18	26.19
4	85.18	26.19	14	87.49	26.14
17	85.15	24.74	15	90.30	25.73
0	84.92	25.11	19	82.62	25.22
13	84.73	24.99	7	85.92	25.16
8	84.23	21.81	0	84.92	25.11
11	84.14	26.22	12	84.00	25.06
12	84.00	25.06	13	84.73	24.99
9	83.96	24.39	5	81.38	24.87
16	83.27	26.20	6	85.92	24.75
19	82.62	25.22	17	85.15	24.74
5	81.38	24.87	9	83.96	24.39
10	81.30	22.88	18	89.15	24.27
3	79.15	28.83	10	81.30	22.88
1	77.11	21.14	8	84.23	21.81
2	68.94	34.23	1	77.11	21.14

Table 3.5: Means and standard deviations of the hourly ozone concentrations ($\mu g \cdot m^{-3}$) of the twenty simulations. The reference simulation is indexed by 0 and the other simulations are indexed as in Table 3.4. On the left, the simulations are sorted by their mean; on the right, they are sorted by their standard deviation. The relative standard deviation of the means is 5.5%, and the relative standard deviation of the standard deviations is 10.4%.

Notice that the standard deviation has a greater spread than the mean. The relative standard deviation[5] of the means and of the standard deviations shown in Table 3.5 are 5.5% and 10.4% respectively.

[5]The relative standard deviation is the standard deviation divided by the mean.

More details are provided by the distribution of each ozone field. The percentiles associated with the simulations are shown in Table 3.6. In addition, Figure 3.15 shows the relative frequency distributions of:

1. simulation 15 (splitting order) which provides the highest concentrations (highest mean and highest percentiles) with a standard deviation close to the reference simulation;
2. simulation 2 (Louis turbulence closure) which has the lowest concentrations and the highest standard deviation;
3. simulation 1 (chemical mechanism RADM 2) which is associated with the lowest standard deviation.

These three simulations exhibit the most extreme behavior and therefore give a good idea of the uncertainty due to the changes in the parameterizations.

#	10th	20th	30th	40th	50th	60th	70th	80th	90th
15	58	69	77	84	91	97	104	111	122
18	60	69	76	83	88	94	101	109	120
14	55	66	74	81	87	94	101	109	121
6	55	66	74	80	86	92	98	106	117
7	55	66	73	80	86	92	98	106	117
4	53	64	72	79	85	91	98	107	118
17	55	65	72	79	85	91	97	105	117
0	54	65	72	79	85	91	97	105	116
13	54	65	72	79	84	90	97	105	116
8	57	66	73	79	84	89	95	102	111
9	54	65	72	78	84	90	96	104	114
11	52	63	71	77	84	90	97	105	117
12	53	64	71	78	84	90	96	104	115
16	52	63	71	77	83	89	96	104	115
19	51	63	70	77	83	89	95	103	114
10	52	63	70	76	82	88	94	100	110
5	50	61	69	75	81	87	94	101	112
3	41	55	65	73	80	87	95	103	115
1	50	61	68	73	78	84	89	95	102
2	28	40	50	58	67	75	85	96	111

Table 3.6: Percentiles of ozone concentrations ($\mu g \cdot m^{-3}$), sorted by the mean of the percentiles.

The comparisons deal with the concentrations computed over the whole (restricted) domain and at all time steps. A finer analysis of the variability deals with the spatial and temporal variabilities.

The spatial variability is estimated from the time average of the spatial standard deviations (the standard deviations computed with the concentrations in all cells and at a given time step). It appears that the spatial variability and the standard deviation on the whole field at once lead to the same conclusions. The correlation between the spatial variability and the global standard deviation of all simulations can be as high as 98.7%.

The temporal variability is estimated from the spatial average of the temporal standard deviations (computed with all concentrations in a given cell). The correlation of this temporal variability with the global standard deviations is still above 98%. The variability can also be

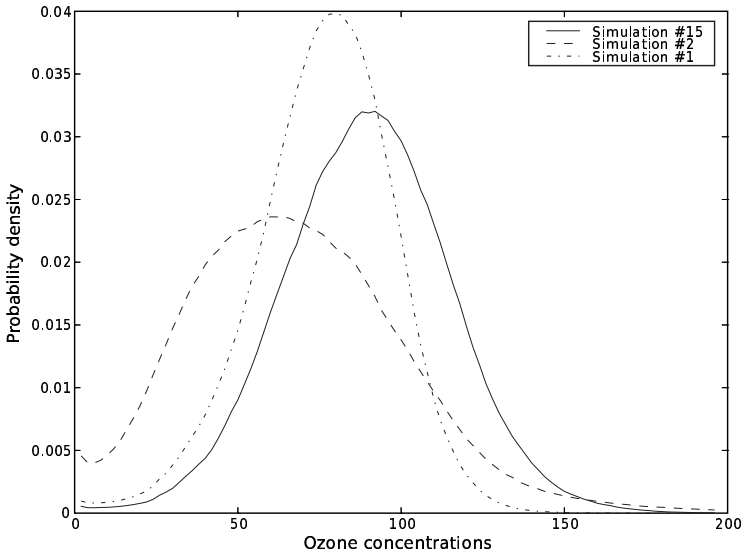


Figure 3.15: Relative frequency distributions of ozone concentrations ($\mu g \cdot m^{-3}$) for the simulations 15, 2 and 1 which show the most extreme behavior (in terms of mean and standard deviation).

estimated daily and then averaged over the days. In this case, the correlation with the global standard deviation is 94%, which is still high. Hence the parameterizations that introduce some variability increase both the temporal and the spatial variabilities. In Section 3.2.4, the most strongly impacted regions are identified.

A key point in ozone forecasts is the daily maximum. Table 3.7 shows the distribution of the means and the standard deviations of the ozone daily peaks. The behavior of the daily maxima differs from the field averages previously analyzed.

While the means of the maxima are less widely spread, the standard deviations are strongly spread and range from $16.75\mu g \cdot m^{-3}$ (simulation 1, RADM 2) to $33.50\mu g \cdot m^{-3}$ (simulation 2, Louis parameterization). On one hand, the highest standard deviation is reached with the Louis closure. Notice that if the Louis closure is only used in stable conditions (simulation 3), the impact on the daily maxima is much lower, which means that the nighttime concentrations have a small influence on the peaks. On the other hand, the lowest standard deviation comes from the chemical mechanism RADM 2. It is also associated with the lowest concentrations, which is consistent with Gross et Stockwell [2003].

Contrary to the concentration averages, the daily maxima are more variable in space than in time, as shown in Figure 3.16. The simulation with RADM 2 is essentially the only simulation for which the temporal variability is similar to the spatial variability. The impact of the parameterizations is also greater on the spatial variability than on the temporal variability: their relative standard deviations are 15.1% and 9.5% respectively.

Comparisons with the Reference Simulation

Now that the global variability has been analyzed, comparisons with the reference simulation allow us to give details about the impact of each change in the parameterizations or in the numerical choices.

Sorted by mean			*Sorted by standard deviation*		
#	Mean	Standard deviation	#	Mean	Standard deviation
15	108.79	22.62	2	95.71	33.50
14	104.40	23.87	16	101.39	24.35
18	104.11	22.87	11	101.04	23.98
4	102.39	23.38	14	104.40	23.87
7	102.03	22.58	4	102.39	23.38
6	101.71	22.05	3	99.73	23.17
16	101.39	24.35	12	99.59	22.94
11	101.04	23.98	18	104.11	22.87
0	100.92	22.62	0	100.92	22.62
17	100.62	22.49	15	108.79	22.62
13	100.52	22.57	7	102.03	22.58
9	99.77	21.68	13	100.52	22.57
3	99.73	23.17	17	100.62	22.49
12	99.59	22.94	19	98.96	22.25
19	98.96	22.25	6	101.71	22.05
5	98.34	21.61	9	99.77	21.68
8	97.90	20.10	5	98.34	21.61
10	96.31	19.32	8	97.90	20.10
2	95.71	33.50	10	96.31	19.32
1	91.49	16.75	1	91.49	16.75

Table 3.7: Means and standard deviations of the ozone daily maxima ($\mu g \cdot m^{-3}$) of the twenty simulations. On the left, the simulations are sorted by their mean; on the right, they are sorted by their standard deviation. The relative standard deviation of the means is 3.6%, and the relative standard deviation of the standard deviations is 13.5%.

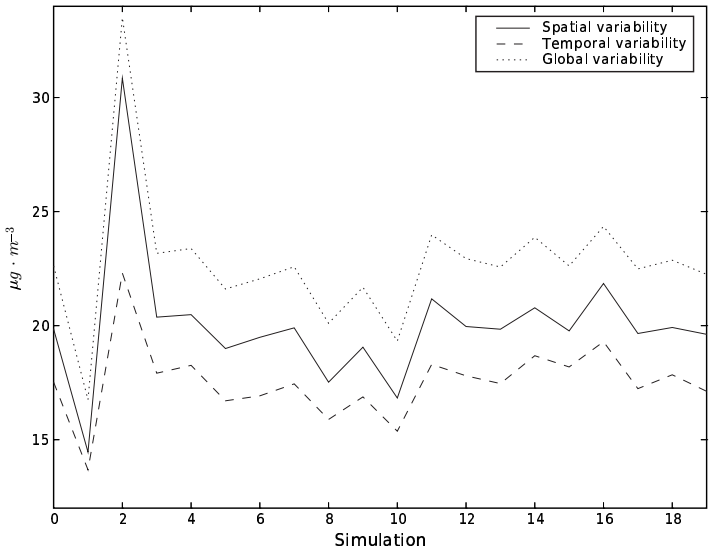


Figure 3.16: Spatial, temporal and global variabilities ($\mu g \cdot m^{-3}$) of ozone daily maxima for the twenty simulations. The spatial variability is estimated with the time average of the spatial standard deviations (the standard deviations computed with the peaks in all cells, for a given day). The temporal variability is estimated with the spatial average of the temporal standard deviations (computed with all daily maxima in a given cell). The global variability is measured by the standard deviation of all daily maxima.

First examined are the bias, the standard deviation of the distance to the reference simulation (namely the difference with the reference simulation) and the correlation with the reference simulation. These values are shown in Table 3.8.

The changes in the turbulence closure (simulation 2 and 3) lead to the largest differences. The chemical mechanism RADM 2 (simulation 1) also has an impact but not as strong as what was seen in the previous subsection with respect to the variability. On the contrary, the fine resolution (0.1°, simulation 16) leads to strong differences with the reference simulation, even if this was not obvious from the previous analyses. It is noteworthy that simulation 17, with a 1.0° resolution, has a lower but still significant impact.

The other main changes are due to the splitting method (simulation 15), the vertical resolution (nine levels, simulation 18), the emission vertical distribution (simulation 8) and the land use coverage used for the biogenic emissions (simulation 10).

For each change in the model, there is an explanation for its low or high impact on the output concentrations. We do not provide such explanations due to the number of simulations and because the purpose of the paper is to describe the global uncertainty due to the parameterizations and the numerical choices. What should be emphasized instead is that the results are sensitive to the physical parameterizations, the input data sets *and* the numerical issues.

The relative frequency distribution of all concentrations of the reference simulation is shown in Figure 3.17 with its uncertainty due to the parameterizations. The concentration distribution is sensitive to the parameterizations. The uncertainty (estimated with the relative standard deviation) in the relative frequency distribution is about 25%–30% for concentrations in $[30\mu g \cdot m^{-3}, 170\mu g \cdot m^{-3}]$. This high uncertainty is consistent with the uncertainty roughly shown in Figure 3.15.

Sorted by bias			*Sorted by standard deviation*			
#	Bias	Standard deviation	#	Bias	Standard deviation	Correlation
15	5.39	6.65	2	-15.97	18.60	0.85
18	4.23	5.10	3	-5.77	10.40	0.93
14	2.57	3.60	16	-1.64	8.95	0.94
7	1.00	1.32	1	-7.81	7.55	0.96
6	1.00	2.63	15	5.39	6.65	0.97
4	0.26	2.69	8	-0.69	5.62	0.98
17	0.24	5.07	18	4.23	5.10	0.98
13	-0.18	0.98	17	0.24	5.07	0.98
8	-0.69	5.62	10	-3.62	5.04	0.98
11	-0.78	2.81	5	-3.54	3.69	0.99
12	-0.92	2.48	14	2.57	3.60	0.99
9	-0.96	2.30	11	-0.78	2.81	0.99
16	-1.64	8.95	4	0.26	2.69	1.00
19	-2.30	2.34	6	1.00	2.63	0.99
5	-3.54	3.69	12	-0.92	2.48	1.00
10	-3.62	5.04	19	-2.30	2.34	1.00
3	-5.77	10.40	9	-0.96	2.30	1.00
1	-7.81	7.55	7	1.00	1.32	1.00
2	-15.97	18.60	13	-0.18	0.98	1.00

Table 3.8: Biases and standard deviations of the distance to the reference simulation ($\mu g \cdot m^{-3}$) for the nineteen simulations. On the left, the simulations are sorted by their bias; on the right, they are sorted by their standard deviation.

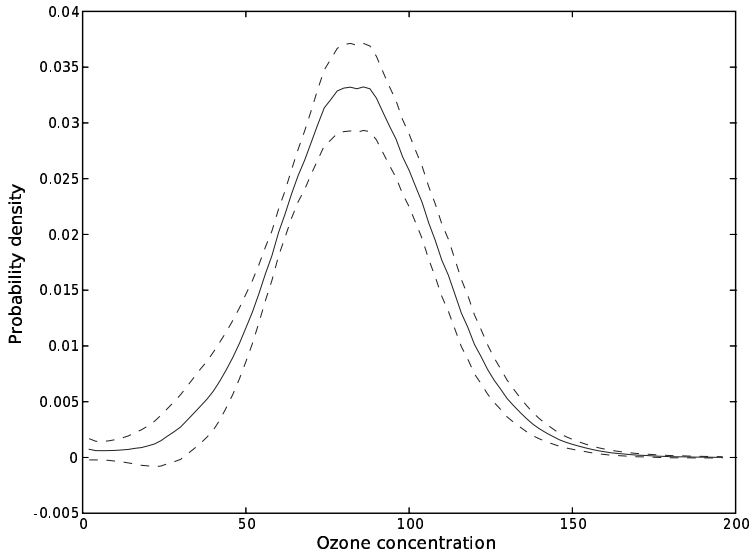


Figure 3.17: Relative frequency distribution $f([O_3])$ of the ozone concentrations ($\mu g \cdot m^{-3}$) of the reference simulation and the functions $f([O_3]) \pm \sigma_f([O_3])$ where $\sigma_f([O_3])$ is the standard deviation of $f([O_3])$ computed from all simulations.

In Figure 3.18, the mean of the daily evolution for ozone over the whole domain and all days is shown for all simulations. This daily evolution is also shown independently for the ensembles generated with changes (1) in the physical parameterizations, (2) in the input data and (3) in the numerical approximations. From these figures, the prominent changes are the turbulence closure and the chemical mechanism. The profile is less sensitive to the input data, but this is an average profile that can hide spatial or temporal variabilities (which are analyzed below). The spread (estimated with the relative standard deviation) is 4% on the peak and 6% for the whole profile. It reaches 9% at 0400 UT, which is high since the impacts are not cumulative (see Section 3.2.4 for the cumulative effects).

One question about the nature of the variability introduced by each change lies in the bias and its nocturnal and diurnal evolutions. The standard deviation of the difference with the reference simulation does not provide this information since it may be low even with systematic biases in the night or in the daytime. In the same way, systematic biases may appear in given regions. In Table 3.9, the amount of negative biases (concentrations less that the reference concentrations) are reported for:

1. the daily biases: the biases (mean over all grid cells of the difference) computed for each day;

2. the daytime biases: the biases computed for each day but only during the daytime (from 0400 UT to 1800 UT);

3. the nocturnal biases: the biases computed for each day but only during the night;

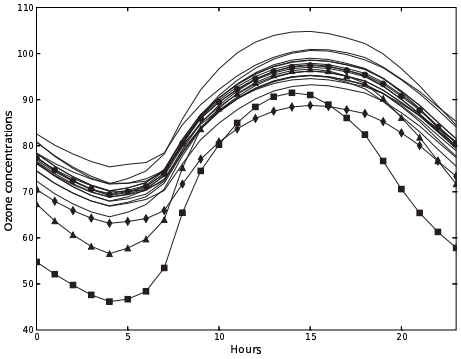
4. the spatial biases: the biases (mean over all time steps of the difference) computed for each grid cell.

It first indicates that the simulations with the largest biases are characterized by a clear trend: they are either above or below the reference simulation at nearly all hours. It also demonstrates that the biases at night and during the daytime can strongly differ. This is true for simulation 8 (emission vertical distribution) due to the fact that the pollutants emitted at night and in the second layer barely influence the ground concentrations. During the daytime, the emissions are mixed in the boundary layer which decreases the impact of the vertical location of the pollutants at release time. In simulation 4 (Wesely's parameterization), the nocturnal concentrations are often below the reference concentrations but only slightly below since the total amount of negative biases is very close to the amount associated with the daytime. The last simulation that shows such differences is simulation 13 (100 s as a time step) for unclear reasons. The active chemical reactions are not the same at night as in the daytime, which may explain why the numerical time step has a different impact.

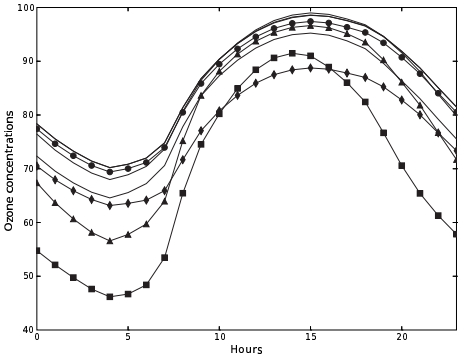
Another point lies in the spatial and temporal differences. The day-by-day bias may hide spatial inhomogeneities of the bias. Simulations 16 and 17 (0.1° and 1.0° horizontal resolution respectively) are good examples. Loosely speaking, choices in the simulation setup may impact the spatial distributions independently from their temporal effects.

Comparisons with Observations

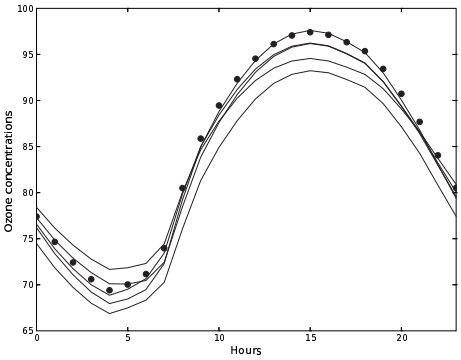
The results of the simulations are compared to ozone peaks which are usually a major concern of forecasting systems. The comparisons are performed with 242 stations over Europe. Each selected station has a reasonable amount of measurements (at least 30 peak measurements during the 126 days of the comparison period). There are 27,000 peak observations and 620,000 hourly observations.



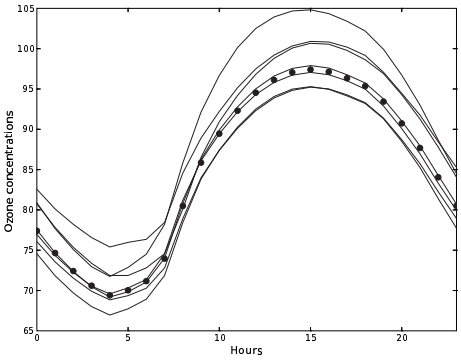
(a) All simulations.



(b) Physical parameterizations.



(c) Input data. The lowest concentrations are reached by simulation 10 (GLCF land use coverage for biogenic emissions).



(d) Numerical approximations. The highest concentrations are reached by simulation 15 (Strang splitting).

Figure 3.18: Ozone daily profile for the twenty simulations and for the three groups related to (b) the physical parameterizations, (c) the input data and (d) the numerical approximations. The dots represent the reference simulation (#0), the squares represent simulation 2 (Louis closure), the triangles represent simulation 3 (Louis closure in stable conditions) and the diamonds represent simulation 1 (RADM 2).

#	Standard deviation ($\mu g \cdot m^{-3}$)	Daily negative bias (%)	Daytime negative bias (%)	Nocturnal negative bias (%)	Spatial negative bias (%)
2	18.60	100	100	100	98
3	10.40	100	100	100	100
16	8.95	100	100	100	79
1	7.55	100	100	100	100
15	6.65	0	0	6	1
8	5.62	70	84	44	64
18	5.10	0	0	0	0
17	5.07	27	25	33	44
10	5.04	100	100	100	100
5	3.69	100	100	100	100
14	3.60	0	13	0	0
11	2.81	88	84	90	76
4	2.69	35	30	60	33
6	2.63	15	17	15	1
12	2.48	81	90	71	91
19	2.34	100	100	100	100
9	2.30	85	85	83	79
7	1.32	0	0	0	0
13	0.98	61	42	95	82

Table 3.9: The amount of negative biases among the daily biases, the daytime biases, the nocturnal biases and the spatial biases. The simulations are sorted by their standard deviation.

Following EPA [1991], we first evaluate the results with the normalized bias (MNBE), the mean normalized gross error (MNGE) and the daily unpaired (in time, paired in space) peak prediction (UPA). A cutoff level of $80\mu g \cdot m^{-3}$ is used and the errors are evaluated as "computed minus observed" (a positive bias represents overestimation). A simulation is assessed through the amount of stations that match the most-restrictive EPA suggested performances: $\pm 5\%$ for the normalized bias, $\pm 30\%$ for the normalized gross error and $\pm 15\%$ for the unpaired peak prediction accuracy. The root mean square (RMS), the correlation and the overall bias of all daily ozone peaks are also reported in Table 3.10.

#	MNBE (%)	MNGE (%)	UPA (%)	RMS (peaks)	Correlation (peaks)	Bias (peaks)
0	43	100	61	23.54	0.71	-4.47
1	7	97	21	29.22	0.65	-14.19
2	13	85	39	24.86	0.73	-8.26
3	32	98	59	23.54	0.72	-5.27
4	52	100	69	22.73	0.73	-2.33
5	24	99	48	24.47	0.71	-7.25
6	45	100	61	24.03	0.69	-3.36
7	48	100	64	23.67	0.70	-3.18
8	35	100	52	25.03	0.69	-6.29
9	36	99	56	24.02	0.71	-5.50
10	21	99	36	26.40	0.67	-9.65
11	43	99	62	23.02	0.73	-4.38
12	29	99	45	24.59	0.70	-6.75
13	41	100	55	23.85	0.71	-4.99
14	57	100	76	22.26	0.74	0.01
15	47	100	64	25.11	0.67	5.39
16	62	100	75	22.83	0.72	-0.39
17	33	99	51	24.12	0.71	-6.60
18	38	98	52	24.48	0.68	-2.70
19	46	100	58	23.72	0.71	-4.67

Table 3.10: Percentages of stations that meet the most-restrictive EPA recommendations on hourly concentrations (cutoff of $80\mu g \cdot m^{-3}$) for the mean bias, the mean gross error and the unpaired (in time, paired in space) peak prediction; the root mean square ($\mu g \cdot m^{-3}$), the correlation and the overall bias ($\mu g \cdot m^{-3}$) for the ozone daily peaks (all stations put together).

The mean normalized gross error is within the EPA limits at almost all stations for all simulations except for the simulation 2 with the Louis closure: for this simulation, the underestimation is indeed too high ($-8.26\mu g \cdot m^{-3}$ on the peaks). Simulations 1 and 10 also have a strong bias on the daily peaks, but they do not underestimate all concentrations above $80\mu g \cdot m^{-3}$ as much as simulation 2. In this case, the MNGE does not distinguish the simulations, even if other indicators give a wide spread. One might consider that the uncertainty due to the parameterizations is below the error that the EPA limit on the MNGE can detect. This is rather speculative since only single changes were introduced in the simulations and, in the next subsection, it is shown that the amount of stations below the MNGE limit can increase if several changes are introduced at the same time.

Meanwhile, there is a high uncertainty in the concentrations above $80\mu g \cdot m^{-3}$ according to the normalized bias statistics. The amount of stations whose bias is acceptable ($\pm 5\%$) ranges

from 7% to 62%. In the opposite way as for the MNGE, the MNBE test may be questionable precisely because of its variability. A conclusion may be that this test is too severe to be relevant: a good percentage of acceptable stations would come mainly from a favorable configuration of the model. The uncertainty in the model is too high to grant a reasonable validity for this test.

The UPA statistics are also well spread but they show a lower variability: their relative standard deviation is 9% against 37% for the MNBE. In comparison, the RMS and the correlations vary slightly with 6% and 3% respectively of relative standard deviation among the simulations. Therefore the correlations do not provide substantial information. Moreover the correlation between RMS and 1 - UPA reaches 90%, which means that it is essentially useless to compute both.

Finally there is a rather high uncertainty in the peak levels: almost $20\mu g \cdot m^{-3}$ of bias between simulations 1 and 15, with mean observed-peaks at $103\mu g \cdot m^{-3}$. Nevertheless the standard deviation of the biases is only $4\mu g \cdot m^{-3}$. This means that an overall bias is not detailed enough to show the uncertainty.

Even if these results can only barely be generalized, defining indicators adequately related to the uncertainty in the models is not an easy task. They are supposed to distinguish simulations and, at the same time, to be robust enough to changes (within the uncertainty range) in the models. None of the previous indicators seem to be balanced enough for this purpose.

Combined Changes

In this section, we try to estimate the impact of combined changes: several parameterizations are changed at the same time. All combinations of the parameterizations and the numerical choices introduced in Table 3.4 cannot be applied because of the computational costs (there would be 184,320 simulations). Hence only four alternatives are kept: the Louis closure, the RADM 2 mechanism, the deposition velocities as computed in Wesely [1989] and the vertically distributed emissions. The first two parameterizations are included due to their strong impact, the third one due its improvements in the results as compared to the observations and the fourth one because of its low variability. Refer to Table 3.11 for the list of the simulations.

#	Emissions	Deposition	Turbulence	Chemistry
a	Ground	Zhang	Troen & Mahrt	RACM
b	Ground	Zhang	Troen & Mahrt	RADM 2
c	Ground	Zhang	Louis	RACM
d	Ground	Zhang	Louis	RADM 2
e	Ground	Wesely	Troen & Mahrt	RACM
f	Ground	Wesely	Troen & Mahrt	RADM 2
g	Ground	Wesely	Louis	RACM
h	Ground	Wesely	Louis	RADM 2
i	Two layers	Zhang	Troen & Mahrt	RACM
j	Two layers	Zhang	Troen & Mahrt	RADM 2
k	Two layers	Zhang	Louis	RACM
l	Two layers	Zhang	Louis	RADM 2
m	Two layers	Wesely	Troen & Mahrt	RACM
n	Two layers	Wesely	Troen & Mahrt	RADM 2
o	Two layers	Wesely	Louis	RACM
p	Two layers	Wesely	Louis	RADM 2

Table 3.11: The 16 simulations set up with the four alternative parameterizations.

Table 3.12 shows the spread on ozone peaks for the new set of simulations. It should be compared to Tables 3.5 and 3.7. The spread is clearly higher in the new set of simulations. As a consequence, the whole uncertainty due to the parameterizations cannot be easily assessed on the basis of single changes in the parameterizations. The results from the previous sections cannot claim more than an estimate of a lower bound on the uncertainty.

All concentrations			*Daily maxima*		
#	Mean	Standard deviation	#	Mean	Standard deviation
g	69.26	35.71	g	97.58	35.03
c	68.94	34.23	c	95.71	33.50
o	70.14	31.94	o	95.32	31.39
k	69.73	30.52	k	93.29	30.17
h	56.17	27.08	e	102.39	23.38
e	85.18	26.19	a	100.92	22.62
d	56.26	26.07	h	80.91	21.75
a	84.92	25.11	m	99.36	20.70
p	58.47	24.49	d	79.80	20.67
l	58.41	23.46	i	97.90	20.10
m	84.49	22.84	p	80.67	19.82
f	77.27	21.96	l	79.29	18.96
i	84.23	21.81	f	92.72	17.12
b	77.11	21.14	b	91.49	16.75
n	77.51	19.30	n	90.85	15.73
j	77.34	18.50	j	89.61	15.50

Table 3.12: Means and standard deviations of the ozone concentrations ($\mu g \cdot m^{-3}$) and their daily peaks for the sixteen simulations. As for the whole concentrations, the relative standard deviation of the means is 15%, and the relative standard deviation of the standard deviations is 20%. As for the daily maxima, the relative standard deviations are 8% and 28% respectively.

The same is true about the error statistics. Even the MNGE limitation ($\pm 30\%$) is not satisfied by more than 90% of the stations for 6 simulations of the 16. Five simulations have a root mean square above $30\mu g \cdot m^{-3}$ whereas none of the previous twenty simulations reaches such a RMS (see Table 3.10). It is obviously due to the combination of changes that individually contribute to decrease the concentrations: the underestimation is then worsened.

The "cumulative underestimations" can be seen in Figure 3.19 (to be compared to Figure 3.18). For instance, the simulation that combines the Louis closure and the chemical mechanism RADM 2 shows low concentrations. The non-linearity even increases this effect. In Figure 3.20 we compare the simulation d (Louis closure and RADM 2) and the linear combination "#c + #b - #a". Both should be equal if the dependences were linear. The concentrations of the simulation d are even lower than the concentrations of the linear combination. This means that the uncertainty is not additive.

The mean spread (relative standard deviation of the ensemble) of the daily-profile concentrations reaches 16% (against 6% with single changes). The highest spread is reached at 0400 UT with 23%. The spread on the peak is 9%. Notice that this spread is a measure of the uncertainty.

Another measure is the relative standard deviation of all concentrations, not of the mean profile. The relative standard deviation is computed for concentrations in $[40\mu g \cdot m^{-3}, 130\mu g \cdot$

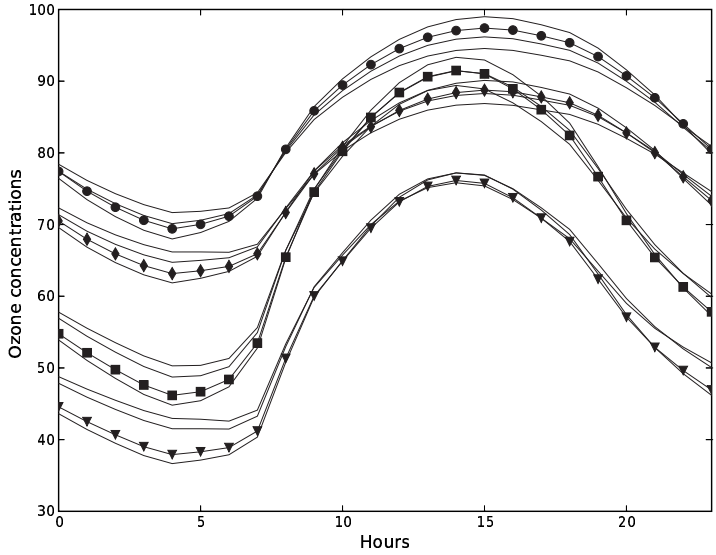


Figure 3.19: Ozone daily profile for the sixteen simulations. The dots represent the reference simulation (#a), the diamonds represent simulation b (RADM 2), the squares represent simulation c (Louis closure) and the triangles represent simulation d that combines the Louis closure and RADM 2.

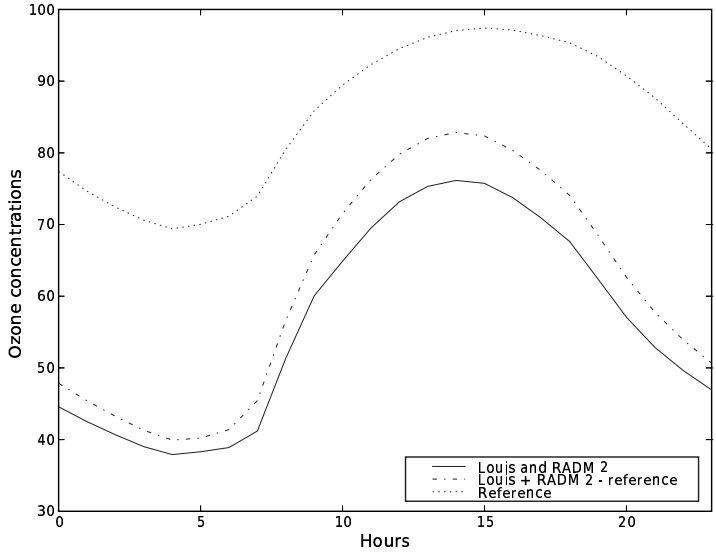


Figure 3.20: Ozone daily profile of the reference simulation a, simulation d (Louis closure and RADM 2) and the linear combination "#c + #b - #a" that adds linearly the effects of simulations b (RADM 2) and c (Louis closure). The non-linearity increases the impact of the parameterizations and the concentrations of simulation d are closer to the reference concentrations than the concentrations of the linear combination. It is especially true for the peak.

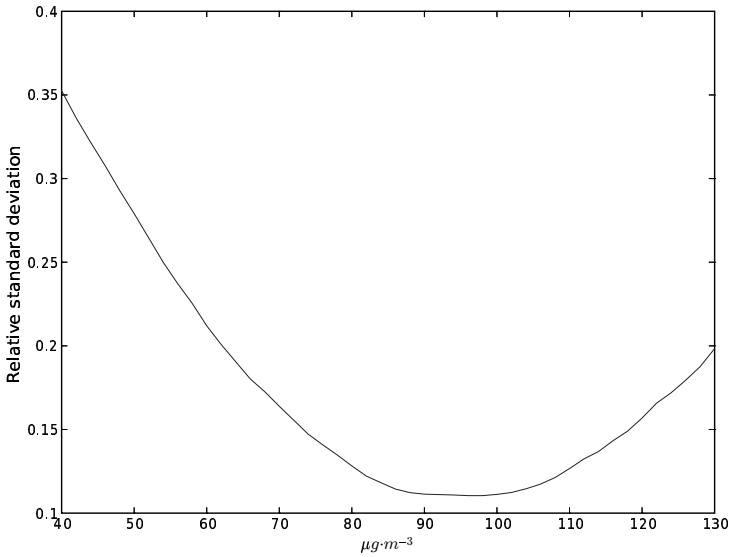


Figure 3.21: Relative standard deviation of the ensemble versus ozone concentrations. The average standard deviation is 17%.

m^{-3}] to include only the main concentrations (refer to the relative frequency distribution shown in Figure 3.17). Figure 3.21 shows the relative standard deviation versus ozone concentrations. The average of the relative standard deviation on this interval is 17%. The lowest concentrations have the largest uncertainty. It means that the processes at night are sensitive to the available parameterizations. The turbulence closure plays an important role at night when the values of the vertical diffusion coefficients are hard to estimate in stable conditions and at the top of the first layer. As for the daily peaks, the relative standard deviation, computed over the whole domain and with all days, reaches 11%. It is difficult to determine the reason why the peaks have a low uncertainty as compared to the other concentrations (Figure 3.21): it may be due to less uncertainty in all parameterizations, or the peaks may be sensitive only to a few leading processes. For instance, the turbulence may be less determinant in well mixed conditions whereas the photochemical activity is strong at the same time.

The ensemble is not equally spread everywhere in the domain as shown in Figure 3.22. The uncertainty measured by the standard deviation (for concentrations in $[40\mu g \cdot m^{-3}, 130\mu g \cdot m^{-3}]$) is high around the coasts and it tends to be high in polluted regions (Southern regions and, due to the emissions, in Great Britain and Poland). In Northern Italy, the Alps are also associated with a high uncertainty. The uncertainty on the peaks has the same spatial distribution. The turbulence closure may notably explain these uncertainties, and the chemical mechanism has probably a strong impact close to the emission locations. However a more detailed study of each process is necessary to properly analyze the spatial inhomogeneities due to each process.

3.2.5 Conclusion

It has been shown that a chemistry-transport model is sensitive to its physical parameterizations, to the associated input data and to the numerical approximations. The turbulent closure and the chemical mechanism introduce the highest uncertainty. The overall uncertainty, measured with the relative standard deviation of an ensemble of sixteen simulations, is estimated at 17% for the common concentration levels and at 11% for the daily peaks. It has been shown

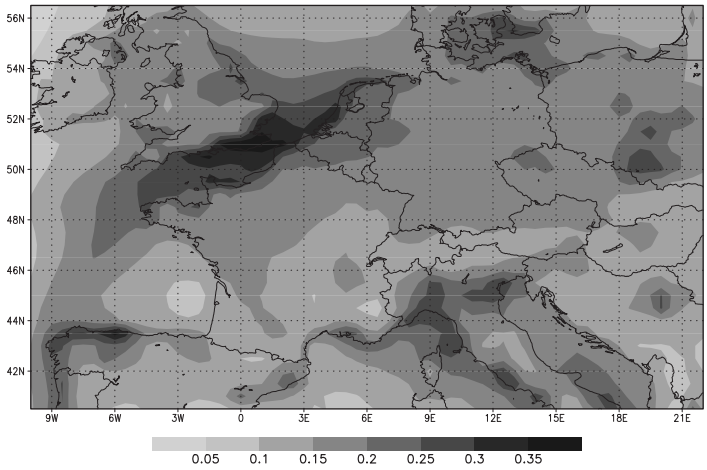


Figure 3.22: Relative standard deviation of the ensemble for concentrations in $[40\mu g{\cdot}m^{-3}, 130\mu g{\cdot}m^{-3}]$. The average standard deviation is 17%. The uncertainty is notably high along the coasts.

that this uncertainty was notably high along the coasts. The uncertainty is too high to let any configuration of the chemistry-transport model fully satisfy the common requirements in comparisons with observations. This low robustness suggests that ensemble approaches are necessary in most applications.

A remaining question is whether these conclusions are limited to the Polyphemus system, even if this system includes commonly used parameterizations. Moreover, this work should be extended to aerosol modeling for which many physical parameterizations and numerical algorithms are also available (hybrid models, nucleation laws, etc.). Another extension deals with the uncertainty due to the input fields to the model such as the meteorological fields or the emissions.

In a next step, one may want to take advantage of the ensemble to provide improved forecasts. The point is to find (actually to forecast) the best combination of the models, as an improvement of the ensemble-mean or ensemble-median approaches.

Statistical measures

Notations Let y be the vector of model outputs and let o be the vector of the corresponding observations. These vectors both have n components. Their means are $\bar{y}$ and $\bar{o}$.

Relative standard deviation

$$\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}{\bar{y}} \tag{3.7}$$

Bias

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(y_i - o_i) \tag{3.8}$$

Root mean square error (RMS)

$$\text{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - o_i)^2} \tag{3.9}$$

Correlation

$$\text{correlation} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (o_i - \bar{o})^2}} \tag{3.10}$$

Mean normalized bias error (MNBE)

$$\text{MNBE} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - o_i}{o_i} \tag{3.11}$$

Mean normalized gross error (MNGE)

$$\text{MNGE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - o_i|}{o_i} \tag{3.12}$$

Unpaired peak prediction accuracy (UPA) For one day:

$$\text{UPA}_{\text{day}} = \frac{y_{\max} - o_{\max}}{o_{\max}} \tag{3.13}$$

The UPA is then averaged over all days.

3.3 Incertitude liée aux données d'entrées

3.3.1 Introduction

Les deux sections précédentes mesurent l'impact des schémas numériques, des approximations numériques, de certaines données d'entrée et de la formulation du modèle. La section précédente recommande l'utilisation de simulations d'ensemble pour la plupart des études tant l'incertitude est élevée. Cette conclusion ne peut qu'être renforcée si l'incertitude due à toutes les données d'entrée est prise en compte. Le but de cette section est de quantifier l'incertitude dont les données d'entrée sont la source.

Des simulations d'ensemble ont déjà été réalisées pour des études de dispersion, avec des variations dans les données météorologiques [Straume, 2001; Warner *et al.*, 2002; Draxler, 2003]. Concernant les simulations avec photochimie, les principaux travaux sont Hanna *et al.* [1998]; Beekmann et Derognat [2003] (efficacité de réductions d'émissions) et Hanna *et al.* [2001]; Hanna et Davis [2002] (estimation de l'incertitude des concentrations simulées). Ces études reposent sur des perturbations sur un grand nombre de champs d'entrée, et sur des simulations Monte Carlo. Les estimations de l'incertitude dans les données d'entrée sont précisément évaluées dans Hanna *et al.* [1998, 2001] et sont adaptées dans l'étude de cette thèse. Cette dernière apporte sa contribution en estimant l'incertitude due aux données d'entrée sur l'Europe et en la comparant à celle due à la formulation du modèle.

Plusieurs données d'entrée du système de simulation sont perturbées selon une incertitude minimale. On estime donc une borne inférieure sur l'incertitude (comme à la section précédente), d'autant plus que tous les champs d'entrée ne sont pas inclus dans l'étude. En particulier, les champs météorologiques ne sont pas perturbés dans l'étude : il semble en effet plus judicieux d'utiliser les prévisions d'ensemble de l'ECMWF.

À la section 3.3.2, les simulations effectuées sont introduites. Les perturbations associées aux données y sont listées. Ensuite, la section 3.3.3 analyse les résultats en deux temps. La convergence des simulations est vérifiée pour plusieurs variables-cibles. Ensuite, l'incertitude est quantifiée et décrite (répartition spatiale, évolution temporelle). Une comparaison à l'incertitude due à la formulation du modèle conclut l'étude.

3.3.2 Simulations Monte Carlo

La configuration de la simulation utilisée est exactement celle présentée à la section 2.5.4 (évaluation du système), sans simplification. Une différence réside simplement dans la période de simulation, du fait du coût très important des simulations Monte Carlo (800 simulations). La simulation se déroule sur onze jours, du 1er juillet au 11 juillet inclus. Sauf mention contraire, les quatre premiers jours sont exclus de sorte à ce que les conditions initiales aient un impact négligeable. De même, trois cellules autour du domaine sont exclues pour atténuer l'impact des conditions aux limites. Sur la semaine restante, la simulation réalise les performances rassemblées dans le tableau 3.13. La comparaison aux observations a une valeur limitée sur une période aussi courte, mais elle montre que le modèle reproduit de manière satisfaisante les phénomènes de la période. L'impact des données d'entrée peut alors être raisonnablement estimé.

Les données d'entrée perturbées sont : les coefficients d'atténuation nuageuse (des constantes photolytiques), les vitesses de dépôt de l'ozone et du dioxyde d'azote, les conditions aux limites d'ozone, les émissions anthropogéniques et biogéniques, et les constantes photolytiques. Outre les données météorologiques, on peut considérer que seules manquent les constantes de réactions (non photolytiques) parmi les données dont l'incertitude a *a priori* un impact fort sur l'ozone. Les conditions aux limites et vitesses de dépôt des autres espèces ont vraisemblablement un impact plus faible ; des études complémentaires devraient être menées pour s'en assurer.

Tab. 3.13 – Statistiques d'erreur de la simulation de référence (sans perturbations) pour les pics journaliers d'ozone.

Indicateur	Réseau EMEP	Réseau Pioneer	Réseau BDQA
Moyenne des observations ($\mu g \cdot m^{-3}$)	107.9	101.0	95.1
Moyenne simulée ($\mu g \cdot m^{-3}$)	109.6	105.5	102.6
RMSE ($\mu g \cdot m^{-3}$)	20.1	21.7	22.8
Corrélation (%)	71.1	72.2	64.2
BF	1.01	1.03	1.06
MNGE[1](%)	14.9	16.0	18.2
Stations réalisant le critère EPA sur MNGE (%)	93.8	92.1	88.8
UPA[1](%)	1.2	2.5	5.7
Stations réalisant le critère EPA sur UPA (%)	75.3	75.4	68.9

[1] Moyenne sur l'ensemble des stations.

Suivant Hanna *et al.* [1998, 2001], chaque champ est perturbé selon une loi lognormale. Toutes les valeurs d'un champ sont perturbées avec un même coefficient, c'est-à-dire que le même coefficient est appliqué dans toutes les mailles et à tous les pas de temps. Pourtant, les incertitudes estimées sur les données d'entrée (voir ci-dessous) ne sont pas fournies pour tout un champ, mais pour une valeur quelconque de ce champ (localisée en temps et en espace). Aucune information sur la corrélation entre les perturbations n'étant connue, on perturbe toutes les valeurs de manière identique (corrélation entre les perturbations égale à 1). Une autre stratégie serait de perturber indépendamment les valeurs de toutes les mailles ou/et à tous les pas de temps. Ceci conduirait à des compensations qui annuleraient en grande partie les effets de l'incertitude. De plus, les champs perdraient leur cohérence spatiale ou/et temporelle.

Pour chaque simulation, tous les champs sont perturbés indépendamment entre eux, sauf mention contraire (voir ci-dessous).

Les amplitudes des incertitudes associées aux données d'entrée sont reportées dans le tableau 3.14. Elles sont inspirées de Hanna *et al.* [1998, 2001]; Schmidt [2002]; Beekmann et Derognat [2003]. Des incertitudes faibles sont choisies car les perturbations s'appliquent à l'ensemble du champ. En effet, l'incertitude « globale » (c'est-à-dire de l'ensemble du champ) diminue à mesure que l'échelle augmente (échelle continentale ici) et que la période de simulation s'allonge (onze jours). Il faut noter que des incertitudes faibles sur les données d'entrée conduisent nécessairement à évaluer une borne inférieure sur l'incertitude.

3.3.3 Analyse de l'incertitude

Afin de donner une indication visuelle de la dispersion des sorties, l'ensemble des profils journaliers moyens d'ozone est reporté à la figure 3.23.

Étude de la convergence

Huit cents simulations sont effectuées. Avant d'analyser les résultats, il convient de s'assurer que la procédure Monte Carlo a convergé. La convergence dépend de la variable analysée ; l'erreur

Tab. 3.14 – Incertitude sur les données d'entrée. L'incertitude d'une variable v est exprimée, de manière approximative, en variation relative $\pm\delta$ sur un intervalle de confiance à 95%. Plus rigoureusement, le percentile à 97.5% est situé à $\overline{v}(1+\delta)$. On en déduit l'intervalle de confiance à 95% de $\ln(v)$ et donc son écart-type. L'écart-type de $\ln(v)$ vaut $\frac{1}{2}\ln(1+\delta)$.

Donnée	Incertitude[a]	Écart-type (LN)[b]
Atténuation nuageuse	±30%	0.131
Vitesses de dépôt (O_3 et NO_2[c])	±30%	0.131
Conditions aux limites (O_3)	±20%	0.091
Émissions anthropogéniques[d]	±50%	0.203
Émissions biogéniques	±100%	0.347
Constantes photolytiques	±30%	0.131

[a] Intervalle de confiance (approximatif) à 95%.
[b] Écart-type sur le logarithme de la variable.
[c] La même perturbation est appliquée à O_3 et NO_2.
[d] La même perturbation est appliquée à NO et NO_2, d'une part ; une autre perturbation est appliquée à tous les composés volatils organiques, d'autre part.

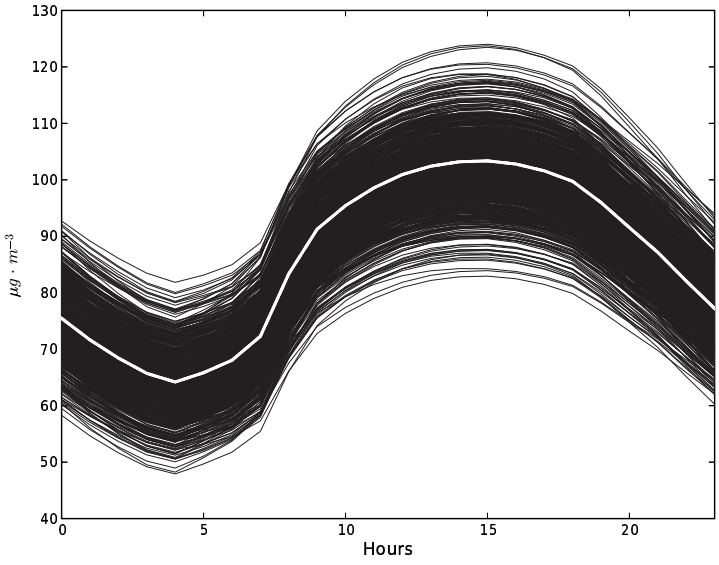


Fig. 3.23 – Profils journaliers moyens d'ozone pour les 800 simulations Monte Carlo. En blanc (et au centre), on représente la simulation de référence (c'est-à-dire sans perturbations).

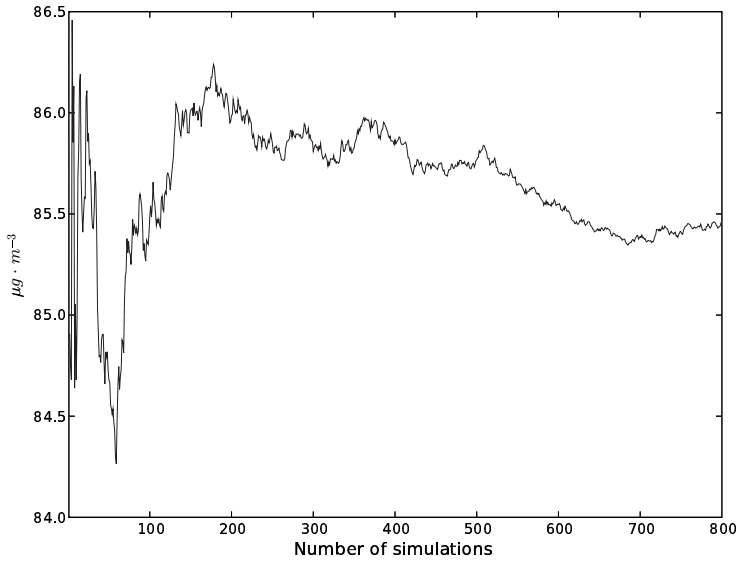


FIG. 3.24 – Espérance (estimée) de la moyenne spatio-temporelle des concentrations d'ozone sur les sept derniers jours simulés. Après quelques fortes oscillations, la moyenne évolue peu. Même si la moyenne s'infléchit nettement à partir de 500 simulations, les variations restent faibles (quelques dixièmes de $\mu g \cdot m^{-3}$) par rapport à la valeur de la moyenne (environ 85.5 $\mu g \cdot m^{-3}$).

est en fait proportionnelle à l'écart-type de la variable de sortie. Plus une variable est moyennée, plus il est vraisemblable que sa variance soit faible. L'espérance (estimée) de la moyenne spatio-temporelle d'ozone doit donc converger rapidement, comme le montre la figure 3.24. De même l'espérance des moyennes spatio-temporelles des concentrations à 4h (minimum) et à 15h (maximum) convergent en moins de 200 simulations (figure 3.25). On ne s'intéresse pas seulement aux moyennes de l'ensemble (espérances) mais surtout à la dispersion, puisque l'objectif est d'estimer l'incertitude. La convergence de l'écart-type des moyennes spatio-temporelles à 4h et à 15h est montrée à la figure 3.26. Il en est de même pour un point précis du domaine (sans moyenne spatiale), voir figure 3.27.

Avec suffisamment de simulations, une densité de probabilité peut aussi être reconstituée. La figure 3.28 montre l'évolution de la densité de probabilité en fonction du nombre de simulations. La densité de probabilité continue à évoluer avec le nombre de modèles. De plus, le nombre de simulations n'est pas assez élevé pour permettre une description fine de la densité, comme l'illustre la figure 3.29.

En conclusion, le nombre de simulations permet d'estimer des moyennes (espérances) et des écarts-types de concentrations de polluants. Les densités de probabilité calculées ne semblent pas fiables.

Analyse de l'incertitude

D'après l'étude de convergence précédente, l'incertitude peut être raisonnablement évaluée avec l'écart-type de l'ensemble. La distribution spatiale et temporelle de l'incertitude, mesurée par la répartition et l'évolution de l'écart-type, est *a priori* inégale puisque la plupart des champs perturbés sont soit localisés spatialement (conditions aux limites), soit très hétérogènes (vitesses de dépôt, émissions), soit très variables (constantes photolytiques).

Pour analyser la répartition spatiale de l'incertitude, la figure 3.30 reporte la moyenne tem-

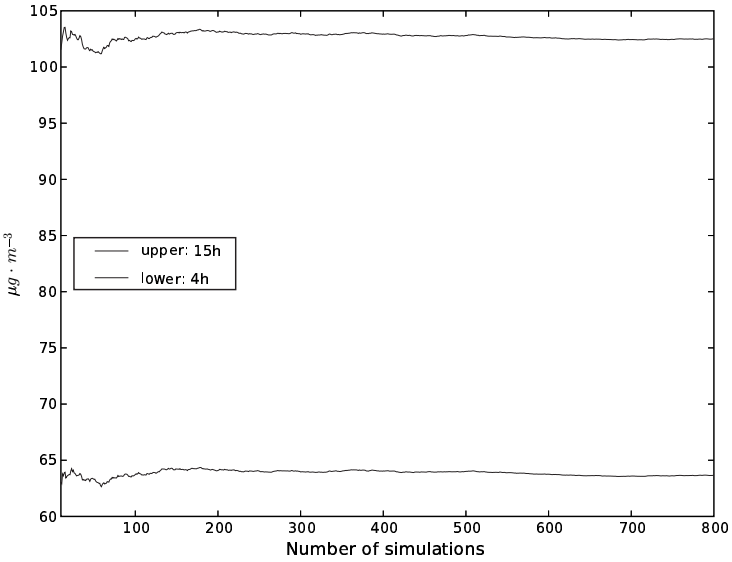


FIG. 3.25 – Espérance (estimée) de la moyenne spatio-temporelle des concentrations d'ozone à 4h et à 15h sur les sept derniers jours simulés. À partir de 150 simulations, on peut considérer que la convergence s'est effectuée.

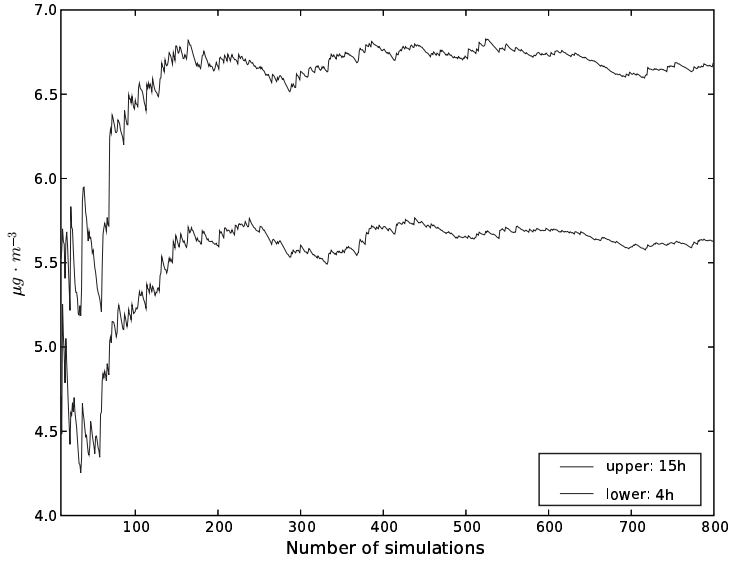
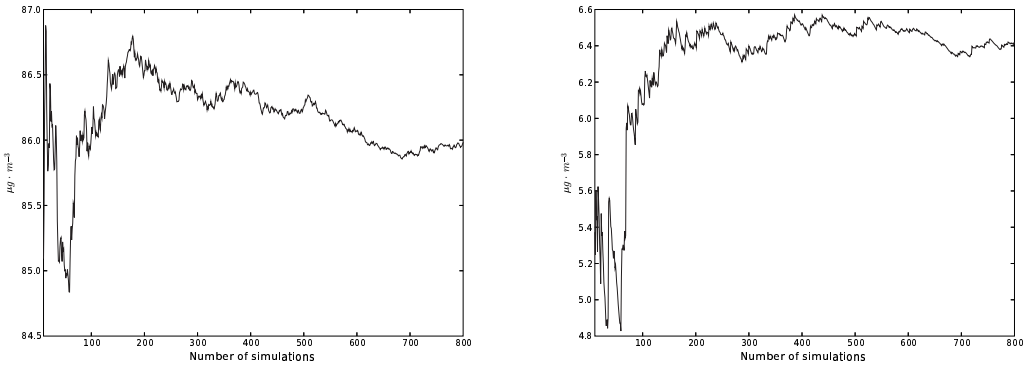


FIG. 3.26 – Écart-type de la moyenne spatio-temporelle des concentrations d'ozone à 4h et à 15h sur les sept derniers jours simulés. À partir de 200 simulations, voire 400 simulations (selon l'exigence), on peut considérer que la convergence s'est effectuée.



(a) Espérance (b) Écart-type

FIG. 3.27 – Espérance et écart-type de la moyenne temporelle des concentrations dans une cellule (au milieu du domaine). On peut raisonnablement conclure à la convergence de la procédure.

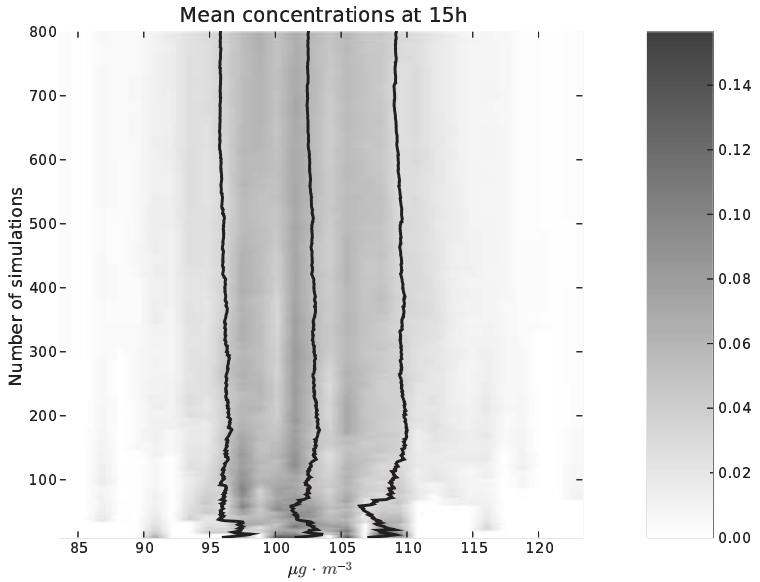
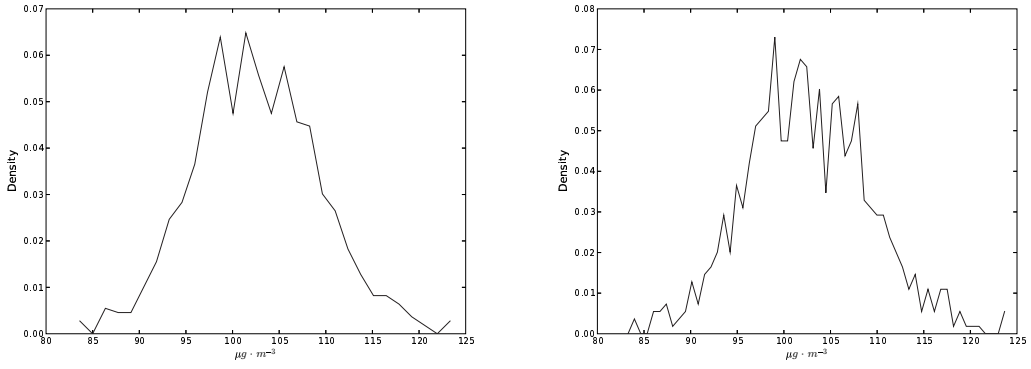


FIG. 3.28 – Évolution de la densité de probabilité en fonction du nombre de modèles. La densité de probabilité est discrétisée avec 30 points. Les lignes superposées représentent la moyenne, et la moyenne à laquelle on ajoute ou on retire l'écart-type.



(a) 30 points de discrétisation

(b) 60 points de discrétisation

FIG. 3.29 – Densité de probabilité de la concentration moyenne (sur tout le domaine et toute la simulation) d'ozone à 15h. La densité est approchée par des fréquences d'apparition dans 30 ou 60 intervalles. Des oscillations apparaissent avec une discrétisation encore grossière.

porelle de la distribution spatiale de l'écart-type relatif (dans chaque cellule). Cet écart-type est moyenné en temps et puis divisé par la concentration moyenne dans la maille. On constate une hétérogénéité forte, avec l'influence probable des conditions aux limites à l'ouest et celle des émissions au-dessus de l'Angleterre et puis à l'est de celle-ci. De manière générale, on constate une augmentation de l'incertitude aux lieux d'émissions fortes. Les émissions y ont en effet leur impact maximal. L'impact restant localisé, on peut penser qu'il s'agit principalement de l'influence du monoxyde d'azote qui réagit rapidement avec l'ozone pour le titrer (voir chapitre 4 pour plus de détails).

On représente à la figure 3.31 l'évolution temporelle de l'incertitude. L'incertitude maximale est atteinte en journée. L'écart-type est corrélé à 85.2% avec les concentrations (moyennes spatiales). Les fortes concentrations d'ozone sont donc associées à une incertitude plus forte. Ceci explique pourquoi l'incertitude varie, de plus, assez nettement d'un jour sur l'autre.

Pour donner une idée des comportements extrêmes, on s'intéresse aux distributions spatio-temporelles de l'ozone. La figure 3.32 montre quatre comportements extrêmes. On remarque la similitude entre les simulations de moyenne élevée (ou faible) et d'écart-type élevé (ou faible). La corrélation entre les moyennes et les écarts-types (de toutes les simulations) vaut près de 58%. Ceci est significatif et confirme que les concentrations élevées sont associées à des incertitudes plus fortes.

Pour conclure cette analyse, il est intéressant de quantifier l'incertitude. La figure 3.33 représente l'incertitude sur le profil journalier moyen d'ozone. L'ecart-type moyen vaut 6.2 $\mu g \cdot m^{-3}$, ce qui est similaire aux l'écarts-types moyens des figures 3.30 et 3.31 (7.5 $\mu g \cdot m^{-3}$). L'incertitude relative moyenne (figure 3.30) vaut 8.7%. Sur le profil journalier moyen, l'incertitude relative, calculée avec la moyenne de l'écart-type divisé par la concentration moyenne à chaque heure, vaut 7.4%.

3.3.4 Conclusion

L'incertitude due aux données d'entrée a été évaluée sur la base de plusieurs données d'entrée et avec des incertitudes minimales. L'étude propose donc une borne inférieure sur l'incertitude, tout comme à la section 3.2. Une incertitude relative d'environ 8% est constatée. En com-

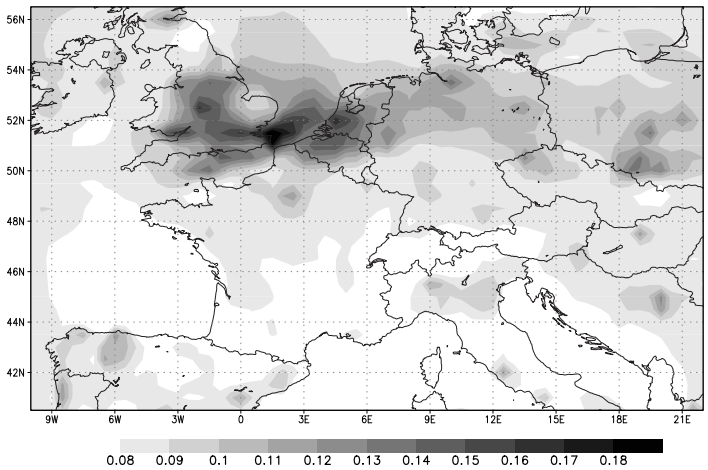


FIG. 3.30 – Répartition spatiale de l'incertitude relative (la moyenne temporelle de l'écart-type divisée par la concentration moyenne dans la cellule). Sur cette figure, la bande d'une largeur de trois cellules (1.5°) n'a pas été retirée autour du domaine.

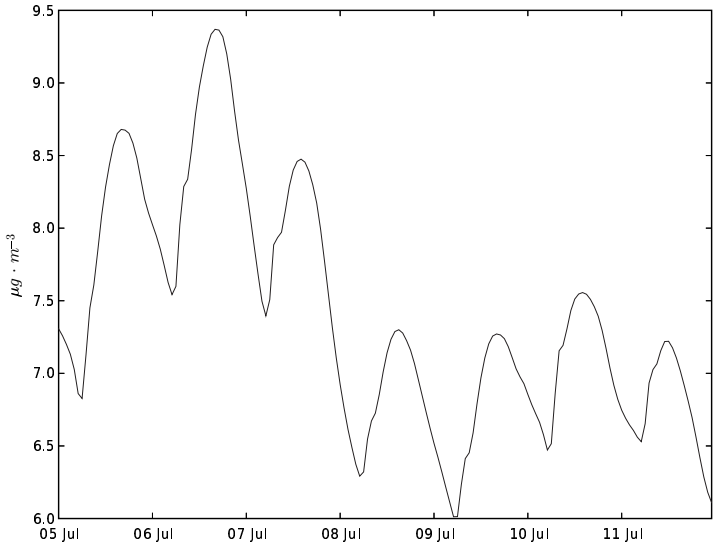


FIG. 3.31 – Évolution temporelle de l'écart-type (moyenne spatiale en $\mu g \cdot m^{-3}$).

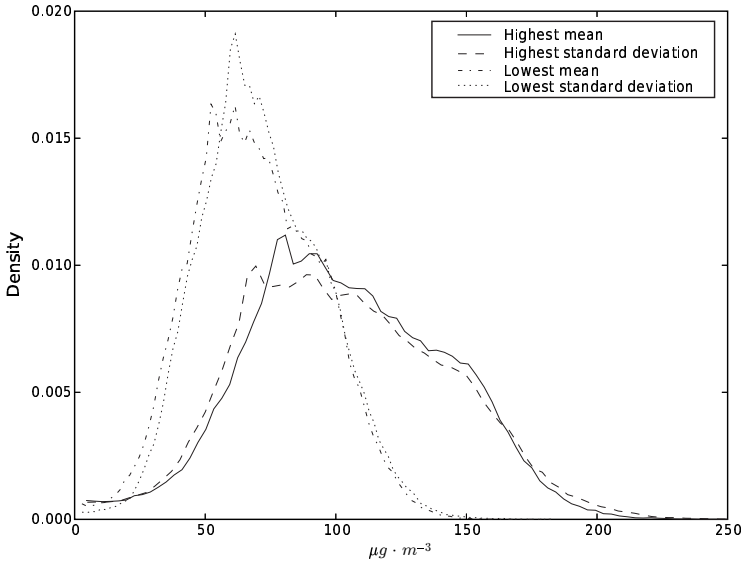


FIG. 3.32 – Distribution (spatio-temporelle) des concentrations d'ozone pour des simulations extrêmes, dont les caractéristiques sont d'avoir la moyenne la plus grande (103.4 $\mu g \cdot m^{-3}$), l'écart-type le plus grand (40.3 $\mu g \cdot m^{-3}$), la moyenne la plus petite (69.2 $\mu g \cdot m^{-3}$) et l'écart-type le plus petit (23.6 $\mu g \cdot m^{-3}$).

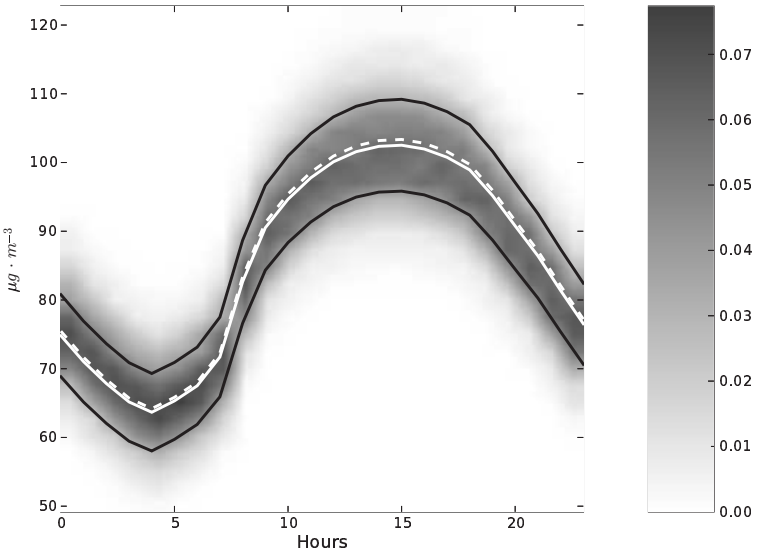


FIG. 3.33 – Profil journalier moyen d'ozone. La densité de probabilité (dégradés) est représentée avec l'espérance (courbe continue en blanc), l'espérance à laquelle l'écart-type est soustrait ou ajouté (courbes en noir) et le profil de la simulation de référence (courbe discontinue en blanc).

paraison, l'incertitude due aux paramétrisations physiques et aux approximations numériques (section 3.2) est environ le double. Il faut noter que les deux estimations ne sont pas faites sur la même période. Si on restreint les simulations de la section 3.2 à la période considérée ici, la différence se confirme, avec une incertitude relative de 16.4% sur le profil journalier pour les paramétrisations (changements multiples) et de 7.4% pour les données d'entrée.

Cependant l'incertitude sur les données d'entrée (considérées ici) n'est pas négligeable. Il s'agit d'une nouvelle source d'incertitude qui renforce la conclusion de la section précédente : des simulations d'ensemble doivent être réalisées aussi souvent que possible tant l'incertitude est importante.

Une étude complémentaire serait l'estimation de l'incertitude due à chaque variable d'entrée. L'impact d'autres sources d'incertitude devrait aussi être évalué, principalement pour les champs météorologiques (via les prévisions d'ensemble de l'ECMWF) et les constantes des réactions chimiques. Enfin, une étude de l'incertitude totale devrait être menée. Il s'agirait se croiser l'incertitude due à la formulation du modèle et celle due aux données d'entrée.

Chapitre 4

Émissions européennes : étude de sensibilité

On considère souvent les émissions comme un élément essentiel des simulations photochimiques. L'étude de sensibilité de ce chapitre permet d'analyser leur impact (structure spatiale, dépendance aux espèces, etc.). Les émissions étant des données très incertaines, elles constituent un élément limitant des simulations de la qualité de l'air. Afin de réduire l'incertitude que les émissions induisent, une piste consiste à affiner leur connaissance par assimilation des données d'observation (modélisation inverse des émissions). Or, le nombre d'observations n'est pas suffisant pour inverser tous les paramètres d'émission. En conséquence, il faut sélectionner les paramètres les plus sensibles. Cette étude prépare donc à la modélisation inverse des émissions à l'échelle européenne en identifiant les paramètres à inverser préférentiellement.

Sommaire

Le chapitre est constitué de
Mallet, V. et Sportisse, B. (2005a). A comprehensive study of ozone sensitivity with respect to emissions over Europe with a chemistry-transport model. *J. Geophys. Res.*, 110(D22)

4.1 Sensibilité des concentrations d'ozone aux émissions

4.1.1 Introduction

Emissions are a key input in air quality models and have therefore motivated many research efforts from the generation of emission data to their evaluation and improvement. In addition emission reductions are undertaken in order to satisfy the requirements of new laws and regulations in air pollution control. In this context the sensitivity of photochemical pollutants to their emitted precursors is of high interest. There are at least three motivations.

First the estimation of the sensitivities to emissions improves the understanding of the chemistry-transport models. It shows the prominent sensitivities and assesses the relative impact of emissions as compared to the known impacts of other processes. Second the sensitivities coupled with the uncertainty in the emissions provide an estimate of the uncertainty in the output concentrations due to the emissions. One may assess the reliability of a model and decide which part of the emissions should be improved as a priority. Third the sensitivity selects the emissions that could be optimized through inverse modeling [e.g., Chang *et al.*, 1997; Mendoza-Dominguez et Russell, 2001; Elbern et Schmidt, 2002; Quélo, 2004]. The aim is then to perform inverse modeling of the emissions to which the measured concentrations are sensitive enough to allow a valuable inversion. The conclusions of this paper are mainly related to the third option.

Inverse modeling of emissions should be performed on the most sensitive emission parameters. Otherwise the inversion would not be able to improve the model outputs (compared to measurements) or it would lead to unrealistic updates in the emissions. For instance the impact of the temporal distribution of emissions is weak [Tao *et al.*, 2004]. This study addresses in details the question of the prominent impacts. It notably ranks the emitted species, the emitting locations and the release time. It estimates the influence scope in space and time of the emissions. It also estimates the a priori quality of the observational network to perform inverse modeling.

There are several methods to estimate the sensitivities and they may be applied to different cases. In Jiang *et al.* [1997], the sensitivities are estimated along a single day and a single trajectory, and with finite differences. Pryor [1998] analyzes the impact on ozone concentrations of the emission changes over eight years. Bastrup-Birk *et al.* [1997] study, over seven years, the sensitivity of ozone exposures to changes in emission scenarios. Using an adjoint model, Menut [2003] addresses, among other sensitivities, the sensitivity to emissions at regional scale, and Schmidt et Martin [2003] deal with European emissions and focus on their impact over Paris area.

This paper proposes a comprehensive study of the sensitivity to European emissions during summer 2001, using differentiated versions of a chemistry-transport model and a basic Monte Carlo method. The chemistry-transport model is differentiated into (1) a tangent linear model that produces the first-order derivatives of all outputs to a given model-input, and (2) an adjoint model that delivers the derivatives of a given model-output to all inputs (emissions, in this study). This way, the sensitivity is estimated from first-order derivatives. It is thus "local" and restricted to the given emission inventory. To obtain a more global picture of the sensitivity, Monte Carlo simulation are performed with perturbations in the emissions. The emissions are associated with log-normal probability density functions as advocated in Hanna *et al.* [1998, 2001]. Three sets of 200 simulations are generated to estimate the uncertainty due to NOx emissions, VOC emissions and biogenic emissions.

The paper is organized as follows. Section 4.1.2 describes the underlying modeling system, its physical components and its chemistry-transport model. The context is also detailed through the simulation of photochemistry, over Europe and during summer 2001, on which the sensitivity study is based. In Section 4.1.3, we expose the sensitivities that are estimated and

the techniques to compute them. The following sections report the results obtained with the tangent linear model (Section 4.1.4), the adjoint model (Section 4.1.5) and the Monte Carlo simulations (Section 4.1.6).

4.1.2 Modeling System

Description

The modeling system is Polyphemus (Mallet *et al.* [2005], available at `http://www.enpc.fr/cerea/polyphemus/`), version 0.2, notably based on the library for atmospheric chemistry and physics AtmoData [Mallet et Sportisse, 2005b] and the Eulerian chemistry-transport model Polair3D [Boutahar *et al.*, 2004].

Many configurations are available in Polyphemus. We have selected the most detailed physical parameterizations. The configuration is not chosen in order to provide the best forecasts but to use a reliable physics. This way the sensitivities will not be affected by artificial adjustments in the model.

The simulation is performed with the following physical parameterizations and input data:

1. meteorological data: the most accurate ECMWF data available for the period (i.e. $0.36° \times 0.36°$, the TL511 spectral resolution in the horizontal, 60 levels, time step of 3 hours, 12 hours forecast-cycles starting from analyzed fields);

2. land use coverage: USGS[1] finest land cover map (24 categories, 1km Lambert)

3. deposition velocities: the revised parameterization proposed in Zhang *et al.* [2003b];

4. vertical diffusion: within the boundary layer, the Troen's and Mahrt's parameterization as described in Troen et Mahrt [1986], with the boundary-layer height provided by the ECMWF; above the boundary layer, the Louis' parameterization [Louis, 1979].

5. boundary conditions: daily means extracted from outputs of the global chemistry-transport model Mozart 2 [Horowitz *et al.*, 2003] run over a typical year.

Since the study deals with emissions, we provide more details about their generation. Anthropogenic emissions are generated with the EMEP[2] expert inventory for 2001. The spatial distribution comes along with the inventory. A typical time distribution of emissions, given for each month, day and hour [GENEMIS, 1994], is applied to each emission sector (called SNAP categories, i.e. sectors from the Selected Nomenclature for Air Pollution). The monthly coefficients also depend on the country, and the time zone of each country is taken into account in the hourly coefficients. As for the chemical distribution, the inventory species are disaggregated into real species using speciation coefficients provided in Passant [2002]. NOx emissions are split into 90% of NO (in mass), 9.2% of NO_2 and 0.8% of HONO. The aggregation into model species (for RACM) is done following Middleton *et al.* [1990].

Biogenic emissions are computed as advocated in Simpson *et al.* [1999]. Isoprene emissions are affected to the model species ISO (isoprene in RACM) and all emissions of terpenes are affected to API (α-pinene and other cyclic terpenes with one double bond in RACM).

Both anthropogenic and biogenic emissions are hourly emissions; the sensitivities are therefore computed with respect to hourly emissions.

As for numerical issues, the advection-diffusion-reaction equation is solved using:

[1] U.S. Geological Survey.

[2] Co-operative Programme for Monitoring and Evaluation of the Long-range Transmission of Air Pollutants in Europe.

1. a first-order operator splitting, the sequence being advection–diffusion–chemistry;

2. a direct space-time third-order advection scheme with a Koren flux-limiter advocated in Verwer *et al.* [1998];

3. a second-order order Rosenbrock method (suited for stiff problems) for diffusion and chemistry.

Test Case

The simulation takes place over Europe in the summer 2001. The domain is $[40.25°N, 10.25°W] \times [56.75°N, 22.25°E]$ (see Figure 4.1), with $0.5° \times 0.5°$ cells, namely 33 cells along latitude and 65 cells along longitude. There are five cells along z whose centers are 25 m, 325 m, 900 m, 1600 m and 2500 m. The top height of the last cell is 3000 m, which is high enough to enclose the planetary boundary layer in most cases. The time step is 600 s. The chemical mechanism is RACM (with 72 species and 237 reactions – see Stockwell *et al.* [1997]).

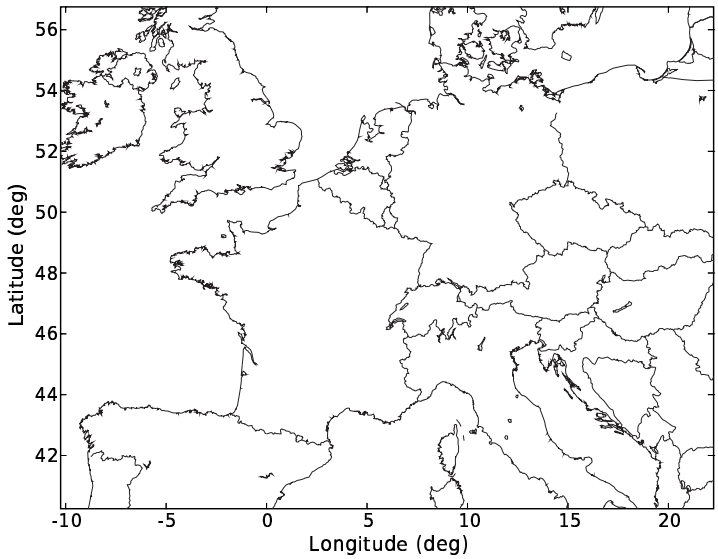


Figure 4.1: Domain $[40.25°N, 10.25°W] \times [56.75°N, 22.25°E]$ of the reference simulation.

The simulation is evaluated with comparisons to measurements from two networks. The EMEP network for 2001 includes 151 stations that provide hourly measurements. The second set of stations provides up to $622,000$ hourly measurements of ozone concentration from the 242 urban, periurban and rural stations over Europe that were used in the Pioneer experiment (`http://euler.lmd.polytechnique.fr/pioneer/`). Table 4.1 shows statistics about comparisons against hourly measurements and concentrations at 1500 UT (always close to the daily maximum for ozone).

Scatter plots (for the two networks) of concentrations at 1500 UT are shown in Figure 4.2 and Figure 4.3. Figures 4.4 and 4.5 show concentrations at 1500 UT at Montgeron (France) and at a station in the Netherlands. The root mean squares at these stations are 23.1 $\mu g \cdot m^{-3}$ and 23.3 $\mu g \cdot m^{-3}$ respectively, which is representative of the overall statistics for the second network.

	EMEP network	Second network
	Hourly concentrations	
RMS	26.0 $\mu g \cdot m^{-3}$	28.7 $\mu g \cdot m^{-3}$
Correlation	57%	66%
Bias	6.7 $\mu g \cdot m^{-3}$	12.6 $\mu g \cdot m^{-3}$
	Concentrations at 1500 UT	
RMS	21.7 $\mu g \cdot m^{-3}$	23.4 $\mu g \cdot m^{-3}$
Correlation	61%	68%
Bias	-3.7 $\mu g \cdot m^{-3}$	2.5 $\mu g \cdot m^{-3}$

Table 4.1: Statistics of the simulation for ozone concentrations over four months.

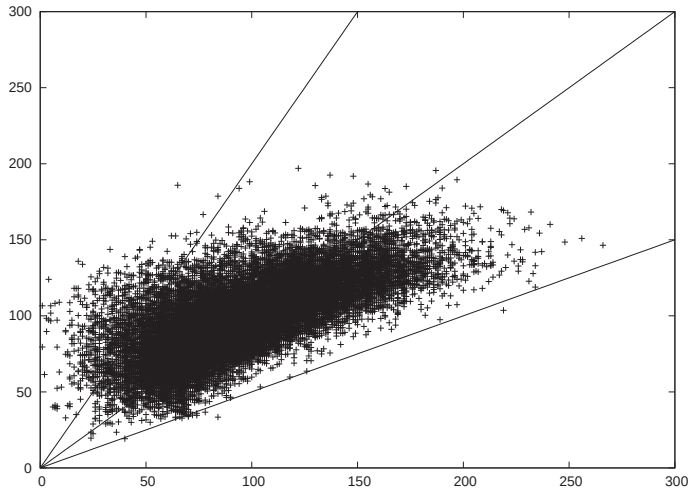


Figure 4.2: Scatter plot of simulated concentrations ($\mu g \cdot m^{-3}$) at 1500 UT versus measurements of the second network. Obviously, the model underestimates the highest concentrations. However the scatter plot confirms the satisfactory results summarized in Table 4.1.

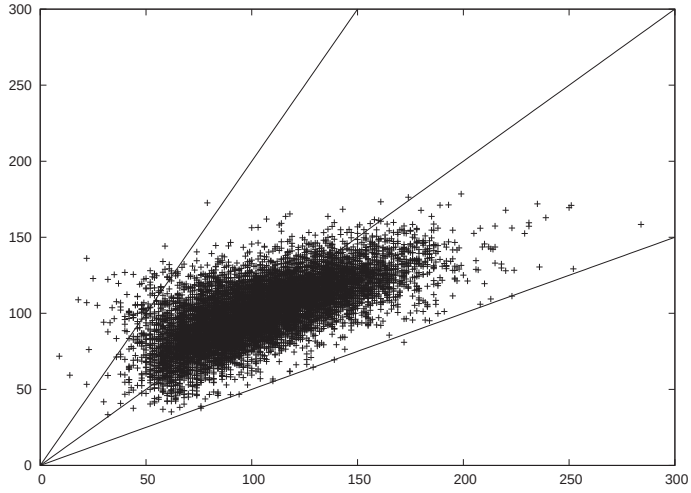


Figure 4.3: Scatter plot of simulated concentrations ($\mu g \cdot m^{-3}$) at 1500 UT versus measurements of the EMEP network. The same analysis as for the other network (Figure 4.2) holds.

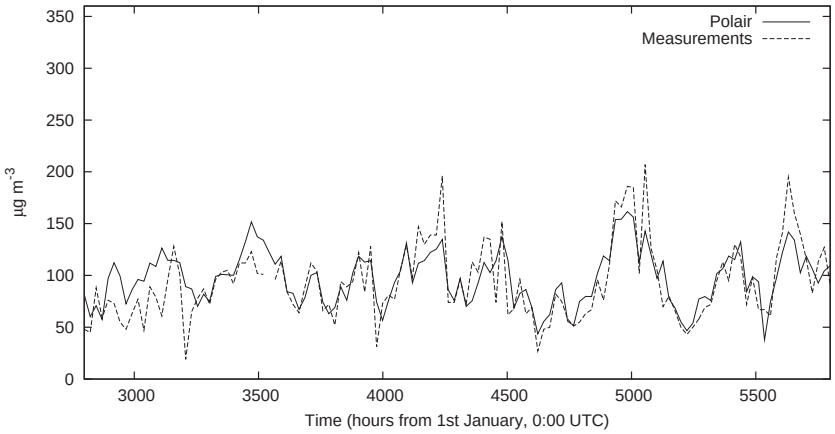


Figure 4.4: Concentrations at 1500 UT at Montgeron (France) from 27 April 2001 to 31 August 2001. The root mean square at this station is 23.1 $\mu g \cdot m^{-3}$.

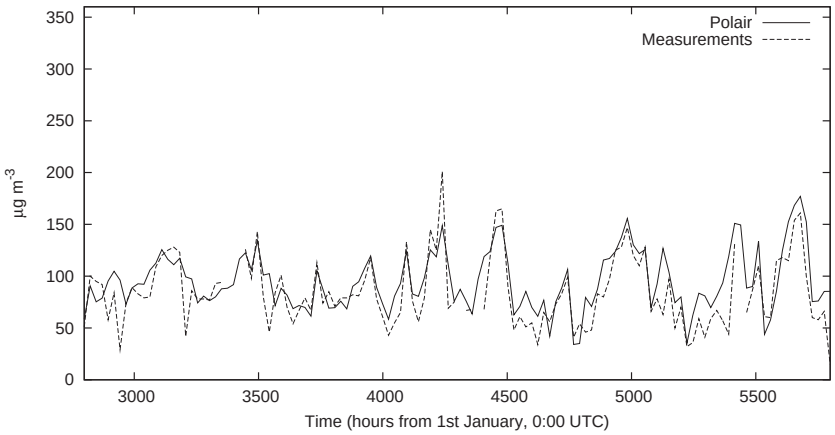


Figure 4.5: Concentrations at 1500 UT at a station in the Netherlands from 27 April 2001 to 31 August 2001. The root mean square at this station is 23.3 $\mu g \cdot m^{-3}$.

According to these comparisons, the system gives satisfactory results in the chosen configuration [Hass *et al.*, 1997; Schmidt *et al.*, 2001] and allows us to perform reliable sensitivity analyses.

4.1.3 Methodology

Sensitivities Selection

Among the sensitivities that may be computed, we study the sensitivities of ozone concentrations because:

1. ozone is always an important concern in air quality modeling;
2. the modeling system performs well for ozone which is a long-range pollutant and for which the physical processes are well detailed;
3. the amount of measurements for ozone is much higher than for other species, which means that inverse modeling of emissions will mainly rely on assimilation of ozone measurements.

When it is necessary to further reduce the output parameters, we select ozone concentrations at network stations, to prepare inverse modeling of emissions. The focus is sometimes put on ozone peaks due to their usual importance.

The sensitivities may also be computed with respect to selected emissions. This limitation comes from computational costs. It is also justified because there are strongly emitting locations (cities) and the impact of changes in the emissions is assumed to be mainly due to these emissions. In the context of inverse modeling, it is natural to focus on the major emission sources due to their prominent impact.

In addition the output sensitivities may be aggregated. For instance, the sensitivities may be aggregated per emitted species so as to rank these species.

Estimated Sensitivities

We mainly compute relative sensitivities, as defined below. Let e be a scalar input (emission) of the model f and c be a scalar output (ozone concentration):

$$c = f(e) \tag{4.1}$$

Since e is uncertain, we assume that it follows a Gaussian law: $e \sim \mathcal{N}(e_0, \sigma^2(e_0))$ where e_0 is the mean value of e and $\sigma(e_0)$ its standard deviation. We define the relative standard deviation as $\sigma^r(e_0) = \frac{\sigma(e_0)}{e_0}$.

We define the absolute sensitivity as:

$$s^a(e) = \frac{\partial f}{\partial e}(e) \tag{4.2}$$

and the relative sensitivity as:

$$s^r(e) = e\frac{\partial f}{\partial e}(e) \tag{4.3}$$

We estimate the sensitivity of the output c with respect to e using the linearized form of equation (4.1). Let:

$$\delta e := e - e_0 \sim \mathcal{N}\left(0, \sigma^2(e_0)\right) \tag{4.4}$$

From equation (4.1), $\delta c = \frac{\partial f}{\partial e}\delta e$ and

$$\delta c \sim \mathcal{N}\left(0, \left(\frac{\partial f}{\partial e}(e_0)\sigma(e_0)\right)^2\right) \tag{4.5}$$

This leads to:

$$c \sim \mathcal{N}\left(f(e_0), (s^r(e_0)\sigma^r(e_0))^2\right) \tag{4.6}$$

Assume that the most probable values for e are in $[e_0 - \sigma_0, e_0 + \sigma_0]$, then the most probable values of c are in $[f(e_0) - s^r\sigma_0^r, f(e_0) + s^r\sigma_0^r]$. Knowing that σ_0^r is usually given [Hanna *et al.*, 1998, 2001], the values that c may reach are determined by the relative sensitivity s^r. For instance, if the emission uncertainty is assumed to be equal to 30%, the output concentration c may be corrected by about $\pm 0.3 s^r$.

Moreover, according to Hanna *et al.* [1998, 2001], we can assume that all emissions have a similar relative standard deviation, with the exception of biogenic emissions. Therefore we directly compare relative sensitivities to identify the most sensitive emissions or emission parameters.

The extension to the vectorial case is straightforward and also shows that the relative sensitivity is a suitable criterion.

Evaluation Techniques

We use three techniques to estimate the sensitivities:

1. the tangent linear model (Section 4.1.4): it is the differentiated version of the model and it returns the derivatives of all output concentrations with respect to a given emission (that is, for a given species, a given location and release time). The tangent linear model is well suited to evaluate the temporal and spatial impact of the major emitting locations.

2. the adjoint model (Section 4.1.5): it provides the derivatives of a given output concentration with respect to all emissions. The adjoint model provides useful information with respect to the spatial extent of the sensitivity and the impact of all emissions at a given location.

3. Monte Carlo simulations (Section 4.1.6): the emissions are perturbed according to a log-normal law (as suggested in Hanna *et al.* [1998, 2001]) to estimate the probability density function of the output concentrations. This technique provides "global" sensitivities, not derivatives.

From the technical point of view, Monte Carlo simulations are handled by a specific module of Polyphemus. The tangent linear model and the adjoint model are obtained through automatic differentiation of Polair3D [Mallet et Sportisse, 2004].

4.1.4 Sensitivity Analysis with the Tangent Linear Model

Experiment Setup

Two periods are analyzed: the first one is 16–17 July 2001 and the second one is 24–25 August 2001. These periods were chosen due to good performances of the model. Moreover the second period is characterized by high ozone concentrations, which is a key situation to be investigated.

The sensitivity of ozone concentration $[\mathrm{O_3}]_{h_c,i_c,j_c,k_c}$ with respect to the emission E_{h_e,i_e,j_e} is:

$$\frac{\partial [\mathrm{O_3}]_{h_c,i_c,j_c,k_c}}{\partial E_{h_e,i_e,j_e}} \tag{4.7}$$

Variable	Possible values	Selected values	Comments
		Input emissions	
i_e	$[\![0, 64]\!]$	–	Major sources
j_e	$[\![0, 32]\!]$	–	Major sources
h_e	$[\![1, \infty[\![$	$[\![1, 24]\!]$	Hourly emissions
		Output concentrations	
i_c	$[\![0, 64]\!]$	$[\![0, 64]\!]$	All cells
j_c	$[\![0, 32]\!]$	$[\![0, 32]\!]$	All cells
k_c	$[\![0, 4]\!]$	$[\![0, 1]\!]$	First two levels
h_c	$[\![0, \infty[\![$	$[\![h_e, h_e + 24]\!]$	Hours from 16 July 2001 0000 UT or from 24 August 2001 0000 UT

Table 4.2: Description of the sensitivities of $[O_3]_{h_c,i_c,j_c,k_c}$ with respect to E_{h_e,i_e,j_e} that are analyzed.

where the pollutant is emitted at the time step h_e, in the cell (i_e, j_e) and the concentration is taken at the time step h_c in the cell (i_c, j_c, z_c).

Table 4.2 shows the values for h_c, i_c, j_c, z_c, h_e, i_e and j_e in this study with the tangent linear model.

For each species, a major emission source (i_e, j_e) is a source whose daily maximum flux is greater than or equal to the half of daily maximum flux (in the whole domain) for the species. The main sources were chosen because they should be associated with the highest relative sensitivities: if the absolute sensitivity slightly varies over the domain, the relative sensitivity will be much higher to the major sources. Moreover if emissions are selected for an inverse modeling experience over Europe, the emissions of the main cities will naturally be included. The number of cells considered as major emission sources is reported in Table 4.3 for each species.

h_e is the emission time step. $h_e = 0$ at 0000 UT on 16 July 2001 (or 24 August 2001), and $h_e = 1$ at 0100 UT on 16 July 2001 (or 24 August 2001) since we deal with hourly emissions.

With the tangent linear model, the sensitivities of all output concentrations are available. One only selects the input emissions with respect to which the sensitivities are computed. However, because of storage constraints, there is still a selection in the output sensitivities. $i_c \in [\![0, 64]\!]$, $j_c \in [\![0, 32]\!]$ and $k_c \in [\![0, 1]\!]$ means that sensitivities in all cells of the first two levels are selected.

h_c represents hours from 16 July 2001 (or 24 August 2001) 0000 UT. For a given h_e, the sensitivity is non-zero only if $h_c \geq h_e$. It cannot be zero for $h_c = h_e$ because, as Polair3D solves the chemistry-transport equation between $h_e - 1$ and h_e, emissions are interpolated (linearly) between $h_e - 1$ and h_e. So the concentrations at time step h_e are sensitive to emissions at time step h_e. The sensitivity is therefore returned from h_e and then for one day (i.e. until $h_e + 24$).

The number of major sources (sum over all species) is 205. Since there are 24 emission steps ($h_e \in [\![1, 24]\!]$), 4920 (one-day) simulations are performed.

Species	Number of cells	Species	Number of cells
ALD	7	ISO	36
API	10	KET	10
CO	7	NO	10
CSL	10	NO2	10
ETE	7	OLI	8
ETH	3	OLT	7
HC3	10	ORA2	6
HC5	9	SO2	11
HC8	12	TOL	10
HCHO	2	XYL	10
HONO	10		

Table 4.3: Number of sources per species with respect to which sensitivities are computed with the tangent linear model. The species are RACM emitted species and are defined precisely in Stockwell *et al.* [1997]. Note that isoprene (ISO) biogenic emissions are diffuse, which leads to a high number of cells.

This sensitivity study aims at determining to which parameters the model is sensitive. We want to rank the most important species, emission time steps, etc. To achieve this goal, maxima and norms of sensitivities are computed.

The analysis is mainly based on aggregated sensitivities. An aggregated sensitivity is a sum of sensitivities with a fixed input or output index and is computed through norms of vectors. For instance, let $S_{h_c}(h)$ be the vector of the sensitivities at the simulation time-step h ($h_c = h$), indexed by $(h_e, i_e, j_e, k_c, E, i_c, j_c)$:

$$S_{h_c}(h) = \left(\frac{1}{N_{\text{cells}}(E)} \frac{\partial \left[\text{O}_3\right]_{h_c=h,i_c,j_c,k_c}}{\partial E_{h_e,i_e,j_e}} \right)_{h_e,i_e,j_e,k_c,E,i_c,j_c} \tag{4.8}$$

where $N_{\text{cells}}(E)$ is the number of emission sources for species E with respect to which a sensitivity has been computed. $\frac{1}{N_{\text{cells}}(E)}$ is therefore a normalization factor needed to affect the same weight to all emitted species. From this vector, we compute two main indicators. $\max S_{h_c}(h)$ is useful to indicate whether emissions could have a strong local impact on concentrations at a given time step h. $\|S_{h_c}(h)\|_1$ denotes the norm one of the vector $S_{h_c}(h)$ and measures the global impact of emissions at the time step h.

These two indicators may be derived for all indices: S_{h_c}, S_{i_c}, S_{j_c}, S_{k_c}, S_E, S_{i_e}, S_{j_e} and S_{h_e}. Actually, i_c and j_c are put together: $S_{(i_c,j_c)}$; in the same way, we define $S_{(i_e,j_e)}$ and $S_{\Delta h=h_c-h_e}$.

Note that we have not provided the units (of the sensitivities) in the following results: the point is to compare the sensitivities. The impact on the concentrations is estimated in later sections (4.1.5 and 4.1.6).

Results and Discussion

Sensitivity to Chemical Species: S_E Tables 4.4 and 4.5 show the sensitivities (maxima and norm one respectively) with respect to all emitted species.

NO is clearly associated with the highest sensitivities as compared to the other species. The norm one identifies NO as the most important emitted species. The sensitivities with respect to NO_2 are about 30 times lower. There is a constant ratio between NO emissions and NO_2

16 – 17 July		24 – 25 August	
Species	$\max S_E$	Species	$\max S_E$
NO	-6.2	NO	-7.9
ISO	$9.9{\cdot}10^{-1}$	ISO	$7.8{\cdot}10^{-1}$
HCHO	$4.0{\cdot}10^{-1}$	API	$3.9{\cdot}10^{-1}$
API	$3.0{\cdot}10^{-1}$	HCHO	$3.6{\cdot}10^{-1}$
NO2	$2.8{\cdot}10^{-1}$	NO2	$2.4{\cdot}10^{-1}$
OLI	$9.5{\cdot}10^{-2}$	OLI	$1.4{\cdot}10^{-1}$
OLT	$8.6{\cdot}10^{-2}$	HONO	$9.2{\cdot}10^{-2}$
XYL	$7.9{\cdot}10^{-2}$	OLT	$6.9{\cdot}10^{-2}$
HONO	$7.9{\cdot}10^{-2}$	XYL	$6.3{\cdot}10^{-2}$
SO2	$4.8{\cdot}10^{-2}$	SO2	$5.5{\cdot}10^{-2}$
ETE	$3.2{\cdot}10^{-2}$	CO	$2.8{\cdot}10^{-2}$
CO	$2.2{\cdot}10^{-2}$	ETE	$2.5{\cdot}10^{-2}$
HC3	$1.9{\cdot}10^{-2}$	HC3	$1.5{\cdot}10^{-2}$
TOL	$1.5{\cdot}10^{-2}$	HC8	$1.3{\cdot}10^{-2}$
HC5	$1.5{\cdot}10^{-2}$	TOL	$1.1{\cdot}10^{-2}$
HC8	$1.2{\cdot}10^{-2}$	HC5	$1.1{\cdot}10^{-2}$
KET	$3.8{\cdot}10^{-3}$	ALD	$3.6{\cdot}10^{-3}$
CSL	$-3.7{\cdot}10^{-3}$	KET	$2.9{\cdot}10^{-3}$
ALD	$2.9{\cdot}10^{-3}$	CSL	$-2.2{\cdot}10^{-3}$
ETH	$2.0{\cdot}10^{-4}$	ETH	$3.2{\cdot}10^{-4}$
ORA2	0.0	ORA2	0.0

Table 4.4: Maximum ozone sensitivity to the emitted species, i.e. $\max S_E$ for all species.

16 – 17 July		24 – 25 August	
Species	$\|S_E\|_1$	Species	$\|S_E\|_1$
NO	$1.9{\cdot}10^{1}$	NO	$2.8{\cdot}10^{1}$
XYL	2.1	ISO	3.4
ISO	1.5	XYL	2.2
HCHO	$9.1{\cdot}10^{-1}$	HCHO	1.8
OLT	$8.4{\cdot}10^{-1}$	API	1.4
API	$6.9{\cdot}10^{-1}$	OLT	1.1
TOL	$6.3{\cdot}10^{-1}$	CO	1.0
ETE	$5.7{\cdot}10^{-1}$	NO2	$9.1{\cdot}10^{-1}$
NO2	$5.6{\cdot}10^{-1}$	ETE	$7.1{\cdot}10^{-1}$
CO	$5.4{\cdot}10^{-1}$	HC3	$6.7{\cdot}10^{-1}$
HC3	$5.2{\cdot}10^{-1}$	TOL	$6.4{\cdot}10^{-1}$
OLI	$3.2{\cdot}10^{-1}$	SO2	$5.7{\cdot}10^{-1}$
SO2	$3.1{\cdot}10^{-1}$	OLI	$4.4{\cdot}10^{-1}$
HC5	$2.7{\cdot}10^{-1}$	HC5	$4.0{\cdot}10^{-1}$
HC8	$2.4{\cdot}10^{-1}$	HC8	$2.7{\cdot}10^{-1}$
KET	$1.3{\cdot}10^{-1}$	HONO	$1.6{\cdot}10^{-1}$
HONO	$1.1{\cdot}10^{-1}$	KET	$1.1{\cdot}10^{-1}$
ALD	$5.0{\cdot}10^{-2}$	ALD	$9.8{\cdot}10^{-2}$
CSL	$2.6{\cdot}10^{-2}$	ETH	$2.3{\cdot}10^{-2}$
ETH	$1.1{\cdot}10^{-2}$	CSL	$1.9{\cdot}10^{-2}$
ORA2	0.0	ORA2	0.0

Table 4.5: Ozone sensitivity to the emitted species (norm one), i.e. $\|S_E\|_1$ for all species.

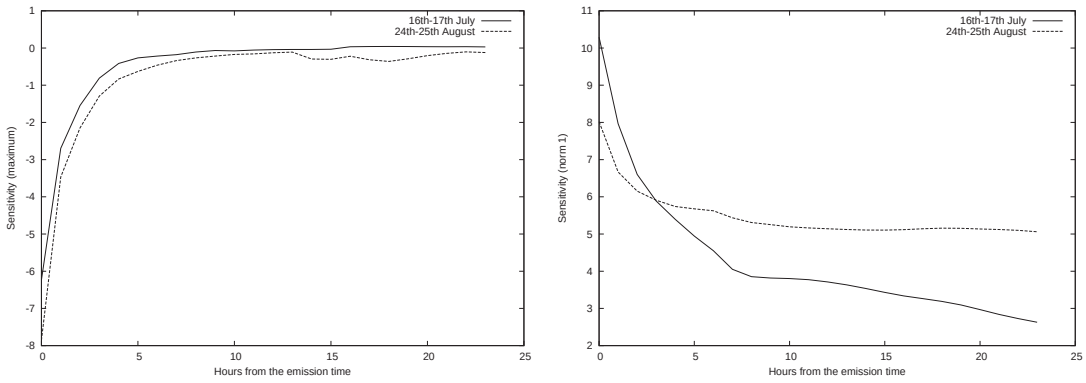
emissions. According to the sensitivities, this ratio may be an important parameter because of the sensitivity to NO emissions. Nonetheless this ratio may have a slight impact if only NO_2 emissions are adjusted. In this context, inverse modeling of this ratio is hard to achieve.

There are highly reactive species such as CSL (cresol and other hydroxy substituted aromatics) that have only a slight influence on ozone concentrations. They have a high absolute sensitivity (not reported here) but the relative sensitivity remains low because of their low emission fluxes. For instance, the maximum absolute sensitivity with respect to CSL is as high as the maximum sensitivity of ISO (isoprene) and its norm one is even significantly higher than ISO. It shows that the absolute sensitivities are not indicators suited in the perspective of inverse modeling.

The results raise the following question: what would be the sensitivity with respect to all volatile organic compounds as compared to NO? A rough idea may be drawn by summing up all the sensitivities. However a finer analysis based on the adjoint model is performed in Section 4.1.5.

Note that ozone concentrations are not sensitive to ORA2 emissions: ORA2 is only a product in RACM.

Temporal Sensitivity: $S_{\Delta h=h_c-h_e}$ It is of high interest to estimate the period over which emissions have some influence. Figure 4.6 shows the time evolution of the sensitivity.



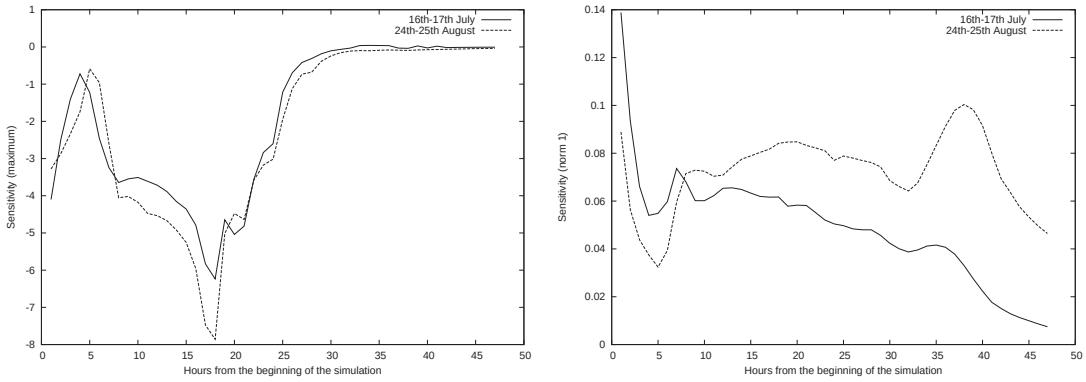
(a) (b)

Figure 4.6: Sensitivity (maximum on left, norm one on right) as function of the number of hours between the emission time and the simulated time. It shows the time evolution of the sensitivity.

The main effect is observed during the first hours. The maximum sensitivity quickly decreases, which tends to demonstrate that the emissions have only a local impact. Nevertheless the norm one of the sensitivity decreases slowly. The sensitivity after a few hours, is rather low but not negligible. As a conclusion, the emissions may have a strong local impact, and a more diffuse effect still lasts for several hours.

Temporal Sensitivity: S_{h_c} Depending on the hour in the day, ozone concentrations may be more or less sensitive to the emissions as shown in Figure 4.7.

For each hour h_c, the norm one is computed with the available sensitivities, namely the sensitivities with respect to the emissions released within 24 hours before the hour h_c. But the number of emission times, before h_c and to which the sensitivities were computed, depends on



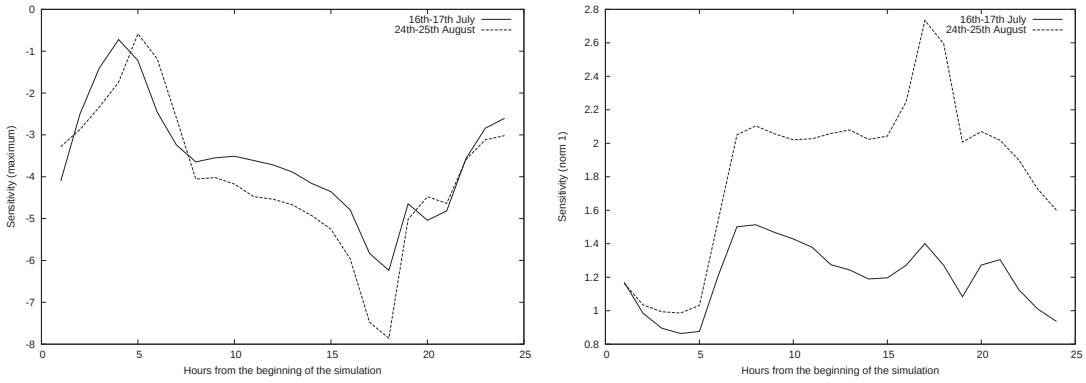
(a) (b)

Figure 4.7: Sensitivity (maximum on left, norm one on right) as function of the hour h_c in the day.

h_c: if $h_c = 2$, only emissions released at $h_e = 1$ and $h_e = 2$ are taken into account; at the end, e.g. $h_c = 40$, the emissions taken into account are released at any $h_e \in [\![40 - 24, 24]\!]$ because the sensitivity to emissions released after $h_e = 24$ are not available (Table 4.2). Hence the sensitivities shown in Figure 4.7 should be carefully analyzed. However if the first hours and the last hours are discarded, the other sensitivities are reliable.

It appears that the concentrations at any time in the day may be influenced by the emissions.

Temporal Sensitivity: S_{h_e} Figure 4.8 shows the sensitivities as function of the emission time. The point is to check whether all emissions in the day have a similar impact on the concentrations. Actually, since we analyze the relative sensitivities, its shows the possible impact of changes in the emissions within their uncertainty range, assumed to be the same relative range for every hour.



(a) (b)

Figure 4.8: Sensitivity (maximum on left, norm one on right) as function of the emission time.

From the norm one, ozone concentrations seem to be more sensitive to the emissions during the daytime. However the sensitivity to the nightly emissions is not negligible as compared to the highest daytime sensitivity. The nightly emissions are strongly lower than the daytime

emissions, but it appears that their absolute sensitivity is at least as high as for the daily emissions. This is probably due to the lower impact of the vertical diffusion during night.

Sensitivity to the Emission Location: $S_{(i_e,j_e)}$ There are 83 locations (i_e, j_e) (among the 205 (E, i_e, j_e) combinations – see Table 4.3). It is not possible to point out clearly which locations lead to the highest sensitivities because (1) the sensitivity depends on the emitted pollutant, and (2) there is no gap in the list of sensitivities. The locations have been sorted from the most sensitive location to the less sensitive one. For the experiment over 16–17 July, the fifteen first sensitivities (norm 1) are: 8.90, 5.92, 5.81, 5.75, 5.13, 4.00, 3.91, 3.89, 3.01, 2.69, 2.69, 2.42, 2.30, 2.25, 1.86. It then decreases slowly down to 0.07 (if the sensitivities to ORA2 are excluded). The 37th location is associated with a sensitivity lower than a tenth of the sensitivity associated with the first location (8.90).

The highest sensitivities are mainly reached at the locations where NO is emitted. It seems more relevant to rank the locations for each species, i.e. to analyze S_{E,i_e,j_e} instead of S_{i_e,j_e}. Then there are too few locations per species to draw reliable conclusions. Nevertheless, for each species, it appears that the sensitivities rapidly decrease among the locations. For most species, associated with about ten cells (see Table 4.3), there is a ratio of about three between the first sensitivity and the last one. Recall that the emission locations were included in the experiment if their emission flux was greater than the half of the highest emission flux. Therefore, if the absolute sensitivities were constant, the ratio between the extreme sensitivities would be two. The actual decrease is higher, which tends to show that ozone concentrations are mainly sensitive to a few emission locations.

Sensitivity to the Concentration Location: $S_{(i_c,j_c)}$ There are 2145 cells that may be sorted in the same way as for the emission locations. The sensitivities also decrease slowly. Again, the cells where NO is strongly emitted (i.e. the 10 cells $(i_{\mathrm{NO}}, j_{\mathrm{NO}})$) lead to high sensitivities in their neighborhood. The 88th cell is associated with a sensitivity lower than a tenth of the highest sensitivity (norm one). Since there are 85 emitting points to which the sensitivity was computed, it means that highest sensitivities are mostly found close to the main sources. This is illustrated by Figure 4.9.

Vertical Profile As previously mentioned, sensitivities are available only in the first two vertical levels. The maximum of S_{k_c} is found in the first level (twice as high as the maximum in the second level), but the norm one is similar in both levels.

Validity of the Study

There are mainly two limitations that may question the validity of the study: (1) the amount of emission points with respect to which the sensitivities are computed, (2) the number of experiments (two periods).

As for the number of emission points, there is no indication, in the previous analyses, that the inclusion of more sources would change the results: the highest relative sensitivities are chiefly due to the highest emission locations. However this experiment with the tangent linear model only claims a validity with respect to the strongest emission sources. In Section 4.1.5, the sensitivities are computed with respect to all locations.

There are only two experiments: one in July and another in August. Both give similar results even if the meteorological conditions and the ozone concentrations are strongly different. Moreover the same experiment has also been performed over 11–12 July 2001, but with an older

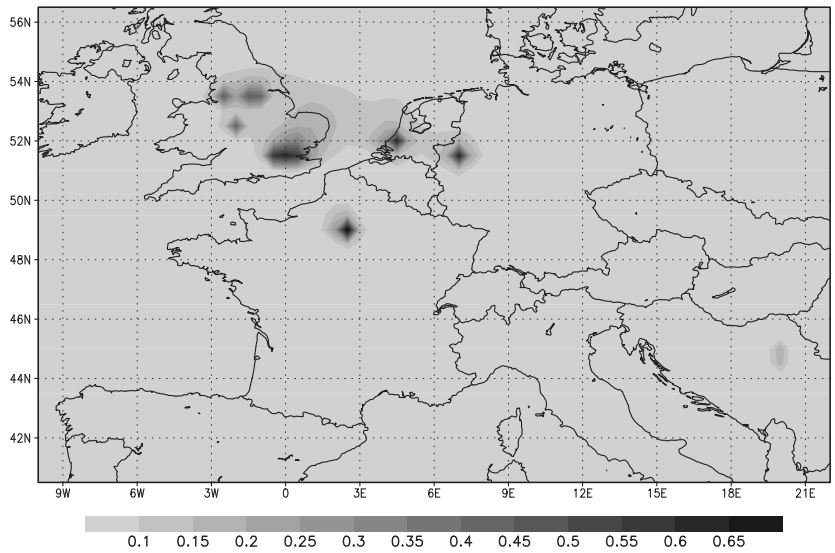


Figure 4.9: Relative sensitivities $S_{(i_c,j_c)}$ of output concentrations (mean over 16–17 July and 24–25 August). The highest sensitivities are mainly reached in the vicinity of the major emission sources.

version of the simulation system. The results are not reported in this paper because the simulation relied on less satisfactory parameterizations (Wesely's parameterization for deposition velocities – Wesely [1989] –, Louis' closure – Louis [1979] – for the vertical diffusion, a rough cloud attenuation scheme) and, for instance, its RMS at 1500 UT was about 25 $\mu g \cdot m^{-3}$. Nevertheless, this experiment led to the same conclusions as the two experiments detailed in this paper. It means that the results are repeatable, even with other models. Finally, a less detailed analysis, but still computing relative sensitivities with the tangent linear model of Polair3D, has been performed at regional scale over Île-de-France. As far as the results may be compared, the conclusions of the regional study are consistent with the analyses at continental scale.

4.1.5 Sensitivity Analysis with the Adjoint Model

Experiment Setup

The sensitivities are computed over the same periods, starting from 16 July 2001 and 24 August 2001. While the tangent linear model requires the choice of a limited set of emissions, applying the adjoint model once provides the sensitivities of a scalar value (i.e. the ozone concentration in a given cell and at a given time) with respect to all emissions. In this study, the sensitivities are computed at all cells that contain an EMEP station. There are 105 such stations in 103 cells. The sensitivity of ozone concentrations at 1500 UT are selected because they are close to the peaks which are a major concern in ozone forecasts. The 206 simulation (103 cells, 2 periods) are performed over 3 days and a half: starting 16 July 2001 and 24 August 2001 at 0000 UT and ending 19 July and 27 August at 1500 UT.

In addition, the time evolution is analyzed at two stations of different nature: one that contains Paris (highly emitting area) and another one containing the EMEP station Montandon. Note that EMEP stations are not located in strongly emitting regions, in order to measure the long-range transboundary pollution.

The sensitivities are not computed with respect to all model emissions but with respect to inventory emissions. Inventory emissions are provided on the EMEP grid (polar-stereographic projection) and are gathered into four inventory species: NOx, NMVOC (volatile organic com-

pounds), SOx and CO (we focus on NOx and NMVOC in this study, due to their prominent impact). The sensitivities are computed with respect to yearly emissions which are the raw data provided by EMEP. The gradients are the derivatives with respect to EMEP yearly emissions, over the period of the simulation. It means that the contributions (to the gradient) coming from emissions released before the beginning of the simulation are discarded. However these discarded contributions are negligible.

Results and Discussion

Distribution over the EMEP Network For each station, we compute the norm one of the sensitivity: the sum over all cells of the relative sensitivities. It estimates the impact that changes in the emissions can have at the stations. The distribution of the norm one over the stations is shown in Figure 4.10. The means over all stations is $17.2\mu g \cdot m^{-3}$ and $10.3\mu g \cdot m^{-3}$ for NO emissions (27 August and 19 July respectively), $3.0\mu g \cdot m^{-3}$ and $5.1\mu g \cdot m^{-3}$ for NMVOC emissions (27 August and 19 July respectively).

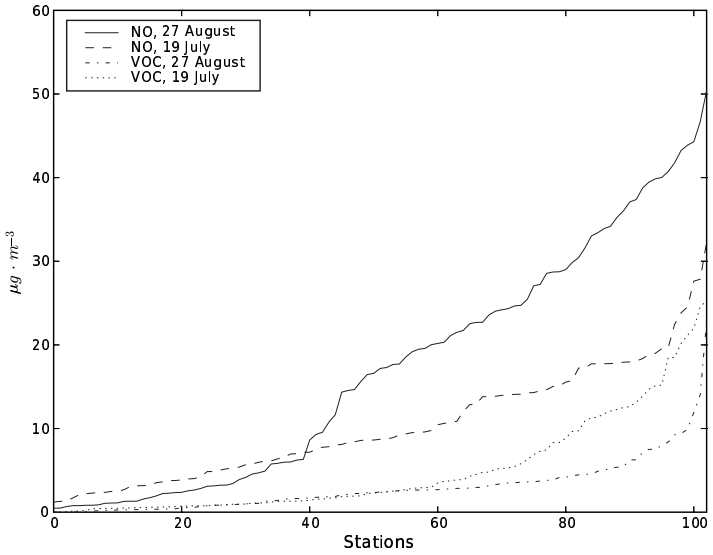


Figure 4.10: Norm one of the relative sensitivity for all stations. The sensitivities of the four cases are sorted independently so that they are increasing functions of the stations. The stations order is not the same one in the four cases.

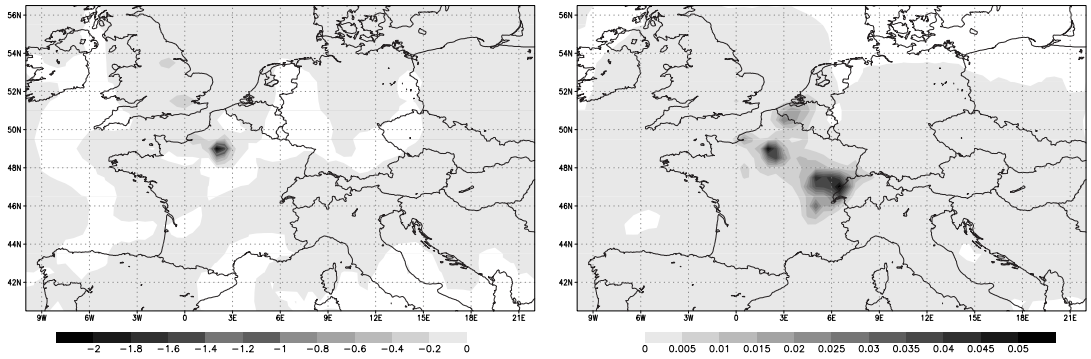
The sensitivity to NO emissions is at least twice as high as the sensitivity to NMVOC emissions, which is also the ratio that may be computed from Table 4.5. With this ratio, inverse modeling of emissions may be performed with both NO and NMVOC emissions as control parameters.

The sensitivities are highly spread among the stations, from stations at which ozone concentrations are not sensitive to emissions to stations with high sensitivities (up to $50.4\mu g \cdot m^{-3}$). The correlation between the sensitivities for 19 July and 27 August is low (about 15%). The sensitivities strongly vary due to the conditions. However if one considers at each station the average of the sensitivities for the two periods, there are still several stations with a low sensitivity. The sensitivities in other meteorological conditions would be useful to confirm this remark.

Spatial Extent For inverse modeling, it is useful to check that the observations are sensitive to emissions of the whole domain. Otherwise measurements would not bring enough information to invert all emissions. The spatial extent of the sensitivities depends on:

1. the meteorological conditions (e.g., wind velocities);
2. the species (chemical reactivity): because of the efficient titration of ozone by NO, the sensitivity to NOx emissions tend to be more local than the sensitivity to NMVOC;
3. the emission locations (the relative sensitivities tend to be high at emission locations): the sensitivities usually have a larger extent when they are computed at stations far from the major emission sources.

It appears that the spatial extent may strongly differ due to these factors. For instance, in Figure 4.11, a local sensitivity is obtained over Paris to NOx emissions and a larger extent is found for the sensitivity of a concentration at Montandon to NMVOC emissions.



(a) (b)

Figure 4.11: Sensitivity (in $\mu g \cdot m^{-3}$) of ozone concentrations at 1500UT (ozone peak) on 26 August 2001. On the left, the figure shows the sensitivity of the ozone concentration at Paris with respect to NOx emissions. On the right, the figure shows the sensitivity of the ozone concentration at the EMEP station Montandon with respect to NMVOC emissions. The extent of the sensitivity may strongly differ depending on the station and the emitted pollutants.

To assess the impact of all stations, we first associate a spatial extent to each station. For each station, we select all emission cells to which the ozone concentration (at 1500 UT) has a significant sensitivity. A sensitivity is assumed to be significant if it is among the highest sensitivities whose sum contributes to 75% of the overall sensitivity (norm one). It is found that the domain is well covered, as shown in Figure 4.12 (example with NO emissions and for 19 July). However it includes the spatial extent of stations with low sensitivities. Figure 4.12 shows the area covered in case where 25% of the stations are discarded because of their low sensitivities (less than $5\mu g \cdot m^{-3}$ – see Figure 4.10). The area covers the main part of the domain. The same is observed for 24 August. As for VOC emissions, the area is even bigger. Providing that the sensitivities are high enough, this is a promising result for inverse modeling.

Time Evolution The sensitivities are computed using the adjoint model which integrates the chemistry-transport equation backward in time. In the previous sections, the integration was performed over three days and a half. It was assumed that emissions released before have

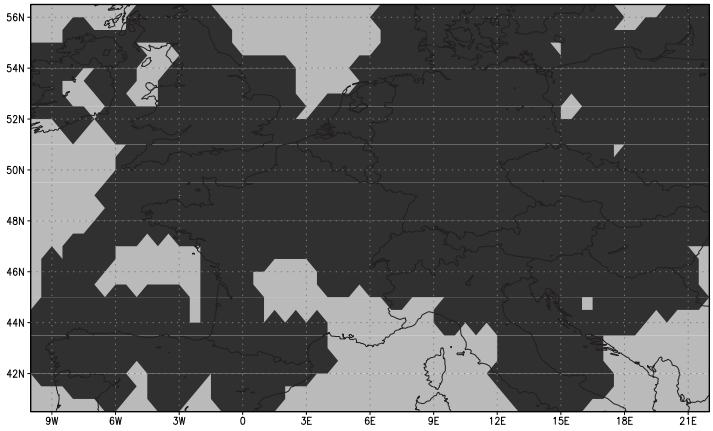


Figure 4.12: Area covered by the significant sensitivities associated with all stations of the EMEP network, for NO emissions and for 19 July.

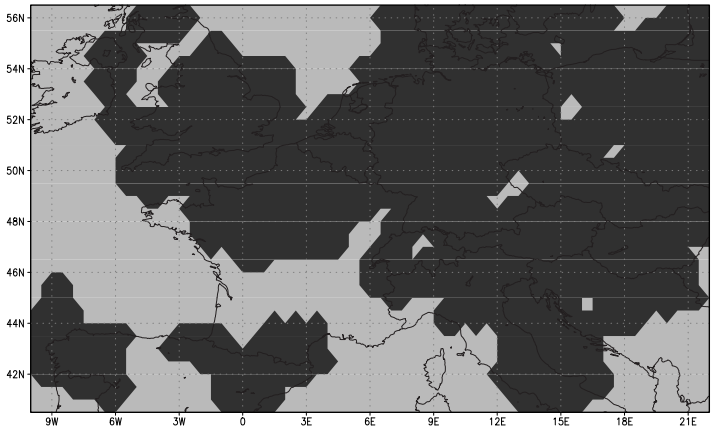


Figure 4.13: Area covered by the significant sensitivities associated with selected stations of the EMEP network, for NO emissions and for 19 July. 75% of EMEP stations are included due to their high sensitivity. This figure can be compared with Figure 4.12 in which all EMEP stations are included.

a negligible impact. The evolution of the computed sensitivities backward in time shows to which hourly emissions the ozone concentrations are sensitive. As shown in Figure 4.14, ozone concentrations (still at 1500 UT) are mainly sensitive to the emission released in the first hours. This is especially true over Paris due to the high local emissions which constitute the main part of the (relative) sensitivity.

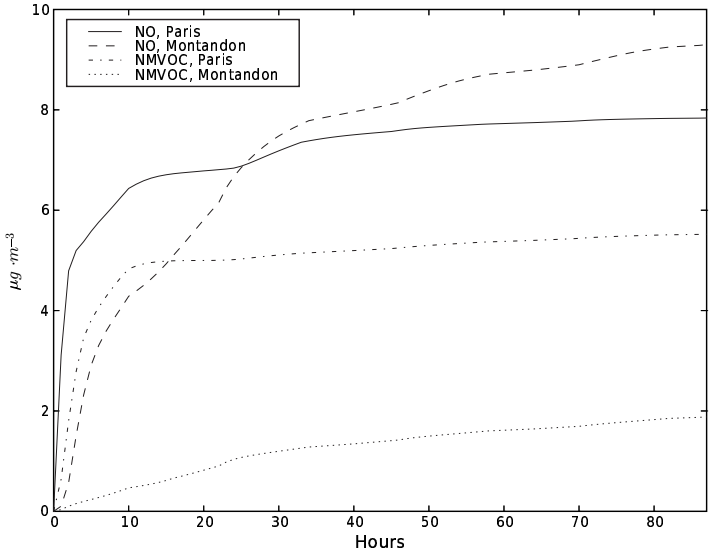


Figure 4.14: Norm one of cumulative relative sensitivities as function of the number of hours over which the backward integration is performed.

4.1.6 Sensitivity Analysis with Monte Carlo Simulations

As previously mentioned, the Monte Carlo simulations are performed in order to estimate "global" sensitivities. Moreover the emissions in the whole domain are perturbed, which provides additional information to the sensitivities to the major emission sources (Section 4.1.4) or to the sensitivities over the EMEP network (Section 4.1.5).

Experiment Setup

The Monte Carlo simulations are performed over one week: from 16 July 2001 to 22 July 2001 (included). The initial conditions are the same ones for all simulations. Three sets of two hundred simulations are performed with perturbations in NO emissions (first set), VOC emissions without biogenic emissions (second set) and biogenic emissions (third set). For each simulation the whole emission field is modified, that is, the emissions at all time steps and in all cells. Note that, for a perturbation in VOC emissions, all VOC species are perturbed with the same coefficient. This is also true for biogenic emissions (ISO and API).

From Hanna *et al.* [2001], we assume that emissions (including biogenic emissions) have a log-normal distribution with standard deviation of the log-transformed data set to 0.203. It roughly means that the emissions mainly vary within ±50% of the reference value. This is the uncertainty proposed in Hanna *et al.* [2001] for major emission points. Other emissions, notably biogenic emissions, are associated with higher uncertainty in Hanna *et al.* [2001]. However the

point of this paper is to estimate the sensitivity, not the uncertainty. All emissions are therefore perturbed in the same way.

Results and Discussion

Sensitivity to Inventory Species The sensitivity is estimated from the standard deviation in an ensemble for output ozone concentrations. In Figure 4.15, the averages over the whole domain of standard deviation of the ensembles (for ozone concentrations in the first layer) are shown for NO, VOC and biogenic emissions. The same results for a restricted set of 100 simulations are also reported to demonstrate that the ensemble has reasonably converged.

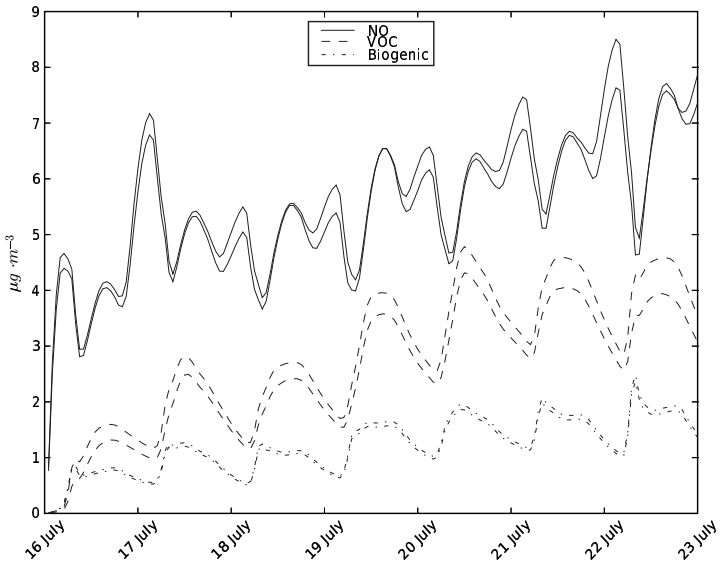


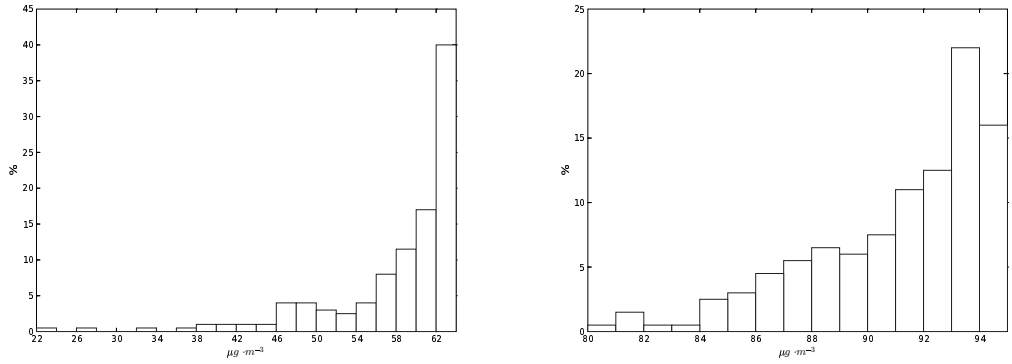
Figure 4.15: Time evolution of the standard deviation (averaged over the whole domain) of the ensemble for first-layer ozone concentrations. It is shown for NO, VOC and biogenic emissions for the full ensemble (200 simulations) and for restricted ensemble (100 simulations).

In Figure 4.15, it appears that the emissions have an impact in a very few hours, especially NO emissions. Then the impact slowly increases due to accumulation in the domain for about four days. The sensitivity is stabilized in the last three days.

NO emissions still imply the highest sensitivity with an average standard deviation of $6.3\mu g \cdot m^{-3}$ for the last three days, to be compared with $3.9\mu g \cdot m^{-3}$ for VOC emissions and $1.5\mu g \cdot m^{-3}$ for biogenic emissions. The sensitivity to VOC and biogenic emissions is close to the sensitivity to NO emissions. A consequence for inverse modeling of emissions is that VOC and NO emissions should be both optimized.

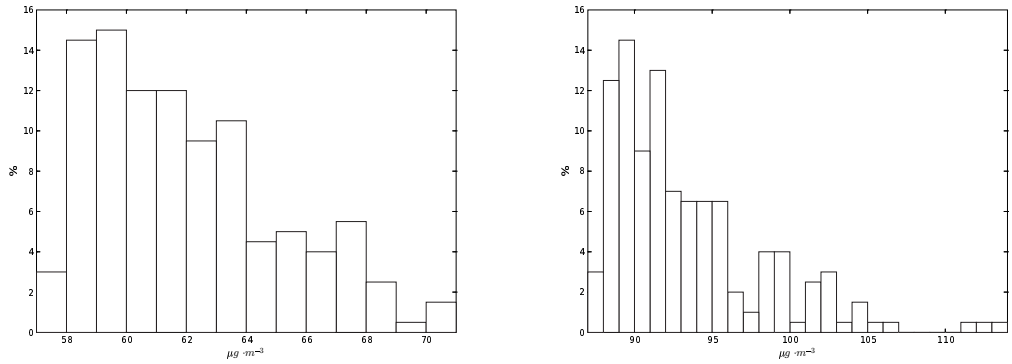
Sensitivity Spatial-Distribution The spatial distribution of the sensitivity is estimated from the distribution of the standard deviation of the ensemble. The sensitivity due to NO emissions is more spread than the sensitivity to VOC emissions. Nevertheless it appears that there is a noteworthy dependence on the meteorological conditions. The sensitivity is high in the regions that are in the plume of the major emission locations, especially of emissions from Great Britain. Otherwise the distribution of the sensitivity strongly varies from one day to another. Simulations over a longer period should be performed to draw further conclusions.

"Global" Sensitivity The probability density function of the concentrations shows how the system is sensitive to the most probable emissions. In Figures 4.16 and 4.17, the distributions of the minima and the maxima of ozone concentrations are shown for changes in NO emissions and changes in VOC emissions, which provides more details than the averaged standard deviations of Section 4.1.6. The distributions roughly show the characteristics of a log-normal distribution. They would be log-normal distributions in case the model is linear since the emissions are assumed to be log-normal. The shape of the distribution (maximum probability on left or on right) is due to the sign of the sensitivity: additional NO emissions titrate ozone concentrations whereas additional VOC emissions lead to ozone production. Another remark is the high sensitivities of daily minima to NO emissions.



(a) (b)

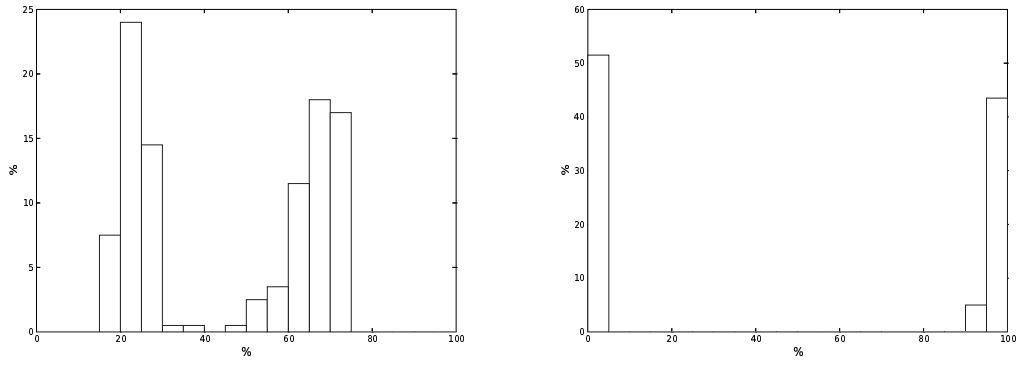
Figure 4.16: Distributions of the means of the daily ozone minima (left) and maxima (right) due to changes in NO emissions. The means are computed over the last three days of the simulation (20–22 July) and over the whole domain from which a band of three cells has been discarded (to minimize the influence of the boundary conditions). The reference simulation is associated with a minimum of $62.7\mu g \cdot m^{-3}$ and a maximum of $92.3\mu g \cdot m^{-3}$.



(a) (b)

Figure 4.17: Same distributions as in Figure 4.16, but due to changes in VOC emissions.

In Figures 4.16 and 4.17, the analyzed concentrations are mean concentrations that may hide compensations: the emissions could increase ozone production in a given region (or at given hours in the day) while it could decrease ozone production somewhere else. To check that there is a clear tendency in the whole domain, the proportion of ozone peaks below those of the reference simulation is computed for each simulation. If this proportion is low for a given simulation, it indicates that the simulation has a tendency to lower ozone peaks. If the proportion is in the vicinity of 50%, there is no tendency. The percentage of simulations that have a proportion (in the above sense) equal to x% is plotted in Figure 4.18. Changes in VOC emissions lead to either lower or higher ozone daily maxima in all cells, as indicated by the two extreme modes in Figure 4.18 (right). NO emissions are not associated with such a clear dependency but there are still two distinct modes in the distribution. Similar conclusions are found for the daily minima. Moreover biogenic emissions show the same behavior as the other VOC emissions.



(a) (b)

Figure 4.18: Percentage of simulations versus the percentage of ozone daily maxima (in all cells in the domain minus a three-cell band at the borders) above the reference concentrations for NO emissions (left) and for VOC emissions (right). Bars whose abscissae are below 50% are associated with simulations in which most peaks are below the reference peaks.

4.1.7 Conclusion

From the previous results, it appears that ozone concentrations are more sensitive to NO emissions than to emissions of any other species. However the sensitivity to all VOC species is not negligible as compared to the sensitivity to NO emissions. The ratio between the sensitivities to NO emissions and to VOC emissions is about two in the studied cases and depends on the meteorological conditions and on the involved locations. Inverse modeling of emissions would therefore be performed on both NO and VOC emissions.

A possible impact of NO emissions is estimated with a relative sensitivity of $6.3\mu g \cdot m^{-3}$, which is rather low knowing that the uncertainty in the emissions is about $\pm 50\%$ and that the error on ozone concentrations (root mean square) is about $20\mu g \cdot m^{-3}$. It should also be compared to the uncertainty due to the physical parameterizations and numerical approximations, e.g. at least $10\mu g \cdot m^{-3}$ on ozone peaks [Mallet et Sportisse, 2006]. It means that inverse modeling of emissions is a difficult task and that results from such an experiment should be carefully checked. The investigation of second-order sensitivities should be performed (see Quélo *et al.*

[2005] for an application at regional scale).

On the other hand, the emissions have a rather local effect in time and space, and the sensitivities can be high close to strongly emitting locations. Inverse modeling of these sources could benefit from these high sensitivities. Moreover concentrations at every hours are sensitive to emissions, and, in the same way, emissions released at any time can be associated with high sensitivities. Depending on the abilities to forecast all concentrations, inverse modeling may take advantage of all observations and may retrieve emissions released at any time. We also found that the spatial distribution of the sensitivities covers the whole simulation domain.

Another issue to be investigated is the opportunity to use observations of other pollutants, especially of NO_2, even if there are much less observations for other species than for ozone.

Chapitre 5

Prévision d'ensemble

Ce chapitre étudie le potentiel de plusieurs méthodes d'ensemble dans le but d'améliorer les prévisions. Plusieurs ensembles, contenant jusqu'à 48 membres (modèles), sont générés avec le système Polyphemus. Les membres diffèrent entre eux par leurs paramétrisations physiques, leurs approximations numériques et leurs données d'entrée. Chaque modèle est évalué sur quatre mois de l'année 2001 avec les observations des centaines de stations des réseaux EMEP, Pioneer et BDQA (comme au chapitre 2). L'étude montre que plusieurs combinaisons linéaires de modèles ont le potentiel d'accroître fortement les performances (lors des comparaisons aux mesures). Les poids optimaux associés aux modèles ne sont robustes ni spatialement ni temporellement. La prévision de ces poids nécessite donc des méthodes adaptées, comme la sélection de périodes d'apprentissage ou l'utilisation d'algorithmes d'apprentissage statistique. Des améliorations significatives des performances sont obtenues avec les combinaisons prévues. Une diminution d'environ 10% de l'erreur quadratique moyenne est atteinte sur les pics d'ozone. Les concentrations horaires d'ozone bénéficient d'améliorations plus nettes encore.

Sommaire

Le chapitre est essentiellement constitué de
Mallet, V. et Sportisse, B. (2005c). Toward ensemble-based air-quality forecasts. *En révision pour publication dans J. Geophys. Res.*

5.1 Introduction

Bien que rarement estimée, l'incertitude dans les modèles de chimie-transport est une limitation majeure en prévision de la qualité de l'air. La source de cette incertitude réside dans les champs d'entrée (émissions, vitesses de dépôt, données de sol, champs météorologiques, etc.), comme reporté dans Hanna *et al.* [1998, 2001], et dans la formulation des modèles elle-même [Russell et Dennis, 2000; Mallet et Sportisse, 2006]. Le chapitre 3 apporte une contribution à cette problématique. L'incertitude est si élevée que la fiabilité des résultats de modèle doit être soigneusement évaluée, et que la prévision d'ensemble est une stratégie pertinente pour cela. Straume *et al.* [1998]; Dabberdt et Miller [2000]; Galmarini *et al.* [2004]; Straume [2001]; Warner *et al.* [2002] ont estimé l'incertitude dans les modèles de dispersion en utilisant la prévision d'ensemble. Concernant l'exposition à l'ozone, Hanna *et al.* [1998]; Beekmann et Derognat [2003] ont pris en compte l'incertitude dans les champs d'entrée des modèles grâce à des simulations Monte Carlo, ceci pour tester l'efficacité de l'impact de réductions d'émissions. Hanna *et al.* [2001]; Hanna et Davis [2002] ont estimé l'incertitude dans les prévisions photochimiques (d'ozone) toujours sur la base de simulations Monte Carlo. Dans le chapitre 3 de cette thèse, ainsi que dans Mallet et Sportisse [2006], une approche multi-modèles a permis d'estimer l'incertitude due à la formulation du modèle.

Concernant les prévisions photochimiques quotidiennes, peu de développements ont été entrepris dans le but d'associer des incertitudes aux prévisions ou dans le but de dépasser les limitations des données et modèles incertains. Les améliorations dans les prévisions de la qualité de l'air ont été recherchées par des développements en modélisation, par l'introduction de données plus fines et par l'accroissement des ressources de calcul. Malheureusement, les performances ne se sont que peu améliorées dans le même temps [Russell et Dennis, 2000]. Vraisemblablement, la raison provient de la forte incertitude qui cache les efforts de modélisation et des modèles qui sont généralement ajustés pour délivrer des prévisions satisfaisantes (ce qui est aussi suggéré dans Russell et Dennis [2000]). Prendre en compte l'incertitude semble un moyen indiqué dans le but d'améliorer les prévisions. Une stratégie prometteuse est l'utilisation de prévisions d'ensemble et la combinaison des membres de l'ensemble.

Une méthode « brutale » est la moyenne d'ensemble qui consiste simplement à prendre pour prévision la moyenne des membres de l'ensemble [Delle Monache et Stull, 2003]. Pour que cette stratégie soit payante, il faut vérifier les deux hypothèses sous-jacentes (et très restrictives) suivantes. Il faut que l'ensemble permette d'approcher raisonnablement la véritable densité de probabilité des concentrations de sortie. Il faut ensuite que l'espérance de cette densité de probabilité soit proche de la vraie valeur. Du fait du nombre limité de modèles, et aussi de la méconnaissance des incertitudes, il est difficile de satisfaire la première hypothèse. De plus, aucune étude ne porte crédit à la seconde hypothèse. Des méthodes plus complexes ont été utilisées dans d'autres domaines, comme les superensembles en météorologie [Krishnamurti *et al.*, 2000] ou les moyennes bayésiennes [Hoeting *et al.*, 1999] dans d'autres domaines.

Dans ce chapitre, on étudie plusieurs méthodes de construction de combinaison optimale des membres de l'ensemble. L'objectif est d'accroître les performances des prévisions quotidiennes, performances mesurées via la comparaison aux observations. Les méthodes sont appliquées aux concentrations horaires et aux pics journaliers d'ozone à l'échelle Européenne (domaine similaire à ceux utilisés dans les chapitres précédents), sur l'été 2001 (voir chapitre 2) et avec les centaines de stations des trois réseaux introduits au chapitre 2. Les ensembles introduits contiennent jusqu'à 48 membres, ce qui permet d'étudier les caractéristiques des ensembles les plus performants. À la section 5.2, de plus amples détails sont fournis sur la configuration des ensembles et de leurs membres. La section 5.3 présente les méthodes étudiées et analyse leur potentiel, c'est-à-dire la qualité *a posteriori* (i.e., connaissant toutes les observations) de leur combinaison. À la

section 5.4, des méthodes permettant d'approcher en prévision les combinaisons optimales sont décrites et testées. La sélection des membres des ensembles est aussi abordée.

5.2 Prévisions d'ensemble utilisées

5.2.1 Simulation de référence

Pour cette étude, le domaine européen est $[40.25°\text{N}, 10.25°\text{W}] \times [56.75°\text{N}, 22.25°\text{E}]$. L'étude couvre la période du 27 avril 2001 au 31 août 2001. On prévoit les concentrations horaires d'ozone ainsi que les pics journaliers. On définit une simulation de référence (modèle de référence ou configuration de référence, pas nécessairement le meilleur modèle lors de comparaisons aux observations) :

1. données météorologiques : champs ECMWF (résolution de $0.36° \times 0.36°$, résolution spectrale horizontale TL511, 60 niveaux, pas de temps de 3 heures, cycles de prévision de 12 heures démarrant à partir de champs analysés) ;
2. occupation des sols : données de l'USGS (donc 24 classes, 1 km Lambert) ;
3. mécanisme chimique : RACM [Stockwell *et al.*, 1997] ;
4. émissions : inventaire EMEP, converti selon Middleton *et al.* [1990] ;
5. émissions biogéniques : estimées selon Simpson *et al.* [1999] ;
6. vitesses de dépôt : issues de la paramétrisation révisée et proposée dans Zhang *et al.* [2003b] ;
7. diffusion verticale : dans la couche limite, on utilise la paramétrisation de Troen et Mahrt introduite dans Troen et Mahrt [1986], avec la hauteur de couche limite fournie dans les données ECMWF ; au-dessus de la couche limite, on repose sur la paramétrisation de Louis [Louis, 1979] ;
8. conditions aux limites : elles sont générées sur la base des sorties, pour une année météorologique typique, du modèle de chimie-transport global Mozart 2 [Horowitz *et al.*, 2003] ;
9. schémas numériques : ce sont les schémas utilisés par défaut dans Polair3D (voir section 2.3), c'est-à-dire, avec une séparation d'opérateur du premier ordre (advection–diffusion–chimie), un schéma d'ordre trois avec limiteur de flux de type Koren et un schéma de Rosenbrock d'ordre 2 pour la diffusion et la chimie [Verwer *et al.*, 2002].

Les sections 1.5, 1.6 et 2.3 détaillent les points précédents. La résolution verticale reste faible pour des problèmes de temps calcul (48 modèles). La première couche est située entre 0 m et 50 m. L'épaisseur des quatre autres couches est d'environ 600 m, avec une hauteur supérieure de 3000 m pour la dernière couche.

5.2.2 Description des ensembles

Trois ensembles sont introduits :

1. l'ensemble 1 est composé de la simulation de référence ainsi que de 21 simulations qui diffèrent de la simulation de référence par un seul changement, soit dans les paramétrisations physiques, soit dans les données d'entrée (à Polyphemus), soit dans les approximations numériques, soit dans les données intermédiaires et incertaines calculées par Polyphemus. Le tableau 5.1 liste tous ces changements.
2. l'ensemble 2 est construit sur la base des changements effectués pour les modèles 17, 8, 4, 2 et 1 (numéros tirés du tableau 5.1). Toutes les combinaisons possibles de ces changements sont incluses dans cet ensemble. Il y a donc 32 membres (modèles) dans l'ensemble 2.

3. l'ensemble 3 rassemble tous les membres des ensembles 1 et 2. Les ensembles 1 et 2 ont six membres en commun (0, 1, 2, 4, 8 et 17) ; l'ensemble est donc composé de 48 membres.

Tab. 5.1 – Paramétrisations physiques, données brutes d'entrée (pour Polyphemus), approximations numériques et données perturbées (dans la chaîne de calcul de Polyphemus), pour la génération de l'ensemble 1. Chaque membre (modèle) de cet ensemble est construit à partir de la simulation de référence à laquelle un unique changement (colonne « alternative ») est appliqué.

n°	Modèle	Référence	Alternative	Commentaire
		Paramétrisations physiques		
1.[a]	Chimie	RACM	RADM 2 [Stockwell *et al.*, 1990]	
2.	Diffusion verticale	Troen & Mahrt	Louis [Louis, 1979]	
3.			Louis pour les conditions stables	Troen & Mahrt conservé en conditions instables
4.	Vitesses de dépôt	Zhang [Zhang *et al.*, 2003b]	Wesely [Wesely, 1989]	
5.	Flux de surface	Flux de chaleur[b]	Flux de moment[b]	Utilisé dans le calcul de la résistance aérodynamique (vitesses de dépôt)
6.	Atténuation nuageuse	méthode RADM [Chang *et al.*, 1987; Madronich, 1987]	Esquif (ESQUIF [2001])	
7.	Humidité relative critique	Fonction de σ	Constante sur deux niveaux	Utilisée dans le calcul de l'atténuation nuageuse dans la méthode RADM
		Données d'entrée brutes		
8.	Distribution verticale des émissions	Toutes dans la première couche	Toutes dans les deux premières couches	
9.	Occupation des sols	USGS	GLCF	Pour le calcul des vitesses de dépôt
10.	Occupation des sols	USGS	GLCF	Pour le calcul des émissions biogéniques
11.	Exposant p dans Troen & Mahrt	2	3	
12.	Constantes photolytiques	JPROC (de Models-3, EPA)	Fonction de l'angle zénithal	
		Approximations numériques		
13.	Pas de temps	600 s	100 s	
14.			1800 s[c]	
15.	Résolution verticale	5 niveaux	9 niveaux	La hauteur de la première couche demeure 50 m
16.	Hauteur de la première couche	50 m	40 m	La hauteur supérieure des autres couches ne change pas
17.	Équation de continuité	$\mathrm{div}\,(V) = 0$	$\mathrm{div}\,(\rho V) = 0$	
		Autres données perturbées		
18.	Hauteur de couche limite	ECMWF	Augmentée de 10%	
19.	Émissions de NO	EMEP	Augmentées de 25%	Émissions biogéniques incluses
20.	Émissions biogéniques	Simpson *et al.* [1999]	Augmentées de 100%	À l'exclusion des émissions biogéniques de NO
21.	Conditions aux limites d'ozone	Mozart 2	Diminuées de 10%	

[a] Le modèle de référence est le modèle n°0.
[b] Calculé grâce aux formules de Louis.
[c] L'advection est intégrée sur des pas de temps sous-multiples de 1800 s de sorte à satisfaire la condition CFL (Courant-Friedrichs-Lewy).

Les ensembles 1 et 2 sont similaires à ceux introduits à la section 3.2 (et dans Mallet et Sportisse [2006]), pour l'estimation de l'incertitude dans les concentrations d'ozone simulées. L'incertitude n'est pas évaluée dans ce chapitre. Une idée de la dispersion des résultats est cependant fournie par les profils moyens et journaliers d'ozone – voir figure 5.1. La moyenne de l'écart-type horaire des concentrations des profils de la figure 5.1 vaut 10.4 $\mu g \cdot m^{-3}$. Afin d'illustrer la distribution spatiale de l'ensemble 3, l'écart-type de l'ensemble 3 est calculé dans chaque cellule et pour chaque heure, et puis la moyenne relative de ces écarts-types est représentée dans la figure 5.2.

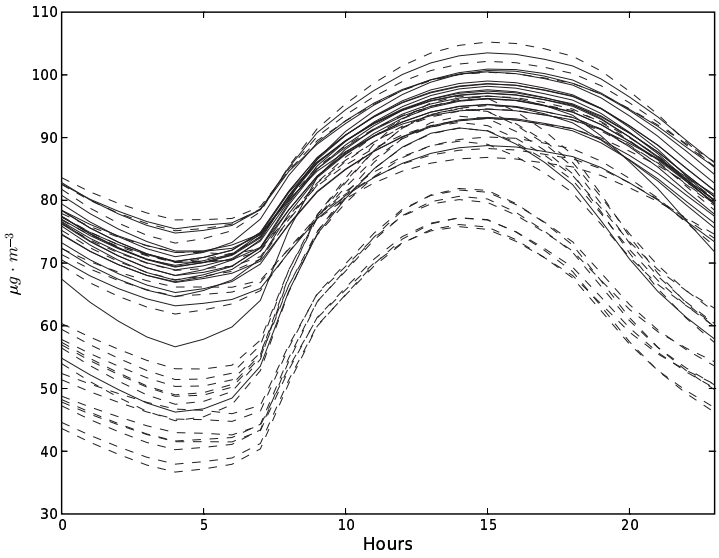


FIG. 5.1 – Profils journaliers d'ozone pour les 48 modèles (ensemble 3). Les lignes discontinues correspondent aux modèles qui ne sont pas dans l'ensemble 1. Les concentrations sont moyennées sur tout le domaine (à l'exclusion d'une bande de trois cellules autour du domaine) et sur les 127 jours de la simulation.

5.2.3 Comparaison aux observations

Les mesures des trois réseaux d'observation, présentés à la section 2.5.2, sont utilisées. Les stations retenues disposent de 30% des observations qui peuvent être faites, pendant les 127 jours simulés, pour les concentrations horaires et pour les pics journaliers. Les réseaux sont :

- Réseau 1 : il s'agit du réseau Pioneer. Il fournit environ 619 000 observations horaires et 27 500 pics.
- Réseau 2 : il correspond au réseau EMEP, avec 240 000 mesures horaires et 10 400 pics journaliers.
- Réseau 3 : il désigne le réseau de la BDQA. Il propose 997 000 observations horaires et 42 000 pics.

On peut se référer à la section 2.5.2 pour plus de détails. Afin de mieux interpréter les résultats qui suivent, il convient de noter que les réseaux 1 et 2 sont spatialement plus étendus que le réseau 3 qui, lui, délivre un nombre important d'observations sur la France. Le réseau 2 permet de tester les méthodes exclusivement sur des stations régionales. On considère les statistiques d'erreur suivantes (déjà reportées dans le tableau 2.1) : l'erreur quadratique moyenne

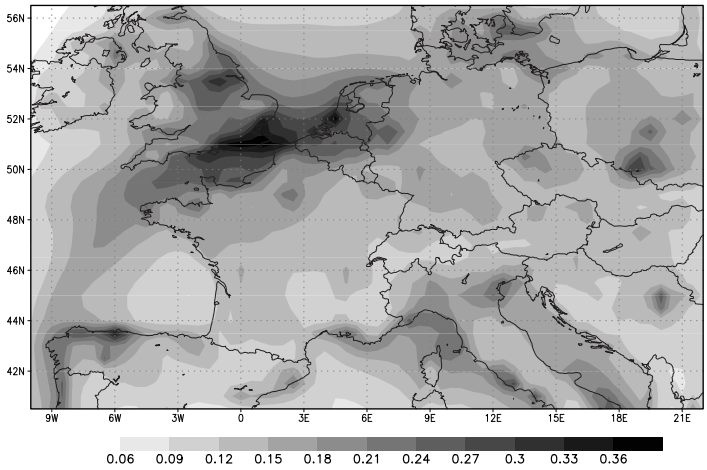


FIG. 5.2 – Distribution spatiale de la dispersion de l'ensemble 3. L'écart-type de l'ensemble est calculé dans chaque cellule et à chaque heure. Ensuite, ces écarts-types sont moyennés (en temps) dans chaque cellule et divisés par la concentration moyenne de la cellule, ce qui rend un écart-type relatif.

(RMSE), la corrélation et le facteur de biais. On rappelle leur définition :

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - o_i\right)^2} \tag{5.1}$$

$$\mathrm{correlation} = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(o_i - \bar{o}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}\sqrt{\sum_{i=1}^{n}\left(o_i - \bar{o}\right)^2}} \tag{5.2}$$

$$\mathrm{biais} = \left|1 - \frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{y}_i}{\tilde{o}_i}\right| \tag{5.3}$$

où y est le vecteur des concentrations simulées, o le vecteur des observations correspondantes. Les deux vecteurs sont de taille n. Leurs moyennes sont notées $\bar{y}$ et $\bar{o}$. Le vecteur $\tilde{o}$ est défini comme le vecteur des observations au-dessus de 40 $\mu\mathrm{g}\cdot\mathrm{m}^{-3}$ et $\tilde{y}$ comme les concentrations simulées correspondantes.

Le tableau 5.2 présente les performances des trois ensembles, mesurées par les indicateurs précédents et sur les trois réseaux.

Pour des prévisions quotidiennes, le modélisateur est généralement capable de choisir un modèle dont les performances seront proches du meilleur modèle. L'objectif est donc de délivrer une prévision présentant des performances plus élevées que le meilleur modèle. En d'autres termes, une méthode permettant d'identifier le meilleur modèle ne serait pas satisfaisante. En conséquence, les membres de l'ensemble doivent être combinés. Sachant que la prévision d'ensemble est coûteuse en temps calcul, une combinaison de modèle est satisfaisante si elle conduit à des améliorations significatives. On considère qu'une diminution de 10% de l'erreur quadratique moyenne du meilleur modèle (c'est-à-dire 2–3 $\mu\mathrm{g}\cdot\mathrm{m}^{-3}$) est requise pour qu'une méthode d'ensemble soit intéressante. Ce seuil est arbitraire, mais il a un sens en prévision. Le meilleur modèle est souvent ajusté pour délivrer de bonnes prévisions ; il correspond à une configuration favorable trouvée par le modélisateur. Améliorer les performances d'un modèle ajusté, dans le but de diminuer l'erreur quadratique moyenne de 10%, n'est pas chose facile, surtout dans le cadre de prévision.

Tab. 5.2 – Performances des ensembles par rapport aux observations des trois réseaux. La RMSE est en $\mu g \cdot m^{-3}$, la corrélation en % et le biais en %. Les meilleurs résultats pour chaque réseau sont en gras. Les statistiques moyennes sont les moyennes des statistiques individuelles des modèles.

	Concentrations horaires			Pics journaliers		
Ensemble	RMSE	Corrélation	Biais	RMSE	Corrélation	Biais
			Réseau 1			
Ensemble 1						
Meilleur modèle	27.0	66.1	**1.8**	**22.7**	73.8	**0.1**
Statistique moyenne	29.0	63.8	11.3	24.2	71.1	2.9
Ensemble 2						
Meilleur modèle	**26.7**	**67.9**	**1.8**	23.0	**74.8**	**0.1**
Statistique moyenne	29.1	64.8	13.4	26.4	69.1	6.2
Ensemble 3						
Statistique moyenne	29.0	64.4	12.6	25.6	69.8	4.9
Plus mauvais modèle	32.1	60.8	27.1	33.5	62.2	17.2
			Réseau 2			
Ensemble 1						
Meilleur modèle	**25.7**	63.6	0.5	**21.5**	69.7	**0.1**
Statistique moyenne	26.8	60.6	7.8	22.6	67.4	2.7
Ensemble 2						
Meilleur modèle	26.3	**63.9**	**0.2**	21.6	**70.2**	0.4
Statistique moyenne	28.9	59.9	12.9	25.4	64.4	6.6
Ensemble 3						
Statistique moyenne	28.1	60.1	11.0	24.4	65.5	5.1
Plus mauvais modèle	35.1	54.4	28.7	32.1	56.7	17.3
			Réseau 3			
Ensemble 1						
Meilleur modèle	29.4	65.5	3.2	**24.9**	72.2	**0.2**
Statistique moyenne	32.5	61.6	15.3	26.5	67.8	2.9
Ensemble 2						
Meilleur modèle	**29.0**	**67.8**	**0.2**	25.1	**74.4**	0.5
Statistique moyenne	31.2	62.9	12.8	29.1	65.4	6.8
Ensemble 3						
Statistique moyenne	31.7	62.4	13.8	28.2	66.2	5.4
Plus mauvais modèle	35.8	58.8	26.0	37.5	55.4	17.7

5.3 Combinaison de modèles : méthodes et potentiels

5.3.1 Notations

On note un ensemble par le symbole $\mathcal{E}$ ou $\mathcal{E}_i$. Par exemple, $\mathcal{E}_3 = \mathcal{E}_1 \cup \mathcal{E}_2$. Un réseau est désigné par $\mathcal{N}$ ou $\mathcal{N}_i$. Le cardinal d'un réseau (nombre de stations) ou d'un ensemble (nombre de modèles) est noté $|\cdot|$. On note les concentrations simulées par un modèle $M_{t,x}$ ou $M_{m,t,x}$ (s'il s'agit du modèle n°m), où t est le pas de temps et x indique la station considérée. Les moyennes temporelle et spatiale sont notées $\overline{M}_x^t$ et $\overline{M}_t^x$ respectivement. La moyenne sur toutes les stations et pour toute la période de simulation est $\overline{M}^{t,x}$. Les observations sont notées $O_{t,x}$ et $C_{t,x}$ sont les concentrations combinées (combinaison linéaire de sorties de modèles).

5.3.2 Introduction aux méthodes de combinaison

Moyenne d'ensemble et ensemble médian

La moyenne d'ensemble est définie par

$$\mathrm{EM}_{t,x} = \frac{1}{|\mathcal{E}|} \sum_{M \in \mathcal{E}} M_{t,x} \tag{5.4}$$

Les concentrations de l'ensemble médian sont

$$\mathrm{EMD}_{t,x} = \mathrm{median}(\{M_{t,x}\}_{M \in \mathcal{E}}) \tag{5.5}$$

S'il y a un nombre pair de modèles, on considère la moyenne des deux modèles « centraux ».

Sélection de modèles

Pour chaque station, le meilleur modèle est sélectionné. Le modèle résultant est noté EB^{s} ('B' signifie « best » et 's' signifie « station »). De la même manière, sélectionner le meilleur modèle à chaque date (mais pour toutes les stations) définit le « méta-modèle » EB^{d} ('d' dénote « date »).

Méthodes avec moindres carrés

La meilleure combinaison linéaire, au sens des moindres carrés (LS, pour l'anglais, « least-squares »), est

$$\mathrm{ELS}_{t,x} = \sum_m \alpha_m M_{m,t,x} \tag{5.6}$$

où α est le vecteur qui minimise sans contrainte la somme

$$\sum_{t,x} \left[O_{t,x} - \sum_m \alpha_m M_{m,t,x} \right]^2 \tag{5.7}$$

Une version débiaisée est

$$\mathrm{EULS}_{t,x} = \overline{O}^{t,x} + \sum_m \alpha_m \left(M_{m,t,x} - \overline{M}_m^{t,x} \right) \tag{5.8}$$

où α minimise (toujours sans contrainte)

$$\sum_{t,x} \left[O_{t,x} - \overline{O}^{t,x} - \sum_m \alpha_m \left(M_{m,t,x} - \overline{M}_m^{t,x} \right) \right]^2 \tag{5.9}$$

EULS est parfois appelé superensemble [suivant Krishnamurti *et al.*, 2000].

Les poids (α) peuvent être calculés pour chaque station et pour chaque pas de temps. Les combinaisons correspondantes sont notées avec les exposants 's' (station) et 'd' (date) ; on peut ainsi obtenir $\mathrm{ELS}^{\mathrm{s}}$ ou $\mathrm{ELS}^{\mathrm{d}}$. Les moyennes sont adaptées à la nouvelle cible ; par exemple :

$$\mathrm{EULS}^{\mathrm{s}}_{t,x} = \overline{O}^{t}_{x} + \sum_{m} \alpha^{\mathrm{s}}_{m,x} \left(M_{m,t,x} - \overline{M}^{t}_{m,x} \right) \tag{5.10}$$

où le vecteur $\alpha^{\mathrm{s}}_{x} = (\alpha^{\mathrm{s}}_{1,x}, \alpha^{\mathrm{s}}_{2,x}, \alpha^{\mathrm{s}}_{3,x}, \ldots)$ minimise

$$\sum_{t} \left[O_{t,x} - \overline{O}^{x}_{t} - \sum_{m} \alpha^{s}_{m,x} \left(M_{m,t,x} - \overline{M}^{t}_{m,x} \right) \right]^{2} \tag{5.11}$$

5.3.3 Potentiel des méthodes

Dans les formules précédentes, les poids sont calculés sur la base de *toutes* les observations. Lors de prévisions opérationnelles, les poids doivent être prévus, c'est-à-dire prédits sur la base des observations *passées*. Cependant, dans cette section, les méthodes sont jugées de part leurs performances *a posteriori* (avec toutes les observations connues). Ceci permet de connaître le potentiel de ces méthodes.

Toutes les statistiques sont présentées dans le tableau 5.3.

Moyenne d'ensemble et ensemble médian

Pour chaque ensemble, les résultats de EM et EMD sont meilleurs que les statistiques moyennes de l'ensemble. Néanmoins, ils ont souvent des performances inférieures à celles du meilleur modèle. Aucune moyenne d'ensemble et aucun ensemble médian n'est associé à une RMSE inférieure à 90% de la meilleure RMSE de l'ensemble considéré. La moyenne d'ensemble et l'ensemble médian conduisent donc à de faibles performances. Ceci contredit les résultats obtenus par Delle Monache et Stull [2003]. Cela dit, cette dernière étude n'impliquait que quatre modèles, sur six jours et avec cinq stations, ce qui limite la fiabilité des résultats, comme les auteurs l'ont d'ailleurs souligné.

Sélection de modèles

Les performances de EB^{s} et EB^{d} sont satisfaisantes, spécialement sur les pics. Les RMSE sont souvent en dessous de 90% de la RMSE du meilleur modèle.

Méthodes avec moindres carrés

Tous les commentaires qui suivent sont valables à la fois pour les méthodes avec moindres carrés et pour leurs versions débiaisées. Leurs performances sont en effet suffisamment proches.

Les méthodes avec moindres carrés et avec une seule combinaison pour tout le réseau et toutes les dates apportent des améliorations significatives. La RMSE est souvent nettement en dessous de 90% de celle du meilleur modèle.

Cependant, les meilleures performances sont de loin atteintes par les méthodes avec moindres carrés par station ou par date. Sur le réseau 2, $\mathrm{EULS}^{\mathrm{d}}$ basé sur l'ensemble 3 atteint même une RMSE de $8\mu g \cdot m^{-3}$ et une corrélation de 96.3% pour les pics journaliers.

Les combinaisons basées sur l'ensemble 3 conduisent logiquement aux meilleurs résultats puisque cet ensemble inclut toutes les simulations. Les combinaisons aux moindres carrés basées sur l'ensemble 2 sont légèrement meilleures que celles associées à l'ensemble 1, ce qui pourrait

Tab. 5.3 – Performances potentielles des combinaisons de modèles, lors de comparaisons aux observations des trois réseaux. La RMSE est en $\mu g \cdot m^{-3}$, la corrélation en % et le biais en %. Pour les réseaux 1 et 3, seules les meilleures combinaisons (pour la RMSE) sont reportées. Les conclusions tirées des résultats sur le réseau 2 sont très similaires pour les deux autres réseaux.

	Concentrations horaires			Pics journaliers		
Ensemble	RMSE	Corrélation	Biais	RMSE	Corrélation	Biais
			Réseau 1			
Ensemble 1						
EULS^{d}	16.7	87.3	2.6	13.5	91.6	1.4
Ensemble 2						
EULS^{d}	16.3	87.9	2.5	13.3	91.9	1.4
Ensemble 3						
EULS^{s}	16.5	87.7	2.0	10.9	94.5	1.0
EULS^{d}	14.5	90.6	2.0	11.6	93.9	1.1
			Réseau 2			
Ensemble 1						
EM	25.9	61.9	6.3	22.0	68.7	0.7
EMD	26.4	60.9	7.7	22.1	68.0	1.0
EB^{s}	23.1	70.6	2.4	19.7	75.3	2.4
EB^{d}	24.2	67.0	2.6	19.9	74.8	2.4
ELS	23.7	68.0	0.8	18.7	78.2	2.5
EULS	23.4	68.8	0.0	18.5	78.7	3.2
ELS^{s}	16.4	86.3	0.7	12.9	90.3	1.2
EULS^{s}	16.0	86.8	0.2	12.5	90.9	1.4
ELS^{d}	17.1	84.8	0.5	12.5	90.9	1.3
EULS^{d}	16.7	85.5	0.2	12.1	91.4	1.4
Ensemble 2						
EM	25.2	64.4	5.5	23.1	70.5	4.6
EMD	25.3	64.0	4.9	23.3	69.6	4.6
EB^{s}	22.4	72.5	0.9	19.1	77.1	2.0
EB^{d}	24.0	67.3	1.3	19.6	75.6	2.1
ELS	24.3	66.2	0.7	19.6	75.8	2.7
EULS	24.0	66.9	0.4	19.4	76.2	3.4
ELS^{s}	17.3	84.6	0.9	12.8	90.4	1.0
EULS^{s}	16.9	85.3	0.1	12.3	91.2	1.4
ELS^{d}	15.9	87.1	0.3	11.4	92.4	1.1
EULS^{d}	15.4	87.9	0.1	11.0	93.1	1.2
Ensemble 3						
EM	24.9	64.2	1.3	22.3	70.7	2.7
EMD	25.7	61.5	4.5	22.2	69.2	0.7
EB^{s}	22.1	73.2	0.8	18.9	77.6	1.7
EB^{d}	23.8	68.1	1.6	19.4	76.2	2.0
ELS	23.5	68.8	0.9	18.3	79.3	2.4
EULS	23.2	69.6	0.0	18.1	79.7	3.0
ELS^{s}	15.5	87.8	0.6	10.5	93.7	0.8
EULS^{s}	15.2	88.3	0.2	10.1	94.1	1.0
ELS^{d}	11.9	93.0	0.1	8.3	96.1	0.6
EULS^{d}	11.6	93.3	0.0	8.0	96.3	0.6
			Réseau 3			
Ensemble 1						
EULS^{d}	16.9	88.3	3.1	13.9	91.9	1.4
Ensemble 2						
EULS^{d}	16.4	89.0	3.0	13.3	92.6	1.3
Ensemble 3						
EULS^{d}	15.0	90.8	2.6	11.9	94.1	1.1

être dû au nombre de membres, à la dispersion plus grande des simulations ou simplement à une configuration favorable. Les combinaisons aux moindres carrés par date réalisent généralement de meilleures performances que les combinaisons par station. Le rapport entre le nombre de stations disponibles par date et le nombre de mesures par station (sur les 127 jours) pourrait être une explication pour les concentrations horaires. Pourtant il est possible que la structure spatio-temporelle des champs calculés joue un rôle important.

À la figure 5.3, les pics journaliers d'ozone en une station représentative (du point de vue des statistiques) illustre les améliorations.

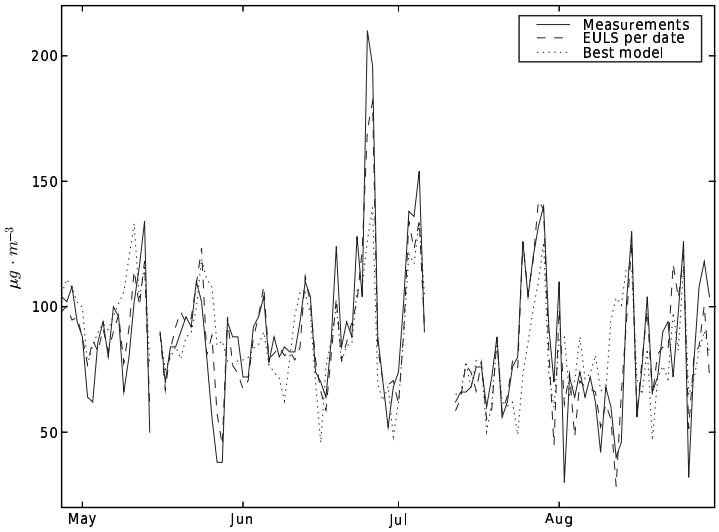


FIG. 5.3 – Pics d'ozone journaliers à la station Harwell (station du réseau 2) pour les 127 jours simulés (120 mesures sont disponibles). Le meilleur modèle est extrait de l'ensemble 1. La combinaison EULS$^{\mathrm{d}}$ est basée sur l'ensemble 1. Le meilleur modèle est associé à une RMSE de 22.0 $\mu\mathrm{g} \cdot \mathrm{m}^{-3}$ et une corrélation de 63.4%. EULS$^{\mathrm{d}}$ est associé à une RMSE de 12.1 $\mu\mathrm{g} \cdot \mathrm{m}^{-3}$ et une corrélation de 90.6%.

5.4 Prévision des combinaisons et sélection des membres

Les résultats précédents montrent un fort potentiel des méthodes aux moindres carrés. L'objectif est donc de les utiliser en prévision, c'est-à-dire de prévoir les poids associés avec chaque modèle sur la base des poids calculés pour les jours précédents. Cette démarche peut être vue comme une procédure d'assimilation de données contrainte par la structure de l'ensemble.

Sauf mention contraire, les tests qui suivent sont réalisés avec l'ensemble 1, sur le réseau 2 et pour les pics journaliers d'ozone.

5.4.1 Stabilité des poids

Puisque ELS$^{\mathrm{s}}$ and ELS$^{\mathrm{d}}$ présentent des performances prometteuses, les combinaisons peuvent être prévues à chaque station (pour une période donnée) ou à chaque pas de temps (et pour toutes les stations). De sorte à faciliter la prévision des poids, les combinaisons dans lesquelles les poids ont une faible variabilité temporelle sont particulièrement intéressantes. Il est aussi

avantageux que les poids soient spatialement robustes, c'est-à-dire qu'ils puissent être appliqués à un autre réseau ou en d'autres cellules du maillage. Avec de tels poids, le champ au sol, dans toutes les cellules, pourrait être prévu, ce qui constitue un apport important des modèles de chimie-transport 3D.

Il est intéressant de noter que (1) il existe des poids constants pour toute la période (127 jours) associés à des combinaisons efficaces (ELS^s), et (2) il existe aussi des poids uniformes (sur tout un réseau) associés avec des combinaisons performantes (ELS^d). La première question est de savoir si ces poids peuvent être prévus.

Méthodes aux moindres carrés par date

L'évolution temporelle des poids associés à ELS^d, pour trois modèles particuliers, est reportée figure 5.4. Ces poids sont extrêmement variables. Même les poids les plus élevés (en valeur absolue), qui contribuent à la majeure partie de la combinaison, sont fortement variables. Ces poids semblent donc difficiles à prévoir.

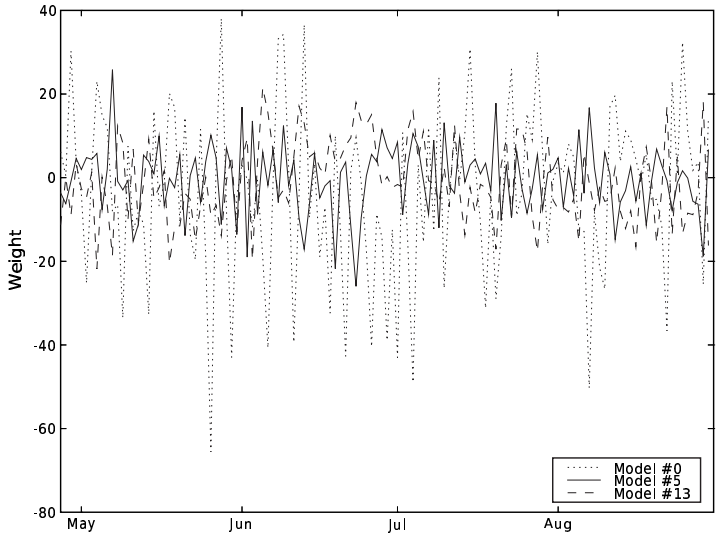


FIG. 5.4 – Évolution temporelle des trois poids les plus instables pour ELS^d (ensemble 1, réseau 2), c'est-à-dire les poids ayant le plus grand écart-type. Les poids associés aux autres modèles sont aussi fortement variables.

Une autre propriété est qu'il est difficile de reporter ces poids sur un autre réseau ou dans d'autres cellules. Appliquer les poids calculés pour le réseau 3 (ensemble 1, pics journaliers d'ozone) au réseau 2 conduit à une RMSE de 55.8 $\mu g \cdot m^{-3}$. Il s'agit du report le moins favorable puisque les deux réseaux contiennent des stations de natures différentes et leurs extensions spatiales diffèrent fortement. Une expérience plus favorable consiste à calculer les poids sur le réseau 2 (Europe) et à les appliquer sur le réseau 3 (France). La RMSE résultante est 24.6 $\mu g \cdot m^{-3}$ (corrélation de 74.7%) ce qui est raisonnable puisque le meilleur modèle a une RMSE de 24.9 $\mu g \cdot m^{-3}$ et une corrélation de 72.2%. Il existe une expérience encore plus favorable. Tout comme le réseau 2, le réseau 1 possède des stations européennes mais inclut aussi des stations régionales et urbaines, y compris des stations du réseau 3. Le réseau 3 est donc plus proche du réseau 1 que du réseau 2. Appliquer les poids calculés sur le réseau 1 au réseau 3 donne de meilleures performances, avec une RMSE de 17.4 $\mu g \cdot m^{-3}$ et une corrélation de 87.1%.

Il s'agit d'un résultat encourageant qui tend à montrer qu'il est possible d'appliquer certains poids dans des cellules sans observations.

Méthodes aux moindres carrés par station

Les poids (pour ELS^s) associés à chaque modèle sont très variables sur le réseau (réseau 2), comme le montre la figure 5.5. De plus, il n'existe pas de sous-ensemble de stations sur lequel les poids sont similaires. Ceci n'est pas étonnant puisqu'associer un unique jeu de poids par modèle (à toutes les stations et à toutes les dates) ne conduit pas aux améliorations les plus remarquables (tableau 5.3, ELS).

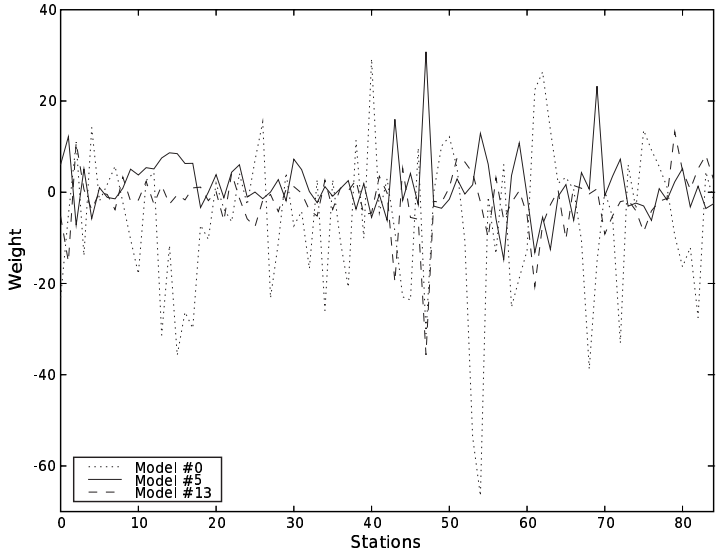


FIG. 5.5 – Distribution, sur les 85 stations du réseau 2, des trois poids les plus instables pour ELS^s (ensemble 1), c'est-à-dire les poids associés aux plus forts écarts-types. Les poids des autres modèles sont aussi fortement variables.

5.4.2 Report des poids d'un jour à l'autre

Une méthode évidente pour la prévision des poids est de reporter les poids calculés sur les jours précédents. Dans cette section, les statistiques sont calculées sur les 96 derniers jours de sorte à ce que les 30 premiers jours puissent servir de période d'apprentissage. Une période d'apprentissage de n jours se réfère aux n jours précédant le jour à « prévoir ». Il s'agit donc d'une fenêtre d'apprentissage glissante.

Méthodes aux moindres carrés par station

Calculer les poids par station sur une période d'apprentissage de 22 à 30 jours (22 est un minimum puisqu'il y a 22 poids) ne permet pas d'améliorer les prévisions. Les meilleurs résultats avec cette méthode sont obtenus avec une période d'apprentissage de 30 jours et conduisent à une RMSE de 40.7 $\mu g \cdot m^{-3}$. Augmenter la période d'apprentissage doit aider (ELS^s fonctionne), mais la période de test (96 jours dans le cas étudié) deviendrait trop courte pour que les résultats soient significatifs. On peut cependant indiquer qu'une période d'apprentissage

de 60 jours permet d'obtenir, sur les 36 jours simulés restant, une RMSE de 22.3 $\mu g \cdot m^{-3}$ (contre 21.6 $\mu g \cdot m^{-3}$ pour le meilleur modèle) et une corrélation de 76.5% (meilleur modèle : 74.6%). La conclusion est que cette méthode n'est pas satisfaisante pour les simulations étudiées. Néanmoins, d'autres tests, sur de plus longues périodes, sont nécessaires.

Méthodes aux moindres carrés par date

À chaque date, les poids sont les mêmes pour toutes les stations. Ils sont calculés sur la base d'une période d'apprentissage allant de 1 jour à 30 jours. La figure 5.6 montre que cette méthode fonctionne bien avec une période d'apprentissage d'environ 5–7 jours. Des périodes d'apprentissage plus longues n'améliorent pratiquement pas les résultats. Les performances sont proches des celles de ELS.

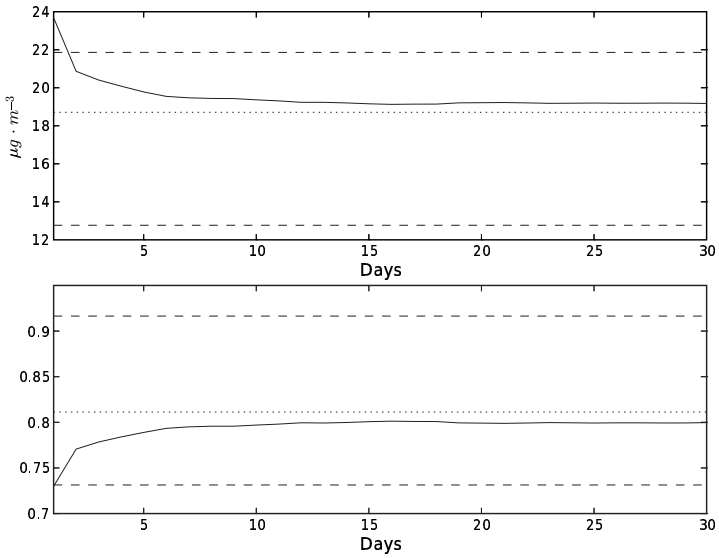


FIG. 5.6 – RMSE (graphique supérieur) et corrélation (graphique inférieur) pour une combinaison reposant sur des poids (identiques pour toutes les stations) calculés à chaque pas de temps par une optimisation aux moindres carrés sur une période d'apprentissage de x jours (abscisse). Les lignes discontinues sont les performances du meilleur modèle de l'ensemble et de ELS^{d} (ensemble 1). La ligne en pointillés est la performance de ELS.

Avec une période d'apprentissage de 30 jours, la RMSE de la combinaison prévue est de 19.2 $\mu g \cdot m^{-3}$ (meilleur modèle : 21.9 $\mu g \cdot m^{-3}$) et la corrélation est de 80.0% (meilleur modèle : 73.3%). Le critère sur la RMSE (en dessous de 90% de la RMSE du meilleur modèle) est donc rempli (pour l'ensemble 1 et le réseau 2). Ce n'est pas le cas avec tous les ensembles et tous les réseaux, comme le montre le tableau 5.4. Cependant, on note toujours des améliorations significatives.

Un point important pour expliquer ces améliorations est l'évolution temporelle des poids. La figure 5.4 montre de fortes variations qui expliquent qu'une période d'apprentissage d'un jour a peu de chance de fonctionner. En effet, la figure 5.6 indique bien qu'une période d'apprentissage d'un jour donne de faibles performances (RMSE à 23.7 $\mu g \cdot m^{-3}$ et corrélation à 73%). Les coefficients calculés sur une période d'apprentissage de 30 jours sont bien plus stables, ce qui est illustré par la figure 5.7.

TAB. 5.4 – Performances sur les pics journaliers d'ozone sur les derniers 96 jours simulés pour ELS^d, ELS, le meilleur modèle (de l'ensemble) et la combinaison avec les poids de la méthode aux moindres carrés calculés sur une période d'apprentissage de 30 jours (précédant le jour de la prévision). Dans chaque colonne, la RMSE en $\mu g \cdot m^{-3}$ est suivie par la corrélation en %.

Ensemble	ELS^d	ELS	Meilleur modèle	Prévision
		Réseau 1		
Ensemble 1	14.1 – 91.7	19.6 – 83.3	22.4 – 78.0	20.5 – 81.7
Ensemble 2	13.9 – 92.0	20.5 – 81.5	22.4 – 78.1	21.3 – 80.0
Ensemble 3	12.0 – 94.1	19.2 – 84.0	22.4 – 78.1	20.2 – 82.2
		Réseau 2		
Ensemble 1	12.8 – 91.6	18.7 – 81.1	21.9 – 73.1	19.2 – 80.0
Ensemble 2	11.6 – 93.1	19.6 – 79.0	21.9 – 73.8	20.4 – 77.2
Ensemble 3	8.4 – 96.4	18.2 – 82.3	21.9 – 73.8	19.0 – 80.4
		Réseau 3		
Ensemble 1	14.6 – 91.8	21.1 – 81.8	24.0 – 76.4	21.8 – 80.6
Ensemble 2	13.9 – 92.5	21.1 – 81.9	23.9 – 76.6	22.1 – 80.0
Ensemble 3	12.4 – 94.1	20.2 – 83.5	23.9 – 76.6	21.2 – 81.6

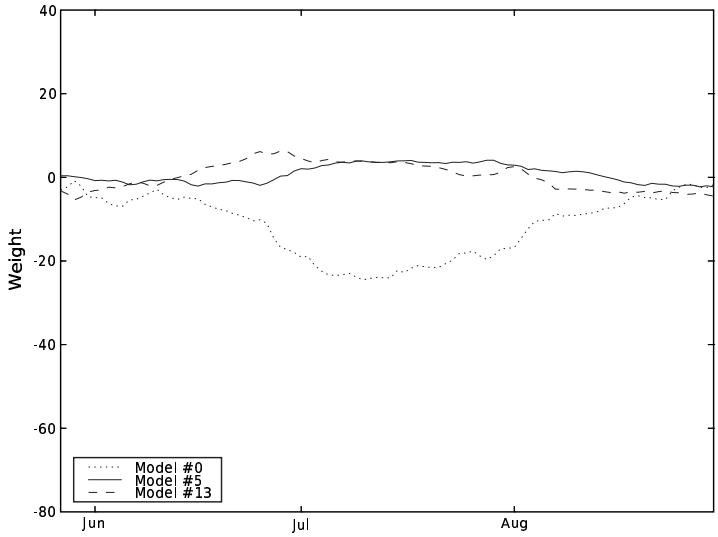


FIG. 5.7 – Évolution temporelle de trois poids calculés avec une période d'apprentissage de 30 jours (précédant chaque prévision), pour l'ensemble 1 et le réseau 2. La figure peut être comparée à la figure 5.4 qui présente des coefficients bien plus variables. Les deux figures ont les mêmes échelles en ordonnée.

Concentrations horaires

Les prévisions horaires peuvent aussi être améliorées grâce à des combinaisons « apprises » sur les 30 jours précédant la prévision, et estimées par date, comme à la section 5.4.2. Afin de prévoir les poids d'une heure h, seules les concentrations et les observations pour l'heure h, lors de la période d'apprentissage, sont retenues. Inclure toutes les concentrations horaires diminue les performances.

Les résultats sont rassemblés dans le tableau 5.5. Les performances sont significativement augmentées, en particulier sur les réseaux 1 et 2. On peut noter que ces performances sont similaires à celles de ELS (dont on peut considérer que la période d'apprentissage est toute la période de simulation).

Tab. 5.5 – Performances pour les concentrations d'ozone horaires sur les 96 derniers jours pour ELS^{d}, ELS, le meilleur modèle (de l'ensemble) et la combinaison aux moindres carrés estimée avec une période d'apprentissage de 30 jours précédant la prévision. Les poids associés à une heure h sont estimés uniquement avec les concentrations et observations aux heures h de la période d'apprentissage. Dans chaque colonne, la RMSE en $\mu\text{g} \cdot \text{m}^{-3}$ est suivie par la corrélation en %.

Ensemble	ELS^{d}	ELS	Meilleur modèle	Prévision ELS^{d}
		Réseau 1		
Ensemble 1	17.2 – 87.3	22.9 – 75.9	26.8 – 68.4	22.7 – 76.6
Ensemble 2	16.8 – 87.9	24.0 – 73.2	26.7 – 69.9	23.3 – 75.2
Ensemble 3	14.9 – 90.6	22.7 – 76.5	26.7 – 69.9	22.5 – 77.1
		Réseau 2		
Ensemble 1	17.3 – 85.5	23.9 – 70.1	25.9 – 65.6	23.6 – 71.0
Ensemble 2	16.1 – 87.7	24.6 – 67.9	26.7 – 65.7	24.6 – 68.0
Ensemble 3	11.9 – 93.4	23.6 – 71.0	25.9 – 65.7	23.4 – 71.5
		Réseau 3		
Ensemble 1	17.2 – 88.4	23.3 – 77.5	28.7 – 68.0	22.9 – 78.4
Ensemble 2	16.7 – 89.2	24.9 – 73.7	28.5 – 69.9	23.7 – 76.8
Ensemble 3	15.3 – 91.0	22.9 – 78.4	28.5 – 69.9	22.8 – 78.7

5.4.3 Apprentissage statistique

Appliquer une combinaison optimale calculée sur une période d'apprentissage peut être performant (voir la section précédente), mais des algorithmes plus sophistiqués ont été développés dans le domaine de l'apprentissage statistique[1]. Un algorithme classique est, par exemple, l'algorithme de descente par gradient[2] [Cesa-Bianchi *et al.*, 1996]. Dans le cas présent, on applique cette méthode indépendamment à chaque station. L'objectif est de minimiser une fonction de perte (terminologie de l'apprentissage statistique) définie par

$$L_t(\alpha_t) = \left(\sum_m \alpha_{m,t} M_{m,t} - O_t \right)^2 \tag{5.12}$$

[1] « machine learning » en anglais

[2] « gradient descent algorithm for on-line regression » en anglais

Les poids $\alpha_{t-1} = (\alpha_{1,t-1}, \alpha_{2,t-1}, \alpha_{3,t-1}, \ldots)$ sont mis à jour ainsi :

$$\alpha_t = \alpha_{t-1} - \eta L'_{t-1}(\alpha_{t-1}) \tag{5.13}$$

η est un taux d'apprentissage. Les résultats y sont sensibles (selon des tests qui ne sont pas reportés ici). On prend $\eta = 5 \cdot 10^{-7}$ pour les ensembles 1 et 2, et $\eta = 2.5 \cdot 10^{-7}$ pour l'ensemble 3. Les résultats sont stables dans le voisinage de ces valeurs (pour des variations de $\pm 50\%$). Les poids initiaux sont pris égaux à $\frac{1}{N}$ où N est le nombre de modèles – cela correspond à la moyenne d'ensemble.

Le tableau 5.6 montre les résultats de cette méthode de descente. Les performances (cinquième colonne) sont légèrement meilleures que les performances de la méthode aux moindres carrés avec les poids calculés à chaque date (troisième colonne ; voir aussi la section 5.4.2). L'algorithme d'apprentissage statistique réussit là où le report des coefficients du jour précédent ne fonctionne pas (section 5.4.2). Sachant qu'il existe plusieurs variantes de tels algorithmes d'apprentissage (avec des mises à jour différant de l'équation 5.13), il y a bien là une piste prometteuse pour des améliorations plus importantes.

Tab. 5.6 – Performances sur les pics journaliers d'ozone sur les 96 derniers jours pour ELS$^{\mathrm{d}}$, ELS$^{\mathrm{d}}$ avec des poids prédits (comme à la section 5.4.2 et dans le tableau 5.4), le meilleur modèle (de l'ensemble) et la combinaison prévue par l'algorithme de descente par gradient. Les 30 premiers jours constituent une période d'apprentissage minimale. C'est la raison pour laquelle les résultats diffèrent légèrement de ceux du tableau 5.4 (dont la cinquième colonne correspond à la troisième colonne de ce tableau). Dans chaque colonne, la RMSE en $\mu\mathrm{g} \cdot \mathrm{m}^{-3}$ est suivie par la corrélation en %.

		Prévision		Descente
Ensemble	ELS$^{\mathrm{d}}$	ELS$^{\mathrm{d}}$	Meilleur modèle	par gradient
		Réseau 1		
Ensemble 1	13.8 – 91.9	20.3 – 81.5	22.4 – 77.5	20.1 – 82.1
Ensemble 2	14.2 – 91.4	21.0 – 80.0	22.4 – 77.7	19.5 – 83.0
Ensemble 3	11.2 – 94.7	20.0 – 82.1	22.4 – 77.7	19.6 – 83.0
		Réseau 2		
Ensemble 1	13.0 – 91.1	18.8 – 80.0	21.8 – 72.5	18.8 – 80.6
Ensemble 2	13.1 – 90.9	20.2 – 76.8	21.8 – 73.5	18.2 – 81.7
Ensemble 3	10.6 – 94.2	18.8 – 80.2	21.8 – 73.5	18.2 – 81.6
		Réseau 3		
Ensemble 1	14.7 – 91.7	21.8 – 80.6	24.2 – 76.2	21.7 – 81.0
Ensemble 2	15.0 – 91.4	22.1 – 80.1	24.1 – 76.4	22.7 – 82.9
Ensemble 3	12.6 – 94.0	21.3 – 81.6	24.1 – 76.4	21.0 – 82.3

5.4.4 Sélection de modèles

Dans le tableau 5.4, l'ensemble 1 montre de meilleures performances que l'ensemble 2, même si l'ensemble 1 a moins de membres (22 contre 32) et qu'il est moins dispersé. Du fait du coût des calculs, il est utile de réduire le nombre de modèles à inclure dans un ensemble. La figure 5.8 montre les performances de ELS$^{\mathrm{d}}$ fonctions du nombre de modèles, où les membres de l'ensemble 3 sont inclus un par un. Même si l'impact de modèles supplémentaires décroît avec le nombre de modèles déjà présents, les performances augmentent toujours significativement.

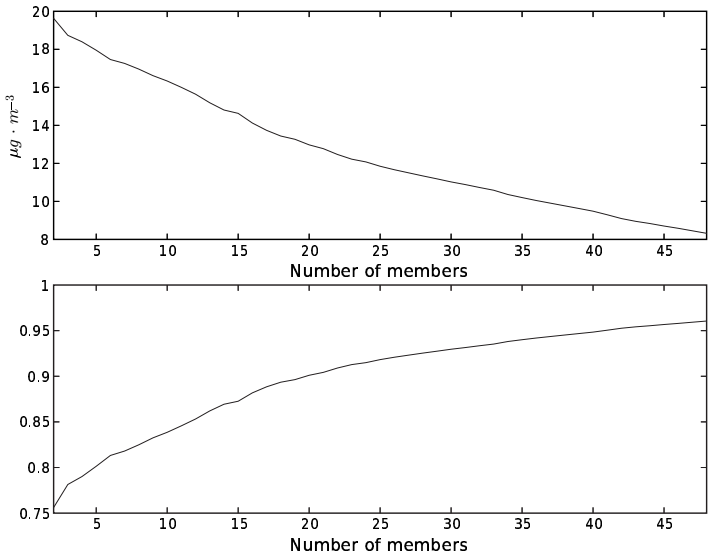


FIG. 5.8 – Performances (RMSE en haut, corrélation en bas) de ELS^d fonction du nombre de modèles dans l'ensemble. Les modèles sont pris dans l'ensemble 3.

Une autre question concerne l'apport de chaque modèle dans l'amélioration des performances. À la figure 5.9, les contributions de plusieurs modèles de l'ensemble 2 (à quatre sous-ensembles basés sur l'ensemble 1, et de tailles 5, 10, 15 et 20) sont reportées. Des différences entre les contributions se distinguent principalement pour les petits ensembles. De même, les différences entre contributions de plusieurs modèles de l'ensemble 1 (à quatre sous-ensembles basés sur l'ensemble 2, et de tailles 5, 10, 15 et 20) sont aussi faibles pour des ensembles de base assez grands (les différences sont même plus estompées que dans le cas précédent). De plus, la corrélation entre les RMSE des modèles et leur contribution à la RMSE de la combinaison vaut moins de 30%. Ceci signifie que les meilleurs modèles n'apportent pas forcément les meilleures contributions à l'ensemble.

On ne peut pas clairement identifier une raison pour laquelle les combinaisons basées sur l'ensemble 2 sont moins performantes que celles basées sur l'ensemble 1. L'ensemble 2 inclut des simulations avec de multiples changements (voir section 5.2) et seulement 5 changements sont impliqués, ce qui pourrait expliquer une certaine pauvreté, en comparaison de l'ensemble 1 qui implique 21 changements.

5.5 Conclusion

Le système de simulation Polyphemus a la capacité de générer un ensemble de prévisions avec une grande dispersion dans les concentrations simulées et avec un grand nombre de modèles. Les combinaisons optimales de modèles montrent un fort potentiel. Alors que la moyenne d'ensemble et l'ensemble médian n'améliorent guère les prévisions, les résultats peuvent être fortement améliorés par des combinaisons linéaires avec des poids optimaux.

On a vu que les poids calculés sur un réseau ne s'appliquent pas forcément à un autre réseau et donc en d'autres points de grille. Cette faible robustesse spatiale des poids devrait être spécifiquement étudiée car les prévisions sur tout un domaine sont une contribution importante des modèles de chimie-transport 3D.

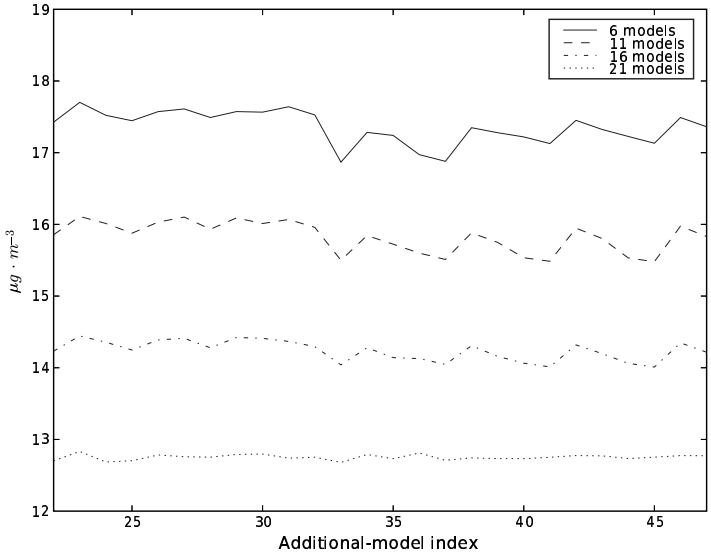


FIG. 5.9 – Quatre ensembles sont construits avec les 5, 10, 15 et 20 premiers membres de l'ensemble 1. Un modèle de l'ensemble 2 (abscisse) est ensuite ajouté à ces ensembles et la RMSE de ELS^d (ordonnée) est calculée. 26 modèles de l'ensemble 2 (tous les modèles qui sont dans l'ensemble 2 sans être dans l'ensemble 1) sont inclus de la sorte. Ceci mesure la contribution de chaque modèle à la performance de la combinaison.

Les prévisions quotidiennes requièrent la prédiction des poids des combinaisons optimales. Les poids sont très instables d'un jour à l'autre ou d'une station à l'autre. Des poids plus stables ont été proposés avec des combinaisons aux poids constants sur une période de 30 jours et sur toutes les stations du réseau. Ces poids peuvent être raisonnablement prédits et les combinaisons associées permettent des améliorations significatives des performances sur les pics journaliers et sur les concentrations horaires. Une diminution d'environ 10% de la RMSE est atteinte sur les pics journaliers. Les concentrations horaires bénéficient d'améliorations encore plus fortes.

De plus, les algorithmes d'apprentissage statistique semblent une voie prometteuse qui ne nécessite pas l'introduction de poids identiques pour les nombreuses stations d'un réseau d'observation. L'algorithme de descente par gradient donne de bons résultats lorsqu'il est appliqué station par station, là où appliquer des poids calculés à chaque station sur une période d'apprentissage de 30 jours ne fonctionne pas.

Un ensemble avec moins de membres et une dispersion plus faible qu'un autre ensemble peut être associé à des combinaisons supérieures. La sélection de modèles a alors été abordée. L'ajout de modèles apporte toujours des améliorations, mais elles ne sont que peu liées aux performances individuelles des modèles.

Des études ultérieures sont nécessaires pour analyser les « bons » ensembles. Des sources d'incertitude supplémentaires pourraient être introduites. Les prévisions d'ensemble météorologiques et des simulations Monte Carlo sur les autres données incertaines sont des étapes *a priori* utiles de ce point de vue. Le coût des calculs sera alors un point crucial. Des méthodes pertinentes sont requises pour l'introduction de méthodes Monte Carlo couplées à des changements discrets (dans la formulation du modèle, comme ce qui est fait dans ce chapitre avec des changements dans les paramétrisations physiques et dans les approximations numériques).

Un travail à effectuer réside manifestement dans la prédiction des poids. Comme ce chapitre

le montre, le potentiel de combinaisons de modèles est très élevé et il est bien plus élevé que celui des combinaisons prédites ici. Des méthodes d'apprentissage statistique pourraient aider à approcher ce potentiel.

La prévision d'ensemble permet aussi de délivrer des prévisions probabilistes. Ce serait un apport d'information notable sur les prévisions. Cela aiderait à estimer l'incertitude des prévisions et permettrait du même coup une intégration raisonnable des modèles de chimie-transport, par exemple pour l'estimation du risque.

Conclusion

Bilan

Cette thèse propose d'abord une estimation de l'incertitude attachée aux modèles de chimie-transport. Les sources d'incertitude étudiées sont :

- Les schémas numériques. Des tests numériques ont montré notamment l'influence du pas de temps et du schéma d'advection. Il ne s'agit cependant pas de l'incertitude prépondérante.
- La formulation du modèle. Des changements notamment dans les paramétrisations physiques, mais aussi dans la discrétisation du modèle et les choix de jeux de données, ont mis en évidence son impact. Il s'agit de la plus grande incertitude identifiée par les études de cette thèse.
- Les données d'entrée du modèle, à l'exclusion (principalement) des données météorologiques et des constantes des réactions chimiques.

L'incertitude première réside dans les paramétrisations physiques. L'incertitude relative est évaluée à 17%[3]. Les données d'entrée sont associées à une incertitude relative d'environ 8%. Les deux estimations sont des bornes inférieures. Les schémas numériques ont moins d'impact sur les simulations. Les incertitudes étant estimées par des écarts-types, il faut considérer que les concentrations les plus probables sont entre la valeur moyenne moins 17% et la valeur moyenne plus 17% (dans le cas de perturbations sur les paramétrisations physiques). Il s'agit d'une incertitude minimale déjà forte. Le cumul des incertitudes (numériques, dans la formulation du modèle, dues à toutes les données d'entrée) est bien sûr plus élevé, et suggère des estimations de l'incertitude pour la plupart des études. En effet, les travaux en qualité de l'air analysent souvent des effets fins (impact d'émissions, ajustement de paramètres physiques, détection de seuils d'alerte, etc.) et l'incertitude est potentiellement plus grande que les effets observés. Hanna *et al.* [1998]; Beekmann et Derognat [2003] donnent deux bons exemples de tels travaux où l'impact de réductions d'émissions est comparé à l'incertitude.

Pour dépasser les limitations de l'incertitude, plusieurs approches sont possibles, parmi lesquelles la modélisation inverse et la prévision d'ensemble. Cette thèse prépare la modélisation inverse d'émissions à l'échelle continentale grâce à une étude de sensibilité détaillée. Dans une telle application, l'incertitude joue un rôle important. Une expérience de modélisation inverse avec prise en compte de l'incertitude a déjà été réalisée à l'échelle régionale [Quélo *et al.*, 2005] : la sensibilité des émissions optimisées à certains paramètres incertains a été estimée.

En prévision, les modèles sont ajustés de sorte à minimiser l'écart aux observations. Pour réduire l'erreur de prévision, il est possible de tirer parti de l'incertitude en combinant plusieurs modèles. Dans cette thèse, des méthodes de prévision d'ensemble, plus élaborées que les moyennes d'ensemble classiques, donnent des résultats encourageants, avec des réductions

[3]Pour la concentration d'ozone, qui est la variable de sortie étudiée dans cette thèse. De plus, les travaux ayant été menés à l'échelle continentale, sur des périodes allant d'une semaine à quatre mois, le bilan chiffré est bien sûr indicatif.

significatives des erreurs de prévision.

Prolongements

Des prolongements « directs » à cette thèse sont les mêmes études avec des variations pertinentes. Par exemple, les études se sont focalisées sur l'ozone, mais l'estimation de l'incertitude, l'amélioration de prévisions ou la modélisation inverse peuvent concerner d'autres polluants (NO_2, SO_2, etc.). L'étude des aérosols peut aussi bénéficier de tels travaux. Pour ces derniers, l'incertitude due aux approximations numériques est vraisemblablement plus forte. De plus, la physique des aérosols étant encore mal connue, de nombreuses incertitudes résident dans les processus eux-mêmes.

Certaines études gagneraient à être menées sur de plus longues périodes, afin d'accroître leur représentativité. Par exemple, des comportements spécifiques s'observent en hiver, et, dans cette thèse, les périodes de simulation sont principalement en été.

Un prolongement important est l'évaluation fine de l'incertitude due aux champs météorologiques. Les prévisions d'ensemble de l'ECMWF (ensembles de 50 membres) peuvent permettre de construire une étude sérieuse.

Concernant la prévision, le système Polyphemus est en phase de test pour inclusion sur la plate-forme Prév'air (plate-forme de prévision opérationnelle de la qualité de l'air, opérée par l'INERIS – `http://www.prevair.org/`). La plate-forme est une opportunité intéressante notamment pour juger des performances de méthodes d'ensemble.

Perspectives

Parmi les perspectives, il y a la caractérisation plus fine de l'incertitude. En particulier, l'incertitude *a posteriori* doit être évaluée. Il s'agit d'estimer la densité de probabilité des concentrations, connaissant les observations. Pour cela, un cadre bayésien doit être proposé. Une difficulté réside dans la description de la vraisemblance des modèles.

Concernant la prévision, les méthodes de combinaison de modèles doivent être approfondies. Il peut s'agir de méthodes d'apprentissage statistiques ainsi qu'il en est donné un aperçu dans cette thèse. Le cadre bayésien peut là aussi intervenir, via l'utilisation de moyennes bayésiennes [Hoeting *et al.*, 1999]. La comparaison entre les performances de ces méthodes et celles de l'assimilation de données « traditionnelle » (assimilations variationnelle et séquentielle) est nécessaire. Pour ce qui concerne les moyennes bayésiennes, un point important est la structure de l'ensemble. Cette dernière doit être contrôlée, par exemple pour conserver une grande dispersion. L'objectif ultime est la génération d'ensembles décrivant précisément l'incertitude, c'est-à-dire dont les densités de probabilité sont fiables, ce qui est une tâche particulièrement ardue.

Enfin, une question ouverte réside dans la combinaison de la prévision d'ensemble et de l'assimilation de données (variationnelle ou séquentielle). L'assimilation pourrait être effectuée sur chaque membre de l'ensemble, ou seulement sur un membre de référence (à déterminer) dont les mises à jour (issues de la procédure d'assimilation) seraient appliquées aux autres membres (par exemple, des émissions optimisées ou des conditions initiales corrigées). Il y a là un cadre mathématique à construire et des méthodes à en déduire. La structure de l'ensemble peut aussi être une information utile dans la procédure d'assimilation de données.

De ce point de vue, la structure de l'erreur mérite notamment d'être étudiée. Elle permettrait d'approcher de manière satisfaisante les matrices de covariance d'erreur de l'état. Elle rendrait possible l'ajout de l'erreur modèle dont la forme est actuellement largement inconnue. Il s'agit d'une alternative à d'autres approches comme la paramétrisation de processus non résolus.

Les développements méthodologiques esquissés précédemment ainsi que les simulations à venir seront accompagnés d'une explosion des temps de calcul. Dans le tableau 5.7, plusieurs avancées prévisibles sont reportées avec un coût calcul estimé. La figure 5.10 illustre les applications qu'il sera possible de cumuler, en supposant un doublement de la puissance de calcul tous les deux ans. Étant donné la puissance nécessaire aux calculs les plus complexes, on comprend bien que les perspectives seront atteintes en s'appuyant à la fois sur des perfectionnements purement numériques et sur des progrès méthodologiques.

Tab. 5.7 – Développements à venir et estimation des coûts associés.

Développement	Coût (facteur)	Commentaire
Discrétisation verticale	2	Sans extension en stratosphère
Aérosols	3	
Résolution horizontale	25	Passage à 0.1°
Assimilation de données	90	15 itérations (assimilation variationnelle)
Ensemble	400	Assimilation sur tous les membres
Impact	10	10 scénarios d'émission

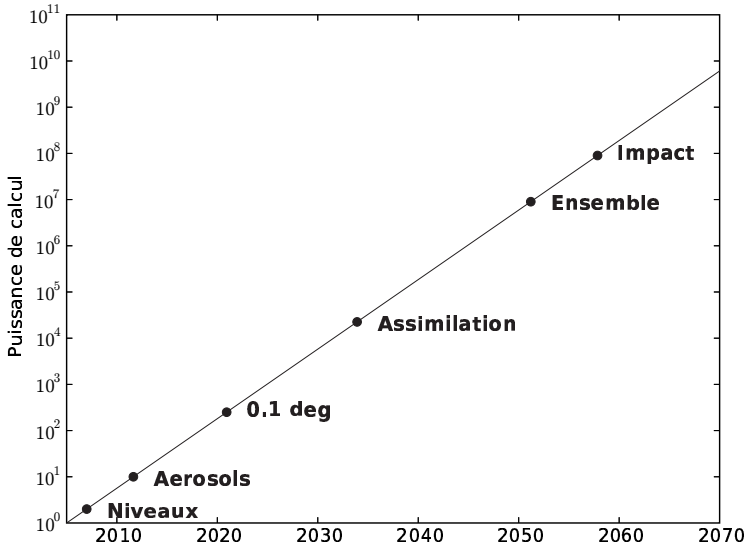


Fig. 5.10 – Nouvelles applications (cumulées) menées à temps de calcul constant en fonction du temps, en supposant un doublement de la puissance de calcul tous les deux ans. L'augmentation de la puissance des processeurs diminue, mais elle devrait être compensée par la généralisation du calcul distribué.

Annexe A

Data processing and parameterizations in atmospheric chemistry and physics : the AtmoData library

Dans le contexte de la chimie atmosphérique, la recherche s'appuie sur des modèles complexes qui nécessitent la manipulation de données multidimensionnelles et la gestion de paramétrisations physiques. Une bibliothèque orientée objet permet d'apporter des structures de données adaptées et de rassembler un nombre extensible de paramétrisations. La bibliothèque AtmoData, écrite en C++, illustre cette approche. Les structures qu'elle définit (pour des données associées à des coordonnées) et les paramétrisations qu'elle propose sont décrits dans cet appendice. La conception de la bibliothèque contribue à la qualité des codes et à leur pérennité. Elle permet aussi de partager facilement les développements.

Sommaire

Cette annexe est constituée de
Mallet, V. et Sportisse, B. (2005b). Data processing and parameterizations in atmospheric chemistry and physics : the AtmoData library. Rapport technique 12, CEREA

A.1 Introduction

Modeling activities in atmospheric chemistry and physics have been supporting the development of complex models dealing with many different applications, from research to operational platforms. The need for well designed software is therefore of high interest to help the researchers minimizing and sharing their development efforts and transferring efficiently their work to operational actors. The library AtmoData[1] was devised to be a valuable tool to this respect. It brings facilities for data processing in the field and a set of state-of-the-art physical parameterizations. Notice that it does not depend on the models to be used and, as such, may be widely distributed. It should be viewed as a toolbox proposing generic facilities such as Blas [Lawson *et al.*, 1979] or Lapack [Anderson *et al.*, 1999] for linear algebra.

This paper first describes the context for which the library AtmoData was undertaken. The second section shows the advantages of a library over other software frameworks. The overall design of AtmoData is explained in the third section. In the following sections, the content and the abilities of AtmoData are detailed. Finally the next steps in the development of the library are drawn.

A.2 Context

Most applications in atmospheric chemistry rely on the so-called 3D chemistry-transport models (CTM). Many CTMs are now available and widely used: Chimere [Schmidt *et al.*, 2001], CMAQ [Byun et Ching, 1999], DEHM [Christensen, 1997], EMEP [Simpson *et al.*, 2003], Eurad [Hass, 1991], Lotos [Builtjes, 1992], Polair3D [Boutahar *et al.*, 2004], among other valuable models. These models are computer codes written to solve at least the gas-phase chemistry-transport equation:

$$\frac{\partial c_i}{\partial t} + \operatorname{div}(V c_i) = \operatorname{div}\left(\rho K \cdot \nabla \frac{c_i}{\rho}\right) + \chi_i(c,t) + E_i - D_i \tag{A.1}$$

where:

- c, the unknown, is the vector of pollutants concentrations and c_i its i-th component;
- V is the wind velocity (three-dimensional vector);
- ρ is the air density;
- K is the 3×3 diffusion matrix, $-\operatorname{div}\left(\rho K \cdot \nabla \frac{c_i}{\rho}\right)$ being a parameterization for the turbulent diffusion flux;
- $\chi_i(c,t)$ stands for the chemistry (production and loss due to chemical reactions), notice its dependence on c that couples the equations of all individual species;
- E is the vector of emission rates;
- D is the vector of deposition fluxes.

The transport is modeled by the advection term $\operatorname{div}(V c_i)$ and the diffusion term $-\operatorname{div}(K \cdot \nabla c_i)$. The chemical term $\chi_i(c,t)$ may be very complex and there still lies important research issues (e.g., for multi-phase processes and aerosol dynamics).

[1] AtmoData is freely available at `http://www.enpc.fr/cerea/atmodata/` under the GNU General Public License.

However the transport terms and the term $\chi_i(c,t)$ as coming from widely used chemical mechanisms (e.g., photochemical mechanisms for ozone such as RACM – Stockwell *et al.* [1997] –, CBM IV – Gery *et al.* [1989] – or SPARC – Carter [1990]) may be satisfactorily discretized and integrated in time. Many numerical schemes are now dedicated to solve the chemistry-transport equation (A.1) and they are often fairly accurate as compared to the reliability of the data introduced in the equation (such as the diffusion matrix K). One may refer to Verwer *et al.* [1998] to get details about the relevant numerical schemes.

A key issue to ensure the quality of the simulations is to use reliable input data, the input data being all the values involved in equation (A.1) but the vector of concentrations c. A few input fields are simply provided by other models such as meteorological models (which notably provide the wind fields) or chemistry-transport model running at a higher scale (to provide boundary conditions). Interpolation of these fields is usually sufficient. There are other fields that have to be computed with physical parameterizations. For instance, the diffusion matrix K is assumed to be diagonal and its third diagonal term K_{zz}, which determines the vertical diffusion, is usually computed with a parameterization [e.g., Troen et Mahrt, 1986] involving raw meteorological fields (provided by the meteorological model) and additional data such as land use coverage.

To generate K_{zz} or any other input data that is computed according to parameterizations, a suited tool is welcome. First, the large amount of *multidimensional data* (five-dimensional fields are common because of the chemical species that add an extra dimension to the 3D space and the time) requires dedicated structures and functionalities. Second, the *parameterizations* themselves should be properly implemented and available for various applications. It makes the simulation process safer and more efficient.

AtmoData is a solution for both data processing and the management of the parameterizations needed in atmospheric chemistry. To satisfy the requirements previously described, AtmoData was designed as a library.

A.3 The need for a library

A tool to fulfill the needs previously shown could be a single program, a set of programs, a library or a combination of these solutions. Notice that a library is useless without programs to call it. However the tool itself may be:

1. a stand-alone library: the user has thus to write its own programs on top of the library.
2. end-user programs: one then uses interfaces (configuration files or graphical user interfaces) to process and generate the data. Notice that a CTM may be such an end-user program. Actually this is a common case since most CTMs compute themselves the data they need.

The review of the advantages brought by a library is clearly in favor of this solution. The advantages relevant in the current context are described below.

A.3.1 Data structures

A library is the only solution that may provide data structures. A program including this library benefits from the framework designed to handle the multidimensional data involved in atmospheric chemistry. Computer languages dedicated to scientific computation such as Fortran 90 provide data structures. Nevertheless other computer languages, that are not specially designed for scientific applications, have not even a matrix structure. For instance, C++ and Python are used for scientific applications with libraries that are not part of the language;

one may cite the Matrix Template Library[2] [Siek et Lumsdaine, 1999] in C++ or Numarray[3] [Greenfield *et al.*, 2003] in Python.

Moreover a data structure may be more sophisticated than a vector or a matrix. A structure dedicated to data processing in atmospheric chemistry has its own requirements such as handling multidimensional data, dealing with gridded data, performing input/output operations in formats used in the field (e.g. NetCDF and Grib). The implementation of such a structure naturally leads to a library.

A.3.2 Parameterizations

Besides data processing, its is useful to gather the parameterizations in a library. One may split the parameterizations in two categories:

- the basic parameterizations: they are simple parameterizations such as the formula to compute the relative humidity from the specific humidity, the temperature and the pressure.
- the complex parameterizations: they are specific parameterizations involving more computations and often some expertise. Examples may be the parameterizations for the vertical diffusion [Louis, 1979; Troen et Mahrt, 1986] or the deposition velocities [Wesely, 1989; Zhang *et al.*, 2003b].

In practice, any program may call the parameterizations of the library. This allows to efficiently share parameterizations. It is also safer than duplicating a given implementation in many different programs. There is a slight drawback since the code in the library may not be exactly written in the way the program needs it. If the library is well designed, it should be the case only seldom.

A library is well suited for the physical parameterizations since they may be updated. They may evolve because of new results from the ongoing research, because of the addition of new abilities or new options in these parameterizations, or simply because of bug fixes. Improving a library is easy and natural. The programs that use this library take advantage of any improvement in the library, without further developments within the programs themselves.

Most of the previous advantages cannot be claimed by a system that would simply rely on programs instead of a library, and it is unfortunately the case for several CTMs that directly incorporate the parameterizations.

A way to improve such a library is to add parameterizations in it. Since most parameterizations are independent from each other, adding a parameterization does not raise any problem. If all parameterizations were implemented in a single program (may it be a CTM) instead, one would have to take into account a new possibility (new data, new options) in the interface (to a CTM for instance). A parameterization added in a library is simply a stand-alone function, which is easy to manage.

A.3.3 Code quality

Within programs, the use of a library makes the code clear. It naturally organizes the code around relevant and powerful data structures provided by the library: the multidimensional data are well handled and come along with convenient functions. For instance, instead of writing nested loops and adding a temporary variable to compute a correlation between two arrays, a single call to a function called `Correlation` in a library may be sufficient. The code

[2] `http://www.osl.iu.edu/research/mtl/`

[3] `http://www.stsci.edu/resources/software_hardware/numarray/`

mainly becomes a set of calls to the library. Its structure and its content are therefore easy to understand.

If the library has no bug, the code is also safer. And this is easier to build a bug-free library than a robust single-program because a library is tested through the numerous programs that use it. In addition, every component of a library may be tested independently and carefully while a single program is more or less tested as a whole.

Finally, a library helps in evolving the programs themselves since, within a program, calls to the new parameterizations may be introduced at no cost (providing the data associated with the new parameterizations are available). For instance, a library may provide an alternative parameterization to compute the critical relative humidity (as function of the pressure and the surface pressure), and a call to this new parameterization, instead of the base one, would enable to include it in a program without further development.

A.3.4 Shared development

Last but not the least, a library may be shared within a research team or even within a given community. The community would save time and energy for the core research activities if it could gather a few parts of its codes – the parts that are otherwise duplicated by all teams. A library should be open to the contributions of all teams and any team could use the parameterizations provided by the library. The flexibility of a library avoids the need for a common model that would be all of a piece and that would therefore be harder to achieve and to manage.

As a conclusion to this quick review of the specific advantages of a library, it is obvious that a library is the right framework to build a tool for data processing and physical parameterizations. Notice that this strategy has shown to be relevant in other fields, e.g. with Blas [Lawson *et al.*, 1979] and Lapack [Anderson *et al.*, 1999] for linear algebra.

A.4 Design

The library may be mainly split into its data processing abilities and its set of parameterizations. Both are linked since the latter uses the structures of the former. Obviously, each parameterization is associated with a function. The content of the data processing part is not so clear and its aims have to be defined. The first requirement is the ability to deal with multidimensional data. Aptitudes to perform input and output operations are also needed because of the amount of data read and written in various formats. Among valuable characteristics are the coding facilities and the robustness of the code; in practice: simplicity, readability, checks and error diagnosis.

Now that the content of the library is roughly determined, the question of the computer language should be addressed. According to the previous requirements, an object-oriented language is necessary. Recall that an object-oriented language manipulates objects and that objects are composed of:

1. attributes: they are the data contained by the object. For instance, if the object is a vector, its attributes can be the data array and the length of the vector.

2. methods: they are functions associated with the object. They usually allow to access to the attributes, to modify the attributes or to perform operations on them. If the object is a vector, the methods may return the length of the vector (method usually called `GetLength`), the value of the vector at a given index, the minimum of the values stored in the vector (e.g. `GetMin`), or they may allow to set a value at a given index or to resize the vector (`Resize` or `Reshape`).

An object is like a black box that stores data (attributes) and that provides the methods to deal with its attributes. It allows to deliver simple and readable codes. Notice that in object-oriented languages, all variables are objects, including the floating-point numbers. In addition, the library provides new types of object, in other words user-defined objects. This way, AtmoData defines new types of object to handle multidimensional data along with convenient methods. The objects ease the management of multidimensional data.

In efficient scientific-computing, five languages are widely used: three in the Fortran series (77, 90 and 95), C and C++. Fortran 77 and C are not object-oriented languages and are therefore discarded. As compared to C++, Fortran 90/95 lacks several features that give a clear advantage to C++ – see Cary *et al.* [1997]. There are notably this two missing features: the genericity and the exceptions. The genericity, through the so-called templates, enables a code to be independent from the type of the variables involved. As a consequence, one may write a function (a parameterization) once and use it with single-precision variables and double-precision variables. In Fortran 90/95, one may have written twice the same function. Exceptions are also missing in Fortran 90/95. Exceptions provide a framework to manage errors: if given options are set, several checks are performed at runtime and any error is reported by a dedicated object. This object may be launched from any function and at any level in the call stack, and it is easily caught and analyzed: there is no need for the propagation of a diagnosis argument in every function (as for the integer `info` in Lapack). As a consequence, the whole code benefits from the exceptions and this is highly valuable for its safety.

The library AtmoData was therefore written in C++. Even if C++ is superior to Fortran 90/95, Fortran 77 has been widely used in the atmospheric chemistry and people tend to learn Fortran 90/95 instead of C++. Nevertheless, from our experience, learning and using C++ together with the right libraries is barely harder than using Fortran 90/95. Thanks to AtmoData the subtleties of C++ are hidden and the user manipulates high-level objects as easily as Fortran arrays. In the end, using C++ even becomes easier thanks to the facilities of the AtmoData objects. The examples shown in the next two sections should partially prove it.

C++ does not provide multidimensional arrays. There are several libraries to solve this issue. Blitz++ [Veldhuizen, 1998] is probably the best library to this respect. It may handle arrays up to eleven dimensions. It ensures the same efficiency as in Fortran 90/95. AtmoData is based on it.

The whole library uses templates (so does Blitz++) so that it could handle single-precision arrays, double-precision arrays or arrays of any other type. For instance, an array could be declared as a single-precision array or as a double-precision array, and all functions available in the library (parameterizations, input/output operations) may handle this array.

Errors are managed with exceptions. In addition, several debugging levels are defined. The user chooses what the library should check. For instance, a debugging level set to four implies that everything is checked: input/output operations, dimensions compatibility, indices validity and memory allocations. This is very useful during the development process. With a debugging level set to two, a program does not check for indices validity to save computation time. This is the right option for a stable program. This mechanism is very useful to ensure the safety or the efficiency of a program. Notice that the mechanism itself has no cost since the tests to be included are chosen at compile time.

The next two sections detail the content of the two main parts of AtmoData: data processing and the physical parameterizations.

A.5 Data processing in AtmoData

In this section, the main features of AtmoData with respect to data processing are mentioned. A few practical examples illustrate that AtmoData is relevant for data processing in the context of atmospheric modeling:

1. the definition of multidimensional and gridded data (listing 1);
2. the interpolation of fields from one grid (e.g. the grid of a meteorological model) to another one (e.g. the grid of the CTM);
3. the input/output operations involving several formats (e.g. one for the meteorological model and one for the CTM).

Most data-processing abilities are not strictly part of AtmoData, but they are provided by the two main underlying libraries Blitz++ and SeldonData. As previously mentioned, the first one introduces multidimensional arrays that are not in the C++ standard library. SeldonData is built on top of Blitz++ and it mainly provides the data structures and the basic input/output facilities used by AtmoData. SeldonData actually gathers what is not bound to the atmospheric chemistry. For the sake of clarity, the abilities described in this section are assumed to be part of AtmoData even if a few are included in it through SeldonData.

A.5.1 Terminology

So that the following lines should be properly understood, a few technical words are explained. An object is a structure that gathers data, called *attributes*, and functions, called *methods*. A few details were provided in section A.4. An object belongs to a *class* which is the type of this object. For example, if `x` is an integer, it is belongs to the class `int` in C++ (even integers are objects in C++); in other words, `x` is a variable of type `int`. Notice that a class may be a user-defined type: one could define a class `Vector` or a class `List`. `x` is also called an *instance* of the class `int`. An instance of a class `Class` is an object of type `Class`.

A method `Method` of a class `Class` is referred as `Class::Method`. If `Instance` is an instance of `Class`, then a call to `Class::Method` is written `Instance.Method()` or `Instance.Method(argument0, argument1, ...)`. For example, assume that `V` is an instance of a class `Vector` and that there is a method `Vector::GetLength` to return the length of a vector. If one wants to set the variable `L` (an integer) to the length of `V`, one would write: `L = V.GetLength()`.

A.5.2 Grids and data declaration

The first step is to devise a suited data structure. It leads to the implementation of a C++ class. In AtmoData, the main class is called `Data` and it manages multidimensional arrays. The multidimensional arrays provided by Blitz++ are used to store the data, but they are not enough featured: the data should also be associated with coordinates. This is the reason why the class `Data` of AtmoData stores the multidimensional arrays together with their coordinates. The same idea is used in modern data files, e.g. NetCDF or Grib files. These formats store data that is often referred as gridded data. In AtmoData, any instance of `Data` is composed of an array (that stores the data itself) and a set of grids (a grid stores the coordinates along a given dimension). For instance, a surface temperature field (depending on the time, the latitude and the longitude) would be defined as an instance of `Data` that would store a three-dimensional data array (the temperatures themselves), and three grids for the time-steps, the latitudes and the longitudes at which the temperatures are given.

A class `Grid` is defined in AtmoData to manage the coordinates along a given dimension. It basically stores the coordinates along a single dimension. It is derived (C++ inheritance) into a simple class `RegularGrid` which could be seen as a vector of coordinates. An example is shown in the listing A.1.

Listing A.1: Use of the class `Data` with simple grids. Grids are managed as easily as vectors. The `Data` object `Field` is more than a data array since it also stores its coordinates, but the access to its values is trivial. Comments are preceded by `//`.

```
// Defines a grid along X with 4 elements starting at 0
// and with a fixed step of 1.2. The grid is stored in
// double precision.
RegularGrid<double> GridX(0., 1.2, 4);
// Defines a grid along Y with 3 elements.
RegularGrid<double> GridY(3);
// Sets the grid along Y to [1, 1.5, 3].
GridY(0) = 1.;  // Notice that the first index is 0.
GridY(1) = 1.5;
GridY(2) = 3.;

// Declares bi-dimensional gridded data: Field,
// which could be a surface temperature or so.
Data<double, 2> Field(GridX, GridY);

// Sets an element in the data array.
Field(1, 2) = -5.2;

// Grids may be accessed through the operator [].
// Field[1] is the second (indices start at 0) grid
// of Field, namely GridY. RegularGrid<double>::Print is
// a method that displays the coordinates on screen.
// The following line therefore displays [1, 1.5, 3]
// on screen.
Field[1].Print();
// Abscissa (Field[0] is GridX) of the element (2, 1)
// of Field: it sets coord to 2.4.
double coord = Field[0].Value(2, 1);
```

Nevertheless it happens that the coordinates along a given dimension have complex dependencies. In the listing A.1, `Field[0].Value(2, 1)` is the same as `Field[0].Value(2, 0)` or `Field[0].Value(2, 2)` because the abscissa does not depend on the index along Y. The same is true for `Field[1]`. But if `GridY` stored the altitudes y_j, it may depend on the horizontal position x as it is usually the case for the altitude (in meters) of the levels of the meteorological models. For instance, from the hydrostatic equation, one gets $y_{j+1} = y_j + \frac{rT_j}{g} \ln\left(\frac{P_j}{P_{j+1}}\right)$ where P_j the pressure at y_j, r the molar mass constant for dry air, T_j the temperature (usually the averaged temperature of the two levels j and $j+1$) and g the standard gravity constant. Hence y_j depends on the pressure and consequently on the position x. Notice that the altitudes should be referred as y_{ij} (i is the index along x) instead of y_j. A more general class, called `GeneralGrid`, is therefore provided to deal with the complex case where a grid depends on the indices along several dimensions. A `GeneralGrid` may be seen as a multidimensional array, the number of dimensions being the number of dependencies (at least one). One may argue that storing the grids along with every data set is a loss of memory. To solve this problem, an option has been added to avoid memory duplication. The `Data` objects may point to existing grids instead of having grids on their own.

A.5.3 Methods

What is stored in a class `Data` has been previously shown: the data array and the grids. `Data` has also several methods among which a few are necessary, e.g. `Data::GetLength(i)` which returns the dimension of the data array along the dimension #i, or `Data::operator(...)` which is the access method. Other methods add useful functionalities such as statistics computation, e.g. `Data::GetMin()`, `Data::GetMaxIndex()`, `Data::Mean()`, `Data::StandardDeviation()`, etc. They are many other methods: `Data::Resize` to resize the data array (and the associated grids), `Data::SwitchDimensions` to switch dimensions (and the grids with it), `Data::Apply` to apply a given function to all elements in the data array, `Data::Threshold` to threshold the data, etc. Using these methods avoids the use of temporary variables and loops; the code is also safer and easier to read. In addition, outside of the class, interpolation functions have also been written. See the listing A.2.

Listing A.2: Use of the methods of the class `Data` and linear interpolation. Notice that the interpolation function only takes the `Data` instances as argument. The coordinates needed for the interpolation are stored within the `Data` instances.

```
// 'Field_in' and 'Field_out' are two instances
// of Data, with the same number of dimensions.
// 'Field_in' is assumed to be already set.

// Use of a method that thresholds the input data.
Field_in.ThresholdMin(0.1);  // Then Field_in >= 0.1.

// Linear interpolation from data defined on a regular
// grid (Field_in) to data defined on a general grid
// (Field_out). Field_out has general grids because,
// for instance, its vertical coordinates may depend
// on the time-step and the horizontal position.
LinearInterpolationRegularToGeneral(Field_in, Field_out);

// Getting the maximum of Field_out.
double maximum = Field_out.GetMax();
```

A.5.4 Input/output operations

The class `Data` is a powerful tool to deal with data sets. Before manipulating the data and after their transformation and the computations, input/output operations on files are performed (for instance, to read the meteorological fields in NetCDF files, to read given coefficients in text files, to write computed deposition-velocities in binary files, etc.). AtmoData provides classes dedicated to input/output operations on files and in given formats. There is a class for every format that is supported. Each class is able to read a file in a given format and to put the data in a `Data` object. Several classes allow to write the `Data` object into a file in the given format. New formats may be added easily; in the version 1.0 of AtmoData, among the available formats, the most well known are: text files, formatted text files (i.e. stored by columns), binary files, Grib files, NetCDF files and MM5[4] files. Each class dedicated to a format provides an intuitive interface. Moreover most classes define the same interface: for instance, the classes for text files and for binary files have the same interface. Finally one may notice that these classes handle both `Data` and `Grid`. See the listing A.3.

[4]PSU/NCAR mesoscale model – `http://www.mmm.ucar.edu/mm5/`.

Listing A.3: Use of input/output operations on objects of type `Data` and `Grid`.

```
// 'Field' is a single-precision (float)
// 10 by 20 by 30 Data object.
// No grid is associated with 'Field' here.
Data<float, 3> Field(10, 20, 30);
// Grids may also be read from files.
RegularGrid<float> GridX(30);

// Objects for input/output operations.
FormatNetCDF<float> Input;
FormatBinary<double> Output; // Double precision
                             // binary file.
FormatText TextOutput(``;''); // Separator set to
                              // the semicolon.

// Reads the surface pressure from the NetCDF File
// 'file.nc'. 'PS' is the name of the surface pressure
// in this file.
Input.Read(``file.nc'', ``PS'', Field);
// Reads the longitude coordinates (lon) in 'file.nc'.
Input.Read(``file.nc'', ``lon'', GridX);

// Writes 'Field' to a binary file (in double precision).
Output.Write(Field, ``file.bin'');
// In the same way, writes the longitudes in a text file.
TextOutput.Write(GridX, ``file.txt'');
```

A.5.5 Error management

The last major issue is the error management. Depending on the options chosen by the user, AtmoData may perform or not a set of tests at run time. For instance, AtmoData can check that no illegal access (with indices out of range) is attempted. It can also check the compatibility of the dimensions involved, the memory allocations or the validity of input/output operations. As explained in the previous section, the user chooses what to check, knowing that checking indices, for instance, is time consuming.

The listing A.4 shows how easy it is to catch errors with the exceptions. An additional point is that the proper use of a debugger enables to browse the call stack that led to the exception.

Listing A.4: The exceptions enable to gracefully catch errors. Assume that the file 'file.nc' is not a NetCDF file. Then `Input.Read` throws an exception in `ReadPS`. The error is nested in the function `ReadPS` but it is still caught by the `try ... catch` structure. Hence `Err.What` is called and it displays on screen: `ERROR! An input/output operation failed in FormatNetCDF<T>::Read(string FileName, Array<TA, N>& A). "file.nc" is not a valid netCDF file.`

```
// Declares a C++ function that reads the surface
// pressure in file 'file.nc'.
void ReadPS(Data<double, 3>& Field)
{
  FormatNetCDF<double> Input;
  Input.Read(``file.nc'', ``PS'', Field);
}

[...]
```

```
10
// In the main program, the code is put in a
// try ... catch structure to catch the
// exceptions.

15 try
{
  // The surface pressure is declared.
  Data<double, 3> Pressure(10, 20, 30);

20  // Calls the previously defined function 'ReadPS'.
  ReadPS(Pressure);
}
catch(Error& Err) // Catches the exceptions, if any,
                  // of type Error.
25 {
  // Err contains the details about the error.
  // Error::What displays on screen the details.
  Err.What();
}
```

A.6 Atmospheric data and physical parameterizations

AtmoData gathers many functions (outside the class `Data`, in addition to its methods). Most of them are parameterizations and a few are useful functions such as coordinate transformations or statistical computations. Statistical measures are available in order to compare model outputs with observations: correlation, root mean square, normalized gross error, etc. They enable to evaluate a model, for instance as proposed in EPA [1991].

As for the parameterizations, they are implemented in functions that take `Data` objects as arguments. The design is clear:

- all functions are template functions so that they may be used with any type (single precision, double precision, etc.);
- input variables are placed first in the argument list, the output variables are placed after, and the optional parameters are placed at the end (so that they may be omitted and set to their default value);
- the functions should take `Data` objects as argument, which means that the loops (over time, space and species dimensions) are performed within the functions.

The last rule may be broken if no `Data` object is involved or if it is necessary to compute the output field cell by cell. The first rule might also be discarded if there is no other choice (e.g. if the function relies on a non-template function – a Fortran function for instance).

Refer to the listing A.5 to understand the way the functions are used. The example involves a very basic function in order to compute the potential temperature.

Listing A.5: Use of the functions of AtmoData, example with the potential temperature. Notice that the `Data` objects are three-dimensional fields but they could also be four-dimensional fields, with the same calls.

```
Data<double, 3> Temperature(2, 2, 2), Pressure(2, 2, 2),
  PotentialTemperature(2, 2, 2);
```

```
// Fills Temperature and Pressure with
// constant fictitious data.
Temperature.Fill(280.0);
Pressure.Fill(90000.0);

// Computes the potential temperature with the
// temperature and the pressure.
ComputePotentialTemperature(Temperature, Pressure,
  PotentialTemperature);
// One may set the reference pressure to 100000 Pa
// with the optional (and last) argument.
ComputePotentialTemperature(Temperature, Pressure,
  PotentialTemperature, 100000.);
```

The content of AtmoData increases in time. In the version 1.0, there are parameterizations for clouds (cloud diagnosis, height, cloudiness, cloud attenuation, etc.), deposition velocities, vertical diffusion and emissions. Miscellaneous meteorological functions (Richardson number, relative humidity, etc.) and useful functions (computing ECMWF[5] model heights, zenith angle, etc.) are also included.

A.7 AtmoData in use

First a complete documentation is available with the AtmoData distribution (`http://www.enpc.fr/cerea/atmodata/`). Recall that, strictly speaking, a part of the previously described abilities are not part of AtmoData but come from underlying libraries (mainly Blitz++ and SeldonData). The underlying libraries have their own documentation. In practice, one should read the documentation for SeldonData and the documentation for AtmoData. Both have a user's guide and a reference documentation that describes all objects, methods and functions available in the libraries. The user's guides are starting points to understand the involved structures and the way the libraries should be used. Then the user selects the functions he wants to use and reads their description in the reference documentation. The reference documentation is a user-friendly HTML documentation generated thanks to Doxygen[6]. Reading the documentations should not require more than one day. The user should then be able to write his own programs.

The listing A.6 shows quotes from a program using AtmoData and dealing with data from the ECMWF.

Listing A.6: Example quoted from a program using AtmoData and dealing with data from the ECMWF. The input ECMWF fields are read and used to compute the cloud fraction. The grids are first defined and then the `Data` instances. Finally several functions are called to compute and interpolate the cloud fraction to a CTM grid.

```
[...]

/*** Grids ***/

// Input grids (ECMWF).
RegularGrid<real> GridT_in(t_min_in, Delta_t_in, Nt_in);
// Vertical levels depend on t, z, y and x.
GeneralGrid<real, 4> GridZ_in(shape(Nt_in, Nz_in,
```

[5] European Centre for Medium-Range Weather Forecasts – `http://www.ecmwf.int/`.

[6] `http://www.stack.nl/~dimitri/doxygen/`

```
                                        Ny_in, Nx_in),
                                  1, shape(0, 1, 2, 3));
GeneralGrid<real, 4> GridZ_interf_in(shape(Nt_in, Nz_in+1,
                                           Ny_in, Nx_in),
                                     1, shape(0, 1, 2, 3));
RegularGrid<real> GridY_in(y_min_in, Delta_y_in, Ny_in);
RegularGrid<real> GridX_in(x_min_in, Delta_x_in, Nx_in);
// To avoid memory duplication.
GridZ_in.SetDuplicate(false);

// Output grids (CTM).
RegularGrid<real> GridT_out(t_min_out, Delta_t_out, Nt_out);
RegularGrid<real> GridZ_out(Nz_out);
RegularGrid<real> GridY_out(y_min_out, Delta_y_out, Ny_out);
RegularGrid<real> GridX_out(x_min_out, Delta_x_out, Nx_out);
// Heights of the interfaces.
RegularGrid<real> GridZ_interf_out(Nz_out + 1);

// Reads output altitudes.
FormatText Heights_out;
// 'vertical_levels' is a text file with the output
// altitudes (for the CTM).
Heights_out.Read(vertical_levels, GridZ_interf_out);
// Sets values at nodes.
for (k = 0; k < Nz_out; k++)
  GridZ_out(k) = (GridZ_interf_out(k)
                  + GridZ_interf_out(k+1)) / 2.0;

/*** Data instances ***/

// Input fields.
Data<real, 4> Temperature(GridT_in, GridZ_in,
                          GridY_in, GridX_in);
Data<real, 4> Pressure(GridT_in, GridZ_in,
                       GridY_in, GridX_in);
Data<real, 4> Pressure_interf(GridT_in, GridZ_interf_in,
                              GridY_in, GridX_in);
Data<real, 3> SurfacePressure(GridT_in, GridY_in, GridX_in);
Data<real, 4> Humidity(GridT_in, GridZ_in,
                       GridY_in, GridX_in);
Data<real, 4> RelativeHumidity(GridT_in, GridZ_in,
                               GridY_in, GridX_in);
Data<real, 3> BoundaryLayerHeight(GridT_in, GridY_in,
                                  GridX_in);
Data<real, 4> CRH(GridT_in, GridZ_in, GridY_in, GridX_in);
Data<real, 4> CloudFraction(GridT_in, GridZ_in,
                            GridY_in, GridX_in);

// Output fields.
Data<real, 4> CloudFraction_out(GridT_out, GridZ_out,
                                GridY_out, GridX_out);

/*** Reads input data ***/

// ECMWF files are Grib files.
FormatGrib InputMeteo;
```

```
65
   // 'file_in' is the ECMWF file. 130 is the Grib number
   // for the temperature.
   InputMeteo.Read(file_in, 130, Temperature);
   // ECMWF data are stored from the top to the bottom.
70 // Reverses the data along Z to fix that.
   Temperature.ReverseData(1);
   // ECMWF data are stored from North to South.
   // Reverses the data along Y to fix that.
   Temperature.ReverseData(2);
75
   // All fields are read in the same way.
   [...]


80 /*** Computes level heights with pressure levels ***/

   // 'alpha' and 'beta' are coefficient previously read
   // in a file.
   // Computes the pressure at interfaces.
85 ComputePressure(alpha, beta, SurfacePressure,
                  Pressure_interf);
   // Computes the altitudes using the hydrostatic assumption
   // (as shown in the section ``Grids and data declaration'').
   ComputeInterfHeight(Pressure_interf, Temperature,
90                     GridZ_interf_in);

   [..]

   /*** Computes the cloud fraction ***/
95
   ComputeRelativeHumidity(Humidity, Temperature, Pressure,
                           RelativeHumidity);
   RelativeHumidity.ThresholdMax(1.);
   ComputeCriticalRelativeHumidity(SurfacePressure,
100                                Pressure, CRH);
   ComputeCloudFraction(BoundaryLayerHeight, RelativeHumidity,
                        CRH, CloudFraction);

   /*** Output field ***/
105
   // Computes the output field.
   // The following interpolation function is designed for
   // output Data objects associated with at most
   // one general grid (here along dimension #1 -- last
110 // argument).
   LinearInterpolationOneGeneral(CloudFraction,
                                 CloudFraction_out, 1);
   // Writes it to a binary file.
   FormatBinary<float> OutputMeteo;
115 OutputMeteo.Write(CloudFraction_out, directory_out
                                       + "CloudFraction.bin");

   [...]
```

A.8 Conclusion and next steps

This paper has shown the main features of AtmoData, from its aims to its use. AtmoData provides a complete solution to deal with the data and the parameterizations involved in atmospheric chemistry and physics. The bases of AtmoData have been utterly satisfactory so far, in both research and operational use (see the modeling system Polyphemus[7], Mallet *et al.* [2005]). Hence the next steps are mainly the implementation of new parameterizations, especially for aerosol modeling. Other functions will also be included for statistics and input/output operations.

Independently from AtmoData itself, the distribution[8] of the library is a promising goal. The library is released under the GNU General Public License[9] and is ready to be widely distributed. Thanks to its flexibility, it may be used for many applications and by several teams. The whole community is welcome to use and to take part in the development of the library. There is no reason why this work should not be shared since the parameterizations implemented inside are well known. It would avoid duplicated implementations of the same parameterizations, which has already been a loss of time and energy in the community. It would improve the reliability of the systems and let all teams benefit from the best parameterizations at low cost.

[7] `http://www.enpc.fr/cerea/polyphemus/`

[8] AtmoData is freely available at `http://www.enpc.fr/cerea/atmodata/`.

[9] `http://www.gnu.org/copyleft/gpl.html`

Notations

Sigles

ADEME Agence de l'environnement et de la maîtrise de l'énergie ..... p. 58

BDQA Banque de données sur la qualité de l'air ..... p. 58

BF « bias factor » ..... p. 57

CMAQ Community Multiscale Air Quality Model ..... p. 36

ECMWF European Centre for Medium-Range Weather Forecasts ..... p. 36

EMEP Co-operative Programme for Monitoring and Evaluation of the Long-range Transmission of Air Pollutants in Europe ..... p. 33

GLCF Global Land Cover Facility ..... p. 36

INERIS Institut national de l'environnement industriel et des risques ..... p. 9

LUC « land use coverage », occupation des sols ..... p. 32

MM5 Fifth-Generation NCAR / Penn State Mesoscale Model ..... p. 36

MNBE « mean normalized bias error », bias normalisé moyen ..... p. 57

MNGE « mean normalized gross error » ..... p. 57

NCAR National Center for Atmospheric Research (États-Unis) ..... p. 37

Prév'air plate-forme de prévision opérationnelle de la qualité de l'air, opérée par l'INERIS – `http://www.prevair.org/` ..... p. 9

RACM regional atmospheric chemistry mechanism ..... p. 28

RADM 2 regional acid deposition model, version 2, mécanisme chimique ..... p. 28

RMS « root mean square », erreur quadratique moyenne ..... p. 57

RMSE « root mean square error », erreur quadratique moyenne ..... p. 57

SNAP classe d'émission, catégorie d'émetteur ..... p. 33

USGS U.S. Geological Survey ..... p. 36

Variables

χ_i terme de production et de perte par réaction chimique pour l'espèce i ..... p. 27

Δ opérateur de différence entre deux niveaux verticaux ..... p. 29

$\mathcal{W}$ module du vent ..... p. 31

ρ densité de l'air ..... p. 27

θ température potentielle . . . . . p. 29

CF fraction nuageuse . . . . . p. 35

CRH humidité relative critique . . . . . p. 34

$D\mathcal{W}_k$ cisaillement du vent . . . . . p. 29

D_s diffusivité moléculaire de l'espèce s . . . . . p. 31

f_s réactivité de l'espèce s . . . . . p. 31

H_s constante de Henry de l'espèce s . . . . . p. 31

K matrice des coefficients de diffusion turbulente . . . . . p. 27

K_z coefficient de diffusion verticale . . . . . p. 29

LMO longueur de Monin-Obukhov . . . . . p. 30

PAR part du rayonnement active pour la photosynthèse . . . . . p. 33

$PBLH$ hauteur de couche limite atmosphérique . . . . . p. 30

R_a résistance aérodynamique . . . . . p. 30

R_b résistance de couche quasi-laminaire . . . . . p. 30

R_c résistance de canopée . . . . . p. 30

Ri nombre de Richardson . . . . . p. 29

T_{srf} température de surface . . . . . p. 32

U vent zonal . . . . . p. 29

u_* vitesse de friction du vent . . . . . p. 30

V vent (trois composantes) ou vent méridional . . . . . p. 29

v_d vitesse de dépôt . . . . . p. 30

z_0 hauteur de rugosité . . . . . p. 29

z_k hauteur du centre de la maille k . . . . . p. 28

$\tilde{z}_k$ hauteur de l'interface supérieure de maille k . . . . . p. 28

z_t hauteur de rugosité associée à la température . . . . . p. 31

Constantes

κ constante de Von Kármán [0.40] . . . . . p. 29

R constante des gaz parfaits [8.314 $\mathrm{J \cdot K^{-1} \cdot mol^{-1}}$] . . . . . p. 33

WT seuil minimal sur le module du vent [0.001 $\mathrm{m \cdot s^{-1}}$] . . . . . p. 29

Bibliographie

Aissaoui, M. (2004). Étude de la propagation d'incertitudes dans un modèle de chimie-transport, POLAIR3D. Mémoire de D.E.A., DEA M2SAP.

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Croz, J. D., Greenbaum, A., Hammarling, S., McKenney, A. et Sorensen, D. (1999). *LAPACK users' guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, troisième édition.

Bastrup-Birk, A., Brandt, J., Zlatev, Z. et Uria, I. (1997). Studying cumulative ozone exposures in Europe during a 7-year period. *J. Geophys. Res.*, 102(D20):23,917–23,935.

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. *J. Geophys. Res.*, 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. *Int. J. Environment and Pollution*, 22(1/2):17–28.

Browning, G. L. et Kreiss, H.-O. (1994). Splitting methods for problems with different timescales. *Mon. Wea. Rev.*, 122(11):2,614–2,622.

Builtjes, P. (1992). The LOTOS – LOng Term Ozone Simulation – project, summary report. Rapport technique R92/240, TNO, Delft, the Netherlands.

Buizza, R., Miller, M. et Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, 125:2,887–2,908.

Byun, D. W. et Ching, J. K. S., éditeurs (1999). *Science algorithms of the EPA models-3 community multiscale air quality (CMAQ) modeling system*. EPA.

Carter, W. P. L. (1990). A detailed mechanism for the gas-phase atmospheric reactions of organic compounds. *Atmos. Env.*, 24A:481–518.

Cary, J. R., Shasharina, S. G., Cummings, J. C., Reynders, J. V. W. et Hinker, P. J. (1997). Comparison of C++ and Fortran 90 for object-oriented scientific programming. *Computer Phys. Comm.*, 105:20–36.

Cesa-Bianchi, N., Long, P. M. et Warmuth, M. K. (1996). Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Trans. Neural Net.*, 7(3):604–619.

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. *J. Geophys. Res.*, 92(D12):14,681–14,700.

Chang, M. E., Hartley, D. E., Cardelino, C., Haas-Laursen, D. et Chang, W.-L. (1997). On using inverse methods for resolving emissions with large spatial inhomogeneities. *J. Geophys. Res.*, 102(D13):16,023–16,036.

Christensen, J. H. (1997). The Danish Eulerian hemispheric model – a three-dimensional air pollution model used for the arctic. *Atmos. Env.*, 31:4,169–4,191.

Dabberdt, W. F. et Miller, E. (2000). Uncertainty, ensembles and air quality dispersion modeling : applications and challenges. *Atmos. Env.*, 34(27):4,667–4,673.

Debry, É. (2004). *Modélisation et simulation numérique de la dynamique des aérosols atmosphériques*. Thèse de doctorat, École nationale des ponts et chaussées.

Delle Monache, L. et Stull, R. B. (2003). An ensemble air-quality forecast over western Europe during an ozone episode. *Atmos. Env.*, 37:3,469–3,474.

Derognat, C., Beekmann, Baeumle, M., Martin, D. et Schmidt, H. (2003). Effect of biogenic volatile organic compound emissions on tropospheric chemistry during the atmospheric pollution over the Paris area (ESQUIF) campaign in the île-de-France region. *J. Geophys. Res.*, 108(D17).

Draxler, R. R. (2003). Evaluation of an ensemble dispersion calculation. *J. Applied Meteor.*, 42:308–317.

Elbern, H. et Schmidt, H. (2001). Ozone episode analysis by four-dimensional variational chemistry data assimilation. *J. Geophys. Res.*, 106(D4):3,569–3,590.

Elbern, H. et Schmidt, H. (2002). 4D-var data assimilation and its numerical implications for case study analyses. *In* Chock, D. P. et Carmichael, G. R., éditeurs : *Atmospheric Modeling*, pages 165–183. IMA, Springer.

EPA (1991). Guideline for regulatory application of the urban airched model. Rapport technique, EPA.

ESQUIF (2001). Étude et simulation de la qualité de l'air en île de France – rapport final.

Faure, C. et Papegay, Y. (1998). O∂yssée user's guide – version 1.7. Rapport technique 0224, INRIA.

Galmarini, S., Bianconi, R., Klug, W., Mikkelsen, T., Addis, R., Andronopoulos, S., Astrup, P., Baklanov, A., Bartniki, J., Bartzis, J. C. *et al.* (2004). Ensemble dispersion forecasting – part I : concept, approach and indicators. *Atmos. Env.*, 38(28):4,607–4,617.

Gardiner, C. W. (1996). *Handbook of Stochastic Methods – For Physics, Chemistry and the Natural Sciences*. Springer.

Garrat, J. R. (1992). *The atmospheric boundary layer*. Cambridge University Press.

GENEMIS (1994). Genemis (generation and evaluation of emission data) annual report 1993. Rapport technique, EUROTRAC.

Gery, M. W., Whitten, G. Z., Killus, J. P. et Dodge, M. C. (1989). A photochemical kinetics mechanism for urban and regional scale computer modeling. *J. Geophys. Res.*, 94:12,925–12,956.

Greenfield, P., Miller, J. T., Hsu, J.-C. et White, R. L. (2003). numarray : a new scientific array package for Python. *In PyCon DC 2003*.

Gross, A. et Stockwell, W. R. (2003). Comparison of the EMEP, RADM2 and RACM mechanisms. *Journal of Atmospheric Chemistry*, 44:151–170.

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. *Atmos. Env.*, 32(21):3,619–3,628.

Hanna, S. R. et Davis, J. M. (2002). Evaluation of a photochemical grid model using estimates of concentration probability density functions. *Atmos. Env.*, 36:1,793–1,798.

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. *Atmos. Env.*, 35(5):891–903.

Hass, H. (1991). Description of the EURAD chemistry-transport-model version 2 (CTM2). Rapport technique 83, Institute of Geophysics and Meteorology, University of Cologne.

Hass, H., Builtjes, P. J. H., Simpson, D. et Stern, R. (1997). Comparison of model results obtained with several European regional air quality models. *Atmos. Env.*, 31(19):3,259–3,279.

Hoeting, J. A., Madigan, D., Raftery, A. E. et Volinsky, C. T. (1999). Bayesian model averaging : a tutorial. *Stat. Sci.*, 14(4):382–417.

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. *Atmos. Env.*, 35:4,159–4,174.

Holton, J. R. (2004). *An introduction to dynamic meteorology.* Academic Press, quatrième édition.

Honoré, C. (2000). *La photochimie de l'ozone à l'échelle urbaine, un système dynamique non-linéaire.* Thèse de doctorat, Université Paris 6.

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. *J. Geophys. Res.*, 108(D24).

Houtemaker, P. L., Lefaivre, L., Derome, J., Ritchie, H. et Mitchell, H. L. (1996). A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, 124(6):1,225–1,242.

Hundsdorfer, W. et Spee, E. (1995). An efficient horizontal advection scheme for the modeling of global transport of constituents. *Mon. Wea. Rev.*, 123(12):3,554–3,564.

ISO/IEC (1998). *International standard 14882 – Programming language C++.* ANSI, première édition.

Jacob, D. J. (1999). *Introduction to atmospheric chemistry*. Princeton University Press.

Jiang, W., Singleton, D. L., Hedley, M. et McLaren, R. (1997). Sensitivity of ozone concentrations to VOC and NO_x emissions in the Canadian Lower Fraser Valley. *Atmos. Env.*, 31(4):627–638.

Jones, E., Oliphant, T., Peterson, P. *et al.* (2001). SciPy : Open source scientific tools for Python.

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., T. LaRow, D. B. et Williford, E. (2000). Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, 13:4,196–4,216.

Lanser, D. et Verwer, J. G. (1999). Analysis of operator splitting for advection-diffusion-reaction problems from air pollution modelling. *J. Comp. Appl. Math.*, 111:201–216.

Lawson, C. L., Hanson, R. J., Kincaid, D. R. et Krogh, F. T. (1979). Basic Linear Algebra Subprograms for Fortran usage. *ACM Trans. on Math. Soft.*, 5(3):308–323.

Louis, J.-F. (1979). A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteor.*, 17:187–202.

Madronich, S. (1987). Photodissociation in the atmosphere : 1. actinic flux and the effects of ground reflections and clouds. *J. Geophys. Res.*, 92(D8):9,740–9,752.

Makar, P. A., Fuentes, J. D., Wang, D., Staebler, R. M. et Wiebe, H. A. (1999). Chemical processing of biogenic hydrocarbons within and above a temperate deciduous forest. *J. Geophys. Res.*, 104(D3):3,581–3,603.

Mallet, V., Quélo, D. et Sportisse, B. (2005). Software architecture of an ideal modeling platform in air quality – A first step : Polyphemus. Rapport technique 11, CEREA.

Mallet, V. et Sportisse, B. (2004). 3-D chemistry-transport model Polair : numerical issues, validation and automatic-differentiation strategy. *Atmos. Chem. Phys. Discuss.*, 4:1,371–1,392.

Mallet, V. et Sportisse, B. (2005a). A comprehensive study of ozone sensitivity with respect to emissions over Europe with a chemistry-transport model. *J. Geophys. Res.*, 110(D22).

Mallet, V. et Sportisse, B. (2005b). Data processing and parameterizations in atmospheric chemistry and physics : the AtmoData library. Rapport technique 12, CEREA.

Mallet, V. et Sportisse, B. (2005c). Toward ensemble-based air-quality forecasts. *En révision pour publication dans J. Geophys. Res.*

Mallet, V. et Sportisse, B. (2006). Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations : an ensemble approach applied to ozone modeling. *J. Geophys. Res.*, 111(D1).

Massman, W. J., Pederson, J., Delany, A., Grantz, D., den Hartog, G., Neumann, H. H., Oncley, S. P., Pearson, R. et Shaw, R. H. (1994). An evaluation of the Regional Acid Deposition Model surface module for ozone uptake at three sites in the San Joaquin Valley of California. *J. Geophys. Res.*, 99:8,281–8,294.

McRae, G. J., Goodin, W. R. et Seinfeld, J. H. (1982). Numerical solution of the atmospheric diffusion equation for chemically reactive flows. *J. Comp. Phys.*, 45:1–42.

Mendoza-Dominguez, A. et Russell, A. G. (2001). Estimation of emission adjustments from the application of four-dimensional data assimilation to photochemical air quality modeling. *Atmos. Env.*, 35:2,879–2,894.

Menut, L. (2003). Adjoint modeling for atmospheric pollution process sensitivity at regional scale. *J. Geophys. Res.*, 108(D17):8,562.

Middleton, P., Stockwell, W. R. et Carter, W. P. L. (1990). Aggregation and analysis of volatile organic compound emissions for regional modeling. *Atmos. Env.*, 24A(5):1,107–1,133.

Njomgang, H., Mallet, V. et Musson-Genon, L. (2005). AtmoData scientific documentation. Rapport technique 10, CEREA.

Nodop, K., éditeur (1997). *ETEX symposium on long-range atmospheric transport, model verification and emergency response.*

Nordeng, T. E. (1986). Parameterization of physical processes in a three-dimensional numerical weather prediction model. Rapport technique 65, Norwegian Meteorological Institute.

Passant, N. R. (2002). Speciation of UK emissions of NMVOC. Rapport technique AEAT/ENV/0545, AEA Technology.

Pielke, R. A. (2002). *Mesoscale meteorological modeling.* Academic Press, seconde édition.

Pourchet, A., Mallet, V., Quélo, D. et Sportisse, B. (2005). Some numerical issues in Chemistry-Transport Models – a comprehensive study with the Polyphemus/Polair3D platform. Rapport technique 26, CEREA.

Pryor, S. C. (1998). A case study of emission changes and ozone responses. *Atmos. Env.*, 32(2):123–131.

Quélo, D. (2004). *Simulation numérique et assimilation de données variationnelle pour la dispersion atmosphérique de polluants.* Thèse de doctorat, École Nationale des Ponts et Chaussées.

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. *J. Geophys. Res.*, 110(D24).

Russell, A. et Dennis, R. (2000). NARSTO critical review of photochemical models and modeling. *Atmos. Env.*, 34:2,283–2,234.

Sandu, A., Potra, F. A., Carmichael, G. R. et Damian, V. (1996). Efficient implementation of fully implicit methods for atmospheric chemical kinetics. *J. Comp. Phys.*, 129:101–110.

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. *Atmos. Env.*, 31(20):3,459–3,472.

Sandu, A., Verwer, J. G., Loon, M. V., 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. *Atmos. Env.*, 31(19):3,151–3,166.

Schmidt, H. (2002). Sensitivity studies with the adjoint of a chemistry transport model for the boundary layer. *In* Sportisse, B., éditeur : *Air pollution modelling and simulation*, pages 400–410. Springer.

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. *Atmos. Env.*, 35:6,277–6,297.

Schmidt, H. et Martin, D. (2003). Adjoint sensitivity of episodic ozone in the Paris area to emissions on the continental scale. *J. Geophys. Res.*, 108(D17):8,561.

Segers, A. (2002). *Data assimilation in atmospheric chemistry models using Kalman filtering.* Thèse de doctorat, Delft University.

Seinfeld, J. H. et Pandis, S. N. (1998). *Atmospheric chemistry and physics : from air pollution to climate change.* Wiley-Interscience.

Siek, J. G. et Lumsdaine, A. (1999). The Matrix Template Library : generic components for high-performance scientific computing. *Computing in Science and Engineering*, 1(6):70–78.

Simpson, D., Fagerli, H., Jonson, J. E., Tsyro, S., Wind, P. et Tuovinen, J.-P. (2003). Transboundary acidification, eutrophication and ground level ozone in Europe – part I : unified EMEP model description. Rapport technique, EMEP.

Simpson, D., Guenther, A., Hewitt, C. N. et Steinbrecher, R. (1995). Biogenic emissions in Europe – 1. estimates and uncertainties. *J. Geophys. Res.*, 100(D11):22,875–22,890.

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. *J. Geophys. Res.*, 104(D7):8,113–8,152.

Sportisse, B. (1999). *Contribution à la modélisation des écoulements réactifs : réduction des modèles de cinétique chimique et simulation de la pollution atmosphérique.* Thèse de doctorat, École polytechnique.

Sportisse, B. (2000). An analysis of operator splitting techniques in the stiff case. *J. Comp. Phys.*, 161(1):140–168.

Sportisse, B. et Mallet, V. (2005). Calcul scientifique pour l'environnement. Cours de deuxième année à l'ENSTA.

Stockwell, W. R., Kirchner, F., Kuhn, M. et Seefeld, S. (1997). A new mechanism for regional atmospheric chemistry modeling. *J. Geophys. Res.*, 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. *J. Geophys. Res.*, 95(D10):16,343–16,367.

Strang, G. (1968). On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5(3):506–517.

Straume, A. G. (2001). A more extensive investigation of the use of ensemble forecasts for dispersion model evaluation. 40:425–445.

Straume, A. G., Koffi, E. N. et Nodop, K. (1998). Dispersion modeling using ensemble forecasts compared to ETEX measurements. *J. Applied Meteor.*, 37:1,444–1,456.

Stull, R. B. (1988). *An introduction to boundary layer meteorology.* Kluwer Academic Publishers.

Sun, P. (1996). A pseudo-non-time-splitting method in air quality modeling. *J. Comp. Phys.*, 127:152–157.

Tao, Z., Larson, S. M., Williams, A., Caughey, M. et Wuebbles, D. J. (2004). Sensitivity of regional ozone concentrations to temporal distribution of emissions. *Atmos. Env.*, 38(37): 6,279–6,285.

Toth, Z. et Kalnay, E. (1993). Ensemble forecasting at NMC : the generation of perturbations. *Bull. Amer. Meteor. Soc.*, 74:2,317–2,330.

Troen, I. et Mahrt, L. (1986). A simple model of the atmospheric boundary layer ; sensitivity to surface evaporation. *Boundary-Layer Meteor.*, 37:129–148.

van Loon, M., Roemer, M. G. M. et Builtjes, P. J. H. (2004). Model inter-comparison – In the framework of the review of the unified EMEP model. Rapport technique 282, TNO.

Veldhuizen, T. L. (1998). Arrays in Blitz++. *In Proceedings of the 2nd International Scientific Computing in Object-Oriented Parallel Environments (ISCOPE'98)*, Lecture Notes in Computer Science. Springer-Verlag.

Verwer, J. G., Blom, J. G. et Hundsdorfer, W. (1996). An implicit-explicit approach for atmospheric transport-chemistry problems. *Appl. Numer. Math.*, 20:191–209.

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 Math. for Indus.*, 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 J. Sci. Comp.*, 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. *J. Applied Meteor.*, 41:488–504.

Wesely, M. L. (1989). Parameterization of surface resistances to gaseous dry deposition in regional-scale numerical models. *Atmos. Env.*, 23:1,293–1,304.

Zhang, L., Brook, J. R. et Vet, R. (2003a). Evaluation of a non-stomatal resistance parameterization for SO_2 dry deposition. *Atmos. Env.*, 37:2,941–2,947.

Zhang, L., Brook, J. R. et Vet, R. (2003b). A revised parameterization for gaseous dry deposition in air-quality models. *Atmos. Chem. Phys.*, 3:2,067–2,082.

Zlatev, Z. (1995). *Computer treatment of large air pollution models*. Kluwer Academic Publishers.