



HAL
open science

Dispersion atmosphérique et modélisation inverse pour la reconstruction de sources accidentelles de polluants

Victor Winiarek

► **To cite this version:**

Victor Winiarek. Dispersion atmosphérique et modélisation inverse pour la reconstruction de sources accidentelles de polluants. Autre. Université Paris-Est, 2014. Français. NNT : 2014PEST1052 . tel-01004505

HAL Id: tel-01004505

<https://pastel.hal.science/tel-01004505>

Submitted on 11 Jun 2014

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

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

UNIVERSITÉ — — PARIS-EST

Thèse présentée pour obtenir le grade de

Docteur de l'Université Paris-Est

Spécialité : Sciences et Techniques de l'Environnement

par

Victor Winiarek

École Doctorale : SCIENCES, INGÉNIERIE ET ENVIRONNEMENT

***Dispersion atmosphérique et modélisation inverse pour la
reconstruction de sources accidentelles de polluants***

Thèse soutenue le 4 mars 2014 devant un jury composé de :

Pr Philippe Bousquet	Université de Versailles Saint Quentin en Yvelines, LSCE	Président, Rapporteur
Pr Olivier Thual	Université de Toulouse, INPT, CNRS, IMFT	Rapporteur
Dr Anne Mathieu	IRSN, BMTA	Examineur
Dr Lionel Soulhac	École Centrale de Lyon, LMFA	Examineur
Pr Marc Bocquet	École des Ponts ParisTech, CEREAA	Directeur de thèse
Dr Bertrand Carissimo	EDF R&D, CEREAA	Co-directeur de thèse

Remerciements

Cette thèse a été réalisée et financée dans le cadre du projet IMMANENT. Je remercie tous les participants de ce projet pour leur collaboration enrichissante dans un cadre multidisciplinaire, et en particulier Frédéric Bourquin et Cyril Nguyen Van Phu pour en avoir assuré la direction.

Je remercie chaleureusement Marc Bocquet d'avoir réussi à me convaincre de réaliser cette thèse et de m'avoir encadré pendant tout le déroulement de celle-ci. Sa disponibilité à toute épreuve, son excellence scientifique et sa patience m'ont été très précieuses. Je remercie également Bertrand Carissimo d'avoir accepté de co-diriger cette thèse, de m'avoir laissé beaucoup de liberté tout en étant présent quand le besoin s'en faisait ressentir.

Je remercie Philippe Bousquet qui m'a fait l'honneur de présider mon jury en plus d'en être un des rapporteurs. Merci à Olivier Thual d'avoir accepté d'être rapporteur. Merci également à Lionel Soulhac pour sa participation au jury et l'intérêt qu'il a montré à mon travail.

Je tiens à remercier chaleureusement ceux qui, à l'IRSN, contribuent aux bonnes relations avec le CEREА, notamment Irène Korsakissok Denis Quélo et Marilyne Tombette. Un remerciement spécial à Olivier Saunier qui m'a transmis le témoin et m'a aidé tout au long de ma thèse. Un très grand merci enfin à Anne Mathieu qui, en plus d'avoir participé au jury, m'a beaucoup aidé et soutenu pendant ces trois années (et un peu plus).

Un grand merci à toutes les personnes que j'ai croisées au CEREА durant ma thèse (mais aussi dans les années précédentes) pour la bonne ambiance et les discussions toujours intéressantes. Merci à Véronique Dehlinger pour l'assistance précieuse dans les démarches administratives. Merci à Bruno Sportisse et Vivien Mallet qui m'ont recruté et accueilli au CEREА, merci à Luc Musson-Genon avec qui ce fut un plaisir de travailler et de discuter.

Enfin je tiens à remercier du fond du cœur tous mes proches, mes amis et ma famille pour tous les merveilleux moments *hors thèse*, les évasions et les projets. Un énorme merci à Justyna pour m'avoir soutenu en permanence et (re-)motiver quand c'était nécessaire et un énorme merci à Charlotte pour être apparue et avoir donné un sens à la fin de thèse.

Résumé

Les circonstances pouvant conduire à un rejet incontrôlé de polluants dans l'atmosphère sont variées : il peut s'agir de situations accidentelles, par exemples des fuites ou explosions sur un site industriel, ou encore de menaces terroristes : bombe sale, bombe biologique, notamment en milieu urbain. Face à de telles situations, les objectifs des autorités sont multiples : prévoir les zones impactées à court terme, notamment pour évacuer les populations concernées ; localiser la source pour pouvoir intervenir directement sur celle-ci ; enfin déterminer les zones polluées à plus long terme, par exemple par le dépôt de polluants persistants, et soumises à restriction de résidence ou d'utilisation agricole.

Pour atteindre ces objectifs, des modèles numériques peuvent être utilisés pour modéliser la dispersion atmosphérique des polluants. Après avoir rappelé les processus physiques qui régissent le transport de polluants dans l'atmosphère, nous présenterons les différents modèles à disposition. Le choix de l'un ou l'autre de ces modèles dépend de l'échelle d'étude et du niveau de détails (topographiques notamment) désiré.

Différentes méthodes de modélisation inverse pour reconstruire des sources accidentelles sont ensuite présentées, de même que des méthodes d'estimations des erreurs *a priori*, auxquelles les résultats de l'inversion sont particulièrement sensibles. Plusieurs cas d'application, en utilisant des données synthétiques et des données réelles, sont proposés, notamment l'estimation de termes sources consécutifs à l'accident de Fukushima en mars 2011. Par nos méthodes, nous avons estimé que les rejets de césium-137 se situent en 12 et 19 PBq, avec une incertitude comprise en 15 et 65%, et que les rejets d'iode-131 se situent entre 190 et 380 PBq avec une incertitude comprise entre 5 et 10%.

En ce qui concerne la localisation d'une source inconnue, deux stratégies sont envisageables : les méthodes dites paramétriques et les méthodes non-paramétriques. Les méthodes paramétriques s'appuient sur le caractère particulier des situations accidentelles dans lesquelles les émissions de polluants sont généralement d'étendue limitée. La source à reconstruire est alors paramétrisée et le problème inverse consiste à estimer ces paramètres, en nombre réduit. Dans les méthodes non-paramétriques, aucune hypothèse sur la nature de la source (ponctuelle, localisée, ...) n'est réalisée et le système cherche à reconstruire un champs d'émission complet (en 4 dimensions). Plusieurs méthodes sont proposées et testées sur des situations réelles à l'échelle urbaine avec prise en compte des bâtiments, pour lesquelles les méthodes que nous proposons parviennent à localiser la source avec une erreur moyenne de 20m (soit 10% de la taille du domaine d'étude), suivant les situations modélisées et les méthodes inverses utilisées.

Mots-clés : Dispersion atmosphérique, pollution accidentelle, modélisation inverse, reconstruction de source.

Abstract

Uncontrolled releases of pollutant in the atmosphere may be the consequence of various situations : accidents, for instance leaks or explosions in an industrial plant, or terrorist attacks such as biological bombs, especially in urban areas. In the event of such situations, authorities' objectives are various : predict the contaminated zones to apply first countermeasures such as evacuation of concerned population ; determine the source location ; assess the long-term polluted areas, for instance by deposition of persistent pollutants in the soil.

To achieve these objectives, numerical models can be used to model the atmospheric dispersion of pollutants. We will first present the different processes that govern the transport of pollutants in the atmosphere, then the different numerical models that are commonly used in this context. The choice between these models mainly depends of the scale and the details one seeks to take into account.

We will then present several inverse modeling methods to estimate the emission as well as statistical methods to estimate prior errors, to which the inversion is very sensitive. Several case studies are presented, using synthetic data as well as real data such as the estimation of source terms from the Fukushima accident in March 2011. From our results, we estimate the Cesium-137 emission to be between 12 and 19 PBq with a standard deviation between 15 and 65% and the Iodine-131 emission to be between 190 and 380 PBq with a standard deviation between 5 and 10%.

Concerning the localization of an unknown source of pollutant, two strategies can be considered. On one hand parametric methods use a limited number of parameters to characterize the source term to be reconstructed. To do so, strong assumptions are made on the nature of the source. The inverse problem is hence to estimate these parameters. On the other hand non-parametric methods attempt to reconstruct a full emission field. Several parametric and non-parametric methods are proposed and evaluated on real situations at a urban scale, with a CFD model taking into account buildings influence on the air flow. In these experiments, some proposed methods are able to localize the source with a mean error of 20 meters (i.e. 10% of the size of the domain), depending on the simulated situations and the inverse modeling methods.

Keywords : Atmospheric dispersion, accidental pollution, inverse modeling, source reconstruction.

Table des figures

2.1	Structure verticale de l'atmosphère	20
2.2	Impact de la turbulence sur la dispersion d'un panache	23
2.3	Schéma des processus mis en jeu dans la dispersion atmosphérique	24
3.1	Carte du domaine de simulation et des deux sous-domaines (France et Finlande)	51
3.2	Vraie source et moyenne des sources reconstruites	56
3.3	Erreur (RMSE) des sources reconstruites	57
3.4	Figure de mérite des champs de concentration prévus	57
3.5	Domaine de simulation et emplacement des points de mesures	61
3.6	Densité de probabilité des hyper-paramètres r et m	65
3.7	Probabilité marginale de l'hyper-paramètre r	66
3.8	Probabilité marginale de l'hyper-paramètre m	67
3.9	L-curve et courbure correspondante	67
3.10	Source reconstruite de césium-137 en statistiques gaussiennes	68
3.11	Source reconstruite de césium-137 en statistiques semi-gaussiennes	70
3.12	Panache réanalysé de césium-137	70
3.13	Comparaison entre les mesures de césium-137 à Tokyo-MITRI et la réanalyse .	71
3.14	Source reconstruite d'iode-131 en statistiques semi-gaussiennes	71
3.15	Panache réanalysé d'iode-131	72
3.16	Comparaison entre les mesures d'iode-131 à Tokyo-MITRI et la réanalyse . . .	73
3.17	Domaines de simulation WRF et POLAIR3D	75
3.18	Sources reconstruites de césium-137 et incertitudes associées en utilisant dif- férentes combinaisons des 3 jeux de données (4 panels)	79
3.19	Cartes des dépôts de césium-137 observées et réanalysées (4 panels)	80
3.20	Comparaison entre les observations de césium-137 et les réanalyses (3 panels) .	81
4.1	Rappel : Carte du domaine de simulation et des deux sous-domaines (France et Finlande)	95
4.2	Identification de la centrale responsable d'un rejet dans l'atmosphère	96
4.3	Évaluation de la robustesse de l'indicateur gaussien	97
4.4	Évaluation synthétique des indicateurs statistiques	98
4.5	Photographies de l'expérience MUST	99
4.6	Localisations des appareils de mesure de concentration	100
4.7	Localisations des sources des différents rejets	101
4.8	Domaine de simulation et son maillage (MUST)	103
4.9	Test de dualité de Code_Saturne	104
4.10	Source reconstruite par méthodes non-paramétriques (moindres carrés)	107
4.11	Source reconstruite par méthodes non-paramétriques (schéma de Krysta)	108
4.12	Illustration de la déviation du nuage de polluants dans la canopée	110

4.13	Localisation des sources vraie et reconstruites pour le cas 2640138	111
4.14	Localisation des sources vraie et reconstruites pour le cas 2682256	111
4.15	Densités de probabilité des paramètres de la source reconstruite pour le cas 2640138	113
4.16	Contours de densité des sources reconstruites pour le cas 2640138	114
4.17	Contours de densité des sources reconstruites pour le cas 2682256	114
4.18	Localisation des sources vraie et reconstruites pour les cas 2640138 et 2682256 avec rotation du vent amont	115
B.1	Convergence d'une chaîne de Markov pour le cas 2640138	136
C.1	Localisation des sources vraie et reconstruites pour le cas 2640138	138
C.2	Localisation des sources vraie et reconstruites pour le cas 2640246	138
C.3	Localisation des sources vraie et reconstruites pour le cas 2671852	139
C.4	Localisation des sources vraie et reconstruites pour le cas 2672213	139
C.5	Localisation des sources vraie et reconstruites pour le cas 2681829	140
C.6	Localisation des sources vraie et reconstruites pour le cas 2681849	140
C.7	Localisation des sources vraie et reconstruites pour le cas 2682256	141
C.8	Localisation des sources vraie et reconstruites pour le cas 2682320	141
C.9	Localisation des sources vraie et reconstruites pour le cas 2682353	142
C.10	Localisation des sources vraie et reconstruites pour le cas 2692054	142
C.11	Localisation des sources vraie et reconstruites pour le cas 2692131	143
C.12	Localisation des sources vraie et reconstruites pour le cas 2692157	143
C.13	Localisation des sources vraie et reconstruites pour le cas 2692223	144
C.14	Localisation des sources vraie et reconstruites pour le cas 2692250	144
C.15	Densités de probabilité <i>a posteriori</i> pour le cas 2640138	145
C.16	Densités de probabilité <i>a posteriori</i> pour le cas 2640246	146
C.17	Densités de probabilité <i>a posteriori</i> pour le cas 2671852	147
C.18	Densités de probabilité <i>a posteriori</i> pour le cas 2672213	148
C.19	Densités de probabilité <i>a posteriori</i> pour le cas 2681829	149
C.20	Densités de probabilité <i>a posteriori</i> pour le cas 2681849	150
C.21	Densités de probabilité <i>a posteriori</i> pour le cas 2682256	151
C.22	Densités de probabilité <i>a posteriori</i> pour le cas 2682320	152
C.23	Densités de probabilité <i>a posteriori</i> pour le cas 2682353	153
C.24	Densités de probabilité <i>a posteriori</i> pour le cas 2692054	154
C.25	Densités de probabilité <i>a posteriori</i> pour le cas 2692131	155
C.26	Densités de probabilité <i>a posteriori</i> pour le cas 2692157	156
C.27	Densités de probabilité <i>a posteriori</i> pour le cas 2692223	157
C.28	Densités de probabilité <i>a posteriori</i> pour le cas 2692250	158
C.29	Contours de densité des sources reconstruites pour le cas 2640138	159
C.30	Contours de densité des sources reconstruites pour le cas 2640246	159
C.31	Contours de densité des sources reconstruites pour le cas 2671852	160
C.32	Contours de densité des sources reconstruites pour le cas 2672213	160
C.33	Contours de densité des sources reconstruites pour le cas 2681829	161
C.34	Contours de densité des sources reconstruites pour le cas 2681849	161
C.35	Contours de densité des sources reconstruites pour le cas 2682256	162
C.36	Contours de densité des sources reconstruites pour le cas 2682320	162
C.37	Contours de densité des sources reconstruites pour le cas 2682353	163
C.38	Contours de densité des sources reconstruites pour le cas 2692054	163
C.39	Contours de densité des sources reconstruites pour le cas 2692131	164

C.40	Contours de densité des sources reconstruites pour le cas 2692157	164
C.41	Contours de densité des sources reconstruites pour le cas 2692223	165
C.42	Contours de densité des sources reconstruites pour le cas 2692250	165

Liste des tableaux

2.1	Échelles spatiales caractéristiques des phénomènes atmosphériques.	21
2.2	Les différentes approches de modélisation et leurs échelles associées.	27
2.3	Valeurs des constantes du modèle $k - \epsilon$ [Launder and Spalding, 1974].	37
3.1	Origines et nombre d'observations de concentration d'activité dans l'air publiquement communiquées au 31 août 2011.	60
3.2	Estimation des hyper-paramètres et de l'activité rejetée de césium-137 (1) . . .	64
3.3	Estimation des hyper-paramètres et de l'activité rejetée de césium-137 (2) . . .	69
3.4	Estimation des hyper-paramètres et de l'activité rejetée d'iode-131	72
3.5	Configuration et paramétrisation des processus physiques du modèle WRF . . .	74
3.6	Estimation des hyper-paramètres et de l'activité rejetée de césium-137 (3) . . .	78
3.7	Estimation des hyper-paramètres et de l'activité rejetée de césium-137 (4) . . .	82
4.1	Présentation des 14 cas MUST sélectionnés	101
4.2	Scores des simulations directes avec Code_Saturne	106
4.3	Scores moyens des 14 simulations directes avec Code_Saturne et avec le modèle de panache gaussien	106
4.4	Comparaison entre les estimateurs et la vraie source (MUST)	109
4.5	Comparaison entre les hauteurs des estimateurs et de la vraie source	110

Table des matières

1	Introduction	15
1.1	Dispersion atmosphérique : des processus à la modélisation numérique	16
1.2	Estimation de sources par modélisation inverse	17
1.3	Localisation de sources accidentelles	17
2	Dispersion atmosphérique : des processus à la modélisation numérique	19
2.1	Dispersion atmosphérique des polluants : Processus	19
2.1.1	Le transport et la turbulence	20
2.1.2	Les transformations physico-chimiques et les processus de perte	22
2.2	Formalisme : Équation de chimie-transport	24
2.3	Modélisation numérique de la dispersion atmosphérique	26
2.3.1	Les modèles eulériens à grande échelle et à méso-échelle	28
2.3.2	Les modèles lagrangiens	31
2.3.3	Les modèles gaussiens	32
2.3.4	Les modèles de mécanique des fluides à l'échelle locale/urbaine	34
3	Estimation de sources par modélisation inverse	39
3.1	Introduction	40
3.2	Cadre bayésien de la modélisation inverse	41
3.3	Statistiques gaussiennes	42
3.3.1	De la formule de Bayes aux moindres carrés	42
3.3.2	L'estimateur BLUE	43
3.4	Statistiques gaussiennes tronquées	44
3.5	Estimation de l'incertitude <i>a posteriori</i>	45
3.6	Estimation des erreurs <i>a priori</i>	45
3.6.1	Dépendance de l'inversion aux erreurs <i>a priori</i>	45
3.6.2	La L-curve	47
3.6.3	Principe de la vraisemblance maximale	48
3.6.4	Schéma itératif de Desroziers	49
3.7	Application I : Surveillance d'un parc de centrales nucléaires	50
3.7.1	Installations nucléaires surveillées et réseaux d'observation	50
3.7.2	Modèles de dispersion atmosphérique	51
3.7.3	Système opérationnel	53
3.7.4	Expériences synthétiques	54
3.7.5	Résultats	56
3.8	Application II : Fukushima (1)	58
3.8.1	Contexte	58
3.8.2	Observations	59
3.8.3	Modèle de dispersion atmosphérique	59

3.8.4	Modélisation inverse	61
3.8.5	Résultats	64
3.9	Application III : Fukushima (2)	72
3.9.1	Observations	73
3.9.2	Modèle	74
3.9.3	Modélisation inverse	76
3.9.4	Résultats	77
4	Localisation de sources accidentelles	85
4.1	Introduction	85
4.2	Méthodes non-paramétriques	87
4.2.1	Statistiques gaussiennes	87
4.2.2	Statistiques non-gaussiennes	87
4.3	Méthodes paramétriques	89
4.3.1	Cas d'une source ponctuelle	91
4.3.2	Généralisation	91
4.4	Application I : Identifier le lieu d'un rejet radiologique	92
4.4.1	Principes et formalisme	92
4.4.2	Validation	94
4.5	Application II : MUST	98
4.5.1	Contexte - présentation de l'expérience	98
4.5.2	Cas d'étude sélectionnés	99
4.5.3	Modèles de dispersion	102
4.5.4	Reconstruction d'une source ponctuelle à partir des données réelles	104
5	Conclusion et perspectives	117
5.1	Conclusion	117
5.1.1	La dispersion atmosphérique de polluant : des processus à la modélisation numérique	117
5.1.2	Estimation de source accidentelle par modélisation inverse	118
5.1.3	Localisation de sources accidentelles	119
5.2	Perspectives	120
5.2.1	Temps de calcul et cadre opérationnel	120
5.2.2	Amélioration des modèles de dispersion	120
5.2.3	Assimilation des débits de dose γ	121
5.2.4	Généralisation pour l'estimation d'émissions non-accidentelles	121
5.2.5	Localisation de sources multiples	122
Annexe A	Calcul de la vraisemblance marginale des observations	131
Annexe B	Présentation de l'algorithme MCMC	135
Annexe C	Ensemble des résultats sur l'expérience MUST	137
Annexe D	Publication I	167
Annexe E	Publication II	185
Annexe F	Publication III	199
Annexe G	Collaborations	217

Chapitre 1

Introduction

Dans l'éventualité d'une pollution atmosphérique d'origine accidentelle ou terroriste, les objectifs des autorités sont notamment :

- Si cela s'avère nécessaire et possible, intervenir sur la source afin d'interrompre, ou à défaut de maîtriser, le rejet de polluant dans l'atmosphère.
- Déterminer les zones impactées à court terme, évaluer leur contamination et décider de l'évacuation éventuelle de populations.
- Déterminer les zones qui restent dangereuses à plus long terme par la présence, notamment dans le sol, de polluants à longue durée de vie.

Des modèles numériques simulant la dispersion du polluant dans l'atmosphère et son dépôt au sol au cours du temps peuvent aider à réaliser les deux derniers points. Néanmoins ces modèles sont fortement dépendant d'une bonne connaissance du terme source, qui caractérise spatialement et par leur évolution temporelle les émissions du polluant dans l'atmosphère. De même le premier point n'est envisageable qu'à condition de connaître l'origine du rejet de polluant, ce qui n'est pas toujours le cas.

Dans le même temps, les réseaux de capteurs se développent, se densifient et sont capables de communiquer en quasi temps réel les concentrations de nombreux polluants. Ces observations peuvent être utilisées pour corriger les résultats fournis par le modèle par des méthodes d'assimilation de données, et/ou pour estimer des paramètres d'entrée du modèle, notamment le terme source, par des méthodes de modélisation inverse.

Les méthodes de modélisation inverse pour l'estimation de source, qui sont le thème principal de ce travail de thèse, s'appuient sur l'équilibre entre des informations *a priori* et de nouvelles informations apportées par les observations et le modèle numérique de dispersion atmosphérique. Ainsi la première étape consiste à étudier la dispersion atmosphérique, de la connaissance des processus physiques mis en jeu jusqu'à sa modélisation numérique.

1.1 Dispersion atmosphérique : des processus à la modélisation numérique

Le chapitre 2 est consacré à la dispersion de polluant dans l'atmosphère. Afin de la modéliser efficacement, il est nécessaire de bien connaître et comprendre les différents processus physiques mis en jeu. Une fois le polluant émis, son séjour dans l'atmosphère peut se résumer à trois processus :

- Le **transport** influencé par les circulations d'air dans l'atmosphère et en particulier dans la couche limite atmosphérique, milieu pouvant être très turbulent.
- Les **transformations physico-chimiques** qui modifient la concentration du polluant au cours de son séjour dans l'atmosphère : la filiation radioactive (pour les radioéléments), les réactions chimiques, éventuellement multi-phases, les processus liés aux particules (coagulation, évaporation/condensation).
- Les **processus de perte par dépôt** : d'une part le dépôt sec et d'autre part le dépôt humide, sous l'influence des précipitations.

Pour une espèce considérée comme une espèce gazeuse ou mono-dispersée, les processus sont formalisés, de la façon la plus simple, par une équation d'advection-diffusion-chimie, présentée dans le chapitre 2.2, qui décrit l'évolution du champ de concentration du polluant. En règle générale, il n'existe pas de solution analytique à cette équation et des modèles numériques sont créés et utilisés pour la résoudre. De nombreux types de modèles existent pour simuler la dispersion de polluant. Ils diffèrent l'un de l'autre par l'échelle de la simulation (de quelques mètres jusqu'à l'échelle globale) et par le niveau de détail désiré ou envisageable avec les performances informatiques actuelles. Les modèles suivants figurent parmi les plus utilisés en contexte accidentel et sont présentés en détail dans le chapitre 2.3 :

- Les **modèles eulériens de chimie-transport** (CTM pour Chemistry Transport Model) à échelle régionale ou méso-échelle peuvent être utilisés pour simuler la dispersion sur des distances allant de quelques dizaines à plusieurs milliers de kilomètres avec des résolutions horizontales s'étalant approximativement entre 2 et 50 km. À ces échelles, l'influence de bâtiments est négligée et les champs de vent sont calculés en amont par des modèles de météorologie dédiés et utilisés comme données d'entrée du modèle de dispersion.
- Les **modèles lagrangiens particuliers** peuvent être également utilisés à l'échelle régionale et à méso-échelle. La formulation directe, respectivement inverse, s'appuie sur le suivi des trajectoires de particules numériques émises depuis les sources, respectivement les récepteurs, et dont le mouvement est composé d'une composante déterministe, rendant compte de l'influence du champ de vent moyen, et d'une composante stochastique rendant compte des fluctuations dues à la turbulence.
- Les **modèles gaussiens** peuvent être utilisés pour simuler la dispersion sur des distances allant de quelques dizaines de mètres à quelques dizaines de kilomètres. À ces échelles et sous certaines hypothèses, une solution analytique du champ de concentration est obtenue sous la forme d'un panache gaussien ou d'une série de bouffées gaussiennes.

- Les **modèles de mécanique des fluides** (CFD pour Computational Fluid Dynamics) sont utilisés à échelle locale, notamment quand l'influence des bâtiments ne peut plus être négligée. Par résolution des équations de Navier-Stokes, ils calculent avec précision les champs d'un écoulement perturbé par le milieu bâti et la dispersion du polluant dans cet écoulement.

1.2 Estimation de sources par modélisation inverse

Les données d'entrée et les paramétrisations utilisées dans les modèles sont sources d'incertitude dans les résultats obtenus. Parmi les principales sources d'incertitude, la précision des champs de vent et des champs de précipitation a une grande influence sur le résultat du modèle. Mais en situation accidentelle, la plus importante source d'incertitude est la mauvaise connaissance du terme source. Cette incertitude sur le terme source est évidente quand le lieu même du rejet n'est pas connu. Néanmoins même dans le cas contraire, il est difficile d'estimer la quantité de polluant rejetée et l'évolution temporelle de ces rejets.

Les techniques de modélisation inverse proposées dans le chapitre 3 permettent d'utiliser objectivement l'information contenue dans les observations et le modèle numérique pour corriger l'information *a priori* sur le terme source. Dans un cadre bayésien, ces différentes informations sont contenues dans les distributions statistiques des grandeurs considérées. En choisissant des statistiques gaussiennes pour représenter les erreurs *a priori* dans le système, l'estimation du terme source passe par la minimisation d'une fonction quadratique mesurant l'écart entre les observations et la simulation. La linéarité des modèles de dispersion relativement à la source permet de plus d'obtenir des algorithmes relativement rapides qu'il est envisageable d'implémenter en contexte opérationnel, ce qui est proposé dans le chapitre 3.7 pour la surveillance d'un parc d'installations nucléaires européen.

Les méthodes proposées, en estimant rigoureusement les erreurs d'observation *a priori*, sont capables d'estimer le terme source en utilisant dans la même inversion différents types de données, ce qui est important en situation accidentelle où le nombre d'observations est en général faible. Elles sont également capables d'estimer l'incertitude *a posteriori* sur le terme source estimé. Nous avons appliqué ces méthodes pour estimer les termes sources de césium-137 et d'iode-131 consécutifs à l'accident de la centrale nucléaire de Fukushima Daiichi en 2011 dans les chapitres 3.8 et 3.9.

1.3 Localisation de sources accidentelles

Dans le cadre général de l'estimation des émissions de polluants atmosphériques par modélisation inverse, la localisation de sources accidentelles présente un certain nombre de particularités qu'il est possible de prendre en compte dans nos formalismes pour proposer des méthodes de modélisation inverse spécifiquement adaptées à ce problème.

La particularité principale vient du fait qu'au cours d'une situation accidentelle, la source est très certainement d'étendue spatiale limitée, voire ponctuelle. À partir de ce constat, deux stratégies sont envisageables pour reconstruire la source :

- Les **méthodes non-paramétriques** ont pour finalité de reconstruire un champ complet d'émission en 4 dimensions. En prolongement du chapitre 3, des statistiques gaussiennes peuvent être envisagées, mais risquent de ne pas contraindre suffisamment le système pour reconstruire une source localisée. Des définitions d'ébauche différentes sont envisagées afin d'introduire dans le système l'information *a priori* sur l'étendue limitée de

la source.

- Les **méthodes paramétriques** reposent sur la paramétrisation de la source potentielle qui imposent au système de reconstruire une source strictement fidèle à cette paramétrisation (au contraire des méthodes non-paramétriques). Par construction, le nombre de paramètres est faible et il est possible, par exemple par des méthodes stochastiques telles que les algorithmes de Monte Carlo par chaînes de Markov (MCMC), d'obtenir un estimateur de chacun d'entre eux, ainsi que leurs statistiques *a posteriori*.

Pour tester les différentes méthodes proposées, nous allons, dans le chapitre 4.5, utiliser les données issues de la campagne expérimentale MUST (Mock Urban Setting Test), et en particulier les données collectées lors de 14 expériences de dispersion d'un gaz traceur. Pour chacun de ces 14 cas, l'objectif est de localiser le lieu du rejet en utilisant un modèle de CFD et les 48 mesures de concentrations.

Dans le chapitre 4.4, des indicateurs statistiques seront proposés pour déterminer, dans un ensemble de quelques sites suspectés, le lieu d'un rejet accidentel. Ces indicateurs seront évalués sur des expériences synthétiques à l'échelle européenne, configurées pour la surveillance d'un parc d'installations nucléaires à cette échelle.

Chapitre 2

Dispersion atmosphérique : des processus à la modélisation numérique

Sommaire

2.1	Dispersion atmosphérique des polluants : Processus	19
2.1.1	Le transport et la turbulence	20
2.1.2	Les transformations physico-chimiques et les processus de perte	22
2.2	Formalisme : Équation de chimie-transport	24
2.3	Modélisation numérique de la dispersion atmosphérique	26
2.3.1	Les modèles eulériens à grande échelle et à méso-échelle	28
2.3.2	Les modèles lagrangiens	31
2.3.3	Les modèles gaussiens	32
2.3.4	Les modèles de mécanique des fluides à l'échelle locale/urbaine	34

2.1 Dispersion atmosphérique des polluants : Processus

À partir de son émission dans l'atmosphère, terme incertain que nous cherchons à estimer, le séjour d'un polluant dans l'atmosphère peut se schématiser en trois processus :

- Le transport sous l'influence des circulations d'air dans l'atmosphère. Dans la partie basse de l'atmosphère, appelée couche limite atmosphérique, l'écoulement d'air peut être très turbulent, ce qui a un impact sur le mélange de polluant au cours de son transport mais également sur la capacité des modèles numériques à rendre compte de toutes les échelles de l'écoulement.
- Les transformations physico-chimiques : condensation/évaporation/coagulation pour les particules, réactions chimiques et photo-chimiques, éventuellement multi-phasiques, et filiation radioactive pour les radio-éléments. Étant donné les polluants généralement mis en jeu et les échelles de temps caractéristiques des situations accidentelles, seule la filiation radioactive est éventuellement prise en compte.
- Les processus de perte par dépôt sec et lessivage par les précipitations. Ces processus appauvrissent l'atmosphère en polluant et il est important de bien les prendre en compte non seulement pour la précision du champ de concentration du polluant dans l'atmosphère, mais aussi pour la bonne évaluation des conséquences à long terme du fait

d'éventuels dépôts d'espèces à longue durée de vie.

2.1.1 Le transport et la turbulence

Structure verticale de l'atmosphère

L'atmosphère terrestre est stratifiée en plusieurs couches dont les limites correspondent aux discontinuités des variations de la température en fonction de l'altitude. Les quatre premières couches, la troposphère, la stratosphère, la mésosphère et la thermosphère sont représentées sur la figure 2.1. La troposphère, qui s'étend du sol jusqu'à une altitude approximative de 7 – 8 km au niveau des pôles et jusqu'à environ 15 km à l'équateur, est la couche où se produisent les phénomènes météorologiques principaux, tels que pluies et nuages par exemple, et les grands mouvements atmosphériques. À l'intérieur de la troposphère, la couche limite atmosphérique (CLA), dont la hauteur varie de quelques centaines de mètres à 2 – 3 km environ suivant les conditions météorologiques, joue un rôle important dans de nombreux processus physiques de l'atmosphère. C'est en effet dans la CLA que sont émis, transportés et transformés la plupart des polluants. De plus, la surface terrestre joue un rôle primordial dans le bilan énergétique de l'atmosphère et les flux turbulents, essentiellement produits par les forçages thermiques et dynamiques du sol, y déterminent les échanges entre la surface et l'atmosphère [Stull, 1988; Garratt, 1992]. La partie de la troposphère située au-dessus de la CLA est appelée "troposphère libre".

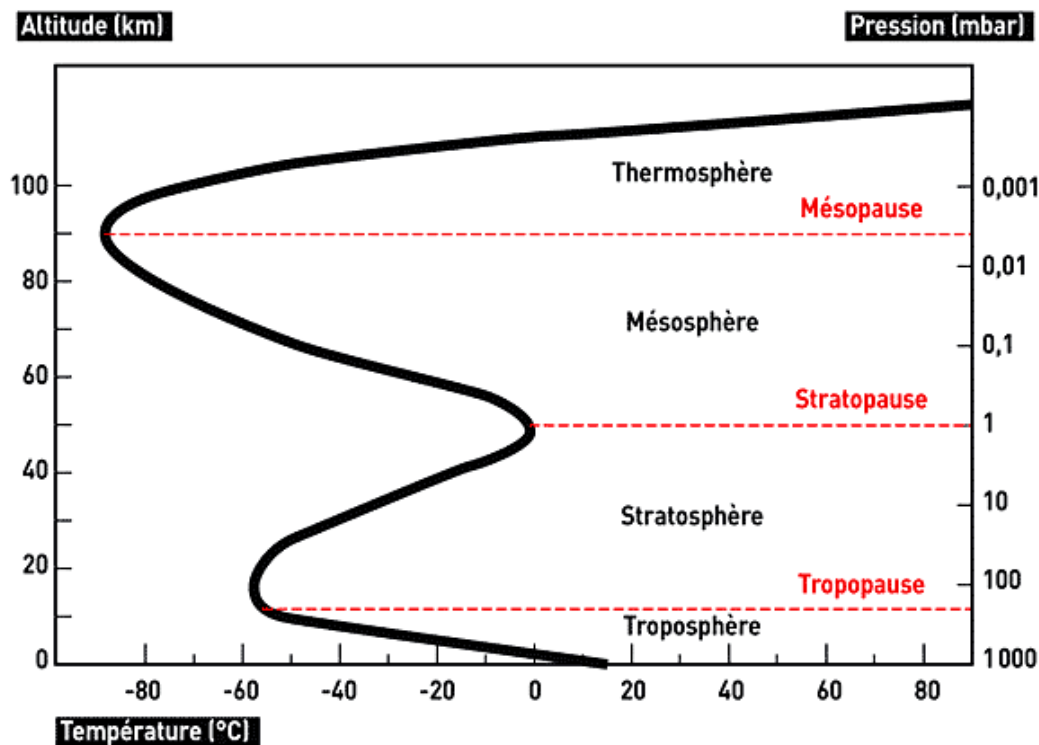


FIG. 2.1: Structure verticale de l'atmosphère (Source : <http://www.cea.fr>).

TAB. 2.1: Échelles spatiales caractéristiques des phénomènes atmosphériques.

Échelle	Taille caractéristique	Phénomènes atmosphériques typiques
Échelle synoptique	10 ⁴ km	Mouvements synoptiques
Méso-échelle	1 km – 1000 km	Mouvements thermiques ascendants, orage
Micro-échelle	< 1 km	Turbulence, influence des obstacles (bâtiments)

Le transport

Le transport des polluants dans l'atmosphère est principalement influencé par la météorologie. Plusieurs échelles spatiales et temporelles peuvent être distinguées et sont caractéristiques de différents phénomènes atmosphériques (tableau 2.1). À grande échelle (*échelle synoptique*), les phénomènes atmosphériques, notamment le vent, sont indépendants de la surface terrestre et ne dépendent plus que de la force de Coriolis et du gradient de pression. Les mouvements synoptiques correspondent au transport de polluants à l'échelle intercontinentale. À méso-échelle, la surface terrestre (orographie, occupation du sol) influence davantage les phénomènes atmosphériques [Perkins et al., 2002]. Le vent vertical est généralement faible et le transport vertical des polluants est principalement régi par la turbulence, c'est-à-dire des mouvements tourbillonnaires couvrant un large spectre d'échelles (voir chapitre 2.1.1). À grande échelle et à méso-échelle, le transport vertical turbulent est principalement lié à l'effet de *flottabilité* qui se traduit dans la CLA par des zones de mouvements turbulents ascendants dus à un gradient de température, dans lesquelles les polluants sont transportés rapidement sur toute la hauteur de la CLA.

À micro-échelle, l'influence de la surface terrestre, de la canopée, voire des bâtiments est importante ; à l'opposé la force de Coriolis devient négligeable. En milieu urbain, du fait de la résistance que les obstacles exercent sur le vent et d'une augmentation de la turbulence, due à une rugosité plus importante des surfaces urbaines, un ralentissement local de l'écoulement est visible. Les bâtiments peuvent dévier horizontalement ou verticalement l'écoulement moyen qui les surplombe [Bornstein, 1987]. Dans le sillage des bâtiments, des zones de recirculation se développent et sont propices au mélange, à la diffusion de quantité de mouvement et à la dissipation de l'énergie turbulente [Oke, 1987]. De plus, une ville a une température plus élevée que le milieu rural qui l'entoure (phénomène d'îlot de chaleur), du fait d'une part des sources de chaleurs anthropiques (chauffage par exemple) et d'autre part du comportement radiatif du milieu bâti lui-même.

La turbulence

La couche limite atmosphérique est un milieu très turbulent dans lequel des mouvements d'air, semblables à des tourbillons, dissipent l'énergie cinétique de l'écoulement. Ces tourbillons participent à la dispersion des polluants dans l'atmosphère, notamment sur la verticale. Un écoulement turbulent se caractérise par un large spectre d'échelles spatiales et temporelles mises en jeu. Les plus petites échelles ne peuvent généralement pas être représentées dans les modèles et sont alors représentées par le mécanisme de *diffusion*, qui est responsable de la dilution des polluants le long de leur parcours dans l'atmosphère et donc de la baisse des concentrations en aval des sources.

Échelles de la turbulence

Pour décrire les écoulements turbulents, on utilise souvent la notion de *cascade d'énergie* : l'énergie cinétique est "transférée" des grandes échelles aux petites échelles quand les grands tourbillons se divisent en tourbillons plus petits. À la plus petite échelle, appelée *échelle de Kolmogorov*, l'énergie est dissipée par la diffusion moléculaire [Kolmogorov, 1991; Lumley and Panofsky, 1964]. Les petites échelles sont caractérisées par le taux de dissipation de l'énergie turbulente ϵ et par la viscosité dynamique de l'air ν . Les échelles de Kolmogorov, de longueur l_k et de vitesse v_k , sont alors définies :

$$l_k \simeq \nu^{3/4} \epsilon^{-1/4}, \quad v_k \simeq \nu^{1/4} \epsilon^{1/4}. \quad (2.1)$$

Une estimation du ratio entre l'échelle spatiale L représentative des grandes échelles de l'écoulement, et l'échelle de Kolmogorov l_k est alors donnée par :

$$\frac{L}{l_k} \simeq Re^{3/4}, \quad (2.2)$$

avec Re le nombre de Reynolds caractéristique de l'écoulement :

$$Re = \frac{U L}{\nu}, \quad (2.3)$$

où U est la vitesse du vent. Pour des valeurs typiques de la CLA ($U = 10 \text{ m.s}^{-1}$ et $\nu \simeq 10^{-5} \text{ m}^2\text{s}^{-1}$), le nombre de Reynolds est de l'ordre de 10^9 , ce qui caractérise un écoulement très turbulent. Le rapport entre les plus grandes échelles et les plus petites est donc très grand et il est difficile pour un modèle de représenter l'ensemble des échelles.

Impact de la turbulence sur la dispersion des polluants

Toutes les échelles de la turbulence n'auront pas le même impact sur la dispersion des polluants. Pour illustrer cette idée, considérons un nuage de polluants émis par une source ponctuelle (figure 2.2). Les tourbillons de grande taille (comparée à la taille du nuage de polluants) ne participent pas à sa diffusion, mais le transportent "en l'état". Les tourbillons de petite taille participent également peu au mélange, même si les petits tourbillons à la limite du nuage participent au mélange entre air frais et air pollué, ce qui crée une dispersion uniforme du nuage. Par contre, les tourbillons dont la taille caractéristique est comparable à celle du nuage de polluants participent fortement à la dispersion du panache, déformé sous l'effet de la turbulence non homogène. Dans l'atmosphère, quelque soit la taille du nuage de polluants, il existera toujours des tourbillons de taille inférieure. En conséquence un nuage de polluants sera toujours dispersé dans l'atmosphère. L'efficacité de ce mélange dépendra de la turbulence et de l'échelle caractéristique du nuage de polluants qui augmente à mesure que celui-ci s'éloigne de sa source.

2.1.2 Les transformations physico-chimiques et les processus de perte

Transformations physico-chimiques

En qualité de l'air, la prise en compte de réactions chimiques et/ou photochimiques est nécessaire pour rendre compte de processus de production/destruction des différentes espèces. Ces processus chimiques sont relativement complexes et généralement non-linéaires et de nombreuses réactions concernant de nombreuses espèces doivent être prises en compte. Néanmoins, dans les situations accidentelles étudiées, les espèces mises en jeu peuvent être considérées

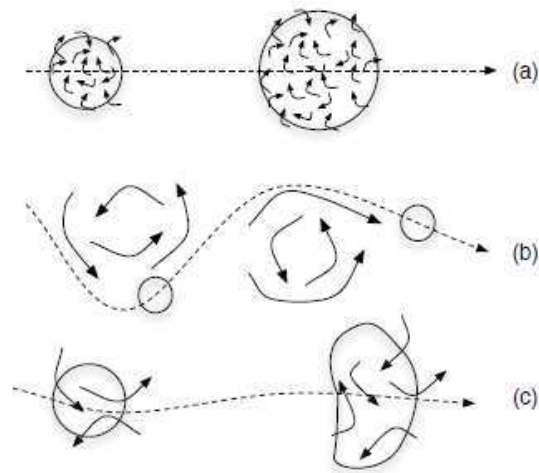


FIG. 2.2: Dispersion d'un nuage de polluants : (a) la taille caractéristique des tourbillons est beaucoup plus petite que celle du nuage, (b) la taille caractéristique des tourbillons est beaucoup plus grande que celle du nuage, (c) les tailles caractéristiques des tourbillons et du nuage de polluants sont comparables (Source : Seinfeld and Pandis [1998]).

comme des espèces passives chimiquement, au moins sur les échelles de temps considérées, ou bien soumises à des processus de filiation radioactive, relativement simples à modéliser.

De plus, si les polluants considérés sont présents dans l'atmosphère sous forme de particules, des processus tels que la condensation, l'évaporation et la coagulation doivent être pris en compte. C'est par exemple le cas de l'iode-131, qui, en cas d'incident ou d'accident sur une centrale nucléaire, peut être rejeté dans l'atmosphère sous forme gazeuse majoritairement, mais qui au cours de son séjour dans l'atmosphère forme en partie des particules, rendant sa modélisation et sa mesure difficiles. Pour les plus grosses particules, la sédimentation gravitationnelle est également un processus d'importance. Ces processus rendent le modèle plus complexe et plus coûteux et, les méthodes d'identification de sources présentées dans ce manuscrit nécessitant l'utilisation intensive de modèles, nous envisagerons dès que possible de les négliger. Un exemple de prise en compte de la sédimentation gravitationnelle est néanmoins visible dans Chazette et al. [2012], dont un résumé est disponible dans l'annexe G, pour la modélisation de la dispersion du panache de cendre émis par le volcan islandais Eyjafjöll en 2010.

Dépôt sec et lessivage

En situation accidentelle, en plus de l'éventuelle décroissance radioactive, le principal processus de perte est le dépôt de polluants sur le sol. On distingue deux types de dépôt :

- le dépôt sec représente le transfert, de l'atmosphère vers le sol, de polluants situés près de la surface, notamment par absorption par les végétaux. Il dépend du type de sol (bâtiments, végétaux) et du polluant considéré. On comprend alors que l'échelle de la variabilité spatiale peut être de quelques mètres seulement. De plus, si l'information sur l'occupation des sols est relativement facile à obtenir, elle contient des incertitudes (mise à jour des informations, variations saisonnières).
- le dépôt humide, ou lessivage, est l'entraînement au sol des polluants par les précipita-

tions, qui peuvent être également très locales. La présence de pluie et de nuages à un endroit précis est difficile à diagnostiquer, ce qui sera également source d'incertitudes dans les modèles.

2.2 Formalisme : Équation de chimie-transport

Considérons une espèce X_i de concentration c_i dans l'atmosphère, modélisée sous forme gazeuse ou comme une particule mono-dispersée. Cette espèce est soumise aux processus détaillés dans la partie précédente et schématiquement rappelés sur la figure 2.3 : transport, diffusion, filiation radioactive, dépôt sec et lessivage.

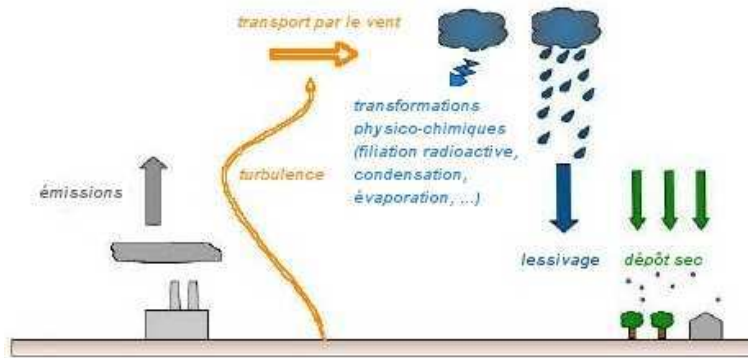


FIG. 2.3: Schéma résumé des processus pris en compte dans la dispersion de polluants en situation accidentelle.

Équation d'advection-diffusion-chimie

Nous faisons l'hypothèse que la dynamique de l'écoulement n'est pas influencée par les polluants, et qu'en conséquence les champs météorologiques (notamment les champs de vent) sont connus (modélisation *offline*, cf. chapitre 2.3). L'évolution de la concentration c_i est donnée par l'équation d'advection-diffusion-chimie, qui découle du principe de conservation de la masse :

$$\underbrace{\frac{\partial c_i}{\partial t}}_{(1)} + \underbrace{\text{div}(\mathbf{u}c_i)}_{(2)} = \underbrace{\text{div}(\mathbf{D}_{m_i}\nabla c_i)}_{(3)} + \underbrace{\chi_i(c)}_{(4)} + \underbrace{\sigma_i}_{(5)} + \underbrace{P_i(c_i)}_{(6)}, \quad (2.4)$$

avec :

- (1) le terme instationnaire,
- (2) le terme d'advection qui représente le transport par l'écoulement d'air, dépendant du champ de vent $\mathbf{u} = (u_x, u_y, u_z)$,
- (3) le terme de diffusion moléculaire, décrite par une loi de Fick. La matrice diagonale \mathbf{D}_{m_i} contient les coefficients de diffusion moléculaire et dépend de l'espèce X_i ,
- (4) le terme de production ou de destruction de X_i par filiation radioactive éventuelle. Si l'espèce n'est pas radioactive, ce terme est nul, sinon il dépend des concentrations $c = (c_1, \dots, c_n)$ de l'ensemble des n espèces dans la chaîne de filiation.

- (5) σ_i représente les émissions (positives),
- (6) $P_i(c_i)$ représente les pertes par dépôt.

Représentation de la turbulence

L'équation 2.4 décrit l'évolution de la concentration c_i au niveau "microscopique". Cependant, comme nous l'avons évoqué dans le chapitre 2.1.1, il est difficile de prendre en compte toutes les échelles de la turbulence dans les modèles. Seule la DNS (Direct numerical simulation) permet de le faire, mais les capacités informatiques actuelles permettent seulement de l'envisager à échelle très locale. Dans les types de modèles généralement utilisés aujourd'hui, une autre approche, la *décomposition de Reynolds*, est privilégiée.

Si Φ représente une variable liée à l'écoulement (vitesse, température, concentration, etc.), on décompose ce champ entre une valeur moyenne, notée $\langle \Phi \rangle$, et une fluctuation turbulente notée Φ' , telles que $\Phi = \langle \Phi \rangle + \Phi'$. Par définition, la moyenne des fluctuations est nulle : $\langle \Phi' \rangle = 0$, et $\langle \Phi_1 \Phi_2 \rangle = \langle \Phi_1 \rangle \langle \Phi_2 \rangle + \langle \Phi_1' \Phi_2' \rangle$. Ainsi en appliquant dans l'équation 2.4 cette décomposition et en moyennant ensuite, on obtient alors après simplification et en prenant en compte la linéarité de la filiation radioactive (et donc de l'opérateur χ_i) :

$$\frac{\partial \langle c_i \rangle}{\partial t} + \text{div}(\langle \mathbf{u} \rangle \langle c_i \rangle) + \text{div}(\langle \mathbf{u}' c_i' \rangle) = \text{div}(\mathbf{D}_{m_i} \nabla \langle c_i \rangle) + \chi_i(\langle c \rangle) + \langle \sigma_i \rangle + \langle P_i \rangle. \quad (2.5)$$

L'équation 2.5 décrit l'évolution des concentrations moyennes à l'échelle du modèle (moment d'ordre 1). En général la modélisation concerne un domaine fini ; pour résoudre l'équation, on prend alors en compte des conditions aux limites, données par les flux entrants aux frontières latérales du domaine simulé, par le dépôt sec et les éventuelles émissions de surface au niveau du sol, et généralement par un flux nul sur la face supérieure du domaine.

Néanmoins il reste à estimer les flux turbulents $\langle \mathbf{u}' c_i' \rangle$, ou plus exactement $\langle u_x' c_i' \rangle$, $\langle u_y' c_i' \rangle$ et $\langle u_z' c_i' \rangle$. Une approche classique est la fermeture de rang 1, ou théorie- \mathbf{K} [Panofsky and Dutton, 1984]. Dans ce schéma de fermeture, les flux turbulents d'une variable Φ sont exprimés en fonction du gradient de cette variable moyennée, à l'instar de la loi de Fick pour la diffusion moléculaire. On écrit alors :

$$\langle \mathbf{u}' \Phi' \rangle = -\mathbf{K}_\Phi \nabla \langle \Phi \rangle, \quad (2.6)$$

ce qui donne pour la fraction molaire $C_i = c_i/\rho$, où ρ est la masse volumique de l'air :

$$\langle \mathbf{u}' C_i' \rangle = -\mathbf{K} \nabla \langle C_i \rangle, \quad (2.7)$$

et donc pour la concentration, en négligeant les fluctuations de densité ($\rho \simeq \langle \rho \rangle$),

$$\langle \mathbf{u}' c_i' \rangle = -\rho \mathbf{K} \nabla \frac{\langle c_i \rangle}{\rho}. \quad (2.8)$$

La matrice de diffusion turbulente \mathbf{K} est supposée la même pour toutes les espèces. En pratique, à partir de la méso-échelle on a également $\mathbf{K} \gg \mathbf{D}_m$ et la diffusion moléculaire est généralement négligée.

2.3 Modélisation numérique de la dispersion atmosphérique

Deux grands types de modèles peuvent être distingués :

- les modèles *online* qui résolvent les lois de conservation et prennent en compte des couplages entre les équations décrivant l'écoulement et l'évolution des concentrations de polluants. Ces modèles sont particulièrement utilisés en qualité de l'air car la constitution de l'atmosphère, en particulier en particules, a une influence sur les circulations d'air et sur les forçages thermiques notamment radiatifs. Par contre en situation accidentelle ce couplage est de moindre importance, voire inexistant.
- les modèles *offline* qui ne résolvent que les lois de conservation de la masse et considèrent donc que l'influence des polluants sur les champs météorologiques est négligeable. Ceux-ci peuvent donc être pré-calculés et utilisés comme données d'entrée du modèle. Les équations décrites dans la partie précédente se placent dans ce cadre, et c'est également dans ce cadre que se situe la suite de ce travail. En effet, en plus du fait que l'influence des polluants sur l'écoulement est faible en situation accidentelle, les méthodes d'estimation de sources présentées ici nécessitent l'usage intensif des modèles de dispersion, typiquement en faisant varier le terme source mais à *écoulement fixé*. En conséquence l'utilisation de modèles *online*, en plus d'être coûteuse, ne se justifie pas.

Parmi les modèles *offline*, diverses approches sont envisageables, suivant les situations (accidentelles, qualité de l'air) et surtout les échelles modélisées, c'est-à-dire la taille du domaine sur lequel les calculs de dispersion sont effectués. On distingue généralement l'échelle locale (jusqu'à quelques dizaines de kilomètres), l'échelle régionale (quelques centaines de kilomètres), l'échelle continentale (quelques milliers de kilomètres) et l'échelle globale. Parmi les différentes approches possibles, nous avons choisi de détailler celles qui, à l'heure actuelle, sont les plus utilisées, notamment en situations opérationnelles, et qui permettent de couvrir les différentes échelles citées précédemment.

- **les modèles eulériens de chimie-transport à grande échelle et méso-échelle** (CTM, Chemistry - Transport Model) : le domaine de simulation est discrétisé en une grille de calcul (ou maillage) sur laquelle est résolue l'équation 2.5 par des méthodes numériques appropriées. Il s'agit d'une approche déterministe, très utilisée en prévision de qualité de l'air, en étude d'impact ou en situation accidentelle et qui peut être utilisée à l'échelle régionale ou continentale, à condition de fournir en entrée du modèle des champs météorologiques pré-calculés à l'échelle correspondante. Cette approche est présentée plus en détails dans le chapitre 2.3.1.
- **les modèles lagrangiens particuliers** : à partir d'une ou de plusieurs sources ponctuelles (ou récepteurs), un grand nombre de particules numériques sont émises. Le modèle suit alors la trajectoire de chaque particule, son mouvement étant composé d'une composante déterministe liée au champ de vent moyen et d'une composante stochastique qui décrit la variabilité due à la turbulence. La concentration de polluants est ensuite évaluée en fonction de la répartition de ces particules numériques. Cette approche est présentée dans le chapitre 2.3.2.
- **les modèles gaussiens** : à l'échelle locale et sous certaines hypothèses simplificatrices, l'équation 2.5 peut conduire à des formules analytiques calculant les concentrations dans

TAB. 2.2: Les différentes approches de modélisation et leurs échelles associées.

Type de modèle	Domaine de simulation	Applications
CFD (<i>online</i> ou <i>offline</i>)		
DNS	échelle très locale $\sim 1\text{km}$	Recherche (turbulence)
LES	échelle très locale $\sim 1\text{km}$	Recherche (turbulence)
RANS	échelle locale $\sim 1 - 10\text{km}$	Environnement complexe, urbain
Modèles <i>offline</i>		
Modèles gaussiens	échelle locale $\sim 1 - 10\text{km}$	Impact local, risque industriel
Modèles lagrangiens	échelle locale $\sim 10\text{km}$ échelle régionale $\sim 100\text{km}$ échelle continentale $\sim 1000\text{km}$	Risque industriel, biologique
Modèles eulériens de chimie-transport	échelle régionale $\sim 100\text{km}$ échelle continentale $\sim 1000\text{km}$	Prévision (photochimie) Impact à grande échelle
Modèles globaux	échelle globale $\sim 10000\text{km}$	Transport intercontinentale
Modèles <i>online</i> à grande échelle		
CTM couplé	échelle continentale	Recherche (photochimie)
GCM (Global Climate Model)	échelle globale	Impact climatique

des *panaches* ou des *bouffées* de polluants. Ces modèles présentent l'avantage d'un coût de calcul très faible et sont présentés dans le chapitre 2.3.3.

- **les modèles de mécanique des fluides** (CFD, Computational Fluid Dynamics) : à l'échelle "très" locale, en particulier en milieu urbain ou plus généralement en présence d'obstacles qui perturbent l'écoulement d'air, la CFD permet de calculer, en résolvant les équations de Navier-Stokes sur un maillage fin et en prenant en compte la présence des obstacles, des écoulements complexes avec une bonne précision. L'évolution de la concentration de polluant est calculée par résolution de l'équation 2.5 sur le même maillage, soit dans le même temps soit par la suite. Cette approche est présentée dans le chapitre 2.3.4.

Les différents types de modèles sont répertoriés dans le tableau 2.2. Les modèles que nous ne présenterons pas en détail sont les modèles "extrêmes", dédiés soit à la modélisation à très petite échelle (DNS pour Direct Numerical Simulation, et LES pour Large Eddy Simulation), du fait des coûts calculs engendrés, soit à la modélisation à très grande échelle spatiale ou temporelle (les CTM couplés avec les modèles de météorologie et les modèles de prévision climatique).

2.3.1 Les modèles eulériens à grande échelle et à méso-échelle

Dans un modèle eulérien, le domaine de calcul est discrétisé en une grille de calcul. Par des méthodes numériques adaptées et en supposant que la concentration au centre de la maille c_i est représentative de la concentration moyenne dans la maille $\langle c_i \rangle$ (hypothèse valable plus largement pour toutes les grandeurs : vitesses de vent, etc.), le modèle résout alors l'équation d'advection-diffusion-chimie suivante, issue des équations 2.5 et 2.8 en abandonnant les symboles $\langle \cdot \rangle$:

$$\frac{\partial c_i(\mathbf{x}, t)}{\partial t} = -\operatorname{div}(\mathbf{u}(\mathbf{x}, t) c_i(\mathbf{x}, t)) + \operatorname{div}\left(\rho(\mathbf{x}, t) \mathbf{K}(\mathbf{x}, t) \nabla \frac{c_i(\mathbf{x}, t)}{\rho(\mathbf{x}, t)}\right) + \chi_i(c(\mathbf{x}, t), \mathbf{x}, t) + P_i(\mathbf{x}, t) + \sigma_i(\mathbf{x}, t). \quad (2.9)$$

Dans cette équation, $c_i(\mathbf{x}, t)$ représente la concentration moyenne de l'espèce X_i en \mathbf{x} et à l'instant t et $\mathbf{u}(\mathbf{x}, t)$ le champ de vent moyen. \mathbf{K} est la matrice de diffusion turbulente. ρ est la masse volumique de l'air. Les champs météorologiques (champs de vent, précipitations, etc.) sont pré-calculés par des modèles adaptés à l'échelle désirée et fournis en entrée du modèle de dispersion.

Les différents termes de l'équation 2.9 sont :

- Le terme correspondant au transport, représenté par deux composantes :
 - le terme d'advection : $\operatorname{div}(\mathbf{u}c_i)$, qui rend compte du transport du polluant par le champ de vent moyen.
 - le terme de diffusion : $\operatorname{div}\left(\rho \mathbf{K} \nabla \frac{c_i}{\rho}\right)$, qui rend compte des fluctuations des mouvements d'air dues aux échelles de turbulence non prises en compte explicitement par le modèle car plus petites que les mailles de celui-ci. À grande échelle, la diffusion moléculaire est négligée devant la diffusion turbulente et la matrice \mathbf{K} est supposée diagonale. La diffusion turbulente horizontale est supposée isotrope, de coefficient K_H constant, dépendant généralement de la résolution horizontale du modèle. Celui-ci est en outre peu sensible à la valeur de ce coefficient étant donné que le transport horizontal est principalement régi par l'advection par le champ de vent moyen et par la présence de diffusion numérique engendrée par les schémas numériques de résolution de l'équation d'advection-diffusion-chimie. Le modèle est par contre sensible au coefficient de diffusion turbulente verticale K_z . Parmi les paramétrisations possibles pour estimer ce coefficient, on peut citer les paramétrisations de Louis [1979] et de Troen and Mahrt [1986], qui reposent sur un diagnostic réalisé à partir des conditions météorologiques.
- Le terme de réactions chimiques : χ_i . En situation accidentelle, étant données les espèces et les échelles de temps considérées, ce terme est soit nul, soit décrit une filiation radioactive. Dans ce cas, on peut écrire :

$$\chi_i(c(\mathbf{x}, t), \mathbf{x}, t) = [\mathbf{F}c(\mathbf{x}, t)]_i, \quad (2.10)$$

où \mathbf{F} représente la matrice de filiation de l'ensemble de la chaîne de filiation et c les concentrations des espèces mises en jeu dans cette chaîne. Si l'espèce étudiée n'est engendrée par aucune autre espèce (cas du césium-137 et de l'iode-131 par exemple), ce

terme se simplifie en :

$$\chi_i(\mathbf{c}(\mathbf{x}, t), \mathbf{x}, t) = -\Lambda_i^d c_i(\mathbf{x}, t), \quad (2.11)$$

où Λ_i^d est le coefficient de décroissance radioactive de l'espèce X_i .

- Le terme rendant compte des processus de perte : P_i . Le dépôt sec sera incorporé au modèle par l'intermédiaire des conditions aux limites (au niveau du sol). Ainsi les processus de perte se résument au dépôt humide. Une paramétrisation simple consiste à écrire :

$$P_i(\mathbf{x}, t) = -\Lambda_i^s(\mathbf{x}, t) c_i(\mathbf{x}, t), \quad (2.12)$$

où Λ_i^s représente un coefficient de lessivage (l'exposant s pour "scavenging"), dépendant des conditions météorologiques. Plusieurs paramétrisations sont possibles pour le calcul de Λ_i^s . Certaines dépendent uniquement des flux de précipitations [Maryon et al., 1991], d'autres prennent en compte l'humidité relative [Pudykiewicz, 1989]. Différents exemples de paramétrisation sont donnés dans les chapitres 3.7, 3.8 et 3.9.

- le terme correspondant au champ d'émission volumique σ_i .

Afin de garantir l'unicité de la solution de l'équation 2.9, des conditions initiales et conditions aux limites sont définies :

- Les conditions initiales à $t = 0$ sont définies par :

$$c_i(\mathbf{x}, 0) = c_{i0}(\mathbf{x}), \quad (2.13)$$

qui est l'état de l'atmosphère au début de la modélisation. En situation accidentelle, ces conditions initiales seront souvent nulles.

- Les conditions aux limites au niveau du sol rendent compte du processus de dépôt sec et d'éventuelles émissions surfaciques :

$$\mathbf{K}(\mathbf{x}_{z=0}, t) \nabla c_i(\mathbf{x}_{z=0}, t) \cdot \mathbf{n} = E_i(\mathbf{x}_{z=0}, t) - v_i^d c_i(\mathbf{x}_{z=0}, t), \quad (2.14)$$

où \mathbf{n} est le vecteur unité orienté du sol vers l'intérieur du domaine. Le terme de gauche représente donc le flux entrant au niveau du sol. v_i^d est la vitesse de dépôt sec et E_i représente les émissions surfaciques éventuelles, généralement nulles en situation accidentelle. Comme pour le coefficient de lessivage, plusieurs paramétrisations sont disponibles pour estimer la vitesse de dépôt sec. La plus simple consiste à prendre une vitesse constante [Quélo et al., 2007; Davoine and Bocquet, 2007; Bocquet, 2012], éventuellement avec une valeur plus faible sur l'océan [Estournel et al., 2012; Korsakissok et al., 2013]. Il est également possible de prendre en compte la nature du sol plus précisément [Wesely, 1989; Zhang et al., 2001].

- Les conditions aux limites latérales sont généralement des conditions de Dirichlet sur les faces avec vent entrant. Si le polluant modélisé n'est pas présent dans l'atmosphère avant la situation accidentelle modélisée et si le domaine de calcul est suffisamment étendu, il est possible de choisir des conditions aux limites nulles. Les conditions aux limites au niveau de la limite supérieure du modèle sont généralement des conditions de flux nul ou de concentrations nulles. Une attention particulière sera portée sur l'étendue verticale du domaine afin de rester dans le cadre de cette hypothèse.

Synthèse et modèle adjoint

Considérons une espèce X , de concentration c , soumise à une décroissance radioactive (mais non engendrée par d'autres espèces) et absente de l'atmosphère en situation normale. La concentration c , supposée calculée sur un domaine très étendu, est donc une fonction continue de l'espace et du temps, solution de l'équation :

$$\begin{aligned} \frac{\partial c(\mathbf{x}, t)}{\partial t} = & - \operatorname{div}(\mathbf{u}(\mathbf{x}, t) c(\mathbf{x}, t)) + \operatorname{div}\left(\rho(\mathbf{x}, t) \mathbf{K}(\mathbf{x}, t) \nabla \frac{c(\mathbf{x}, t)}{\rho(\mathbf{x}, t)}\right) \\ & - \Lambda^d c(\mathbf{x}, t) - \Lambda^s(\mathbf{x}, t) c(\mathbf{x}, t) + \sigma(\mathbf{x}, t), \end{aligned} \quad (2.15)$$

avec :

– pour conditions initiales :

$$c(\mathbf{x}, 0) = 0. \quad (2.16)$$

– pour conditions aux limites au niveau du sol :

$$\mathbf{K}(\mathbf{x}_{z=0}, t) \nabla c(\mathbf{x}_{z=0}, t) \cdot \mathbf{n} = -v^d c(\mathbf{x}_{z=0}, t). \quad (2.17)$$

– pour conditions aux limites latérales et supérieure du domaine :

$$c(\mathbf{x} \in \partial \mathcal{D}_{z>0}^+, t) = 0. \quad (2.18)$$

où $\partial \mathcal{D}_{z>0}^+$ représente les limites du domaine spatial (noté \mathcal{D}) autres que le sol et où la direction du vent est dirigée vers l'intérieur du domaine.

Une observation μ peut alors être exprimée en fonction du champ de concentration c et d'une fonction d'échantillonnage π :

$$\mu = \int_{\Omega} c(\mathbf{x}, t) \pi(\mathbf{x}, t) \, d\mathbf{x} \, dt, \quad (2.19)$$

où $\Omega = \mathcal{D} \times [0, T]$ représente le domaine spatio-temporel de simulation.

Par ailleurs, en multipliant l'équation 2.15 par un champ scalaire ϕ et en intégrant sur tout le domaine spatio-temporel Ω , on obtient la forme *faible* du modèle :

$$\int_{\Omega} \phi \left(\frac{\partial c}{\partial t} + \operatorname{div}(\mathbf{u}c) - \operatorname{div}\left(\rho \mathbf{K} \nabla \frac{c}{\rho}\right) + \Lambda^d c + \Lambda^s c - \sigma \right) \, d\mathbf{x} \, dt = 0. \quad (2.20)$$

À l'aide d'intégrations par parties, on obtient [Pudykiewicz, 1998; Roustan and Bocquet, 2006] :

$$\begin{aligned} \int_{\Omega} c \left(-\frac{\partial \phi}{\partial t} - \operatorname{div}(\mathbf{u}\phi) - \operatorname{div}\left(\rho \mathbf{K} \nabla \frac{\phi}{\rho}\right) + \Lambda^d \phi + \Lambda^s \phi \right) \, d\mathbf{x} \, dt = \\ - \int_{\mathcal{D}} (\phi(T) c(T) - \phi(0) c(0)) \, d\mathbf{x} - \int_{\partial \mathcal{D}_{z>0} \times [0, T]} (\phi \mathbf{c} \mathbf{u}) \cdot d\mathbf{S} \, dt \\ - \int_{\partial \mathcal{D}_{z=0} \times [0, T]} (c \mathbf{K} \nabla \phi - \phi \mathbf{K} \nabla c) \cdot d\mathbf{S} \, dt + \int_{\Omega} \phi \sigma \, d\mathbf{x} \, dt, \end{aligned} \quad (2.21)$$

L'équation 2.21 est valable pour tout champ scalaire ϕ . On peut alors définir un scalaire défini par des propriétés particulières et l'introduire dans cette équation. Considérons le champ scalaire c^* solution de l'équation :

$$\begin{aligned} \frac{\partial c^*(\mathbf{x}, t)}{\partial (-t)} = & - \operatorname{div}(-\mathbf{u}(\mathbf{x}, t) c^*(\mathbf{x}, t)) + \operatorname{div}\left(\rho(\mathbf{x}, t) \mathbf{K}(\mathbf{x}, t) \nabla \frac{c^*(\mathbf{x}, t)}{\rho(\mathbf{x}, t)}\right) \\ & - \Lambda^d c^*(\mathbf{x}, t) - \Lambda^s(\mathbf{x}, t) c^*(\mathbf{x}, t) + \pi(\mathbf{x}, t), \end{aligned} \quad (2.22)$$

Par sa proximité avec l'équation 2.15, cette équation peut être vue comme une équation de dispersion atmosphérique d'un *rétro-panache*, avec une évolution du temps et un champ de vent inversés. On définit de plus les conditions finales et aux limites suivantes :

– Pour conditions finales :

$$c^*(\mathbf{x}, T) = 0. \quad (2.23)$$

– Pour conditions aux limites au niveau du sol :

$$\mathbf{K}(\mathbf{x}_{z=0}, t) \nabla c^*(\mathbf{x}_{z=0}, t) \cdot \mathbf{n} = -v^d c^*(\mathbf{x}_{z=0}, t). \quad (2.24)$$

– Pour conditions aux limites latérales et supérieures du domaine :

$$c^*(\mathbf{x} \in \partial \mathcal{D}_{z>0}^-, t) = 0. \quad (2.25)$$

où $\partial \mathcal{D}_{z>0}^-$ représente les limites du domaine spatial autres que le sol et où la direction du vent \mathbf{u} est dirigée vers l'extérieur du domaine.

Dans ce cas, les conditions initiales et aux limites de c associées aux conditions finales et aux limites de c^* se simplifient dans l'équation 2.21 qui devient :

$$\int_{\Omega} c \left(-\frac{\partial c^*}{\partial t} - \operatorname{div}(\mathbf{u}c^*) - \operatorname{div}\left(\rho \mathbf{K} \nabla \frac{c^*}{\rho}\right) + \Lambda^d c^* + \Lambda^s c^* \right) d\mathbf{x} dt = \int_{\Omega} c^* \sigma d\mathbf{x} dt. \quad (2.26)$$

Finalement en utilisant l'équation 2.22 qui décrit l'évolution de c^* , on obtient :

$$\mu = \int_{\Omega} c(\mathbf{x}, t) \pi(\mathbf{x}, t) d\mathbf{x} dt = \int_{\Omega} c^*(\mathbf{x}, t) \sigma(\mathbf{x}, t) d\mathbf{x} dt. \quad (2.27)$$

On dispose donc de deux méthodes pour calculer l'observation μ , l'une utilise la concentration c , solution du modèle direct décrit par les équations 2.15, 2.16, 2.17 et 2.18 ; l'autre méthode utilise la concentration adjointe c^* , solution du modèle adjoint décrit par les équations 2.22, 2.23, 2.24 et 2.25.

2.3.2 Les modèles lagrangiens

Dans un modèle eulérien, on observe, dans une grille fixe, l'évolution des concentrations d'un polluant porté par un fluide en mouvement. Le point de vue des modèles lagrangiens est différent : il s'agit maintenant de suivre la trajectoire de particules numériques émises à partir des sources dans le modèle. Au cours de son déplacement et au temps t , une particule k est caractérisée par sa position $\mathbf{x}_k(t)$ et sa masse $m_k(t)$. L'évolution de la position est décrite par l'équation :

$$\mathbf{x}_k(t + \Delta t) = \mathbf{x}_k(t) + \mathbf{v}_k(\mathbf{x}_k(t)) \Delta t, \quad (2.28)$$

où \mathbf{v}_k est la vitesse de la particule, composée d'une composante moyenne, notée $\langle \mathbf{v}_k \rangle$, et d'une composante turbulente, notée \mathbf{v}'_k . La variation de cette composante dans la i -ème direction est donnée par l'équation de Langevin :

$$\Delta \mathbf{v}'_{ki}(t) = a_i(\mathbf{x}_k, \mathbf{v}'_k, t) \Delta t + b_{ij}(\mathbf{x}_k, \mathbf{v}'_k, t) \Delta \mathbf{w}_j. \quad (2.29)$$

Dans cette équation, $\Delta \mathbf{w}_j$ est la composante incrémentale du processus stochastique de Wiener. Sa valeur moyenne est 0 et sa variance est égale à Δt . a_i et b_{ij} sont respectivement les termes d'advection et de diffusion (voir Stohl et al. [2005]).

Les processus de perte, c'est-à-dire la décroissance radioactive et les dépôts sec et humide, sont pris en compte dans l'évolution de la masse m_k , qui est décrite par :

$$m_k(t + \Delta t) = m_k(t) e^{-\zeta \Delta t}. \quad (2.30)$$

Le coefficient ζ est la somme des contributions des différents processus de perte :

- la décroissance radioactive

$$\zeta = \frac{\ln(2)}{T_{1/2}}, \quad (2.31)$$

où $T_{1/2}$ est la demi-vie de l'espèce modélisée,

- le dépôt humide

$$\zeta = \Lambda^s, \quad (2.32)$$

- le dépôt sec

$$\zeta = \frac{v^d(h_{\text{ref}})}{2h_{\text{ref}}}, \quad (2.33)$$

où h_{ref} représente une hauteur de référence (15m dans Stohl et al. [2005]) et v^d est la vitesse de dépôt sec.

Finalement, la concentration moyenne de polluant à la position \mathbf{x} et au temps t est donnée par l'estimation de la masse totale de particules dans un volume $\Delta V(\mathbf{x})$ autour de la position \mathbf{x} par la formule lagrangienne suivante :

$$c(\mathbf{x}, t) = \frac{1}{|\Delta V(\mathbf{x})|} \sum_{k=1}^N \delta_k(\mathbf{x}, t) m_k(t). \quad (2.34)$$

$\delta_k(\mathbf{x}, t)$ vaut 1 si la particule k de masse m_k est située dans le volume $\Delta V(\mathbf{x})$ à l'instant t et 0 sinon. N est le nombre total de particules. L'équation 2.34 montre que la concentration, dans un modèle lagrangien, est une fonction discrète de l'espace et du temps.

2.3.3 Les modèles gaussiens

Les modèles gaussiens permettent de représenter la dispersion à l'échelle locale (dans les premières dizaines de kilomètres autour d'une source). Ces modèles fournissent une formulation analytique de la concentration de polluant et sont donc simples et peu coûteux à mettre en œuvre. Deux types de modèles gaussiens existent :

- le modèle de panache, ou gaussien stationnaire, modélise la dispersion dans une situation météorologique supposée uniforme et stationnaire. Le panache émis par une source ponctuelle est modélisé par une distribution gaussienne dans deux directions (horizontale orthogonale au vent, et verticale).
- le modèle à bouffées modélise une émission instantanée par une bouffée gaussienne dans les trois directions et dont le centre est "transporté" par un champ de vent éventuellement non-uniforme et non-stationnaire. Un rejet de polluant est donc modélisé par un ensemble de bouffées gaussiennes qui rendent compte de la dispersion du polluant.

Modèle gaussien de panache

Dans le cadre de la modélisation par une distribution gaussienne d'un panache issu d'une source ponctuelle, nous faisons les hypothèses suivantes :

- La source d'émission est une source ponctuelle continue de débit constant Q_s .
- Les conditions météorologiques sont uniformes et constantes.
- Le vent est suffisamment important pour que la diffusion turbulente dans la direction du vent soit négligeable en comparaison de l'advection.

Dans ce cadre, en plaçant l'axe x dans la direction du vent, la concentration de polluant est donnée par la formule analytique suivante [Korsakissok, 2009] :

$$c(x, y, z) = \frac{Q_s}{2\pi\sigma_y\sigma_z\bar{u}} \exp\left(-\frac{(y-y_s)^2}{2\sigma_y^2}\right) \left[\exp\left(-\frac{(z-z_s)^2}{2\sigma_z^2}\right) + \exp\left(-\frac{(z+z_s)^2}{2\sigma_z^2}\right) \right], \quad (2.35)$$

où (x_s, y_s, z_s) représente les coordonnées de la source et \bar{u} la vitesse du vent moyen. Cette formule prend en compte l'influence du sol, mais pas les dépôts. Les écart-types σ_y et σ_z sont estimés par des formules empiriques reposant sur la stabilité de l'atmosphère telles les formules de Briggs, de Pasquill [1961] ou de Doury [1976].

Modèle gaussien à bouffées

Dans le cas (plus général) d'un modèle instationnaire, le panache est représenté sous la forme d'une série de bouffées gaussiennes dans les trois directions. Dans ce cas, la concentration d'une bouffée peut s'écrire :

$$c(x, y, z) = \frac{Q}{(2\pi)^{3/2} \sigma_x \sigma_y \sigma_z} \exp\left(-\frac{(x-x_c)^2}{2\sigma_x^2}\right) \exp\left(-\frac{(y-y_c)^2}{2\sigma_y^2}\right) \times \left[\exp\left(-\frac{(z-z_c)^2}{2\sigma_z^2}\right) + \exp\left(-\frac{(z+z_c)^2}{2\sigma_z^2}\right) \right]. \quad (2.36)$$

Q est la quantité totale émise pour constituer la bouffée. (x_c, y_c, z_c) sont les coordonnées du centre de la bouffée. Les formules de dispersion pour σ_y et σ_z sont en général les mêmes que dans le cas du panache gaussien. Pour les paramétrisations qui ne donnent pas de formulation spécifique pour σ_x (dans la direction du vent), celle de σ_y est utilisée.

Ainsi en discrétisant une source continue de quantité totale Q par une série de bouffée telle que :

$$Q = \Delta t_{\text{puff}} \sum_{i=0}^{N_{\text{puff}}} Q_s^i, \quad (2.37)$$

où Q_s^i est le débit massique de la source entre l'instant $t_i = t_0 + i \Delta t_{\text{puff}}$ et t_{i+1} , alors la concentration en un point est donnée par la somme des contributions de toutes les bouffées

issues de cette source :

$$c(x, y, z) = \sum_{i=0}^{N_{\text{puff}}} \frac{Q_s^i \times \Delta t_{\text{puff}}}{(2\pi)^{3/2} \sigma_x^i \sigma_y^i \sigma_z^i} \exp\left(-\frac{(x - x_c^i)^2}{2\sigma_x^{i2}}\right) \exp\left(-\frac{(y - y_c^i)^2}{2\sigma_y^{i2}}\right) \times \left[\exp\left(-\frac{(z - z_c^i)^2}{2\sigma_z^{i2}}\right) + \exp\left(-\frac{(z + z_c^i)^2}{2\sigma_z^{i2}}\right) \right]. \quad (2.38)$$

Les coordonnées du centre de la bouffée i émise au temps t_i sont (x_c^i, y_c^i, z_c^i) et évoluent au cours du temps en fonction du champ de vent. Les écarts types associés σ_x^i , σ_y^i et σ_z^i dépendent de l'âge de la bouffée. Si plusieurs sources émettent le même polluant, il est alors nécessaire de prendre en compte toutes les bouffées correspondantes.

Ce type de modèle est parfois appelé *modèle semi-lagrangien*. En effet, le calcul de l'évolution d'une bouffée au cours du temps est, comme dans un modèle lagrangien, décomposé entre un mouvement déterministe lié à l'advection par le champ de vent moyen (calcul des coordonnées du centre de la bouffée) et une composante qui rend compte de l'impact de la turbulence (ici distribution gaussienne dans les trois directions autour de ce centre).

2.3.4 Les modèles de mécanique des fluides à l'échelle locale/urbaine

Les modèles de mécanique des fluides (CFD) se situent dans le cadre de la modélisation eulérienne, déjà évoquée dans le chapitre 2.3.1 pour la modélisation à grande échelle. À l'échelle urbaine, l'influence des différents processus est différente. Afin de prendre en compte pleinement l'influence des obstacles dans la dispersion de polluant, il est nécessaire de considérer des modèles capables de calculer les écoulements complexes induits par la géométrie. Les écoulements dans la couche de surface sont décrits par les équations fondamentales de la dynamique des fluides [Stull, 1988; Garratt, 1992], adaptées à l'échelle considérée.

L'équation de conservation de la masse (sous hypothèse d'anélasticité) s'écrit, en suivant la convention d'Einstein :

$$\frac{\partial \rho u_i}{\partial x_i} = 0. \quad (2.39)$$

L'équation de conservation de la quantité de mouvement est décrite par les équations de Navier-Stokes adaptées à la CLA à échelle locale, c'est-à-dire que les forces de Coriolis sont négligées, ainsi que les termes de sphéricité :

$$\underbrace{\frac{\partial u_i}{\partial t}}_{(1)} + \underbrace{u_j \frac{\partial u_i}{\partial x_j}}_{(2)} = \underbrace{-\frac{1}{\rho} \frac{\partial p}{\partial x_i}}_{(3)} + \underbrace{g_i}_{(4)} + \underbrace{\frac{1}{\rho} \frac{\partial \tau_{ij}}{\partial x_j}}_{(5)}, \quad (2.40)$$

où

- (1) est le terme d'inertie,
- (2) est le terme d'advection,
- (3) est le terme gradient de pression,
- (4) est le terme de gravité,
- (5) représente les effets visqueux de l'air, considéré comme un fluide newtonien, et s'écrit alors (sous hypothèse de viscosité homogène spatialement) :

$$\frac{1}{\rho} \frac{\partial \tau_{ij}}{\partial x_j} = \nu \frac{\partial}{\partial x_j} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} - \frac{2}{3} \frac{\partial u_k}{\partial x_k} \delta_{ij} \right), \quad (2.41)$$

où ν est la viscosité cinématique de l'air.

L'équation de conservation de l'énergie s'écrit :

$$\frac{\partial \theta}{\partial t} + u_j \frac{\partial \theta}{\partial x_j} = \frac{1}{\rho C_p} \frac{\partial}{\partial x_j} \left(\lambda \frac{\partial \theta}{\partial x_j} \right) + S_\theta, \quad (2.42)$$

où θ est la température potentielle de l'air, C_p sa capacité calorifique, λ la conductibilité thermique moléculaire et S_θ un terme source ou puits.

Enfin, l'équation de transport de la concentration c de polluant, considéré passif chimiquement :

$$\frac{\partial c}{\partial t} + u_j \frac{\partial c}{\partial x_j} = D_m \frac{\partial^2 c}{\partial x_j^2} + P + \sigma, \quad (2.43)$$

où D_m est le coefficient de diffusion moléculaire massique, P traduit les processus de perte (dépôts) et σ un terme source.

Prise en compte de la turbulence

Plusieurs approches de simulation numérique sont envisageables pour résoudre les équations précédentes. La *simulation numérique directe* (Direct Numerical Simulation, DNS) consiste à résoudre ces équations en tenant compte de toutes les échelles du mouvement. D'après la théorie de Kolmogorov présentée dans la chapitre 2.1.1, le rapport entre les plus petites échelles et les plus grandes est de l'ordre de $10^5 - 10^7$, ce qui signifie que les mailles les plus petites du modèles doivent avoir une résolution de 0.1 – 1 mm. On atteint alors rapidement un nombre de mailles très élevé. Même si les performances informatiques actuelles permettent d'envisager la DNS à très petites échelles, les temps de calcul restent très importants pour nos échelles d'étude et donc peu adaptés à notre problématique.

Afin de pallier ces limites, dans la *simulation des grandes échelles* (Large Eddy Simulation, LES), seules les grandes échelles sont explicitement résolues alors que les plus petites échelles sont représentées par des modèles simples. Cela permet de réduire les temps de calcul par rapport à la DNS, mais ceux-ci restent tout de même trop élevés pour notre application.

Une troisième approche pour modéliser les écoulements turbulents est l'approche statistique de Reynolds appliquée aux équations de Navier-Stokes (Reynolds Averaged Navier-Stokes, RANS). Dans cette approche déjà évoquée dans le chapitre 2.2, les variables sont décomposées entre valeur moyenne et fluctuations (voir Milliez [2006]). En ce qui concerne la masse volumique, on néglige les fluctuations de ρ sauf pour les termes de flottabilité (approximation de Boussinesq). L'approche RANS présente l'avantage de temps de calcul plus faibles que les approches LES et DNS tout en conservant une bonne précision.

L'équation de conservation de la masse devient alors :

$$\frac{\partial \langle \rho \rangle \langle u_i \rangle}{\partial x_i} = 0. \quad (2.44)$$

Les équations de conservation de la quantité de mouvement s'écrivent :

$$\frac{\partial \langle u_i \rangle}{\partial t} + \langle u_j \rangle \frac{\partial \langle u_i \rangle}{\partial x_j} = -\frac{1}{\langle \rho \rangle} \frac{\partial \langle p \rangle}{\partial x_i} + g_i + \frac{1}{\langle \rho \rangle} \frac{\partial \langle \tau_{ij} \rangle}{\partial x_j} - \frac{\partial \langle u'_i u'_j \rangle}{\partial x_j}. \quad (2.45)$$

La prise en compte de la turbulence fait apparaître des termes supplémentaires : $\langle u'_i u'_j \rangle$, qui représentent les contraintes de Reynolds.

L'équation de conservation de l'énergie devient :

$$\frac{\partial \langle \theta \rangle}{\partial t} + \langle u_j \rangle \frac{\partial \langle \theta \rangle}{\partial x_j} = \frac{1}{\langle \rho \rangle C_p} \frac{\partial}{\partial x_j} \left(\lambda \frac{\partial \langle \theta \rangle}{\partial x_j} - \langle \rho \rangle C_p \langle u'_j \theta' \rangle \right) + \langle S_\theta \rangle. \quad (2.46)$$

Le terme $\langle \rho \rangle C_p \langle u'_j \theta' \rangle$ est le terme de flux turbulent de chaleur.

Enfin l'équation de transport d'un scalaire devient :

$$\frac{\partial \langle c \rangle}{\partial t} + \langle u_j \rangle \frac{\partial \langle c \rangle}{\partial x_j} = D_m \frac{\partial^2 \langle c \rangle}{\partial x_j^2} - \frac{\partial \langle u'_j c' \rangle}{\partial x_j} + \langle P \rangle + \langle \sigma \rangle. \quad (2.47)$$

Fermetures turbulentes

Afin d'estimer les flux turbulents $\langle u'_i u'_j \rangle$, $\langle u'_j \theta' \rangle$ et $\langle u'_j c' \rangle$, un schéma de fermeture du premier ordre est utilisé. Les flux turbulents sont exprimés en fonction des gradients moyens :

$$\langle u'_i u'_j \rangle = -\nu_t \left[\frac{\partial \langle u_i \rangle}{\partial x_j} + \frac{\partial \langle u_j \rangle}{\partial x_i} - \frac{2}{3} \delta_{ij} \frac{\partial \langle u_k \rangle}{\partial x_k} \right], \quad (2.48)$$

$$\langle u'_j \theta' \rangle = -\frac{\lambda_t}{\langle \rho \rangle C_p} \frac{\partial \langle \theta \rangle}{\partial x_j}, \quad (2.49)$$

$$\langle u'_j c' \rangle = -K_t \frac{\partial \langle c \rangle}{\partial x_j}. \quad (2.50)$$

ν_t est la viscosité turbulente cinématique, λ_t la conductivité thermique turbulente avec $\lambda_t / (\langle \rho \rangle C_p) = \nu_t / Pr_t$, Pr_t étant le nombre de Prandtl turbulent, et K_t la diffusivité turbulente avec $K_t = \nu_t / Sc_t$, Sc_t étant le nombre de Schmidt turbulent.

Les équations décrivant l'évolution des quantités moyennes deviennent alors, en abandonnant la notation $\langle \cdot \rangle$:

$$\frac{\partial \rho u_i}{\partial x_i} = 0, \quad (2.51)$$

$$\begin{aligned} \frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = & -\frac{1}{\rho} \frac{\partial p}{\partial x_i} + g_i - \frac{2}{3} \frac{\partial}{\partial x_i} \left[(\nu + \nu_t) \frac{\partial u_k}{\partial x_k} \right] \\ & + \frac{\partial}{\partial x_j} \left[(\nu + \nu_t) \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \right], \end{aligned} \quad (2.52)$$

$$\frac{\partial \theta}{\partial t} + u_j \frac{\partial \theta}{\partial x_j} = \frac{1}{\rho C_p} \frac{\partial}{\partial x_j} \left[\left(\lambda + \rho C_p \frac{\nu_t}{Pr_t} \right) \frac{\partial \theta}{\partial x_j} \right] + S_\theta, \quad (2.53)$$

$$\frac{\partial c}{\partial t} + u_j \frac{\partial c}{\partial x_j} = \frac{\partial}{\partial x_j} \left[\left(D_m + \frac{\nu_t}{Sc_t} \right) \frac{\partial c}{\partial x_j} \right] + P + \sigma. \quad (2.54)$$

Il reste donc à estimer la viscosité turbulente introduite ν_t . Plusieurs approches existent, que l'on ne détaillera pas dans ce manuscrit : modèle de longueur de mélange (Prandtl, 1925), modèle de Louis [Louis, 1979], modèle $k-l$ [Garratt, 1992]. L'approche qui sera utilisée dans ce travail est le modèle $k-\epsilon$ [Launder and Spalding, 1974].

Dans le modèle $k-\epsilon$, Deux grandeurs sont introduites : k est l'énergie cinétique turbulente et ϵ le taux de dissipation turbulente. La viscosité turbulente est alors égale à :

$$\nu_t = C_\mu \frac{k^2}{\epsilon}, \quad (2.55)$$

où C_μ est une constante empirique. Les grandeurs k et ϵ sont soumises aux équations de transport suivantes :

$$\begin{aligned} \frac{\partial k}{\partial t} + u_j \frac{\partial k}{\partial x_j} = & \frac{g_j}{\theta_v} \frac{\nu_t}{Pr_t} \frac{\partial \theta_v}{\partial x_j} + \frac{\partial}{\partial x_j} \left[\left(\nu + \frac{\nu_t}{\sigma_k} \right) \frac{\partial k}{\partial x_j} \right] \\ & + \nu_t \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \frac{\partial u_i}{\partial x_j} - \epsilon, \end{aligned} \quad (2.56)$$

TAB. 2.3: Valeurs des constantes du modèle $k - \epsilon$ [Launder and Spalding, 1974].

σ_k	σ_ϵ	C_μ	$C_{1\epsilon}$	$C_{2\epsilon}$
1	1.3	0.09	1.44	1.92

$$\frac{\partial \epsilon}{\partial t} + u_j \frac{\partial \epsilon}{\partial x_j} = \frac{\partial}{\partial x_j} \left[\left(\nu + \frac{\nu_t}{\sigma_\epsilon} \right) \frac{\partial \epsilon}{\partial x_j} \right] + C_{1\epsilon} \frac{\epsilon}{k} P_c - C_{2\epsilon} \frac{\epsilon^2}{k}, \quad (2.57)$$

où θ_v est la température potentielle virtuelle et P_c est le terme de production par le cisaillement moyen et les forces de flottabilité donné par :

$$P_c = \nu_t \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \frac{\partial u_i}{\partial x_j} + C_{3\epsilon} \frac{g_j}{\theta_v} \frac{\nu_t}{Pr_t} \frac{\partial \theta_v}{\partial x_j}, \quad (2.58)$$

et σ_k , σ_ϵ , $C_{1\epsilon}$, $C_{2\epsilon}$ et $C_{3\epsilon}$ sont des constantes. Afin de ne tenir compte que de la contribution positive des termes de flottabilité, $C_{3\epsilon}$ vaut 0 en atmosphère stable et 1 en atmosphère instable [Viollet, 1988]. Les différentes constantes du modèle $k - \epsilon$ sont celles proposées par Launder and Spalding [1974] dont les valeurs sont indiquées dans le tableau 2.3. En pratique, dans les couches basses de l'atmosphère, la diffusion turbulente représentée par le terme $\frac{\nu_t}{Sc_t}$ est prépondérante devant la diffusion moléculaire D_m . Néanmoins le modèle de CFD utilisé dans ce travail de thèse (Code_Saturne) prend en compte explicitement les deux composantes et nous avons donc choisi de maintenir cette formulation complète dans les équations présentées dans cette section.

Synthèse et modèle adjoint

Comme dans le cadre des modèles CTM, la simulation numérique à l'échelle locale de la dispersion atmosphérique de polluant peut être divisée en deux étapes successives :

1. Le calcul des caractéristiques de l'écoulement par les équations 2.51, 2.52, 2.53, 2.55, 2.56 et 2.57.
2. Le calcul de l'évolution des concentrations de polluants par l'équation 2.54, en utilisant les grandeurs calculées par l'étape précédente (champs de vent et viscosité turbulente notamment), et en utilisant des conditions initiales et aux limites du domaine identiques à celles proposées dans le chapitre 2.3.1.

Cette séparation en deux étapes permet également de mettre en œuvre aisément un modèle adjoint, qui calcule l'évolution de la concentration adjointe c^* , solution de l'équation :

$$\frac{\partial c^*}{\partial (-t)} - u_j \frac{\partial c^*}{\partial x_j} = \frac{\partial}{\partial x_j} \left[\left(D_m + \frac{\nu_t}{Sc_t} \right) \frac{\partial c^*}{\partial x_j} \right] + P + \pi. \quad (2.59)$$

Chapitre 3

Estimation de sources par modélisation inverse

Sommaire

3.1	Introduction	40
3.2	Cadre bayésien de la modélisation inverse	41
3.3	Statistiques gaussiennes	42
3.3.1	De la formule de Bayes aux moindres carrés	42
3.3.2	L'estimateur BLUE	43
3.4	Statistiques gaussiennes tronquées	44
3.5	Estimation de l'incertitude <i>a posteriori</i>	45
3.6	Estimation des erreurs <i>a priori</i>	45
3.6.1	Dépendance de l'inversion aux erreurs <i>a priori</i>	45
3.6.2	La L-curve	47
3.6.3	Principe de la vraisemblance maximale	48
3.6.4	Schéma itératif de Desroziers	49
3.7	Application I : Surveillance d'un parc de centrales nucléaires	50
3.7.1	Installations nucléaires surveillées et réseaux d'observation	50
3.7.2	Modèles de dispersion atmosphérique	51
3.7.3	Système opérationnel	53
3.7.4	Expériences synthétiques	54
3.7.5	Résultats	56
3.8	Application II : Fukushima (1)	58
3.8.1	Contexte	58
3.8.2	Observations	59
3.8.3	Modèle de dispersion atmosphérique	59
3.8.4	Modélisation inverse	61
3.8.5	Résultats	64
3.9	Application III : Fukushima (2)	72
3.9.1	Observations	73
3.9.2	Modèle	74
3.9.3	Modélisation inverse	76
3.9.4	Résultats	77

3.1 Introduction

Les modèles numériques introduits dans le chapitre 2 sont soumis à un certain nombre d’incertitudes : incertitudes sur les champs météorologiques, sur les paramétrisations des processus physiques ou sur la représentativité des observables par exemple. En situation accidentelle, la principale source d’incertitude est la précision du terme source [Abida and Bocquet, 2009].

L’utilisation d’observations disponibles permet de corriger les paramètres du modèle, notamment le terme source. Pour reconstruire le terme source consécutif à un rejet accidentel de polluant dans l’atmosphère, les méthodes de modélisation inverse sont une alternative à une approche “simulation/correction” utilisant des modèles numériques directs. Les techniques de modélisation inverse réalisent une estimation objective du terme source en utilisant les informations issues des observations et du modèle numérique qui simule la dispersion. Celui-ci est généralement linéaire dans les situations accidentelles (pas de processus chimiques non-linéaires), ce qui permet d’écrire, après discrétisation du terme source, la relation entre les observations et le vecteur source :

$$\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\sigma} + \boldsymbol{\epsilon}, \quad (3.1)$$

où $\boldsymbol{\mu}$ dans \mathbb{R}^d représente le vecteur des observations, $\boldsymbol{\sigma}$ dans \mathbb{R}^N la source discrétisée et \mathbf{H} est la matrice jacobienne du modèle numérique, avec l’opérateur d’observation. Le vecteur $\boldsymbol{\epsilon}$ est le vecteur des erreurs d’observation qui prennent en compte les erreurs de mesures, de représentativité et une partie des erreurs de modèle.

Dans un contexte accidentel, le nombre d’observations est très souvent limité. De plus, les conditions météorologiques peuvent mener à une situation d’observabilité faible d’une partie de la source. Dans ces cas, la relation source-observations fournie par l’équation 3.1 constitue un problème inverse mal-posé [Enting, 2002].

Afin de s’affranchir du manque de contraintes dans le système, une possibilité est d’implémenter des méthodes dites paramétriques. Dans ces méthodes, la source à reconstruire est réduite à un nombre très limité de paramètres à estimer. Le problème inverse est *de facto* régularisé et peut être résolu en utilisant différentes techniques, par exemple des techniques d’échantillonnage stochastique [Delle Monache et al., 2008; Yee et al., 2008]. Néanmoins, si la vraie source n’est pas compatible avec la paramétrisation choisie, l’inversion peut échouer ou conduire à des résultats non-physiques. De telles méthodes sont présentées dans le chapitre 4.

Une autre stratégie consiste à utiliser des méthodes non-paramétriques pour reconstruire un champ complet d’émission. Ce champ peut être au final discrétisé, par exemple pour correspondre à la discrétisation du modèle, mais le nombre de variables à estimer reste élevé, éventuellement plus élevé que le nombre d’observations disponibles. Les méthodes non-paramétriques sont à la fois robustes et souples puisqu’elles ne s’appuient pas sur des hypothèses *a priori* fortes sur la nature de la source, mais elles ont leurs propres limites. Si les statistiques des erreurs d’observation sont supposées gaussiennes, ce qui est un cadre relativement classique, l’inversion s’appuie sur la minimisation d’une fonction coût

$$\mathcal{J}_q(\boldsymbol{\sigma}) = \frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}), \quad (3.2)$$

où \mathbf{R} est la matrice de covariance des erreurs d’observation $\mathbf{R} = E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]$. Le choix le plus simple est de négliger les corrélations entre les erreurs et de prendre \mathbf{R} diagonale. Si toutes les observations sont de même nature, il est possible de prendre en compte une unique variance des erreurs r^2 , telles que $\mathbf{R} = r^2\mathbf{I}_d$, où \mathbf{I}_d est la matrice identité dans \mathbb{R}^d . Ce choix est pertinent en situation accidentelle où les appareils de mesures sont suffisamment éloignés les uns des autres pour limiter les corrélations entre les erreurs d’observation. Du fait du faible nombre d’observations et/ou de la mauvaise observabilité d’une partie de la source, la minimisation

de la fonction coût de l'équation 3.2 donne généralement une infinité de solutions. Il est donc nécessaire de régulariser le problème inverse. Un terme de *Tikhonov* [Tikhonov and Arsenin, 1977] est généralement ajouté à la fonction coût :

$$\mathcal{J}_q(\boldsymbol{\sigma}) = \frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) + \frac{1}{2} (\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)^T \mathbf{B}^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}^b). \quad (3.3)$$

L'unicité de la solution au problème inverse est maintenant garantie, mais deux paramètres supplémentaires (un vecteur et une matrice) ont été introduits : $\boldsymbol{\sigma}^b$ est l'ébauche de la source et \mathbf{B} est la matrice de covariance d'erreur d'ébauche, $\mathbf{B} = \text{E} [(\boldsymbol{\sigma}^b - \boldsymbol{\sigma}^t) (\boldsymbol{\sigma}^b - \boldsymbol{\sigma}^t)^T]$ où $\boldsymbol{\sigma}^t$ représente la vraie source. Comme pour \mathbf{R} , il est généralement pertinent en situation accidentelle de prendre $\mathbf{B} = m^2 \mathbf{I}_N$, m^2 représentant la variance des erreurs d'ébauche. Cette hypothèse forte est discutée en détails dans le chapitre 3.6.

Davoine and Bocquet [2007], dans le cas de Tchernobyl, mais aussi Stohl et al. [2012], puis nos propres études dans le cas de Fukushima, ont montré que la source reconstruite était très sensible aux matrices \mathbf{R} et \mathbf{B} , donc aux hyper-paramètres r et m . C'est pourquoi l'estimation rigoureuse des hyper-paramètres est nécessaire. Dans Davoine and Bocquet [2007]; Krysta et al. [2008]; Saide et al. [2011], la technique de la L-curve de Hansen [1992] a été utilisée avec succès pour estimer le rapport r/m ou les deux paramètres. Récemment, d'importants développements méthodologiques pour l'assimilation de données dans le domaine des prévisions météorologiques se sont intéressés à l'estimation des hyper-paramètres définissant les erreurs *a priori*. La plupart de ces développements reposent sur des techniques de validation croisée (General Cross-Validation) ou sur le principe de vraisemblance maximale [Mitchell and Houtekamer, 1999; Chapnik et al., 2004, 2006; Anderson, 2007; Li et al., 2009]. De telles techniques ont également été implémentées dans le contexte de la chimie atmosphérique, en utilisant par exemple la méthode du χ^2 [Ménard et al., 2000; Elbern et al., 2007; Davoine and Bocquet, 2007], le principe de la vraisemblance maximale [Michalak et al., 2004] ou des diagnostics statistiques [Schwinger and Elbern, 2010].

Dans ce chapitre, nous présentons tout d'abord (section 3.2) le cadre général de la modélisation inverse. Ce cadre est bayésien, ce qui signifie que la modélisation inverse repose sur l'équilibre entre des informations *a priori* (ébauche de la source, statistiques des erreurs d'observation et des erreurs d'ébauche) et de nouvelles informations apportées par l'acquisition de nouvelles observations et l'utilisation d'un modèle de dispersion. Puis dans un cadre de statistiques gaussiennes (section 3.3) ou gaussiennes tronquées (section 3.4) pour les erreurs d'observation et d'ébauche et après avoir défini les hyper-paramètres qui caractérisent ces erreurs, nous proposons trois méthodes pour estimer rigoureusement ces hyper-paramètres dans la section 3.6. Finalement trois études d'application sont proposées dans les sections 3.7, 3.8 et 3.9. La première application consiste à implémenter un algorithme semi-automatique de surveillance d'un parc d'installations nucléaires et à l'évaluer sur un grand nombre d'expériences synthétiques. Les deux études suivantes ont pour objectif l'estimation des termes sources de césium-137 et d'iode-131 consécutifs à l'accident de la centrale de Fukushima Daiichi en mars 2011.

3.2 Cadre bayésien de la modélisation inverse

La modélisation inverse repose sur l'équilibre objectif entre une information *a priori* (généralement formulée par le biais d'hypothèses) et l'apport d'information nouvelle par l'assimilation d'observations et l'utilisation d'un modèle. Cela peut notamment se traduire mathématiquement

quement par l'utilisation de la relation de Bayes

$$p(\boldsymbol{\sigma}|\boldsymbol{\mu}) = \frac{p(\boldsymbol{\mu}|\boldsymbol{\sigma}) p(\boldsymbol{\sigma})}{p(\boldsymbol{\mu})}, \quad (3.4)$$

qui lie les fonctions de densité de probabilité (pdf) conditionnelles de la source $\boldsymbol{\sigma}$ et du vecteur d'observations $\boldsymbol{\mu}$.

L'inférence conduit à estimer la probabilité *a posteriori* $p(\boldsymbol{\sigma}|\boldsymbol{\mu})$ (c'est-à-dire après acquisition des observations $\boldsymbol{\mu}$). La pdf $p(\boldsymbol{\mu}|\boldsymbol{\sigma})$, appelée fonction de vraisemblance, est la probabilité *a priori* du jeu de données $\boldsymbol{\mu}$ étant donnée une source $\boldsymbol{\sigma}$. Elle peut être interprétée comme la probabilité *a priori* des erreurs d'observation définies par l'équation 3.1 et notée $p(\boldsymbol{\mu}|\boldsymbol{\sigma}) = p_\epsilon(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})$. La pdf $p(\boldsymbol{\sigma}) = p_b(\boldsymbol{\sigma})$ représente la probabilité *a priori* de la source $\boldsymbol{\sigma}$ (c'est-à-dire avant acquisition de $\boldsymbol{\mu}$). Enfin la vraisemblance marginale (ou évidence) des observations $p(\boldsymbol{\mu})$ ne dépendant pas de la source, on peut également écrire

$$\begin{aligned} p(\boldsymbol{\sigma}|\boldsymbol{\mu}) &\propto p(\boldsymbol{\mu}|\boldsymbol{\sigma}) p(\boldsymbol{\sigma}) \\ &\propto p_\epsilon(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) p_b(\boldsymbol{\sigma}). \end{aligned} \quad (3.5)$$

La pdf $p(\boldsymbol{\sigma}|\boldsymbol{\mu})$ complète contient toute l'information sur la distribution *a posteriori* de la source. Néanmoins, cette information est parfois délicate à manipuler et un estimateur fiable de $\boldsymbol{\sigma}$ est plus généralement recherché. Une méthode pour déterminer un tel estimateur, que nous noterons $\boldsymbol{\sigma}^a$, repose sur la maximisation de la probabilité *a posteriori* $p(\boldsymbol{\sigma}|\boldsymbol{\mu})$:

$$\boldsymbol{\sigma}^a = \underset{\boldsymbol{\sigma}}{\operatorname{argmax}} p(\boldsymbol{\sigma}|\boldsymbol{\mu}). \quad (3.6)$$

Le problème inverse est souvent transformé en la minimisation d'une fonction coût \mathcal{L} , soit $\boldsymbol{\sigma}^a = \underset{\boldsymbol{\sigma}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\sigma})$ avec :

$$\mathcal{L}(\boldsymbol{\sigma}) = -\ln(p(\boldsymbol{\sigma}|\boldsymbol{\mu})) = -\ln(p(\boldsymbol{\mu}|\boldsymbol{\sigma})) - \ln(p(\boldsymbol{\sigma})) + \ln(p(\boldsymbol{\mu})). \quad (3.7)$$

Comme la vraisemblance marginale des observations $p(\boldsymbol{\mu})$ ne dépend pas de la source, le problème inverse se simplifie encore en $\boldsymbol{\sigma}^a = \underset{\boldsymbol{\sigma}}{\operatorname{argmin}} \mathcal{J}(\boldsymbol{\sigma})$ avec :

$$\mathcal{J}(\boldsymbol{\sigma}) = -\ln(p_\epsilon(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})) - \ln(p_b(\boldsymbol{\sigma})). \quad (3.8)$$

3.3 Statistiques gaussiennes

3.3.1 De la formule de Bayes aux moindres carrés

L'estimateur qui minimise la fonction coût définie par l'équation 3.8 est bien évidemment dépendant des pdfs *a priori* : $p_\epsilon(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})$ et $p_b(\boldsymbol{\sigma})$. $p_\epsilon(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})$ décrit les statistiques *a priori* des erreurs d'observation. Ces erreurs, définies par l'équation 3.1, rendent compte à la fois des erreurs de mesures, des erreurs de représentativité de l'observation et également d'une classe d'erreur modèle. $p_b(\boldsymbol{\sigma})$ décrit les statistiques *a priori* des erreurs d'ébauche. Ces erreurs rendent compte de l'écart entre la source $\boldsymbol{\sigma}$ et une ébauche $\boldsymbol{\sigma}^b$ de la source (appelée *first guess* en anglais). Les distributions réelles des erreurs d'observation et d'ébauche ne sont généralement pas connues. Nous sommes donc amenés à réaliser des hypothèses quant à leur nature. Un choix très fréquent consiste à retenir des statistiques gaussiennes.

Dans ce cadre, les erreurs d'observation $\boldsymbol{\epsilon} = \boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}$ sont supposées non biaisées, soit $E[\boldsymbol{\epsilon}] = \mathbf{0}$. Nous supposons de plus qu'elles suivent une loi normale $\mathcal{N}(\mathbf{0}, \mathbf{R})$ de matrice de

covariance $\mathbf{R} = E[\epsilon\epsilon^T]$. La pdf correspondante s'écrit alors :

$$p_\epsilon(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) = \frac{e^{-\frac{1}{2}\boldsymbol{\epsilon}^T\mathbf{R}^{-1}\boldsymbol{\epsilon}}}{\sqrt{(2\pi)^d|\mathbf{R}|}} = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma})^T\mathbf{R}^{-1}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma})}}{\sqrt{(2\pi)^d|\mathbf{R}|}}. \quad (3.9)$$

Les erreurs d'ébauche $\boldsymbol{\sigma} - \boldsymbol{\sigma}^b$, également supposées non biaisées ($E[\boldsymbol{\sigma} - \boldsymbol{\sigma}^b] = \mathbf{0}$), suivent une loi normale $\mathcal{N}(\mathbf{0}, \mathbf{B})$ de matrice de covariance $\mathbf{B} = E[(\boldsymbol{\sigma}^b - \boldsymbol{\sigma}^t)(\boldsymbol{\sigma}^b - \boldsymbol{\sigma}^t)^T]$. La pdf correspondante s'écrit :

$$p_b(\boldsymbol{\sigma}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^b)^T\mathbf{B}^{-1}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^b)}}{\sqrt{(2\pi)^N|\mathbf{B}|}}. \quad (3.10)$$

L'équation 3.8 peut donc s'écrire

$$\begin{aligned} \mathcal{J}(\boldsymbol{\sigma}) &= \frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) + \frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)^T \mathbf{B}^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b) \\ &\quad + \frac{1}{2} \ln \left((2\pi)^{d+N} |\mathbf{R}||\mathbf{B}| \right). \end{aligned} \quad (3.11)$$

Pour des raisons de simplicité, \mathbf{R} et \mathbf{B} sont généralement choisies indépendantes des paramètres à retrouver, ici $\boldsymbol{\sigma}$ (un choix différent sera néanmoins fait dans la section 3.7). Dans ce cas, le dernier terme du membre de droite dans la précédente équation n'influence pas la minimisation de la fonction coût qui peut se simplifier en une forme quadratique

$$\mathcal{J}_q(\boldsymbol{\sigma}) = \frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) + \frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)^T \mathbf{B}^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b). \quad (3.12)$$

La maximisation de la probabilité *a posteriori* de la source est équivalente à la minimisation de la fonction coût \mathcal{J}_q . La formulation bayésienne est donc, dans le cas de statistiques gaussiennes, équivalente à une méthode des moindres carrés, relativement classique en assimilation de données. Dans cette terminologie, le dernier terme du membre de droite peut être vu comme un terme de régularisation de *Tikhonov* [Tikhonov and Arsenin, 1977]. Cette régularisation assure l'existence d'une solution unique au problème (pour une configuration particulière, c'est-à-dire pour un choix de $\boldsymbol{\sigma}^b$, \mathbf{R} , \mathbf{B} , \mathbf{H} , $\boldsymbol{\mu}$) y compris quand le problème inverse défini par l'équation 3.1 est mal posé.

3.3.2 L'estimateur BLUE

La fonction coût définie par l'équation 3.12 est une forme quadratique définie et dérivable sur \mathbb{R}^N . Afin de trouver son minimum, il est donc possible de calculer son gradient en tout point :

$$\nabla \mathcal{J}_q(\boldsymbol{\sigma}) = \mathbf{H}^T \mathbf{R}^{-1}(\mathbf{H}\boldsymbol{\sigma} - \boldsymbol{\mu}) + \mathbf{B}^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b), \quad (3.13)$$

soit

$$\nabla \mathcal{J}_q(\boldsymbol{\sigma}) = (\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1})(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b) - \mathbf{H}^T \mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^b). \quad (3.14)$$

L'estimateur $\boldsymbol{\sigma}^a$ qui minimise la fonction coût \mathcal{J}_q (et donc qui annule son gradient) est appelé BLUE (Best linear unbiased estimator) et sera noté $\boldsymbol{\sigma}^{\text{BLUE}}$:

$$\boldsymbol{\sigma}^a = \boldsymbol{\sigma}^{\text{BLUE}} = \boldsymbol{\sigma}^b + (\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1})^{-1} \mathbf{H}^T \mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^b). \quad (3.15)$$

La matrice Hessienne \mathcal{M}_q de \mathcal{J}_q peut également être estimée. Elle s'écrit :

$$\mathcal{M}_q(\boldsymbol{\sigma}) = \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1}. \quad (3.16)$$

\mathcal{J}_q peut alors s'écrire :

$$\mathcal{J}_q(\boldsymbol{\sigma}) = \mathcal{J}_q(\boldsymbol{\sigma}^{\text{BLUE}}) + \frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^{\text{BLUE}})^T \mathcal{M}_q(\boldsymbol{\sigma} - \boldsymbol{\sigma}^{\text{BLUE}}). \quad (3.17)$$

Ainsi la distribution postérieure $p(\boldsymbol{\sigma}|\boldsymbol{\mu}) \propto e^{-\mathcal{J}_q(\boldsymbol{\sigma})}$ de la source est également gaussienne ; après normalisation, la pdf s'écrit

$$p(\boldsymbol{\sigma}|\boldsymbol{\mu}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^{\text{BLUE}})^T \mathbf{P}_{\text{BLUE}}^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^{\text{BLUE}})}}{\sqrt{(2\pi)^N |\mathbf{P}_{\text{BLUE}}|}}. \quad (3.18)$$

$\mathbf{P}_{\text{BLUE}} = \mathcal{M}_q^{-1}$ est la matrice de covariance des erreurs d'analyse et s'écrit :

$$\mathbf{P}_{\text{BLUE}} = (\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1})^{-1}. \quad (3.19)$$

Dans la pratique, l'étape la plus coûteuse (en ressources informatiques) pour le calcul de l'estimateur et de la matrice de covariance d'erreur associée est l'inversion de la matrice $\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1}$ dans \mathbb{R}^N . Dans le cas où le nombre de paramètres à retrouver est supérieur au nombre d'observations ($N \geq d$), l'utilisation de la formule de Sherman-Morrison-Woodbury permet aisément de transformer les formules du BLUE en :

$$\boldsymbol{\sigma}^{\text{BLUE}} = \boldsymbol{\sigma}^b + \mathbf{B} \mathbf{H}^T (\mathbf{H} \mathbf{B} \mathbf{H}^T + \mathbf{R})^{-1} (\boldsymbol{\mu} - \mathbf{H} \boldsymbol{\sigma}^b), \quad (3.20)$$

et pour la matrice de covariance d'erreur d'analyse :

$$\mathbf{P}_{\text{BLUE}} = \mathbf{B} - \mathbf{B} \mathbf{H}^T (\mathbf{H} \mathbf{B} \mathbf{H}^T + \mathbf{R})^{-1} \mathbf{H} \mathbf{B}. \quad (3.21)$$

Dans ce cas la matrice à inverser est $\mathbf{H} \mathbf{B} \mathbf{H}^T + \mathbf{R}$ dans \mathbb{R}^d . Les équations 3.15 et 3.19 sont strictement équivalentes aux équations 3.20 et 3.21. Les premières sont utilisées dans les cas $N < d$ (plus d'observations que d'inconnues). Dans le cas contraire ($N \geq d$) la seconde formulation est à privilégier.

3.4 Statistiques gaussiennes tronquées

Les hypothèses gaussiennes ne contraignent pas la positivité de la source reconstruite. En général, mais c'est particulièrement le cas en situation accidentelle, le nombre d'observations peut se révéler trop faible pour contraindre suffisamment la source. Pour imposer la positivité de la source, choisir des statistiques gaussiennes tronquées pour les erreurs d'ébauche est une possibilité. La pdf correspondant à une telle distribution, notée $\mathcal{TN}(\boldsymbol{\sigma}^b, \mathbf{B}, \mathbf{0})$, s'écrit

$$p_b(\boldsymbol{\sigma}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)^T \mathbf{B}^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)}}{\sqrt{(2\pi)^N |\mathbf{B}|} \Phi(\boldsymbol{\sigma}^b, \mathbf{B}, \mathbf{0})} \mathbb{I}_{\boldsymbol{\sigma} \geq \mathbf{0}}. \quad (3.22)$$

$\mathbb{I}_{\boldsymbol{\sigma} \geq \mathbf{0}}$ vaut 1 si tous les termes $\sigma_i \geq 0$ ($i = 1, \dots, N$), sinon sa valeur est 0. $\Phi(\boldsymbol{\sigma}^b, \mathbf{B}, \mathbf{0})$ est la fonction de distribution cumulative complémentaire de $\mathcal{N}(\boldsymbol{\sigma}^b, \mathbf{B})$ évaluée en $\mathbf{0}$, soit :

$$\Phi(\boldsymbol{\sigma}^b, \mathbf{B}, \mathbf{0}) = \int_{\boldsymbol{\sigma} \geq \mathbf{0}} \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)^T \mathbf{B}^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)}}{\sqrt{(2\pi)^N |\mathbf{B}|}} d\boldsymbol{\sigma}. \quad (3.23)$$

Le problème inverse est alors résolu par la minimisation de la même fonction coût \mathcal{J}_q (équation 3.12), mais cette fois sous la contrainte $\boldsymbol{\sigma} \geq \mathbf{0}$. Il n'existe alors plus de formule analytique. La minimisation est réalisée par un algorithme de Quasi-Newton [Byrd et al., 1995].

3.5 Estimation de l'incertitude a posteriori

Dans le cadre de statistiques gaussiennes, les incertitudes *a posteriori* sont estimées par la matrice \mathbf{P}_{BLUE} qui est la matrice de covariance des erreurs d'analyse :

$$\begin{aligned}\mathbf{P}_{\text{BLUE}} &= \mathbf{B} - \mathbf{B}\mathbf{H}^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} \mathbf{H}\mathbf{B} \\ &= (\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1})^{-1} .\end{aligned}\quad (3.24)$$

Dans le cadre de statistiques gaussiennes tronquées, l'incertitude *a posteriori* est estimée en réalisant une simulation de Monte Carlo. À chaque tirage, de nouvelles observations $\tilde{\boldsymbol{\mu}} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$ et une nouvelles ébauche $\tilde{\boldsymbol{\sigma}}^b \sim \mathcal{TN}(\boldsymbol{\sigma}^b, \mathbf{B}, \mathbf{0})$ sont définies. Un nouvel estimateur $\boldsymbol{\sigma}^a$ est alors calculé par $\boldsymbol{\sigma}^a = \underset{\boldsymbol{\sigma} \geq \mathbf{0}}{\text{argmin}} \tilde{\mathcal{J}}_q$ avec :

$$\tilde{\mathcal{J}}_q(\boldsymbol{\sigma}) = \frac{1}{2} (\tilde{\boldsymbol{\mu}} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1} (\tilde{\boldsymbol{\mu}} - \mathbf{H}\boldsymbol{\sigma}) + \frac{1}{2} (\boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}^b)^T \mathbf{B}^{-1} (\boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}^b) . \quad (3.25)$$

La dispersion statistique de l'ensemble des sources reconstruites, par exemple l'estimation de l'écart type, permet d'estimer l'incertitude *a posteriori* de la source reconstruite sans perturbation.

3.6 Estimation des erreurs a priori

3.6.1 Dépendance de l'inversion aux erreurs a priori

Que ce soit dans un cadre de statistiques gaussiennes par les équations du BLUE ou dans un cadre de statistiques gaussiennes tronquées par minimisation de la fonction coût \mathcal{J}_q sous contrainte de positivité, il est clair que l'estimateur est dépendant non seulement des observations et du modèle, mais également des matrices de covariance des erreurs *a priori* \mathbf{R} et \mathbf{B} . Le même constat est valable pour l'estimation de l'incertitude *a posteriori* de cet estimateur. En situation accidentelle, il est difficile d'estimer ces matrices et la solution généralement retenue passe par la définition de modèles statistiques qui reposent sur des paramètres, appelés par la suite *hyper-paramètres*, qu'il convient d'estimer rigoureusement. Étant donné le nombre relativement faible d'observations disponibles dans de telles situations, ces modèles doivent être le plus simple possible. En particulier le nombre d'hyper-paramètres doit rester faible.

Une hypothèse simple consiste à choisir des matrices \mathbf{R} et \mathbf{B} scalaires :

$$\mathbf{B} = m^2 \mathbf{I}_N, \quad \mathbf{R} = r^2 \mathbf{I}_d. \quad (3.26)$$

\mathbf{B} est diagonale, ce qui signifie que les erreurs d'ébauche sont supposées non corrélées. Dans un contexte accidentel, par exemple pour reconstruire une explosion ou une série d'explosions indépendantes, cette hypothèse est parfaitement réaliste. Elle sera particulièrement vérifiée si la résolution de l'espace de contrôle n'est pas excessivement fine. Dans le cas contraire (résolution fine de l'espace de contrôle ou événements corrélés), il est possible de complexifier légèrement le modèle afin de prendre en compte une corrélation dans les erreurs. Citons par exemple le modèle exponentiel :

$$\mathbf{B}_{ij} = m^2 e^{-\frac{l_{ij}}{L} - \frac{t_{ij}}{T}}, \quad (3.27)$$

où l_{ij} représente la distance entre les termes $\boldsymbol{\sigma}_i^b$ et $\boldsymbol{\sigma}_j^b$, t_{ij} l'intervalle de temps entre ces deux termes, L et T respectivement la longueur de corrélation et le temps de corrélation. Le modèle

de Balgovind peut également être envisagé :

$$\mathbf{B}_{ij} = m^2 \left(1 + \frac{l_{ij}}{L}\right) \left(1 + \frac{t_{ij}}{T}\right) e^{-\frac{l_{ij}}{L} - \frac{t_{ij}}{T}}. \quad (3.28)$$

Ces modèles ont pour inconvénient principal d'introduire deux hyper-paramètres supplémentaires : la longueur de corrélation L et le temps de corrélation T .

L'hypothèse supplémentaire d'homoscédasticité, qui se traduit par une même variance sur la diagonale de \mathbf{B} , signifie quant à elle que nous ne disposons pas d'information *a priori* sur l'incertitude des paramètres de l'ébauche les uns par rapport aux autres. Dans le cas contraire, il est également possible d'introduire plusieurs variances correspondant à des événements, lieux ou périodes d'émission connus ou plus fortement suspectés. En situation d'urgence, il est peu probable que de telles informations *a priori* soient disponibles. De plus, le nombre d'hyper-paramètres à estimer augmenterait en conséquence.

De même, par ce choix de matrice \mathbf{R} , les erreurs d'observation sont supposées non corrélées. Il s'agit certainement d'une approximation : un instrument peut réaliser une erreur systématique sur les mesures, ce qui se traduit par une corrélation temporelle des erreurs. De plus les erreurs d'observation rendent aussi compte d'erreurs de modèle ; si le modèle échoue à bien modéliser un événement de contamination particulier (par exemple une erreur dans le champs de vent), toutes les observations de la zone concernée seront entachées d'erreurs ayant la même origine, ce qui induit une corrélation au minimum spatiale (mais certainement aussi temporelle). Comme évoqué pour les erreurs d'ébauche, une alternative consisterait à choisir un modèle de corrélation pour les erreurs d'observation. Néanmoins afin de garder un nombre d'hyper-paramètres faible, nous choisirons en général de considérer les erreurs d'observation non corrélées. En situation accidentelle, la relativement faible distribution des équipements de mesures justifie généralement ce choix.

L'hypothèse d'homoscédasticité des erreurs d'observation est généralement acceptable dans le cas où toutes les observations sont de même nature. Mais dans le cas où l'inversion est réalisée en assimilant simultanément différents types de données, cette hypothèse n'est plus envisageable. Il convient alors de définir une variance r_i^2 spécifique à chaque jeu de données i . Ainsi, en conservant \mathbf{R} diagonale et en notant \mathbf{R}_i le bloc diagonal de \mathbf{R} relatif au jeu de données i , on note :

$$\mathbf{B} = m^2 \mathbf{I}_N, \quad \mathbf{R}_i = r_i^2 \mathbf{I}_{d_i}, \quad (3.29)$$

où d_i est le nombre d'observations dans le jeu de données i . Si N_d représente le nombre de jeux de données différents, on a donc $\sum_{i=1}^{N_d} d_i = d$.

Dans le cas où un seul type de données est utilisé pour l'inversion, il est donc nécessaire d'estimer 2 hyper-paramètres : la variance des erreurs d'observation et la variance des erreurs d'ébauche. Nous proposons par la suite trois méthodes pour réaliser rigoureusement cette estimation. La première méthode repose sur le principe de la L-curve de Hansen [1992], les deux autres sur le principe de vraisemblance maximale [Dee, 1995].

Dans le cas de N_d jeux de données assimilés simultanément, il est nécessaire d'estimer $N_d + 1$ hyper-paramètres. Le vecteur des hyper-paramètres sera noté $\boldsymbol{\theta} = (r_1, \dots, r_{N_d}, m)^T$. Si le principe de la L-curve est théoriquement extensible à l'estimation simultanée de plusieurs hyper-paramètres, son utilisation en situation réelle est délicate. À l'opposé, les méthodes reposant sur le principe de vraisemblance maximale peuvent être efficaces dans ces situations.

3.6.2 La L-curve

Cas de deux hyper-paramètres

Dans le cas où un seul type de données est utilisé pour l'inversion et en utilisant les définitions des matrices \mathbf{R} et \mathbf{B} données par les équations 3.26, la fonction coût de l'équation 3.12 peut s'écrire :

$$\mathcal{J}_q(\boldsymbol{\sigma}) = \frac{1}{r^2} \left[\frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) + \frac{\lambda^2}{2} (\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)^T (\boldsymbol{\sigma} - \boldsymbol{\sigma}^b) \right], \quad (3.30)$$

avec $\lambda = \frac{r}{m}$.

L'estimateur $\boldsymbol{\sigma}^a = \underset{\boldsymbol{\sigma}}{\operatorname{argmin}} \mathcal{J}_q$ (en statistiques gaussiennes) ou $\boldsymbol{\sigma}^a = \underset{\boldsymbol{\sigma} \geq 0}{\operatorname{argmin}} \mathcal{J}_q$ (en statistiques gaussiennes tronquées) ne dépend que de λ . Quand λ est élevé, l'inversion est fortement contrainte par le terme de régularisation et $\boldsymbol{\sigma}^a$ reste proche de l'ébauche $\boldsymbol{\sigma}^b$. À l'opposé, quand λ est faible, l'inversion est principalement contrainte par l'écart aux observations (cas d'*over-fitting*). La L-curve est une méthode semi-empirique qui permet de déterminer l'équilibre entre régularisation et *over-fitting*. Elle s'applique aussi bien dans le cadre de statistiques gaussiennes que dans celui de statistiques gaussiennes tronquées. La courbe paramétrique définie par :

$$\begin{cases} x(\lambda) = \ln \|\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^a(\lambda)\|_2 \\ y(\lambda) = \ln \|\boldsymbol{\sigma}^a(\lambda) - \boldsymbol{\sigma}^b\|_2, \end{cases} \quad (3.31)$$

où $\|\cdot\|_2$ désigne la norme euclidienne dans \mathbb{R}^N ou \mathbb{R}^d , est en forme de L. Le point de courbure maximale (le "coin" du L) détermine le coefficient λ d'équilibre [Hansen, 1992].

En situation accidentelle, cette technique a notamment été mise en œuvre par Davoine and Bocquet [2007] pour l'estimation du terme source consécutif à l'accident de Tchernobyl, puis par Krysta et al. [2008] pour la reconstruction de l'expérience ETEX. Dans notre situation, la L-curve permet "seulement" d'estimer le rapport $\frac{r}{m}$, ce qui est suffisant pour calculer l'estimateur de la source $\boldsymbol{\sigma}^a$, mais pas les incertitudes associées (que ce soit sous la forme de la matrice de covariance d'erreur d'analyse P_{BLUE} ou par une simulation Monte Carlo). Pour estimer à la fois r et m , un diagnostic du χ^2 est pertinent pour compléter la L-curve. Quand \mathbf{R} et \mathbf{B} sont correctement estimées, la variable aléatoire

$$(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) + (\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)^T \mathbf{B}^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}^b) = 2\mathcal{J}_q(\boldsymbol{\sigma}), \quad (3.32)$$

qui représente le niveau d'erreur totale dans le système, suit une loi du χ^2 et son espérance mathématique, égale à $2\mathcal{J}_q(\boldsymbol{\sigma}^a)$, est égale au nombre d'observations d (voir Ménard et al. [2000] ou Tarantola [2005] par exemple). Ainsi, une fois le rapport $\frac{r}{m}$ fixé par la méthode de la L-curve, les hyper-paramètres r et m sont estimés pour satisfaire la relation :

$$\mathcal{J}_q(\boldsymbol{\sigma}^a) = \frac{d}{2}. \quad (3.33)$$

Extension à $N_d + 1$ hyper-paramètres

Il est possible d'étendre le concept de la L-curve, toujours couplé à un diagnostic du χ^2 , pour l'estimation de $N_d + 1$ hyper-paramètres. En utilisant les définitions des matrices \mathbf{R} et \mathbf{B} données par les équations 3.29, la fonction coût peut s'écrire :

$$\mathcal{J}_q(\boldsymbol{\sigma}) = \frac{1}{m^2} \left[\frac{1}{2} \sum_{i=1}^{N_d} \frac{1}{\lambda_i^2} (\boldsymbol{\mu}_i - \mathbf{H}_i\boldsymbol{\sigma})^T (\boldsymbol{\mu}_i - \mathbf{H}_i\boldsymbol{\sigma}) + \frac{1}{2} (\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)^T (\boldsymbol{\sigma} - \boldsymbol{\sigma}^b) \right], \quad (3.34)$$

où $\boldsymbol{\mu}_i$ représente le vecteur des observations du jeu de données i , \mathbf{H}_i est la sous-matrice de \mathbf{H} relative à ce jeu de donnée et $\lambda_i = \frac{r_i}{m}$. Nous noterons par la suite $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{N_d})^T$ le vecteur des coefficients λ_i .

Considérons l'hypersurface (dans \mathbb{R}^{N_d+1}) paramétrique définie par :

$$\begin{cases} x_1(\boldsymbol{\lambda}) = \ln \|\boldsymbol{\mu}_1 - \mathbf{H}_1 \boldsymbol{\sigma}^a(\boldsymbol{\lambda})\|_2 \\ \dots \\ x_{N_d}(\boldsymbol{\lambda}) = \ln \|\boldsymbol{\mu}_{N_d} - \mathbf{H}_{N_d} \boldsymbol{\sigma}^a(\boldsymbol{\lambda})\|_2 \\ x_{N_d+1}(\boldsymbol{\lambda}) = \ln \|\boldsymbol{\sigma}^a(\boldsymbol{\lambda}) - \boldsymbol{\sigma}^b\|_2 \end{cases} \quad (3.35)$$

La localisation du point de cette hypersurface où la courbure gaussienne est maximale permet de déterminer le vecteur $\boldsymbol{\lambda}$ d'équilibre. Finalement, la valeur de chaque hyper-paramètre est estimée par un diagnostic du χ^2 pour satisfaire l'équation 3.33.

Le cadre théorique de cette extension du principe de la L-curve est pertinent. Néanmoins dans la pratique, le calcul de la courbure gaussienne d'une hypersurface à partir d'un ensemble (discret) de points est un exercice difficile. Une alternative proposée par Belge et al. [2002] permet de remplacer le calcul de la courbure gaussienne par le calcul d'une distance entre un point origine et un point de l'hypersurface. Ils ont démontré que si ce point origine est judicieusement choisi, le problème est équivalent à la minimisation de cette distance et ont également proposé, dans le cas gaussien, un schéma itératif qui converge vers le point d'équilibre.

3.6.3 Principe de la vraisemblance maximale

La seconde méthode d'estimation des hyper-paramètres repose sur le principe de la vraisemblance maximale [Dee, 1995]. La vraisemblance marginale des observations $p(\boldsymbol{\mu})$ est une fonction dépendant des hyper-paramètres. On peut l'écrire

$$\begin{aligned} p(\boldsymbol{\mu}) &= p(\boldsymbol{\mu}|\boldsymbol{\theta}) = \int p(\boldsymbol{\mu}|\boldsymbol{\sigma}) p(\boldsymbol{\sigma}) d\boldsymbol{\sigma} = \int p(\boldsymbol{\mu}|\boldsymbol{\sigma}, \boldsymbol{\theta}) p(\boldsymbol{\sigma}|\boldsymbol{\theta}) d\boldsymbol{\sigma} \\ &= \int p_\epsilon(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) p_b(\boldsymbol{\sigma}) d\boldsymbol{\sigma} = \int p_\epsilon(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}|\boldsymbol{\theta}) p_b(\boldsymbol{\sigma}|\boldsymbol{\theta}) d\boldsymbol{\sigma}. \end{aligned} \quad (3.36)$$

Le principe de la vraisemblance maximale est le suivant : Le vecteur $\boldsymbol{\theta}$ d'hyper-paramètres qui maximise la vraisemblance marginale des observations est le vecteur d'hyper-paramètres le plus probable étant données les observations.

Statistiques gaussiennes

En statistiques gaussiennes, la vraisemblance marginale des observations (le calcul est détaillé dans l'annexe A) s'écrit :

$$p(\boldsymbol{\mu}|\boldsymbol{\theta}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^b)^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^b)}}{\sqrt{(2\pi)^d |\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}|}}. \quad (3.37)$$

Il est donc possible avec cette formule de procéder à une exploration de l'espace des hyper-paramètres afin de trouver le vecteur $\boldsymbol{\theta}$ qui maximise la vraisemblance marginale des observations. Nous nommerons par la suite cette méthode "méthode de la vraisemblance maximale" ou MVM. Dans le cas où les observations sont plus nombreuses que les inconnues à retrouver ($N < d$) le calcul peut être accéléré en utilisant la formule de Sherman-Morrison-Woodbury

pour écrire :

$$\begin{aligned} (\mathbf{HBH}^T + \mathbf{R})^{-1} &= \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})^{-1}\mathbf{H}^T\mathbf{R}^{-1} \\ |\mathbf{HBH}^T + \mathbf{R}| &= |\mathbf{R}||\mathbf{B}||\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1}|. \end{aligned} \quad (3.38)$$

Statistiques gaussiennes tronquées

Dans le cadre de statistiques gaussiennes tronquées pour les erreurs d'ébauche, la vraisemblance marginale des observations devient (cf. annexe A) :

$$\begin{aligned} p(\boldsymbol{\mu}|\boldsymbol{\theta}) &= \frac{e^{-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^b)^T(\mathbf{HBH}^T + \mathbf{R})^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^b)}}{\Phi(\boldsymbol{\sigma}^b, \mathbf{B}, \mathbf{0})\sqrt{(2\pi)^d|\mathbf{HBH}^T + \mathbf{R}|}} \\ &\quad \times \int \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^{\text{BLUE}})^T\mathbf{P}_{\text{BLUE}}^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^{\text{BLUE}})}}{\sqrt{(2\pi)^N|\mathbf{P}_{\text{BLUE}}|}} \mathbb{I}_{\boldsymbol{\sigma} \geq \mathbf{0}} d\boldsymbol{\sigma}. \end{aligned} \quad (3.39)$$

L'exploration de l'espace des hyper-paramètres est toujours envisageable, mais plus délicate. L'estimation de l'intégrale d'une distribution normale multivariée tronquée est un problème difficile. Le calcul numérique de cette intégrale est néanmoins possible en utilisant des méthodes stochastiques [Hajivassiliou et al., 1996], mais le coût en temps de calcul est plus élevé que dans le cas gaussien.

Il est intéressant de noter que les formules de la vraisemblance marginale sont indépendantes des modèles choisis pour les matrices \mathbf{R} et \mathbf{B} . La méthode de la vraisemblance maximale n'est donc absolument pas limitée à l'estimation de variances d'erreurs, mais peut être appliquée dans un cadre plus général pour l'estimation de différents types d'hyper-paramètres (tels que des longueurs ou des temps de corrélation par exemple). *In fine* dans nos cas d'étude, nous allons utiliser les formules de la vraisemblance pour estimer des variances d'erreurs telles que définies par les équations 3.26 et 3.29.

Il est également à noter que l'estimation des hyper-paramètres est réalisée avant le calcul de l'estimateur, dans un mode *offline*.

3.6.4 Schéma itératif de Desrozières

La troisième méthode que nous proposons s'inscrit également dans le cadre de la vraisemblance maximale mais cherche à éviter l'exploration systématique de l'espace des hyper-paramètres, qui peut se révéler coûteuse. Une méthode plus rapide consiste à proposer un schéma numérique itératif qui convergerait rapidement vers la combinaison d'hyper-paramètres qui maximise la vraisemblance. Cette méthode est inspirée de Desrozières and Ivanov [2001] en prévision météorologique et sera appelée par la suite schéma de Desrozières. Le point de départ est l'expression de la log-vraisemblance marginale des observations dans le cas gaussien (issue de l'équation 3.37) :

$$\begin{aligned} \ln p(\boldsymbol{\mu}|\boldsymbol{\theta}) &= -\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^b)^T(\mathbf{HBH}^T + \mathbf{R})^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^b) \\ &\quad -\frac{1}{2}\ln|\mathbf{HBH}^T + \mathbf{R}| + C, \end{aligned} \quad (3.40)$$

où C est une constante qui ne dépend pas de $\boldsymbol{\theta}$.

Dans le cas de $N_d + 1$ variances d'erreur à estimer, définies par les équations 3.29, chercher à maximiser la log-vraisemblance (et donc à annuler son gradient) conduit aux formules

suivantes :

$$\begin{aligned} m^2 &= \frac{2 J'_b(\boldsymbol{\sigma}^a)}{N - \text{tr}(\mathbf{P}_{\text{BLUE}} \mathbf{B}^{-1})}, \\ r_i^2 &= \frac{2 J'_{oi}(\boldsymbol{\sigma}^a)}{d_i - \text{tr}(\mathbf{H}_i \mathbf{P}_{\text{BLUE}} \mathbf{H}_i^T \mathbf{R}_i^{-1})}, \end{aligned} \quad (3.41)$$

où \mathbf{H}_i et \mathbf{R}_i représentent les sous-matrices respectivement de \mathbf{H} et \mathbf{R} relatives au jeu de données i dont le vecteur d'observation est noté $\boldsymbol{\mu}_i$. \mathbf{P}_{BLUE} est définie par l'équation 3.19 et J'_b et J'_{oi} sont définis par :

$$\begin{aligned} J'_b(\boldsymbol{\sigma}) &= \frac{1}{2} \|\boldsymbol{\sigma} - \boldsymbol{\sigma}^b\|_2^2, \\ J'_{oi}(\boldsymbol{\sigma}) &= \frac{1}{2} \|\boldsymbol{\mu}_i - \mathbf{H}_i \boldsymbol{\sigma}\|_2^2. \end{aligned} \quad (3.42)$$

Dans le cas de statistiques gaussiennes, $\boldsymbol{\sigma}^a = \boldsymbol{\sigma}^{\text{BLUE}}$ défini par l'équation 3.15. Les équations 3.41 peuvent être utilisées dans un schéma itératif. À chaque itération, l'estimateur $\boldsymbol{\sigma}^a$, la matrice de covariance d'erreur d'analyse \mathbf{P}_{BLUE} , puis les variances sont mises à jour (estimation *online*). Ce schéma converge (rapidement) vers un point-fixe qui est un maximum local de la log-vraisemblance [Chapnik et al., 2006]. Sa capacité à converger vers le maximum global dépend des hyper-paramètres initiaux choisis. Dans ce cas, le schéma de Desroziers et la MVM convergent vers la même solution.

Dans le cas de statistiques gaussiennes tronquées, les mêmes équations 3.41 seront utilisées, avec cette fois $\boldsymbol{\sigma}^a = \underset{\boldsymbol{\sigma} \geq 0}{\text{argmin}} \mathcal{J}_q$ (néanmoins \mathbf{P}_{BLUE} reste la matrice de covariance d'erreur d'analyse du BLUE définie par l'équation 3.19). Ce schéma itératif converge vers un point fixe, qui n'est ni le maximum de la vraisemblance de l'équation 3.39, ni même celui de la log-vraisemblance de l'équation 3.40. Dans ce cas, le schéma de Desroziers sera considéré comme une approximation, dont l'avantage principal est la rapidité de convergence. À mesure que le nombre d'observations augmente, nous pouvons nous attendre à ce que les résultats de ce schéma et de la MVM convergent.

3.7 Application I : Surveillance d'un parc de centrales nucléaires

L'objectif de cette étude consiste à proposer un algorithme semi-automatique de surveillance d'un parc d'installations nucléaires européen, capable en cas de situation d'urgence d'assimiler les observations, de reconstruire le terme source provenant des rejets accidentels éventuels, et de prédire à court terme les zones impactées. Réalisée en collaboration avec l'Institut finlandais de météorologie (Finnish Meteorology Institute, FMI), cette étude peut permettre en outre d'évaluer l'efficacité des réseaux d'observation, en particulier le réseau OPERA-AIR proposé par Saunier et al. [2009]. Cette étude constitue la première partie d'un article publié dans Atmospheric Environment [Winiarek et al., 2011] et présenté en Annexe D.

3.7.1 Installations nucléaires surveillées et réseaux d'observation

En ce qui concerne la France, toutes les installations nucléaires civiles sont surveillées, ce qui représente 19 centrales et l'usine de retraitement de La Hague. Le réseau de mesures considéré est le réseau proposé par Saunier et al. [2009], renommé OPERA-AIR. Ce réseau a été configuré en utilisant des méthodes de *network design*, basées sur l'optimisation de l'observabilité des accidents potentiels. Ce réseau est constitué de 100 stations de mesures des concentrations d'activité dans l'air de plusieurs radionucléides sur une plage $10^{-6} - 10^9 \text{ Bq m}^{-3}$.

Concernant la Finlande, 6 sites sont sous surveillance : 2 en Finlande, 2 en Suède et 2 en Russie. Le réseau d'observation considéré est l'actuel réseau "Uljas" qui mesure les doses équivalentes du rayonnement γ . Mais par simplification nous considérerons dans nos expériences synthétiques que ce réseau mesure directement des concentrations d'activité dans l'air. Les mesures de dose équivalente de rayonnement γ sont en effet plus délicates à assimiler car elles rendent compte de la contribution de plusieurs radionucléides présents dans le panache, mais également déjà déposés au sol. Les centrales nucléaires finlandaises sont situées sur les côtes et les centrales suédoises et russes sont assez éloignées du réseau de surveillance finlandais. Cette situation peut impliquer des difficultés possibles dans le processus d'inversion. Cette situation est différente de la configuration sur la France, dans laquelle les installations sous surveillance, bien que réparties non-uniformément sur le territoire français, sont à l'intérieur de l'enveloppe convexe du réseau d'observation.

Cette configuration est représentée sur la figure 3.1 sur laquelle sont représentés les installations surveillées et les réseaux d'observation.

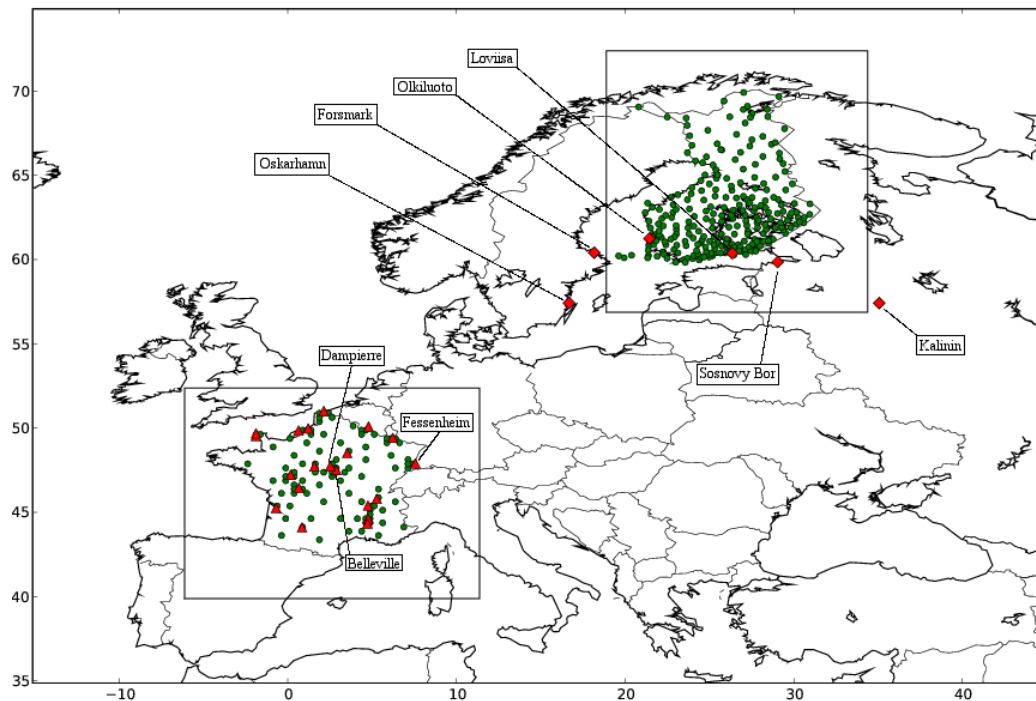


FIG. 3.1: Carte du domaine de simulation et des deux sous-domaines d'étude (France et Finlande). Les diamants rouges représentent les installations nucléaires surveillées par le système et les points verts les stations de mesure.

3.7.2 Modèles de dispersion atmosphérique

Deux modèles de dispersion atmosphérique à échelle régionale sont utilisés.

Le premier est POLAIR3D, le modèle eulérien de la plateforme POLYPHEMUS, développée par le CERE. En ce qui concerne la modélisation du transport de radionucléides, ce modèle a été validé sur l'expérience européenne de traceur (ETEX), l'incident d'Algésiras et l'accident de Tchernobyl [Quélo et al., 2007].

Le césium-137 a été choisi comme espèce d'étude. Le modèle calcule la concentration c de

^{137}Cs , suivant l'équation de transport :

$$\frac{\partial c}{\partial t} + \text{div}(\mathbf{u}c) = \text{div}\left(\rho \mathbf{K} \nabla \left(\frac{c}{\rho}\right)\right) - \Lambda^s c - \Lambda^d c + \sigma. \quad (3.43)$$

\mathbf{u} représente le champ de vent, \mathbf{K} la matrice de diffusion turbulente, considérée diagonale. La composante verticale K_z est calculée suivant la paramétrisation de Louis [1979]. La composante horizontale K_H est prise constante égale à zéro. Λ^s est le coefficient de lessivage (dépôt humide), Λ^d représente la décroissance radioactive et σ la source du radionucléide.

Le césium-137 est modélisé comme un traceur passif avec une décroissance radioactive, de demi-vie égale à 11000 jours. Le dépôt sec est modélisé en utilisant un schéma simple avec une vitesse de dépôt constante : $v^d = 0.2 \text{ cm s}^{-1}$. En ce qui concerne le lessivage, nous avons opté pour un schéma de type *below-cloud*, basé sur les précipitations :

$$\Lambda^s = a \left(\frac{p}{p_0}\right)^b, \quad (3.44)$$

où p représente le taux de précipitation et $p_0 = 1 \text{ mm.h}^{-1}$, a et b étant deux constantes respectivement égales à $8 \cdot 10^{-5} \text{ s}^{-1}$ et 0.8 [Baklanov and Sørensen, 2001].

Pour des raisons de coût calcul, nous avons opté pour ce modèle relativement simple plutôt qu'un modèle plus complexe, par exemple de particules résolues en taille (tel SIREAM dans POLAIR3D). Les méthodes de modélisation inverse s'appuient sur l'utilisation intensive du modèle et l'utilisation d'un modèle relativement simple permet de conserver des temps de calcul compatibles avec le caractère opérationnel de l'algorithme proposé.

Le second modèle est SILAM, le modèle de dispersion atmosphérique développé par le FMI. Ce modèle incorpore à la fois les approches eulérienne et lagrangienne. Le noyau lagrangien est décrit dans Sofiev et al. [2006]. Le noyau eulérien repose sur le schéma de Galperin [2000] pour l'advection et Sofiev [2002] pour la diffusion. Le dépôt sec est calculé suivant une analogie de résistances (Wesely [1989] pour les espèces gazeuses et Zhang et al. [2001] pour les aérosols). Enfin, le dépôt humide suit la paramétrisation de Sofiev [2000] pour les gaz et Sofiev et al. [2006] pour les aérosols.

Les deux modèles sont intégrés sur un domaine couvrant l'Europe à une résolution de $0.25^\circ \times 0.25^\circ$. Le nombre de points de grille est 240×160 . Le modèle POLAIR3D est configuré avec 9 niveaux verticaux s'étendant entre 0 et 3300 m ; le modèle SILAM utilise 7 niveaux jusqu'à 6300 m. Les résultats des modèles sont ensuite sauvegardés sur des sous-domaines couvrant la France et la Finlande. Cette configuration est représentée sur la figure 3.1. Les indicateurs de performance pour la restitution des panaches seront évalués sur ces sous-domaines.

Trois périodes de simulation seront considérées :

- L'année 2007 complète, utilisée pour évaluer les performances de l'algorithme pour une centrale nucléaire (Belleville en France), avec des conditions météorologiques variées.
- De plus, pour évaluer l'efficacité de l'algorithme sur toutes les installations nucléaires surveillées, deux périodes plus courtes (une semaine) seront utilisées. La première semaine débute le 2 décembre 2007, caractérisée par des conditions de vent fort.
- La seconde semaine débute le 8 juin 2007 et est caractérisée par des vents faibles.

Les champs météorologiques utilisés en entrée des deux modèles de dispersion sont les champs fournis par le Centre européen pour les prévisions météorologiques à moyen-terme (CEPMMT,

<http://www.ecmwf.int>). Leur résolution spatiale est de $0.25^\circ \times 0.25^\circ$ et leur résolution temporelle est de 3 heures.

3.7.3 Système opérationnel

Statistiques *a priori* des erreurs - Fonction coût

Dans le système d'estimation de source proposé dans cette section, les erreurs d'observation sont supposées distribuées suivant une loi normale $\mathcal{N}(\mathbf{0}, \mathbf{R})$. \mathbf{R} est choisie diagonale et

$$\mathbf{R}_{ii} = r^2 [\mathbf{H}\boldsymbol{\sigma}]_i^2 + r^2 \mu_{thr}^2, \quad (3.45)$$

où $r^2 \mu_{thr}^2$ représente une variance minimale prise en compte dans le système. Nous avons choisi $r = 0.5$ et $\mu_{thr} = 0.01 \text{ Bq.m}^3$, ce qui constitue des valeurs réalistes pour ce type de situations [Abida and Bocquet, 2009].

En contexte accidentel, le choix $\boldsymbol{\sigma}^b = \mathbf{0}$ est pertinent pour plusieurs raisons [Davoine and Bocquet, 2007; Winiarek et al., 2012] :

- La situation normale est celle où les émissions sont nulles. Aussi, beaucoup des paramètres de $\boldsymbol{\sigma}$ seront probablement nuls.
- Il est difficile de déterminer un terme d'ébauche pertinent, qui soit indépendant des observations utilisées par la suite dans l'inversion.
- Il est encore plus délicat d'estimer son incertitude.

Les erreurs d'ébauche sont supposées distribuées suivant une loi normale $\mathcal{N}(\mathbf{0}, \mathbf{B})$, avec $\mathbf{B} = m^2 \mathbf{I}_N$. Avec une ébauche nulle, m peut être vue comme un facteur d'amplitude, ou facteur de masse, qui empêche la source reconstruite de prendre des valeurs trop élevées et qui se révéleraient non physiques. Les paramètres r et m peuvent être déterminés par les méthodes présentées précédemment, mais dans cette section, du fait de l'antériorité de cette étude, leurs valeurs seront fixés *a priori*. Deux applications de l'estimation des hyper-paramètres sont présentées dans les sections 3.8 et 3.9.

Le système repose sur la minimisation de la fonction coût de l'équation 3.11, qui devient avec nos hypothèses :

$$\mathcal{J}(\boldsymbol{\sigma}) = \sum_{i=1}^d \left(\ln([\mathbf{H}\boldsymbol{\sigma}]_i + \mu_{thr}) + \frac{1}{2r^2} \frac{([\mathbf{H}\boldsymbol{\sigma}]_i - \mu_i)^2}{([\mathbf{H}\boldsymbol{\sigma}]_i + \mu_{thr})^2} \right) + \frac{1}{2m^2} \boldsymbol{\sigma}^T \boldsymbol{\sigma}. \quad (3.46)$$

Dans le cas où le terme source ne serait pas suffisamment contraint par les observations pour rester positif, des statistiques semi-gaussiennes pourraient être envisagées pour les erreurs d'ébauche, ce qui imposerait formellement la positivité de la source. Dans ce cas, la minimisation de la fonction coût devrait être réalisée sous contrainte de positivité. Dans les deux cas, aucune solution analytique n'est disponible et la minimisation est réalisée par un algorithme de Quasi-Newton [Byrd et al., 1995].

Algorithme

L'algorithme proposé s'appuie sur une hypothèse importante : l'installation nucléaire qui rencontre une situation d'urgence est connue (mais pas nécessairement l'instant t_0 de survenue

de l'accident). Cette hypothèse est réaliste, en particulier dans des pays tels que la France ou la Finlande dans lesquels les responsables d'une telle installation doivent prévenir les autorités compétentes en cas d'écart à la situation normale. Cette hypothèse réduit fortement le nombre d'inconnues N ; le problème inverse consiste dans ce cas à estimer un profil temporel des émissions, soit N débits d'émission.

En situation d'urgence, au temps t_n , l'algorithme séquentiel proposé contient trois étapes :

1. *Étape préliminaire* : Toutes les observations disponibles sur l'intervalle $[t_{n-1}, t_n]$ sont collectées et le vecteur des observations $\boldsymbol{\mu}$, qui contient toutes les observations disponibles entre t_0 et t_n , est mis à jour. En parallèle, la matrice \mathbf{H} est également mise à jour, par le calcul, à l'aide du modèle de dispersion atmosphérique direct, de $\Delta t_a / \Delta t_{source}$ nouvelles colonnes (où $\Delta t_a = t_n - t_{n-1}$ est l'intervalle d'assimilation et Δt_{source} est la résolution de la source reconstruite $\boldsymbol{\sigma}^a$) et d'autant de nouvelles lignes que de nouvelles observations.
2. *Étape d'analyse* : L'estimation du terme source $\boldsymbol{\sigma}^a$ est réalisée par la minimisation de la fonction coût \mathcal{J} (équation 3.46).
3. *Étape de prévision* : Une réanalyse entre t_0 et t_n , puis une prévision de Δt_f heures entre t_n et $t_n + \Delta t_f$ est réalisée avec le modèle de dispersion atmosphérique utilisé entre t_0 et $t_n + \Delta t_f$. Le terme source utilisé pour cette simulation est $\boldsymbol{\sigma}^a$ entre t_0 et t_n et une estimation entre t_n et $t_n + \Delta t_f$. Nous avons opté pour une hypothèse de persistance, ce qui signifie que le terme source est supposé constant à partir de t_n (égal au dernier terme estimé de $\boldsymbol{\sigma}^a$).

L'étape préliminaire peut être implémentée comme une tâche de fond y compris en situation normale, ce qui permettrait d'avoir une matrice \mathbf{H} et un vecteur d'observations $\boldsymbol{\mu}$ déjà à jour au déclenchement de l'urgence. En situation normale, les lignes et colonnes de \mathbf{H} les plus anciennes, ainsi que les observations correspondantes dans $\boldsymbol{\mu}$, peuvent être éliminées régulièrement.

3.7.4 Expériences synthétiques

Pour évaluer l'efficacité de l'algorithme proposé, nous allons procéder à un grand nombre d'expériences synthétiques, en faisant varier soit la date d'occurrence de l'accident, soit l'installation concernée et en générant à chacune de ces expériences des erreurs d'observation stochastiques.

Préparation d'une expérience - Vraie source - Vraies erreurs

Le profil temporel de la vraie source $\boldsymbol{\sigma}^t$, fourni par l'IRSN, correspond à une fusion de cœur sans rupture de confinement. Un second pic d'émission est présent 24 heures après le début de l'accident, dû à un rejet intentionnel destiné à abaisser la pression à l'intérieur de la structure de confinement du réacteur. Ce profil est représenté sur la figure 3.2.

Les mesures synthétiques $\boldsymbol{\mu}$ fournis au système d'assimilation de données sont générées à partir des résultats du modèle, supposé parfait, en y ajoutant des erreurs $\boldsymbol{\epsilon}^t$, appelées *erreurs vraies*, distribuées suivant une loi log-normale :

$$\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\sigma}^t + \boldsymbol{\epsilon}^t, \quad (3.47)$$

avec $\epsilon_i^t = e_i [\mathbf{H}\boldsymbol{\sigma}^t]_i$, $\ln(e_i)$ étant distribué suivant la loi normale $\mathcal{N}(1, 0.5)$.

Seules ces mesures (bruitées) $\boldsymbol{\mu}$ sont fournies au système de reconstruction de source ; les vraies erreurs $\boldsymbol{\epsilon}^t$ ne sont pas connues, ni même leur distribution statistique. Au contraire,

nous avons fait l'hypothèse d'erreurs distribuées suivant une loi normale, ce qui a conduit à la fonction coût de l'équation 3.46.

Évaluation de l'algorithme

Pour chaque expérience réalisée, nous avons considéré différents indicateurs statistiques pour évaluer l'efficacité de l'algorithme proposé :

- Pour l'écart entre la vraie source et la source reconstruite, la racine carrée de l'erreur quadratique moyenne (RMSE) :

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^N ([\sigma^t]_n - [\sigma^a]_n)^2}, \quad (3.48)$$

et le coefficient de corrélation de Pearson :

$$R = \frac{\sum_{n=1}^N [\sigma^a - \langle \sigma^a \rangle]_n [\sigma^t - \langle \sigma^t \rangle]_n}{\sqrt{\left(\sum_{n=1}^N [\sigma^a - \langle \sigma^a \rangle]_n^2 \right) \left(\sum_{n=1}^N [\sigma^t - \langle \sigma^t \rangle]_n^2 \right)}}, \quad (3.49)$$

sont considérés. $\langle \cdot \rangle$ désigne la moyenne temporelle (entre t_1 et t_N).

- De même, l'écart entre les "vraies" observations ($\mathbf{H}\sigma^t$) et les observations réanalysées et/ou prévues est évalué avec la RMSE :

$$\text{RMSE} = \sqrt{\frac{1}{d} \sum_{n=1}^d ([\mathbf{H}\sigma^t]_n - [\mathbf{H}\sigma^a]_n)^2}, \quad (3.50)$$

et le coefficient de corrélation de Pearson :

$$R = \frac{\sum_{n=1}^d [\mathbf{H}\sigma^a - \langle \mathbf{H}\sigma^a \rangle]_n [\mathbf{H}\sigma^t - \langle \mathbf{H}\sigma^t \rangle]_n}{\sqrt{\left(\sum_{n=1}^d [\mathbf{H}\sigma^a - \langle \mathbf{H}\sigma^a \rangle]_n^2 \right) \left(\sum_{n=1}^d [\mathbf{H}\sigma^t - \langle \mathbf{H}\sigma^t \rangle]_n^2 \right)}}. \quad (3.51)$$

Ici, $\langle \cdot \rangle$ désigne la moyenne sur l'ensemble des d observations prises en compte.

- Enfin, il est également possible de comparer les panaches complets "vrais", c'est-à-dire calculés par le modèle en connaissant la vraie source, et réanalysés ou prévus en utilisant la source reconstruite. Pour cela un indicateur spécifique, appelé *figure de mérite* et noté fm , est défini à un temps τ par :

$$\text{fm} = \frac{\sum_{h \in S} \min([c^a]_h, [c^t]_h)}{\sum_{h \in S} \max([c^a]_h, [c^t]_h)}, \quad (3.52)$$

où S désigne l'ensemble des indices des points de grille du sous-domaine d'évaluation. c^t désigne le champ (discrétisé sur la grille du modèle de dispersion) de concentrations vraies au temps τ et c^a le champ réanalysé ou prévu au même instant.

3.7.5 Résultats

Performance sur une année complète pour la centrale nucléaire de Belleville (France)

Le premier ensemble d'expériences a pour objectif d'évaluer le comportement de l'algorithme soumis à des conditions météorologiques diverses. Pour cela, la centrale nucléaire de Belleville, qui occupe une place centrale en France, est sélectionnée et l'année 2007 dans son intégralité est considérée. 8760 expériences synthétiques sont définies, une par heure de début d'accident possible. Chaque expérience a une durée d'1 semaine (168 heures). Nous choisissons $\Delta t_a = \Delta t_{source} = 1$ h et $\Delta t_f = 3$ h.

La première évaluation consiste à reconstruire la source *a posteriori*, c'est-à-dire une fois toutes les observations disponibles, $d = 168 \times 100 = 16800$. Avec une résolution de reconstruction $\Delta t_{source} = 1$ h, le nombre d'inconnues N est égal à 168. La source reconstruite moyenne (sur les 8760 expériences synthétiques) est représentée sur la figure 3.2 en comparaison de la vraie source. Les courbes pointillées de part et d'autre représentent les écarts types autour de la moyenne. En valeur absolue, les erreurs les plus fortes apparaissent autour des pics d'émission.

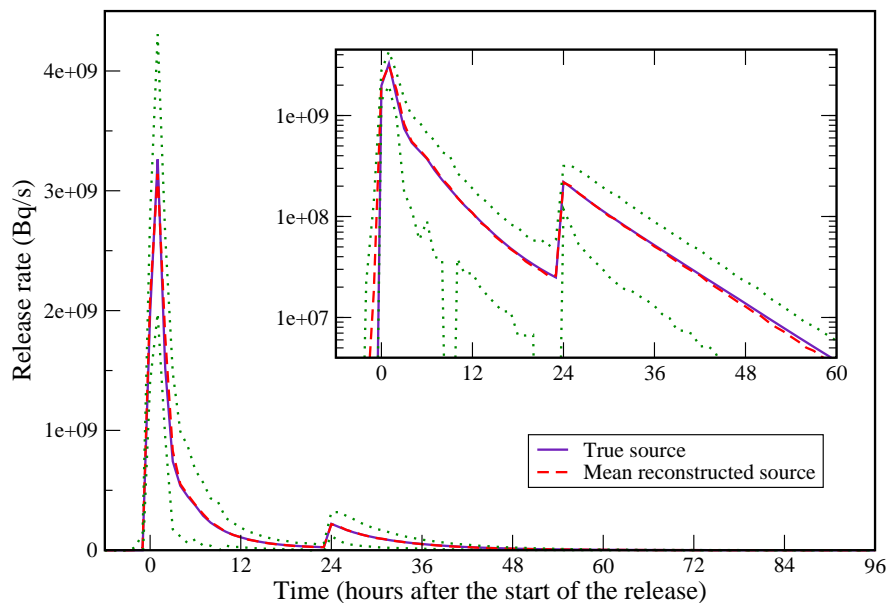


FIG. 3.2: Vraie source et moyenne des sources reconstruites. Les lignes pointillées représentent les écarts types autour de la moyenne.

Une seconde évaluation consiste à utiliser l'algorithme de façon séquentielle, ce qui correspondrait à son utilisation en situation d'urgence. Chaque heure, de nouvelles observations sont disponibles. Une nouvelle analyse (estimation du terme source) est alors réalisée en utilisant toutes les observations disponibles depuis le début de la situation d'urgence. Enfin une prévision à 3 heures est réalisée. Il est alors possible de calculer l'évolution des indicateurs statistiques au fur et à mesure de l'acquisition de nouvelles données. L'évolution de la RMSE concernant la source reconstruite est ainsi représentée sur la figure 3.3. Aucun impact du second pic n'est visible car les erreurs en valeur absolue sont principalement influencées par les erreurs faites sur le premier pic.

En ce qui concerne la qualité de la prévision, la figure de mérite des champs de concentration prévus à 3 h est représentée sur la figure 3.4. L'impact du second pic d'émission est

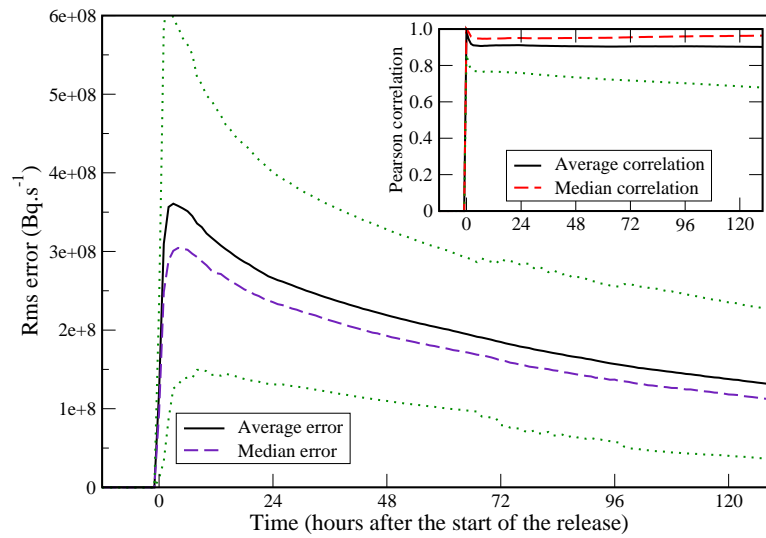


FIG. 3.3: Médiane, moyenne et écarts types autour de la moyenne de la RMSE des sources reconstruites. Dans l'encart : coefficient de corrélation de Pearson (médiane, moyenne et écarts types autour de la moyenne).

clairement visible, avec une baisse de performance durant environ 4 h.

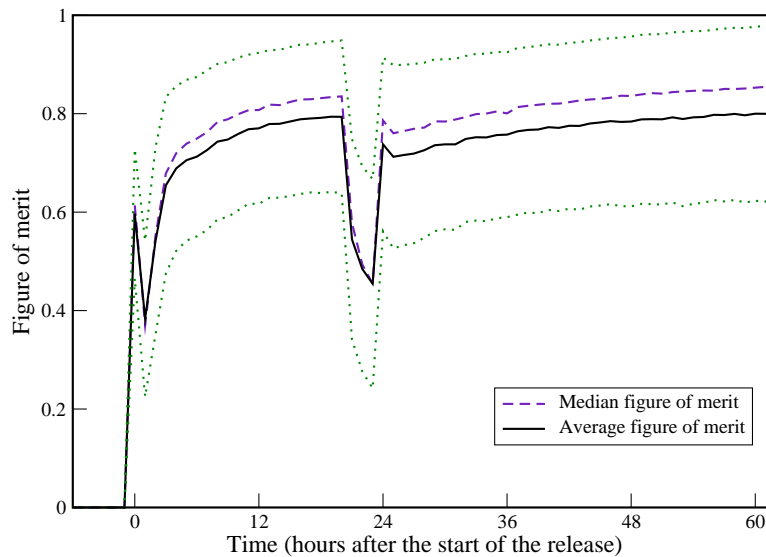


FIG. 3.4: Figure de mérite des champs de concentration prévus à 3 h (médiane, moyenne et écarts types à la moyenne).

Performance globale

Le second ensemble d'expériences synthétiques a pour objectif d'évaluer l'algorithme proposé en faisant varier la centrale responsable du rejet. Dans ce but, toutes les installations nucléaires surveillées sont testées sur deux périodes différentes d'une semaine. La première période est une semaine de décembre 2007, caractérisée par des vents forts. La seconde est

une semaine de juin 2007 et est caractérisée par des vents faibles. Pour chaque installation, 168 accidents (un accident par heure pendant une semaine) sont simulés, ce qui fait un total de 4368 expériences synthétiques pour chacune des deux semaines considérées. Les performances moyennes de l’algorithme sont ensuite évaluées en ce qui concerne la reconstruction de la source et la qualité de la prévision. Comme dans le paragraphe précédent, nous choisissons $\Delta t_a = \Delta t_{source} = 1$ h et $\Delta t_f = 3$ h.

Pour les deux périodes d’intérêt, les résultats sont très similaires à ceux de l’étude sur une année complète : mêmes ordres de grandeurs et mêmes évolutions pour les RMSE, corrélations et figures de mérite. Comme précédemment, la solution médiane est toujours meilleure que la solution moyenne, ce qui montre que le système est en général efficace, sauf sur quelques situations (environ 5%) sur lesquelles il échoue complètement, dégradant la performance moyenne mais peu la médiane. De plus le système présente des meilleurs résultats globaux sur la situation française que sur la situation finlandaise, les situations d’échec étant généralement associées à des installations situées aux bords ou en dehors de l’enveloppe convexe du réseau d’observations.

3.8 Application II : Fukushima (1)

3.8.1 Contexte

Le 11 mars 2011, à 5h46 UTC, un séisme sous-marin de magnitude 9.0 M_W se produit dans l’océan Pacifique au large du Japon, suivi d’un tsunami extrêmement destructeur qui frappa les côtes japonaises environ une heure plus tard. Ces événements ont entraîné l’arrêt automatique de quatre centrales nucléaires japonaises. Des générateurs diesel étaient alors prévus pour alimenter les centrales en électricité et permettre le refroidissement des réacteurs et des piscines de stockage de combustible usagé. Dans la centrale de Fukushima Daiichi, ces systèmes de secours n’ont pas fonctionné, principalement du fait des dégâts occasionnés par le tsunami.

La centrale nucléaire de Fukushima Daiichi possède six réacteurs. Au moment du tremblement de terre, le combustible du réacteur 4 avait déjà été retiré et les réacteurs 5 et 6 étaient en situation d’arrêt à froid pour maintenance.

Dans les heures qui ont suivi, la situation est rapidement devenue critique. Les réacteurs 1, 2 et 3 ont subi des fusions au moins partielles ainsi que des dégagements de fumées et des explosions d’hydrogène. De plus, les barres de combustible usagé stockées dans des piscines prévues à cet effet dans le bâtiment de chaque réacteur sont montés en température à mesure que le niveau d’eau dans les piscines baissait.

Tous ces événements ont causé des rejets massifs de produits radioactifs dans l’atmosphère qu’il est important d’estimer afin de déterminer les zones impactées. Quelques études ont tenté de répondre à cette problématique. Chino et al. [2011], en utilisant une méthode basée sur l’identification préalable d’épisodes de rejet, ont réalisé une estimation des termes sources de ^{137}Cs et de ^{131}I en utilisant seulement quelques observations de concentration d’activité dans l’air au Japon. Mathieu et al. [2012], par une méthode “modèle direct / correction”, ont estimé les termes sources de différents radionucléides en utilisant de nombreuses données : des informations “installations” (événements connus, mesures de température et de pression dans les réacteurs), des mesures sur site (dose équivalente de rayonnement γ) et des observations plus lointaines au Japon. Stohl et al. [2012] ont estimé les termes sources de ^{137}Cs et de ^{133}Xe par modélisation inverse en utilisant des concentrations d’activité dans l’air et des mesures de dépôt dans tout l’hémisphère Nord. Il semble néanmoins que leur estimation était fortement contrainte par leur terme d’ébauche. Enfin dans un second temps, Saunier et al. [2013] ont pro-

posé une méthode pour utiliser les mesures de dose équivalente de rayonnement γ pour estimer le terme source par modélisation inverse. Les mesures de dose équivalente de rayonnement γ sont généralement les plus abondantes et les mieux distribuées mais elles rendent compte de l'influence de plusieurs radionucléides contenus dans le panache et aussi déjà déposés sur le sol, ce qui rend leur assimilation plus délicate. Dans Winiarek et al. [2014] (présenté dans la section 3.9 et visible en annexe F), nous avons utilisé des données de concentration d'activité dans l'air et des données de dépôt pour estimer le terme source de ^{137}Cs en estimant rigoureusement les erreurs *a priori* des différents jeux de données et de l'ébauche.

L'étude présentée dans cette section, réalisée en 2011, a fait l'objet d'un article publié en 2012 dans *Journal of Geophysics Research Atmospheres* [Winiarek et al., 2012], visible en annexe E. L'objectif est l'estimation des termes sources de ^{137}Cs et de ^{131}I en utilisant les mesures de concentrations d'activité dans l'air disponibles.

3.8.2 Observations

Pour réaliser cette première estimation, nous avons utilisé les mesures de concentration d'activité dans l'air rendues publiques avant le 31 août 2011. À cette date, toutes les données de l'International Monitoring System (IMS) du CTBTO (Comprehensive Nuclear-Test-Ban Treaty Organization) n'étaient pas accessibles y compris pour la communauté scientifique. Néanmoins, un certain nombre de pays (parmi lesquels le Japon, le Canada, les États-Unis et les Philippines) avaient déjà fait le choix de la transparence.

Au Japon, les données de la station JPP38 de l'IMS ont été rendues publiques. De plus, des mesures ont été réalisées par le KEK (High Energy Accelerator Research Organization) et le NIES (National Institute for Environmental Studies) à Tsukuba, et par le MITRI (Metropolitan Industrial Technology Research Institute) à Tokyo. Dans la préfecture de Fukushima, le MEXT (Ministère de l'éducation, de la culture, du sport, des sciences et technologies) a réalisé et diffusé un nombre assez important de mesures (mais généralement après le 18 mars).

Aux États-Unis, l'USNDC (United States National Data Center) a communiqué les données des stations de l'IMS : USP70 en Californie, et USP77, USP78, USP79 et USP80 sur des îles situées dans l'Océan Pacifique Nord. De plus, l'EPA (Environmental Protection Agency) a également réalisé des mesures sur son réseau domestique RadNet, ainsi que le Ministère de la Santé au Canada. Enfin, le PNRI (Philippine Nuclear Research Institute) a également rapidement communiqué les mesures de la station PHP52 du réseau IMS.

Au total, 267 mesures de concentration d'activité dans l'air ont été utilisées pour l'estimation de la source de césium-137 (104 situées au Japon et 163 hors du Japon), et 428 mesures pour l'estimation de la source d'iode-131 (233 au Japon et 195 hors du Japon). Les informations sur la provenance des observations sont synthétisées dans le Tableau 3.1 et leur localisation sur la Figure 3.5.

3.8.3 Modèle de dispersion atmosphérique

Le modèle de dispersion atmosphérique utilisé, adapté aux observations à disposition, est le modèle POLAIR3D utilisé sur un domaine suffisamment étendu pour couvrir à la fois le Japon, l'Océan Pacifique Nord et une grande partie des États-Unis et du Canada (voir Figure 3.5). La résolution spatiale horizontale est de $0.25^\circ \times 0.25^\circ$, représentant 652×256 points de grille. Le modèle est configuré avec 15 niveaux verticaux, s'étendant de 0 à 8000m.

Le modèle calcule le champ de concentration c de ^{137}Cs ou de ^{131}I , considérées comme des espèces gazeuses passives, en résolvant l'équation de transport suivante :

$$\frac{\partial c}{\partial t} + \text{div}(\mathbf{u}c) = \text{div}\left(\rho \mathbf{K} \nabla \left(\frac{c}{\rho}\right)\right) - \Lambda^s c - \Lambda^d c + \sigma. \quad (3.53)$$

TAB. 3.1: Origines et nombre d'observations de concentration d'activité dans l'air publiquement communiquées au 31 août 2011.

Source	^{137}Cs	^{131}I
KEK-NIES - Tsukuba	10	10
MITRI - Tokyo	58	160
JPP38 (station CTBTO)	20	9
MEXT - (Préfecture de Fukushima)	16	54
Total Japon	104	233
USNDC - CTBTO - États-Unis	77	82
RadNet - EPA - États-Unis	44	69
Ministère de la Santé - Canada	29	31
PNRI - PHP52 (station CTBTO)	13	13
Total Hors Japon	163	195
Total	267	428

\mathbf{u} représente le champ de vent, \mathbf{K} la matrice de diffusion turbulente, considérée diagonale. La composante verticale K_z est calculée suivant la paramétrisation de Louis [1979]. La composante horizontale K_H est prise constante (égale à $25000 \text{ m}^2 \text{ s}^{-1}$). Λ^s est le coefficient de lessivage, Λ^d représente la décroissance radioactive et σ la source du radionucléide.

Le domaine est suffisamment étendu pour éviter les recirculations de radionucléides non-prises en compte à l'extérieur du domaine ; ainsi, excepté au niveau du sol, les conditions aux limites sont des conditions de Dirichlet (concentrations nulles). Au sol la condition aux limites est :

$$K_z \nabla c \cdot \mathbf{n} = -v^d c, \quad (3.54)$$

où \mathbf{n} est le vecteur unitaire orienté vers l'intérieur du domaine, et v^d est la vitesse de dépôt sec du radionucléide.

Le césium-137 et l'iode-131 sont modélisés comme des traceurs gazeux passifs avec une décroissance radioactive. Leurs demi-vies sont égales respectivement à 11000 et 8,04 jours. Le dépôt sec est modélisé en utilisant un schéma simple avec une vitesse de dépôt constante : $v^d = 0.2 \text{ cm s}^{-1}$ pour ^{137}Cs et $v^d = 0.5 \text{ cm s}^{-1}$ pour ^{131}I .

En ce qui concerne le dépôt humide, une paramétrisation sous-maille, basée sur l'estimation de l'humidité relative, est choisie :

$$\Lambda^s = \max \left(0, \Lambda_{\max}^s \left(\frac{RH - RH_t}{RH_s - RH_t} \right) \right). \quad (3.55)$$

$\Lambda_{\max}^s = 3.5 \times 10^{-5} \text{ s}^{-1}$, RH représente l'humidité relative, RH_t une valeur seuil ($RH_t = 80\%$) et RH_s la valeur de saturation (= 100%) [Pudykiewicz, 1989; Brandt et al., 2002].

Les champs météorologiques utilisés en entrée du modèle de dispersion POLAIR3D sont les champs fournis par le CEPMMT. Leur résolution spatiale est de $0.25^\circ \times 0.25^\circ$ et leur résolution temporelle est de 3 heures.

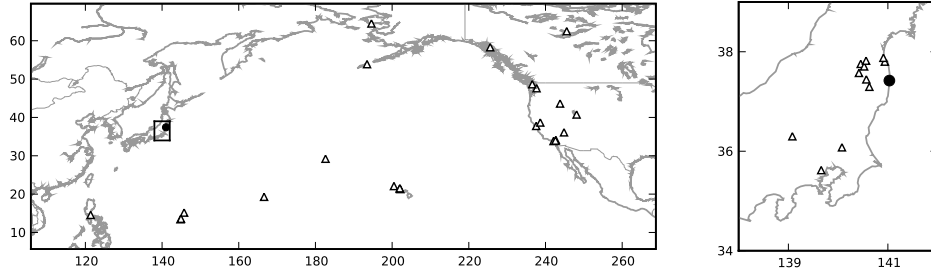


FIG. 3.5: À gauche : domaine de simulation. Les triangles représentent les stations de mesures à l'extérieur du Japon. À droite : zoom sur le Japon. Les triangles représentent les stations de mesures au Japon. Sur les deux figures, le point noir représente la centrale nucléaire de Fukushima Daiichi.

3.8.4 Modélisation inverse

3.8.4.1 Vecteur source, matrice Jacobienne et fonction coût

L'origine géographique des émissions de radionucléides est connue : il s'agit de la centrale nucléaire de Fukushima Daiichi (les émissions ultérieures dans l'atmosphère issues des effluents liquides sécoulant de la centrale étant négligeables comparées aux émissions directes dues aux explosions et fuites des bâtiments). Étant donnée la nature de l'accident, nous supposons de plus que la totalité des radionucléides sont émis dans la première couche verticale du modèle, c'est-à-dire entre 0 et 40m (contrairement par exemple à l'accident de Tchernobyl où une gigantesque explosion, puis un incendie, ont projeté des polluants radioactifs à plusieurs milliers de mètres d'altitude). Cette situation réduit très fortement le nombre d'inconnues dans le système, la source étant ponctuelle et localisée. Le problème inverse consiste donc à retrouver un profil temporel, de résolution 1h, entre le 11 mars 2011 et le 26 mars 2011 inclus, ce qui représente $N = 384$ débits d'émission à estimer.

Indépendamment du nombre d'observations disponibles ($d = 267$ pour le césium-137 et $d = 428$ pour l'iode-131), nous avons choisi de calculer la matrice Jacobienne \mathbf{H} colonne par colonne en réalisant N simulations directes. Une fois que les champs de concentrations calculés par ces N simulations ont été stockés, ce choix nous permet, dans le cas de nouvelles observations disponibles, de mettre à jour la matrice Jacobienne sans réaliser de simulations supplémentaires.

Comme dans la section 3.7, l'ébauche est choisie nulle, $\sigma^b = \mathbf{0}$.

En statistiques gaussiennes, l'estimateur est donné par la formule du BLUE qui devient avec nos hypothèses sur l'ébauche :

$$\sigma^{\text{BLUE}} = (\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1})^{-1} \mathbf{H}^T \mathbf{R}^{-1} \boldsymbol{\mu} = \mathbf{B} \mathbf{H}^T (\mathbf{H} \mathbf{B} \mathbf{H}^T + \mathbf{R})^{-1} \boldsymbol{\mu}. \quad (3.56)$$

En statistiques gaussiennes tronquées, l'estimation du profil temporel des émissions σ est réalisée par la minimisation sous contrainte de positivité de la fonction coût :

$$\mathcal{J}_q(\sigma) = \frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\sigma)^T \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\sigma) + \frac{1}{2} \sigma^T \mathbf{B}^{-1} \sigma. \quad (3.57)$$

3.8.4.2 Estimation des erreurs *a priori*

Étant donné que seules des concentrations d'activité dans l'air sont utilisées pour l'inversion, et du fait du faible nombre d'observations, un modèle à deux hyper-paramètres est choisi

pour caractériser les erreurs *a priori*. Ainsi les matrices de covariance d'erreur \mathbf{R} et \mathbf{B} sont définies par :

$$\mathbf{B} = m^2 \mathbf{I}_N, \quad \mathbf{R} = r^2 \mathbf{I}_d. \quad (3.58)$$

Les statistiques gaussiennes et gaussiennes tronquées (qui avec nos choix d'ébauche et de matrice \mathbf{B} sont maintenant des statistiques semi-gaussiennes) sont considérées, même si nous pouvons anticiper des termes négatifs dans les sources reconstruites dans le cadre de statistiques gaussiennes. Les trois méthodes présentées dans le Chapitre 3.6 sont utilisées pour estimer les variances des erreurs.

L-curve

L'algorithme pour estimer les hyper-paramètres et le terme source est le suivant :

1. La fonction coût considérée est :

$$\mathcal{J}_\lambda(\boldsymbol{\sigma}) = \frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) + \frac{\lambda^2}{2} \boldsymbol{\sigma}^T \boldsymbol{\sigma}. \quad (3.59)$$

2. La L-curve est tracée suivant la courbe paramétrique :

$$\begin{cases} x(\lambda) = \ln \|\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^a(\lambda)\|_2 \\ y(\lambda) = \ln \|\boldsymbol{\sigma}^a(\lambda)\|_2 \end{cases} \quad (3.60)$$

Dans le cas de statistiques gaussiennes :

$$\boldsymbol{\sigma}^a(\lambda) = \boldsymbol{\sigma}^{\text{BLUE}}(\lambda) = \mathbf{H}^T (\mathbf{H}\mathbf{H}^T + \lambda^2 \mathbf{I}_d)^{-1} \boldsymbol{\mu}. \quad (3.61)$$

Dans le cas de statistiques semi-gaussiennes, $\boldsymbol{\sigma}^a(\lambda)$ est donné par la minimisation sous contrainte de positivité de la fonction coût \mathcal{J}_λ de l'équation 3.59.

3. Le paramètre λ d'équilibre, noté λ_e , est déterminé en localisant le point de la L-curve de courbure maximale. Pour cela, la courbe est dans un premier temps lissée à l'aide de splines cubiques suivant un algorithme proposé par De Boor [2001]. La courbure κ est ensuite estimée par la formule :

$$\kappa(\lambda) = \frac{x'y'' - x''y'}{(x'^2 + y'^2)^{\frac{3}{2}}}. \quad (3.62)$$

4. L'estimateur $\boldsymbol{\sigma}^a(\lambda_e)$ constitue alors la source estimée. La grandeur $\mathcal{J}_{\lambda_e}(\boldsymbol{\sigma}^a(\lambda_e))$ est calculée et les hyper-paramètres sont finalement estimés par :

$$\begin{aligned} r^2 &= \frac{2 \mathcal{J}_{\lambda_e}(\boldsymbol{\sigma}^a(\lambda_e))}{d}, \\ m^2 &= \frac{2 \mathcal{J}_{\lambda_e}(\boldsymbol{\sigma}^a(\lambda_e))}{\lambda_e^2 d}. \end{aligned} \quad (3.63)$$

Maximisation de la vraisemblance

En statistiques gaussiennes, l'estimation des variances des erreurs est réalisée par la maximisation de la vraisemblance marginale des observations :

$$p(\boldsymbol{\mu}|r, m) = \frac{e^{-\frac{1}{2} \boldsymbol{\mu}^T (m^2 \mathbf{H}\mathbf{H}^T + r^2 \mathbf{I}_d)^{-1} \boldsymbol{\mu}}}{\sqrt{(2\pi)^d |m^2 \mathbf{H}\mathbf{H}^T + r^2 \mathbf{I}_d|}}. \quad (3.64)$$

En statistiques semi-gaussiennes, la vraisemblance marginale des observations devient :

$$p(\boldsymbol{\mu}|r, m) = \frac{e^{-\frac{1}{2}\boldsymbol{\mu}^T(m^2\mathbf{H}\mathbf{H}^T+r^2\mathbf{I}_d)^{-1}\boldsymbol{\mu}}}{\sqrt{(2\pi)^d|m^2\mathbf{H}\mathbf{H}^T+r^2\mathbf{I}_d|}} \times \int \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^{\text{BLUE}})^T\mathbf{P}_{\text{BLUE}}^{-1}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^{\text{BLUE}})}}{\sqrt{(\pi/2)^N|\mathbf{P}_{\text{BLUE}}|}} \mathbb{I}_{\boldsymbol{\sigma}\geq\mathbf{0}} d\boldsymbol{\sigma}, \quad (3.65)$$

avec

$$\boldsymbol{\sigma}^{\text{BLUE}} = \mathbf{H}^T \left(\mathbf{H}\mathbf{H}^T + \frac{r^2}{m^2}\mathbf{I}_d \right)^{-1} \boldsymbol{\mu}, \quad (3.66)$$

$$\mathbf{P}_{\text{BLUE}} = m^2 \left(\mathbf{I}_N - \mathbf{H}^T \left(\mathbf{H}\mathbf{H}^T + \frac{r^2}{m^2}\mathbf{I}_d \right)^{-1} \mathbf{H} \right). \quad (3.67)$$

Pour estimer l'intégrale, un algorithme de Geweke-Hajivassiliou-Keane (GHK) est utilisé. Il s'agit d'une méthode stochastique par échantillonnage, détaillée dans Winiarek et al. [2012] (cf. Annexe E).

Une fois les hyper-paramètres r et m estimés, la source est reconstruite, soit par la formule du BLUE (équation 3.66) dans le cadre de statistiques gaussiennes, soit par minimisation sous contrainte de positivité de la fonction coût :

$$\mathcal{J}_q(\boldsymbol{\sigma}) = \frac{1}{2r^2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) + \frac{1}{2m^2} \boldsymbol{\sigma}^T \boldsymbol{\sigma}. \quad (3.68)$$

Schéma de Desroziers

Le schéma itératif de Desroziers suivant est utilisé pour estimer simultanément les hyper-paramètres et la source :

1. Si r_0^2 et m_0^2 sont les estimations des variances d'erreur à ce stade, l'estimateur $\boldsymbol{\sigma}^a$ est calculé, en statistiques gaussiennes :

$$\boldsymbol{\sigma}^a = \mathbf{H}^T \left(\mathbf{H}\mathbf{H}^T + \frac{r_0^2}{m_0^2}\mathbf{I}_d \right)^{-1} \boldsymbol{\mu}, \quad (3.69)$$

et en statistiques semi-gaussiennes, $\boldsymbol{\sigma}^a = \underset{\boldsymbol{\sigma}\geq\mathbf{0}}{\operatorname{argmin}} \mathcal{J}_q$ avec :

$$\mathcal{J}_q(\boldsymbol{\sigma}) = \frac{1}{2r_0^2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) + \frac{1}{2m_0^2} \boldsymbol{\sigma}^T \boldsymbol{\sigma}. \quad (3.70)$$

2. De nouvelles variances d'erreur r_1^2 et m_1^2 sont alors estimées :

$$r_1^2 = \frac{(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^a)^T (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^a)}{r_0^2 d - \operatorname{tr}(\mathbf{H}\mathbf{P}_{\text{BLUE}}\mathbf{H}^T)} r_0^2, \quad (3.71)$$

$$m_1^2 = \frac{\boldsymbol{\sigma}^{aT} \boldsymbol{\sigma}^a}{m_0^2 N - \operatorname{tr}(\mathbf{P}_{\text{BLUE}})} m_0^2.$$

avec $\mathbf{P}_{\text{BLUE}} = m_0^2 \left(\mathbf{I}_N - \mathbf{H}^T \left(\mathbf{H}\mathbf{H}^T + \frac{r_0^2}{m_0^2}\mathbf{I}_d \right)^{-1} \mathbf{H} \right)$.

3. Retour à l'étape 1 en remplaçant r_0 par r_1 et m_0 par m_1 .

Ce schéma converge dans ce contexte en 3-4 itérations.

TAB. 3.2: Estimation des hyper-paramètres et des activités totales rejetées de césium-137 dans le cadre de statistiques gaussiennes

paramètre	méthode	avec observations au Japon (104)	avec toutes les observations (267)
r (Bq m ⁻³)	L-curve + χ^2	1.05	1.03
	Schéma de Desroziers	}1.0	0.56
	Vraisemblance maximale		
m (Bq s ⁻¹)	L-curve + χ^2	1.5×10^{12}	1.5×10^{12}
	Schéma de Desroziers	} 1.5×10^{12}	2.2×10^{12}
	Vraisemblance maximale		
Activité rejetée (Bq)	L-curve + χ^2	2.8×10^{16}	6.9×10^{16}
	Schéma de Desroziers	} 2.8×10^{16}	2.1×10^{17}
	Vraisemblance maximale		

3.8.4.3 Estimation des incertitudes *a posteriori*

Dans le cas gaussien, la diagonale de \mathbf{P}_{BLUE} est utilisée pour estimer les incertitudes sur la source reconstruite, les corrélations des erreurs *a posteriori* (les termes non-diagonaux de \mathbf{P}_{BLUE}) étant très faibles par rapport aux termes diagonaux.

Dans le cas semi-gaussien, une simulation de Monte Carlo de 2×10^4 tirages est réalisée. À chaque tirage, les observations sont générées suivant la loi $\mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$ et l'ébauche suivant la loi $\mathcal{TN}(\mathbf{0}, \mathbf{B}, \mathbf{0})$. L'écart type estimé de l'ensemble des sources reconstruites représente l'incertitude de la source estimée sans perturbation.

3.8.5 Résultats

Les observations mesurées au Japon et hors du Japon sont très différentes par nature. Tout d'abord, les distances entre la source et les stations de mesures sont telles qu'il y a plusieurs ordres de grandeur entre les valeurs mesurées : par exemple pour le césium-137, les observations japonaises sont toutes dans l'intervalle $[10^{-1}, 10^2]$ Bq m⁻³ alors que toutes les observations hors du Japon sont dans l'intervalle $[10^{-6}, 10^{-3}]$ Bq m⁻³. De plus l'impact des erreurs modèles (vents et processus de perte) s'accumule le long du trajet du panache. Nous pouvons ajouter encore que les mesures effectuées hors du Japon sont généralement intégrées sur 24 h alors qu'elles sont représentatives de durées plus courtes au Japon (typiquement 1 h ou moins). Pour ces raisons, nous avons choisi de réaliser les inversions en utilisant les observations japonaises seules dans un premier temps (tout en étant conscients que la source estimée dans ce cas constitue une limite inférieure puisque le panache directement émis vers l'Océan Pacifique n'est pas pris en compte), puis toutes les observations afin d'étudier l'impact de ces différences.

Estimation du terme source de césium-137 - Statistiques gaussiennes

Les résultats concernant l'estimation des hyper-paramètres ainsi que l'estimation de l'activité totale rejetée sont indiqués dans le tableau 3.2. Dans le cadre de statistiques gaussiennes, les trois méthodes donnent des résultats très proches. En particulier, nous avons vérifié que le schéma de Desroziers converge vers la solution exacte de la MVM.

Le calcul exact de la vraisemblance pour un assez grand nombre de valeurs de r et de m permet de tracer la densité de probabilité de ces hyper-paramètres (Fig. 3.6), et également leurs probabilités marginales (Fig. 3.7 et Fig. 3.8). Il est alors possible de distinguer deux maxima locaux. Étant donné nos hyper-paramètres initiaux, le schéma de Desroziers a convergé vers le maximum global, mais avec un autre choix d'hyper-paramètres initiaux, il aurait pu converger vers l'autre maximum.

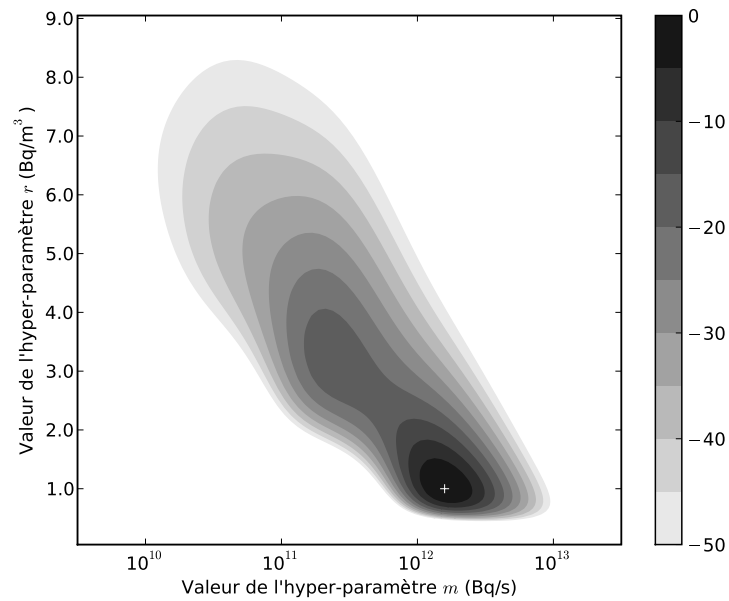


FIG. 3.6: Densité de probabilité des hyper-paramètres r et m estimée par le calcul de la vraisemblance marginale des observations de césium-137 dans le cas gaussien. La croix blanche localise le maximum de vraisemblance.

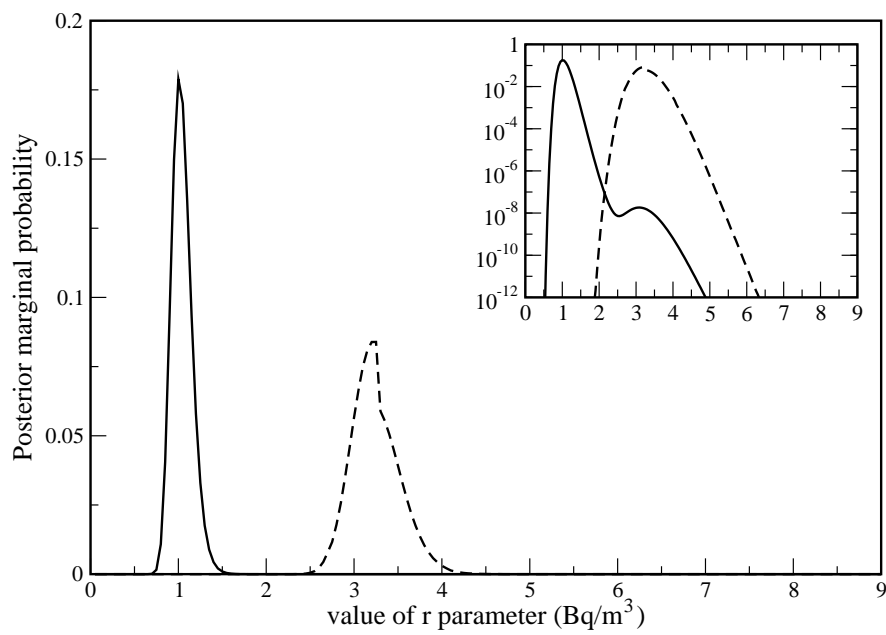


FIG. 3.7: Probabilité marginale de l'hyper-paramètre r . Ligne continue : dans le cas gaussien. Ligne pointillée : dans le cas semi-gaussien.

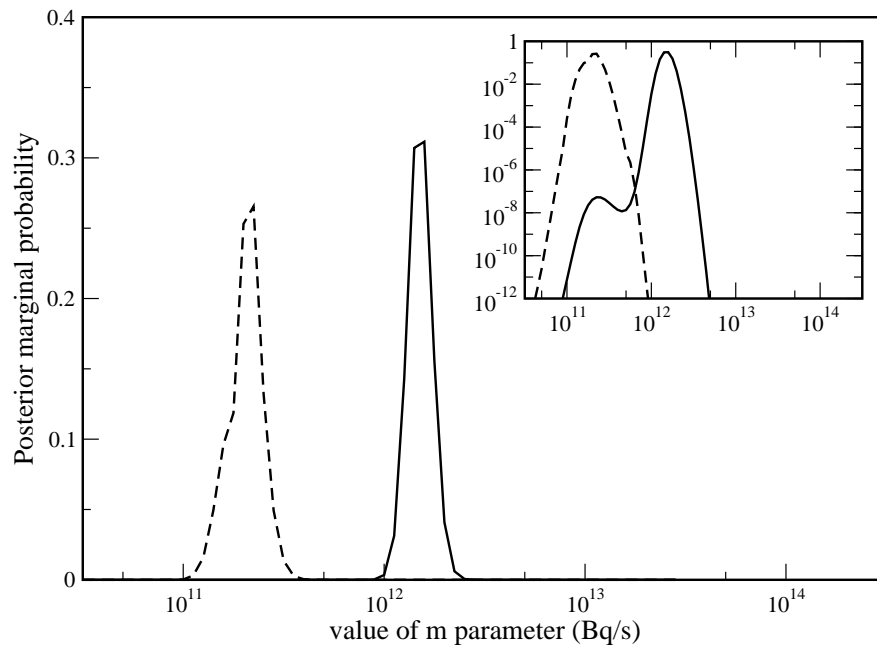


FIG. 3.8: Probabilité marginale de l'hyper-paramètre m . Ligne continue : dans le cas gaussien. Ligne pointillée : dans le cas semi-gaussien.

La L-curve pour cette situation est représentée sur la figure 3.9. Il est intéressant de noter que, notamment à cause du faible nombre d'observations, la L-curve est très arrondie, ce qui rend difficile l'identification du paramètre d'équilibre.

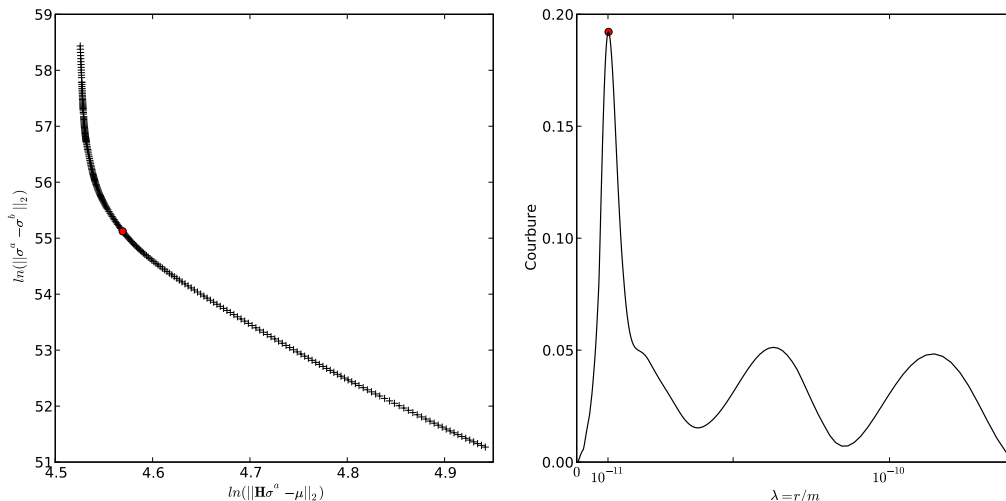


FIG. 3.9: À gauche : L-curve tracée pour l'inversion de la source de césium-137 en statistiques gaussiennes. À droite : courbure correspondante. Le point rouge signale le point d'équilibre (courbure maximale).

Le profil temporel de la source reconstruite est représenté sur la figure 3.10 (méthode MVM pour l'estimation des hyper-paramètres). Il est évident que le nombre d'observations n n'est pas suffisant pour contraindre la positivité de la source. Cela prouve que l'utilisation de statis-

tiques gaussiennes tronquées apporte une information importante au système dans ce cas. Il est néanmoins intéressant de noter que l'estimation d'activité totale rejetée (28 PBq en utilisant les observations japonaises uniquement) est cohérente avec d'autres estimations : 8.8 PBq pour Terada et al. [2012], 12 PBq pour Chino et al. [2011], 16 PBq pour Saunier et al. [2013], 19 PBq pour Mathieu et al. [2012] et 36 PBq pour Stohl et al. [2012]. Sur la figure 3.10, l'incertitude estimée par les termes diagonaux de la matrice de covariance des erreurs d'analyse \mathbf{P}_{BLUE} est également tracée. Les conditions météorologiques ont transporté la plupart du temps le panache radioactif directement vers l'Océan Pacifique. En conséquence, trois périodes d'observabilité, pendant lesquelles l'incertitude est réduite, sont visibles : la première s'étend approximativement du 14 au 15 mars, la seconde autour des 20-21 mars et la dernière, moins évidente, autour du 25 mars. Ces périodes correspondent en général à une orientation du vent vers le sud-ouest, où sont localisées les stations de mesures de Tokyo, de Tsukuba et JPP38.

Les observations hors-Japon, qui théoriquement devraient permettre d'observer le panache émis vers l'Océan Pacifique, notamment les stations situées aux États-Unis et au Canada, apportent peu d'information au système pour plusieurs raisons :

- La grande distance entre la centrale de Fukushima et les stations de mesures engendre des erreurs de modèle très grandes.
- Les observations hors-Japon sont toutes intégrées sur 24 h.
- La nature diffusive du transport atmosphérique lui-même [Enting, 2002] ne permet pas de faire remonter suffisamment d'information.

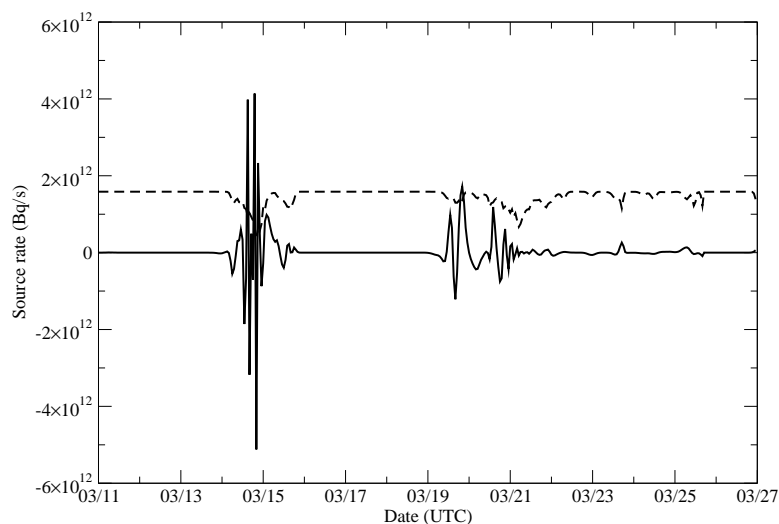


FIG. 3.10: Source reconstruite de césium-137 dans le cas de statistiques gaussiennes en utilisant la méthode de vraisemblance maximale pour l'estimation des hyper-paramètres. Ligne tiretée : incertitude *a posteriori* estimée par les termes diagonaux de \mathbf{P}_{BLUE} .

TAB. 3.3: Estimation des hyper-paramètres et des activités totales rejetées de césium-137 dans le cadre de statistiques semi-gaussiennes

paramètre	méthode	avec observations au Japon (104)	avec toutes les observations (267)
r (Bq m ⁻³)	L-curve + χ^2	4.55	2.88
	Schéma de Desroziers	5.41	2.96
	Vraisemblance maximale	3.25	1.7
m (Bq s ⁻¹)	L-curve + χ^2	3.2×10^{11}	2.0×10^{11}
	Schéma de Desroziers	5.3×10^{10}	1.3×10^{11}
	Vraisemblance maximale	2.0×10^{11}	3.5×10^{11}
Activité rejetée (Bq)	L-curve + χ^2	1.2×10^{16}	1.3×10^{16}
	Schéma de Desroziers	3.3×10^{15}	1.0×10^{16}
	Vraisemblance maximale	1.2×10^{16}	1.9×10^{16}

Estimation du terme source de césium-137 - Statistiques semi-gaussiennes

Les résultats pour l'estimation des hyper-paramètres et des activités totales rejetées sont indiqués dans le tableau 3.3. Les trois méthodes donnent des résultats légèrement différents, mais qui restent cohérents, en particulier quand le nombre d'observations augmente. Les probabilités marginales de r et de m sont également tracées sur les figures 3.7 et 3.8. Quand toutes les données disponibles sont utilisées, l'activité totale rejetée est estimée entre 10 PBq et 19 PBq. En utilisant les méthodes qui nous semblent les plus fiables dans ce cas (L-curve et MVM) et en assimilant uniquement les observations japonaises pour éviter de trop larges erreurs de modèle, nous obtenons une valeur de 12 PBq, ce qui est toujours en accord avec d'autres estimations indépendantes citées précédemment.

Le profil temporel de la source reconstruite de césium-137, en utilisant la MVM pour l'estimation des hyper-paramètres et les observations japonaises seulement, est tracé sur la figure 3.11. Son incertitude, estimée par une simulation de Monte Carlo, est également tracée. Les mêmes trois périodes d'observabilité que dans le cas gaussien apparaissent, légèrement plus larges.

Une fois le terme source estimé, il est possible de modéliser la dispersion du panache, notamment pour connaître les zones impactées par son passage. Sur la figure 3.12, l'évolution du panache réanalysé de césium-137 est représentée à quatre dates différentes entre le 15 et le 21 mars. Le terme source considéré est celui reconstruit en utilisant toutes les observations disponibles et la MVM pour estimer les hyper-paramètres. Sur la figure 3.13, l'évolution temporelle de la concentration d'activité simulée de césium-137 au niveau de la station d'observation de Tokyo-MITRI (qui est celle où nous disposons du plus grand nombre de données) est représentée en comparaison des observations effectivement relevées. Un très bon accord observations/simulations est visible. Les différences visibles, en particulier sur les pics des 21 et 22 mars, proviennent en général du fait que d'autres observations, non-représentées sur cette figure, apportent des informations différentes, parfois contradictoires, et que la reconstruction, et donc la réanalyse sont des compromis de toutes ces informations.

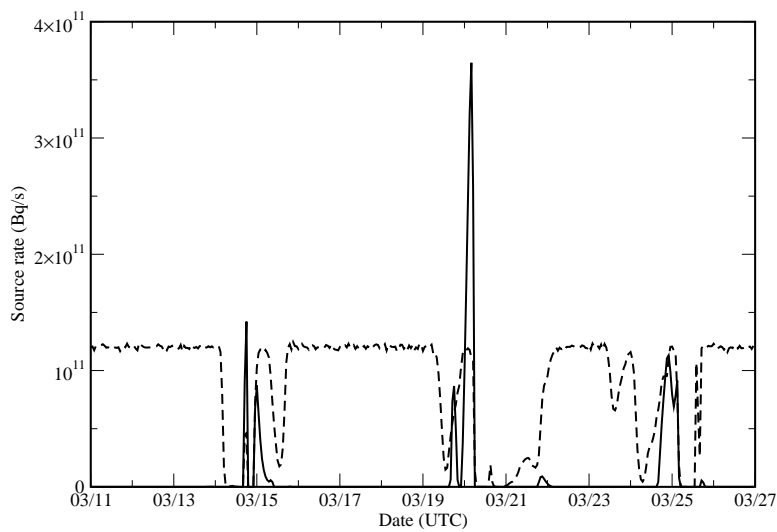


FIG. 3.11: Source reconstruite de césium-137 dans le cas de statistiques semi-gaussiennes en utilisant la méthode de vraisemblance maximale pour l'estimation des hyper-paramètres. Ligne tiretée : incertitude *a posteriori* estimée par une simulation de Monte Carlo.

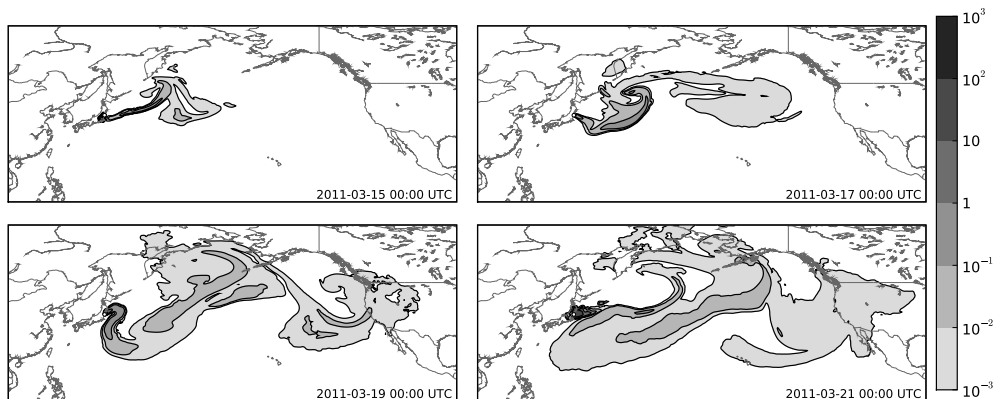


FIG. 3.12: Panache réanalysé de césium-137 - en $\text{Bq}\cdot\text{m}^{-3}$ (méthode de vraisemblance maximale en statistiques semi-gaussiennes pour l'estimation des hyper-paramètres).

Estimation du terme source d'iode-131 - Statistiques semi-gaussiennes

Étant donné les résultats obtenus pour le terme source de césium-137, seules les statistiques semi-gaussiennes sont considérées pour reconstruire le terme source d'iode-131. Les résultats concernant l'estimation des hyper-paramètres et l'activité totale rejetée sont indiqués dans le tableau 3.4. En utilisant les méthodes de la L-curve et la MVM, et les observations japonaises, nous obtenons une estimation pour l'activité totale entre 190 PBq et 380 PBq, ce qui est comparable, bien que légèrement supérieur, avec les estimations de l'Agence japonaise de sécurité nucléaire (NISA) : 150 PBq pour Chino et al. [2011], ainsi qu'avec d'autres estimations : 106 PBq pour Saunier et al. [2013], 124 PBq pour Terada et al. [2012] ou 197 PBq pour Mathieu et al. [2012].

Le profil temporel de la source reconstruite en utilisant seulement les observations japonaises est visible sur la figure 3.14 (MVM pour l'estimation des hyper-paramètres). L'incertitude

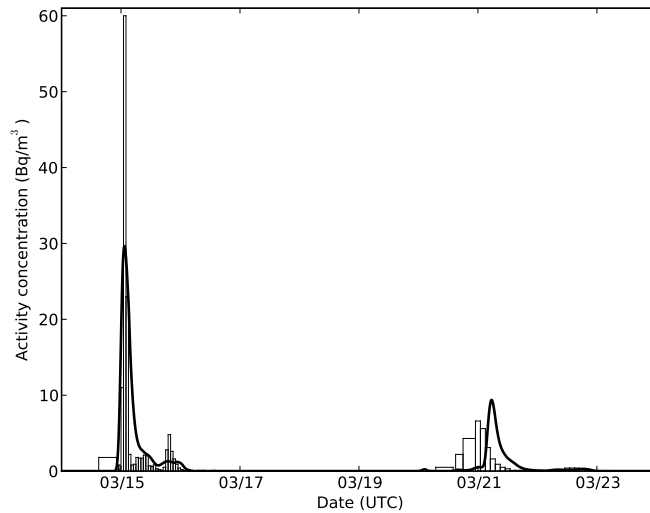


FIG. 3.13: Comparaison entre les observations (créneaux) et la concentration d'activité simulée pour le césium-137, en utilisant une source reconstruite par modélisation inverse (MVM pour les hyper-paramètres).

titude est également tracée sur cette figure. Comparée à la situation précédente pour ^{137}Cs , l'observabilité entre le 19 et le 26 mars est améliorée, certainement du fait d'un plus grand nombre d'observations, notamment de la station de Tokyo-MITRI.

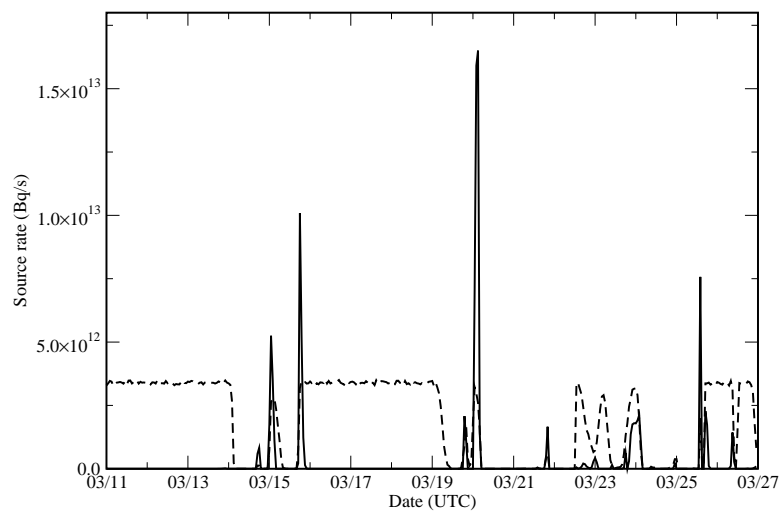


FIG. 3.14: Source reconstruite d'iode-131 dans le cas de statistiques semi-gaussiennes en utilisant la méthode de vraisemblance maximale pour l'estimation des hyper-paramètres. Ligne tiretée : incertitude *a posteriori* estimée par une simulation de Monte Carlo.

La source reconstruite en utilisant toutes les données disponibles a été utilisée pour simuler le transport atmosphérique d'iode-131. L'évolution du panache, à quatre dates différentes entre le 15 et le 21 mars, est représentée sur la figure 3.15 et la comparaison entre les observations à Tokyo-MITRI et les concentrations simulées sur la figure 3.16. Un assez bon accord est visible

TAB. 3.4: Estimation des hyper-paramètres et des activités totales rejetées d'iode-131 dans le cadre de statistiques semi-gaussiennes

paramètre	méthode	avec observations au Japon (233)	avec toutes les observations (428)
r (Bq m^{-3})	L-curve + χ^2	14.0	10.5
	Schéma de Desroziers	18.4	10.6
	Vraisemblance maximale	5.60	4.01
m (Bq s^{-1})	χ^2 + L-curve	2.7×10^{12}	2.0×10^{12}
	Schéma de Desroziers	2.0×10^{11}	1.9×10^{12}
	Vraisemblance maximale	5.6×10^{12}	7.1×10^{12}
Activité rejetée (Bq)	χ^2 + L-curve	1.9×10^{17}	2.0×10^{17}
	Schéma de Desroziers	1.6×10^{16}	1.9×10^{17}
	Vraisemblance maximale	3.8×10^{17}	7.0×10^{17}

entre la réanalyse et les observations.

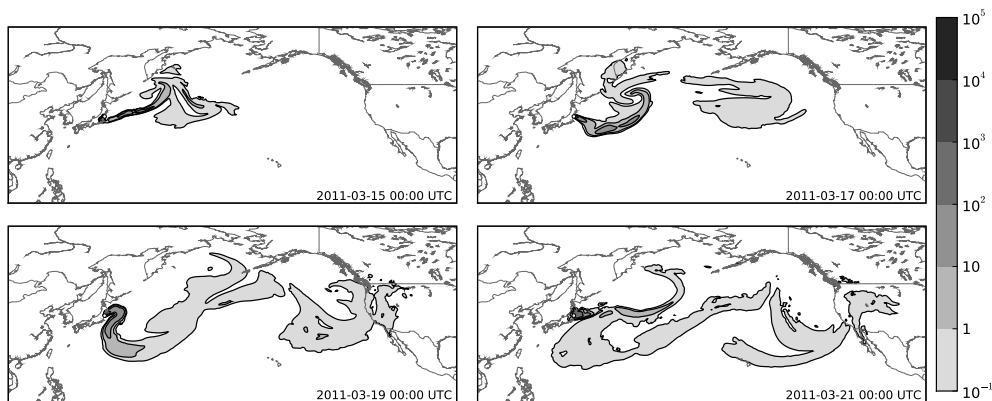


FIG. 3.15: Panache réanalysé d'iode-131 - en Bq.m^{-3} (méthode de vraisemblance maximale en statistiques semi-gaussiennes pour l'estimation des hyper-paramètres).

3.9 Application III : Fukushima (2)

Les premières estimations de termes sources consécutifs à l'accident de la centrale nucléaire de Fukushima Daiichi ont mis en lumière le manque général d'observations. Dans le même temps, il est apparu crucial d'être capable d'assimiler toutes les observations disponibles simultanément. Les méthodes proposées dans ce chapitre permettent, en estimant rigoureusement les erreurs *a priori*, d'équilibrer le poids respectif des différents types de données utilisées. L'étude présentée dans cette section a fait l'objet d'un article publié dans Atmospheric Environment [Winiarek et al., 2014], visible en annexe F. L'objectif est l'estimation du terme source de ^{137}Cs en utilisant simultanément des mesures de concentration d'activité dans l'air et des mesures de dépôt.

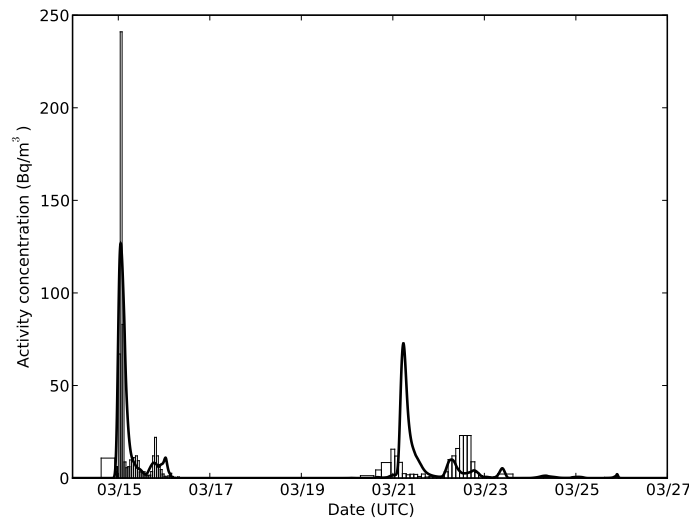


FIG. 3.16: Comparaison entre les observations (créneaux) et la concentration d’activité simulée pour l’iode-131, en utilisant une source reconstruite par modélisation inverse (MVM pour les hyper-paramètres).

3.9.1 Observations

Nous avons appliqué les méthodes présentées précédemment pour estimer le terme source de césium-137 en utilisant $N_d = 3$ jeux de données :

- Les concentrations d’activité de ^{137}Cs dans l’air mesurées au Japon, présentées dans la section 3.8.2. Ce jeu de données (noté DATA1 par la suite) contient $d_1 = 104$ observations.
- Des mesures journalières de dépôt de ^{137}Cs réalisées dans 22 préfectures¹ à partir du 18 mars 2011. Ce jeu de données (noté DATA2) contient $d_2 = 198$ observations.
- Des mesures de dépôt cumulé de ^{137}Cs réalisées dans une zone proche de la centrale (environ 100 km maximum) plusieurs semaines après l’accident. Parmi ces 2180 mesures réalisées par MEXT², 16 ont été retirées car étant situées trop près de la centrale et donc impactées par des effets de champ proche non pris en compte par le modèle de dispersion atmosphérique à méso-échelle. Ce troisième jeu de données “brutes” (noté DATA3a par la suite) contient donc $d_3 = 2164$ observations. Ces observations, représentées sur la figure 3.19(a), sont densément distribuées en espace ; à l’opposé, l’information en temps est intégrée sur une longue période et est donc difficile à exploiter.

¹<http://www.mext.go.jp/english/incident/1305529.htm>

²http://www.mext.go.jp/b_menu/shingi/chousa/gijyutu/017/shiryo/___icsFiles/afieldfile/2011/09/02/1310688_1.pdf

TAB. 3.5: Configuration et paramétrisation des processus physiques du modèle WRF. Un nesting two-way est utilisé entre les domaines 1 et 2.

Domaine	1	2
Résolution spatiale	18 km	6 km
Nombre de points de grille	340 × 250	241 × 241
Nombre de niveaux verticaux	27	27
Pas de temps numérique	60 s	20 s
Pas de temps de sauvegarde	3600 s	3600 s
Planetary boundary layer	Yonsei University	Yonsei University
Micro-physics	Kessler	WRF Single Moment 3
Cumulus physics	Grell-Devenyi	Grell-Devenyi
Longwave radiation	RRTM	RRTM
Shortwave radiation	Dudhia	Dudhia
Surface layer	MM5 similarity	MM5 similarity
Land surface	Noah LSM	Noah LSM
Nudging	Grid nudging	Grid nudging

3.9.2 Modèle

La configuration des observations disponibles nous a amenés à considérer un domaine méso-échelle, plus précis que dans la section 3.8, mais couvrant “seulement” le Japon.

3.9.2.1 Champs météorologiques

Les champs fournis par le CEPMMT ont une résolution spatiale de $0.25^\circ \times 0.25^\circ$ et temporelle de 3 h, ce qui n’est pas suffisamment précis pour cette application. Nous avons donc généré des champs météorologiques à l’aide du modèle numérique WRF (Weather Research and Forecasting, Skamarock et al. [2008]). L’objectif principal est d’obtenir des champs d’une résolution spatiale d’environ 5 km et temporelle d’environ 1 h. Ces champs météorologiques seront fournis en entrée du modèle de dispersion atmosphérique. Les paramétrisations des processus physiques et les caractéristiques des domaines de simulation sont résumés dans le tableau 3.5 et les domaines de simulation sont représentés sur la figure 3.17. Une caractéristique essentielle de ces simulations WRF est l’utilisation de milliers d’observations météorologiques pour contraindre les champs grâce à des techniques de nudging [Stauffer and Seaman, 1994].

3.9.2.2 Modèle de dispersion atmosphérique

Le modèle POLAIRD3D est à nouveau utilisé pour modéliser la dispersion atmosphérique du césium-137. Rappelons que le modèle calcule la concentration d’activité c de ^{137}Cs en résolvant l’équation :

$$\frac{\partial c}{\partial t} + \text{div}(\mathbf{u}c) = \text{div}\left(\rho \mathbf{K} \nabla \left(\frac{c}{\rho}\right)\right) - \Lambda^s c - \Lambda^d c + \sigma. \quad (3.72)$$

\mathbf{u} représente le champ de vent, \mathbf{K} la matrice de diffusion turbulente, considérée diagonale. La composante verticale K_z est calculée suivant la paramétrisation de Louis [1979]. La compo-

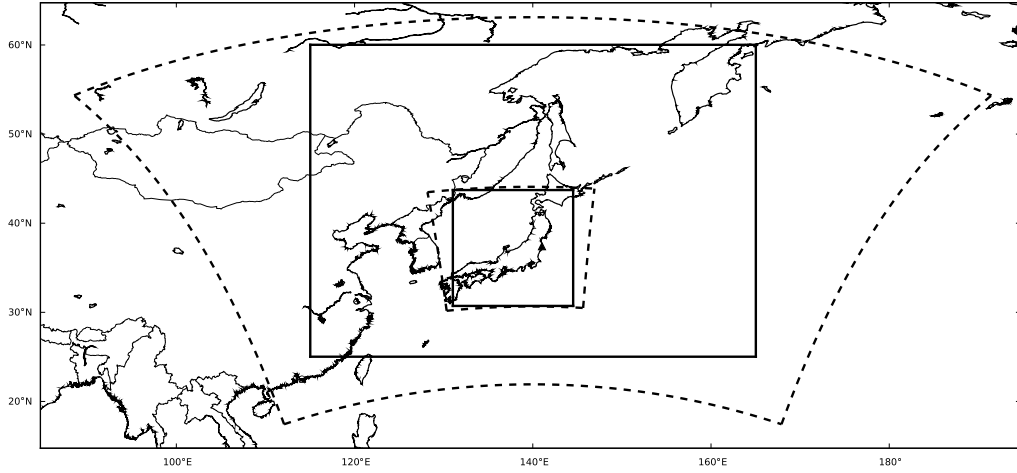


FIG. 3.17: Domaines de simulation. Ligne pointillée : simulation WRF (nesting two-way). Ligne continue : simulation POLAIR3D (nesting one-way)

sante horizontale K_H est prise constante et égale à $1000 \text{ m}^2 \text{ s}^{-1}$ (la résolution du modèle ayant été multipliée par 25 par rapport au modèle de la section 3.8, de $0.25^\circ \times 0.25^\circ$ à $0.05^\circ \times 0.05^\circ$, la valeur de K_H est divisée par 25). Λ^s est le coefficient de lessivage, Λ^d représente la décroissance radioactive et σ la source du radionucléide. La condition à la limite au niveau du sol est :

$$K_z \nabla c \cdot \mathbf{n} = -v^d c, \quad (3.73)$$

où \mathbf{n} est le vecteur unitaire orienté positif vers l'intérieur du domaine et v^d est la vitesse de dépôt sec du ^{137}Cs .

Deux domaines imbriqués de simulation sont considérés. Le plus petit, qui présente la résolution la plus fine, couvre approximativement le Japon, de 131.03°E à 144.53°E et de 30.72°N à 43.72°N avec une résolution spatiale de $0.05^\circ \times 0.05^\circ$. Le nombre de points de grille dans ce domaine est de 270×260 . Étant donnée la taille de ce domaine, il y a un risque important de recirculation de polluant qui pourrait revenir dans le domaine par les bords de celui-ci. Pour pouvoir prendre en compte ce phénomène, il est nécessaire d'utiliser un domaine plus large, à la résolution plus grossière, destiné uniquement au calcul des conditions aux limites du domaine plus fin, par une technique de nesting (one-way). Ce domaine couvre une région s'étendant de 115.03°E à 165.03°E et de 25.02°N à 60.02°N avec une résolution de $0.25^\circ \times 0.25^\circ$. Cette configuration est représentée sur la figure 3.17. Pour les deux domaines, le modèle POLAIR3D est configuré avec 15 niveaux verticaux s'étendant de 0 à 8000 m.

Le césium-137 est modélisé comme une espèce gazeuse passive, avec une demi-vie de 11000 jours. Le dépôt sec est modélisé en utilisant un schéma simple définie par deux vitesses de dépôt constantes : $v^d = 0.01 \text{ cm.s}^{-1}$ au-dessus de l'océan [Estournel et al., 2012] et $v^d = 0.15 \text{ cm.s}^{-1}$ partout ailleurs [Bocquet, 2012]. En ce qui concerne le lessivage, nous avons opté pour un schéma de type *below-cloud* :

$$\Lambda^s = a \left(\frac{p}{p_0} \right)^b, \quad (3.74)$$

où p représente le taux de précipitation (calculé par la simulation WRF) et $p_0 = 1 \text{ mm.h}^{-1}$, a et b étant deux constantes respectivement égales à $8.4 \cdot 10^{-5} \text{ s}^{-1}$ et 0.79 [Maryon et al., 1991].

Ce schéma est préféré à un schéma sous-maille basé sur l'humidité relative, utilisé dans la section 3.8, car ce type de schéma n'a été validé qu'à des résolutions plus grossières [Pudykiewicz, 1989].

3.9.3 Modélisation inverse

3.9.3.1 Configuration

Entre nos travaux présentés dans la section 3.8 et ceux présentés ici, de nouvelles estimations ont mis en évidence de possibles émissions après le 27 mars, autour du 30 [Katata et al., 2012]. Pour cette raison, nous avons étendu notre fenêtre d'inversion jusqu'à la fin du mois de mars 2011. Le problème inverse consiste donc à retrouver un profil temporel, de résolution 1h, entre le 11 mars 2011 et le 31 mars 2011 inclus, ce qui représente $N = 504$ débits massiques à estimer.

La matrice Jacobienne \mathbf{H} est calculée colonne par colonne en réalisant N simulations directes et, comme dans la section 3.8, une ébauche nulle est choisie ($\boldsymbol{\sigma}^b = \mathbf{0}$).

Seules les statistiques gaussiennes tronquées seront considérées, l'estimation du profil temporel des émissions $\boldsymbol{\sigma}$ est donc réalisée par la minimisation sous contrainte de positivité de la fonction coût :

$$\mathcal{J}_q(\boldsymbol{\sigma}) = \frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) + \frac{1}{2} \boldsymbol{\sigma}^T \mathbf{B}^{-1} \boldsymbol{\sigma}. \quad (3.75)$$

3.9.3.2 Estimation des erreurs *a priori*

Étant données les données utilisées pour l'inversion, un modèle à $N_d + 1 = 4$ hyper-paramètres est choisi pour caractériser les erreurs *a priori*. Ainsi les matrices de covariance d'erreur \mathbf{R} et \mathbf{B} sont définies par :

$$\mathbf{B} = m^2 \mathbf{I}_N, \quad \mathbf{R}_i = r_i^2 \mathbf{I}_{d_i}, \quad (3.76)$$

où \mathbf{R}_i représente le bloc diagonal de \mathbf{R} relatif au jeu de données i . m^2 est la variance des erreurs d'ébauche et r_i^2 la variance des erreurs d'observation du jeu de données i .

Deux méthodes présentées dans le Chapitre 3.6 sont utilisées pour estimer les variances des erreurs : la méthode de vraisemblance maximale (MVM) et le schéma de Desroziers. La MVM repose sur le calcul exact de la vraisemblance marginale et sera considérée comme la méthode de référence. Le schéma de Desroziers, qui est une approximation dans le cadre de statistiques semi-gaussiennes, présente l'avantage de la rapidité et il est particulièrement intéressant d'évaluer son efficacité, notamment quand le nombre d'observations augmente.

Maximisation de la vraisemblance

L'estimation des variances d'erreurs est réalisée par la maximisation de la vraisemblance marginale :

$$p(\boldsymbol{\mu}|\boldsymbol{\theta}) = \frac{e^{-\frac{1}{2} \boldsymbol{\mu}^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} \boldsymbol{\mu}}}{\sqrt{(2\pi)^d |\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}|}} \times \int \frac{e^{-\frac{1}{2} (\boldsymbol{\sigma} - \boldsymbol{\sigma}^{\text{BLUE}})^T \mathbf{P}_{\text{BLUE}}^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}^{\text{BLUE}})}}{\sqrt{(\pi/2)^N |\mathbf{P}_{\text{BLUE}}|}} \mathbb{I}_{\boldsymbol{\sigma} \geq \mathbf{0}} d\boldsymbol{\sigma}, \quad (3.77)$$

avec

$$\begin{aligned}\boldsymbol{\sigma}^{\text{BLUE}} &= \mathbf{B}\mathbf{H}^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} \boldsymbol{\mu}, \\ \mathbf{P}_{\text{BLUE}} &= \mathbf{B} - \mathbf{B}\mathbf{H}^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} \mathbf{H}\mathbf{B}.\end{aligned}\quad (3.78)$$

Pour estimer l'intégrale, l'algorithme de GHK est utilisé (cf. Annexe E).

Schéma de Desroziers

Le schéma de Desroziers suivant est utilisé pour estimer dans le même temps les hyper-paramètres et la source :

1. Si \mathbf{R}_0 et \mathbf{B}_0 sont les estimations à ce stade des matrices de covariance des erreurs, avec $\mathbf{R}_{0i} = r_{0i}^2 \mathbf{I}_{d_i}$ le bloc diagonal de \mathbf{R}_0 relatif au jeu de données i et $\mathbf{B}_0 = m_0^2 \mathbf{I}_N$, l'estimateur $\boldsymbol{\sigma}^a = \underset{\boldsymbol{\sigma} \geq \mathbf{0}}{\operatorname{argmin}} \mathcal{J}_q$ est calculé avec :

$$\mathcal{J}_q(\boldsymbol{\sigma}) = \frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}_0^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) + \frac{1}{2} \boldsymbol{\sigma}^T \mathbf{B}_0^{-1} \boldsymbol{\sigma}. \quad (3.79)$$

2. De nouvelles variances d'erreur $(r_{1i}^2)_{(1 \leq i \leq N_d)}$ et m_1^2 sont estimées :

$$\begin{aligned}r_{1i}^2 &= \frac{(\boldsymbol{\mu}_i - \mathbf{H}_i \boldsymbol{\sigma}^a)^T (\boldsymbol{\mu}_i - \mathbf{H}_i \boldsymbol{\sigma}^a)}{r_{0i}^2 d_i - \operatorname{tr}(\mathbf{H}_i \mathbf{P}_{\text{BLUE}} \mathbf{H}_i^T)} r_{0i}^2, \\ m_1^2 &= \frac{\boldsymbol{\sigma}^{aT} \boldsymbol{\sigma}^a}{m_0^2 N - \operatorname{tr}(\mathbf{P}_{\text{BLUE}})} m_0^2.\end{aligned}\quad (3.80)$$

avec $\mathbf{P}_{\text{BLUE}} = \mathbf{B}_0 - \mathbf{B}_0 \mathbf{H}^T (\mathbf{H} \mathbf{B}_0 \mathbf{H}^T + \mathbf{R}_0)^{-1} \mathbf{H} \mathbf{B}_0$. $\boldsymbol{\mu}_i$ représente le vecteur, de dimension d_i , des observations du jeu de données i et \mathbf{H}_i est la sous-matrice de \mathbf{H} relative à ce jeu de données.

3. Retour à l'étape 1 en remplaçant r_{0i} par r_{1i} , pour $1 \leq i \leq N_d$, et m_0 par m_1 .

3.9.3.3 Estimation des incertitudes *a posteriori*

Une simulation de Monte Carlo de 2×10^4 tirages est réalisée. À chaque tirage, les observations sont générées suivant la loi $\mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$ et l'ébauche suivant la loi $\mathcal{JN}(\mathbf{0}, \mathbf{B}, \mathbf{0})$, définies par les équations 3.9 et 3.22. L'écart type estimé de l'ensemble des sources reconstruites représente l'incertitude de la source estimée sans perturbation.

3.9.4 Résultats

3.9.4.1 Avec les concentrations d'activité dans l'air seulement

Les formules présentées dans la section 3.9.3.2 sont valables pour le cas particulier $N_d = 1$. Dans ce cas, elles sont strictement équivalentes à celles utilisées dans la section 3.8. Nous avons donc voulu réaliser dans un premier temps l'inversion en utilisant uniquement les observations de DATA1. L'objectif est de vérifier que les résultats sont cohérents avec notre précédente estimation, bien que les champs météorologiques et le modèle de dispersion atmosphérique soient différents.

Les hyper-paramètres estimés, ainsi que les estimations d'activité totale rejetée, sont reportés dans le tableau 3.6. L'écart entre le schéma de Desroziers, qui est une approximation dans

TAB. 3.6: Estimation des hyper-paramètres et de l'activité totale rejetée de ^{137}Cs correspondante en utilisant uniquement les concentrations d'activité dans l'air (DATA1).

paramètre	méthode	Modèle régional Section 3.8	Modèle méso-échelle
r (Bq m^{-3})	Schéma de Desroziers	5.4	2.1
	Vraisemblance maximale	3.3	1.9
m (Bq s^{-1})	Schéma de Desroziers	5.3×10^{10}	8.9×10^{10}
	Vraisemblance maximale	2.0×10^{11}	1.6×10^{11}
Activité rejetée (Bq)	Schéma de Desroziers	3.3×10^{15}	7.2×10^{15}
	Vraisemblance maximale	1.2×10^{16}	1.1×10^{16}

le cas de statistiques semi-gaussiennes, et la MVM, qui repose sur le calcul exacte de la vraisemblance marginale des observations, est plus faible ici que dans la section 3.8. L'utilisation d'un modèle méso-échelle avec une meilleure résolution spatiale et temporelle améliore certainement la cohérence générale dans le système. En particulier, les plus faibles valeurs estimées pour l'hyper-paramètre r , qui représente l'amplitude des erreurs d'observation notamment les erreurs de représentativité et de modèle, étaient attendues.

Étant donné que seules les observations japonaises sont utilisées, le terme source reconstruit représente une "limite inférieure" de la vraie source, certaines émissions, immédiatement portées vers l'Océan Pacifique par le vent, n'étant pas observables dans cette configuration. Néanmoins, avec la MVM, l'estimation de l'activité totale rejetée est de 11 PBq, ce qui est cohérent avec d'autres estimations : 12 PBq dans la section 3.8 [Winiarek et al., 2012] et pour Chino et al. [2011], 8.8 PBq pour Terada et al. [2012], 16 PBq pour Saunier et al. [2013], 21 PBq pour Mathieu et al. [2012] ou 36 PBq pour Stohl et al. [2012].

La source reconstruite et son incertitude sont représentées sur la figure 3.18(a)³. L'incertitude sur l'activité totale rejetée est d'environ 65%. Les fenêtres d'observabilité sont visibles dans la forme de l'incertitude. Celle-ci diminue durant trois périodes particulières : autour des 14 et 15 mars, puis entre le 19 et le 22 mars et enfin entre le 24 et le 26 mars.

En ce qui concerne la cinétique des émissions reconstruites, nous pouvons nous référer à d'autres travaux d'estimation du terme source, en particulier à Mathieu et al. [2012] qui, en plus des premières observations disponibles sur le Japon, ont également utilisé des observations in-situ (dose équivalente de rayonnement γ par exemple) et des mesures dans les réacteurs (température, pression). Du fait de l'utilisation d'une ébauche nulle et de l'observabilité imparfaite du fait du faible nombre d'observations et de la configuration géographique de l'accident, le terme source estimé dans cette section ne reconstruit pas quelques événements évoqués par ces études : les premières explosions d'hydrogène dans l'unité 1 le 12 mars, les rejets volontaires (pour abaisser la pression à l'intérieur des bâtiments) sur l'unité 3 le 13 mars et les événements (pas nécessairement identifiés) concernant les unités 2 et 3 les 16 et 18 mars. À l'opposé, les multiples événements des 14 et 15 mars sont présents dans la source reconstruite. Néanmoins cette reconstruction est certainement incomplète car les dernières émissions, probablement entre 7 h00 UTC et 12 h00 UTC, ne sont pas présentes. Ces émissions ont fortement contribué à l'empreinte Nord-Ouest visible sur la carte des dépôts sur la figure 3.19(a) et absente sur notre réanalyse (figure 3.19(b)), car aucune observation de concentration d'activité dans l'air n'était disponible dans cette zone pour permettre de reconstruire ces émissions.

³Une version en échelle logarithmique est visible dans l'annexe F.

Des émissions les 20, 21 et 23 mars sont également reconstruites (plus visibles sur le tracé en échelle logarithmique). Les émissions retrouvées le 19 mars ne sont pas mentionnées dans Mathieu et al. [2012], mais semblent compatibles avec les mesures in-situ de dose équivalente de rayonnement γ communiquées par l'opérateur TEPCO⁴, notamment sur la face Nord du bureau principal ou près du portail Ouest de la centrale. Ces émissions sont également retrouvées par Stohl et al. [2012] et correspondent aux heures précédant les tentatives de refroidissement par largage d'eau en utilisant les moyens mis à disposition par les Services d'intervention incendie de Tokyo. Finalement, des émissions le 30 mars sont également reconstruites en accord avec Terada et al. [2012] et Katata et al. [2012].

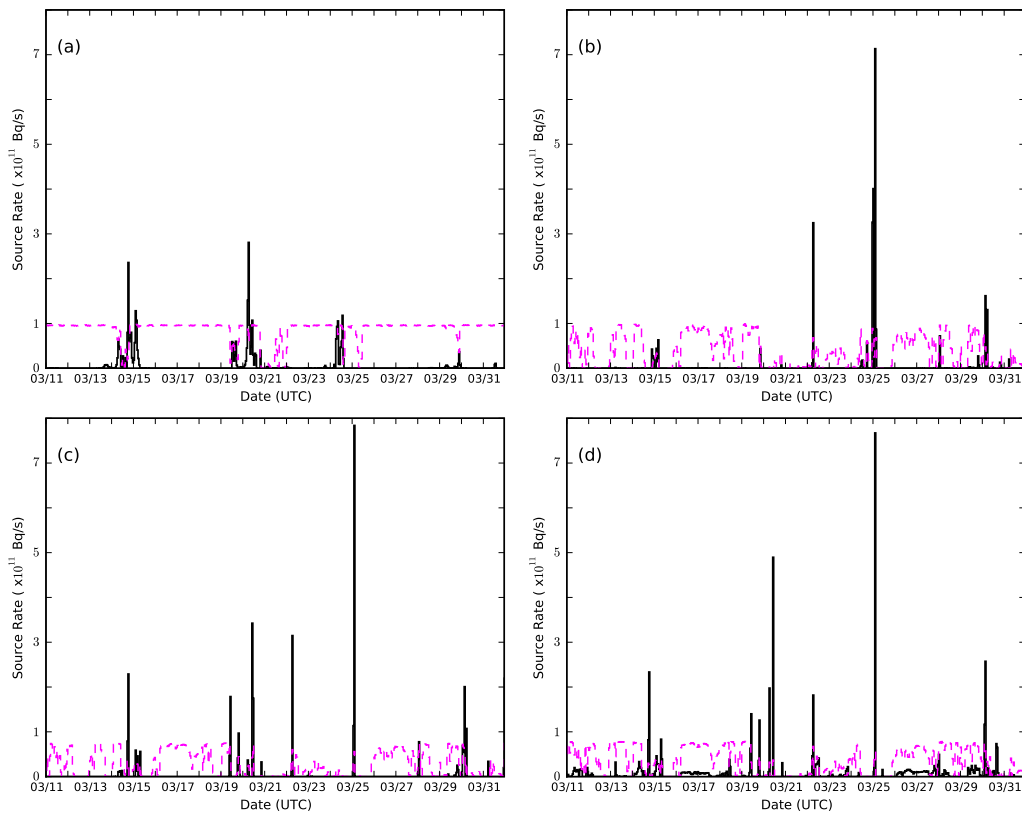


FIG. 3.18: Source reconstruites (lignes pleines) et leurs incertitudes *a posteriori* (lignes tiretées). (a) Avec les concentrations d'activité dans l'air seules (DATA1). (b) Avec les dépôts cumulés seuls (DATA3a). (c) Avec les 3 jeux de données brutes (DATA1, DATA2 et DATA3a). (d) Avec 3 jeux de données, dont les super-observations pour le troisième jeu de données (DATA1, DATA2 et DATA3b). Les mêmes figures en échelle logarithmique apparaissent dans l'annexe F.

Une fois la source estimée, celle-ci est utilisée pour calculer les observations réanalysées par le système. Les nuages de points réanalyses/observations (y compris les observations de DATA2 et DATA3a non utilisées pour l'inversion) sont représentés sur la figure 3.20(a). Il est intéressant de noter que les réanalyses des observations non utilisées pour l'inversion (mesures de dépôt) ne présentent pas de biais par rapport aux observations, ce qui montre la cohérence générale du modèle, en particulier concernant les processus de perte (dépôts sec et humide).

⁴<http://www.tepco.co.jp/en/nu/monitoring/index-e.html>

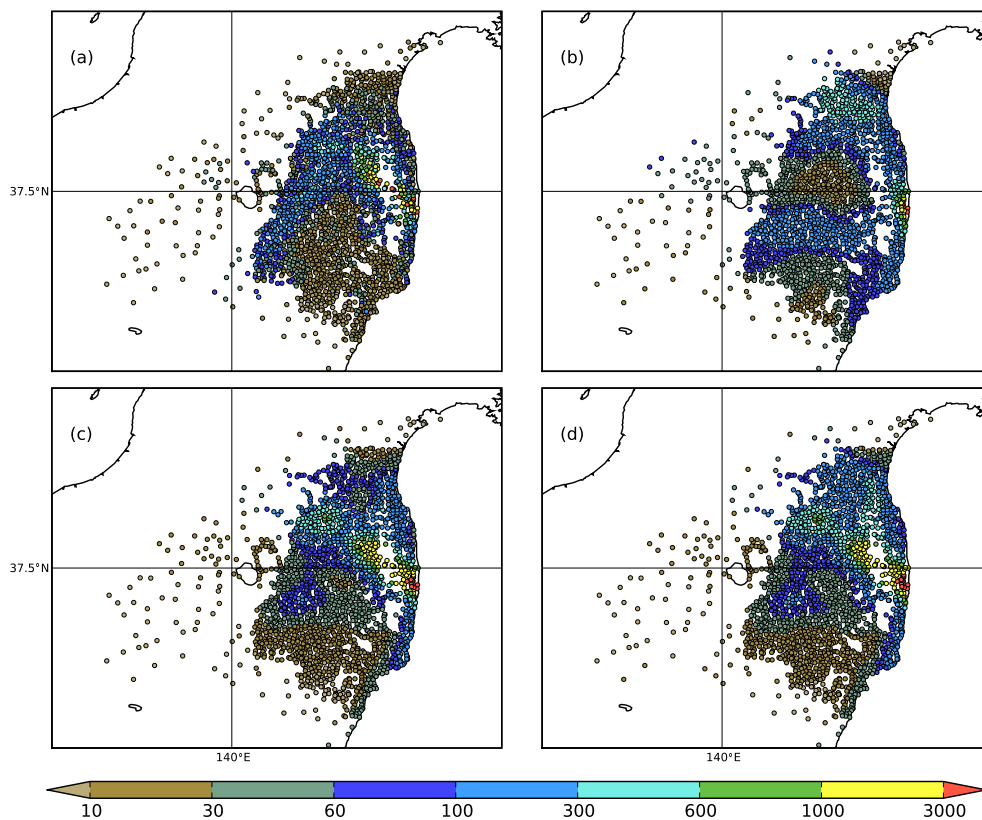


FIG. 3.19: Cartes de dépôt (en $\text{kBq}\cdot\text{m}^{-2}$) observée et réanalysées. (a) Observations. (b) Ré-analyses avec la source reconstruite en utilisant les concentrations d'activité dans l'air seules (DATA1). (c) Avec la source reconstruite en utilisant les 3 jeux de données brutes (DATA1, DATA2 et DATA3a). (d) Avec la source reconstruite en utilisant 3 jeux de données, dont les super-observations pour le troisième jeu de données (DATA1, DATA2 et DATA3b).

3.9.4.2 Avec les données brutes ($N_d = 3$)

Les observations de DATA3a sont des mesures réalisées plusieurs semaines après l'accident, en majorité aux mois de mai et juin 2011. Elles représentent le dépôt cumulé sur toute la période de l'accident et l'information en temps, intégrée, est donc très difficile à obtenir. Si seules ces observations sont utilisées pour l'inversion, l'activité totale estimée est cohérente (entre 11 et 12 PBq), mais le profil temporel, visible sur la figure 3.18(b), n'est pas conforme à ce qui est connu de l'accident. Seules des émissions le 25, et dans une moindre mesure les 23 et 30 avril, sont clairement visibles. De fait, le système a trop de liberté pour se conformer aux données, étant seulement contraint par le terme d'ébauche qui, dans notre cas, apporte seulement une information sur l'amplitude maximale des pics d'émission. C'est pourquoi il apparaît nécessaire d'utiliser ces données, certes abondantes, avec les autres données disponibles qui, bien que moins nombreuses, contiennent de l'information temporelle (DATA1 et DATA2). Les méthodes que nous proposons, en estimant simultanément les erreurs *a priori* de chaque jeu de données, doivent permettre d'équilibrer l'information dans le système.

Les hyper-paramètres et les activités totales estimés sont indiqués dans le tableau 3.7. Comme attendu, du fait du plus grand nombre d'observations dans le système, le schéma de Desroziers donne des résultats proches de ceux donnés par la MVM, que ce soit au niveau de

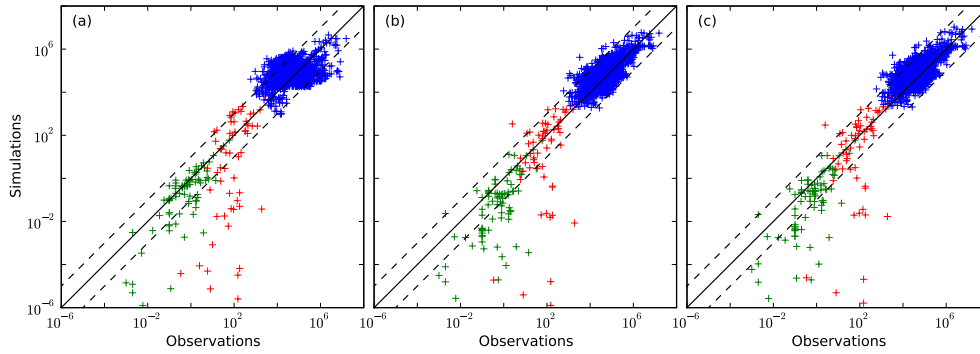


FIG. 3.20: Scatter plot des données réanalysées. Croix vertes : concentration d’activité dans l’air (en Bq.m^{-3}). Croix rouges : dépôt journalier (en Bq.m^{-2}). Croix bleues : dépôts cumulés (en Bq.m^{-2}). (a) Avec la source reconstruite en utilisant les concentrations d’activité dans l’air seules (DATA1). (b) Avec la source reconstruite en utilisant les 3 jeux de données brutes (DATA1, DATA2 et DATA3a). (c) Avec la source reconstruite en utilisant 3 jeux de données, dont les super-observations pour le troisième jeu de données (DATA1, DATA2 et DATA3b).

l’estimation des hyper-paramètres ou de l’activité totale rejetée : entre 12 et 13 PBq.

Le profil temporel de la source reconstruite est visible sur la figure 3.18(c). L’observabilité de l’accident est améliorée : les fenêtres temporelles où l’incertitude diminue sont plus larges, en particulier avant le 14 mars et entre le 19 et le 26 mars. L’incertitude sur l’activité totale estimée a été réduite de 65% en utilisant seulement DATA1 pour l’inversion à 25% dans cette inversion avec DATA1, DATA2 et DATA3a.

Les événements qui étaient retrouvés dans la première estimation le sont toujours. De plus, le système a reconstruit des émissions plus tardives le 15 mars, entre 7h00 UTC et 9h00 UTC, ce qui lui permet, sur des réanalyses (figure 3.19(c)), de modéliser un lobe Nord-Ouest de contamination, même si ce lobe n’est pas uniquement la conséquence de cette contribution dans notre système.

3.9.4.3 Avec des super-observations ($N_d = 3$)

Comme nous l’avons mentionné, les mesures de DATA3a sont très densément distribuées. Cette proximité induit certainement des corrélations dans les erreurs (principalement dans les erreurs de modèle) qui ne sont pas complètement prises en compte dans notre système.

Pour prendre en compte les corrélations dans les erreurs d’observation, la définition d’une matrice \mathbf{R} non-diagonale serait envisageable. Ce choix augmenterait d’une part le coût de l’algorithme puisque l’inverse de la matrice \mathbf{R} est nécessaire pour minimiser la fonction coût \mathcal{J}_q . D’autre part, de nouveaux hyper-paramètres, caractérisant les corrélations, devraient être estimés, ce qui pourrait se révéler difficile étant donné le faible nombre d’observations.

Une autre possibilité pour traiter des données densément distribuées est le “thinning” ou *sélection optimale* des données. De telles techniques sont déjà utilisées pour l’assimilation de données satellitaires en prévision météorologique [Liu and Rabier, 2002].

Une autre méthode consiste à créer des “super-observations” en moyennant toutes les observations dans un voisinage, par exemple dans une même maille du modèle de dispersion. Nous avons opté pour cette méthode : à partir des 2180 données brutes de DATA3a, nous avons créé 523 super-observations parmi lesquelles 4 ont été retirées car situées trop près de la centrale nucléaire. Les 519 observations restantes sont alors utilisées comme troisième jeu de

TAB. 3.7: Estimation des hyper-paramètres et des activités totales rejetées de ^{137}Cs en utilisant les concentrations d'activité dans l'air (DATA1), des mesures journalières de dépôt (DATA2) et des mesures de dépôt cumulé (DATA3a ou DATA3b). r_1 , r_2 et r_3 désignent respectivement les variances d'erreurs d'observation attachées à chacun de ces jeux de données.

paramètre	méthode	avec DATA1	
		DATA2 et DATA3a	DATA2 et DATA3b
r_1 (Bq m^{-3})	Schéma de Desroziers	3.8	2.9
	Vraisemblance maximale	3.3	2.3
r_2 (Bq m^{-2})	Schéma de Desroziers	580	540
	Vraisemblance maximale	210	240
r_3 (Bq m^{-2})	Schéma de Desroziers	325000	240000
	Vraisemblance maximale	320000	230000
m (Bq s^{-1})	Schéma de Desroziers	1.2×10^{11}	1.3×10^{11}
	Vraisemblance maximale	1.0×10^{11}	1.0×10^{11}
Activité rejetée (Bq)	Schéma de Desroziers	1.3×10^{16}	1.9×10^{16}
	Vraisemblance maximale	1.2×10^{16}	1.8×10^{16}

données, noté DATA3b, dans l'inversion (avec DATA1 et DATA2).

Les estimations d'hyper-paramètres et d'activité totale rejetée sont visibles dans le tableau 3.7. En utilisant ces 3 jeux de données, l'activité totale rejetée est estimée entre 18 et 19 PBq.

Les profils temporels de la source reconstruite et de son incertitude sont représentés sur la figure 3.18(d). L'incertitude sur l'activité totale estimée est ici de 15%.

Tous les événements mentionnés dans les sections précédentes sont maintenant retrouvés avec ces jeux de données :

- Autour du 12 mars : explosion d'hydrogène sur l'unité 1.
- Le 13 mars : ouverture de vanne (émission volontaire) sur l'unité 3.
- Les 14 et 15 mars : de multiples ouvertures de vannes et explosions d'hydrogène, concernant principalement les unités 2 et 3. Des émissions tardives le 15 mars sont maintenant retrouvées (entre 7h00 UTC et 9h00 UTC).
- Le 16 mars : événements non identifiés qui correspondent à des baisses de pression dans les unités 2 et 3.
- Le 18 mars : événements non identifiés concernant probablement l'unité 3.
- Le 19 mars : événements non identifiés, mais qui correspondent à des augmentations de rayonnement γ mesuré sur site. Les tentatives de refroidissement par largage d'eau ont commencé juste après ces émissions.
- Le 20 mars : événements non identifiés concernant au minimum les unités 2 et 3.
- Du 21 au 23 mars : événements correspondant à l'apparition de fumées émises depuis les unités 2 et 3.
- Le 25 mars : événement non identifié concernant peut-être l'unité 2.
- Le 30 mars : événement non identifié.

Par rapport à ce qui est généralement admis, les émissions du 15 mars semblent légèrement

sous-estimées alors que les émissions du 25 mars et, dans une moindre mesure, du 20 mars semblent sur-estimées. Une discussion exhaustive concernant les amplitudes des pics d'émission peut être lue dans la section 3.6 de Winiarek et al. [2014] en annexe F.

Si les amplitudes exactes des pics principaux sont toujours incertaines, il est néanmoins possible d'affirmer que l'utilisation conjointe des différents types de données et l'estimation simultanée des variances d'erreurs ont permis de mieux contraindre la source, comparé à l'inversion en utilisant uniquement les observations de dépôt cumulé. Les estimations de l'activité totale rejetée sont parfaitement en accord avec les autres estimations et le profil général est également cohérent.

Chapitre 4

Localisation de sources accidentelles

Sommaire

4.1	Introduction	85
4.2	Méthodes non-paramétriques	87
4.2.1	Statistiques gaussiennes	87
4.2.2	Statistiques non-gaussiennes	87
4.3	Méthodes paramétriques	89
4.3.1	Cas d'une source ponctuelle	91
4.3.2	Généralisation	91
4.4	Application I : Identifier le lieu d'un rejet radiologique	92
4.4.1	Principes et formalisme	92
4.4.2	Validation	94
4.5	Application II : MUST	98
4.5.1	Contexte - présentation de l'expérience	98
4.5.2	Cas d'étude sélectionnés	99
4.5.3	Modèles de dispersion	102
4.5.4	Reconstruction d'une source ponctuelle à partir des données réelles	104

4.1 Introduction

En contexte accidentel, au-delà de l'estimation de la quantité de polluant rejetée et de son évolution au cours du temps, un problème qui peut se poser rapidement aux autorités en charge est la localisation de la source du rejet. En effet, dans le cas d'une attaque terroriste, mais aussi parfois en situation "purement" accidentelle, l'origine du rejet de polluant peut ne pas être connue. À grande échelle, on peut citer bien évidemment le cas de l'accident de Tchernobyl en 1986 au cours duquel les capteurs situés en Europe de l'Ouest (en particulier en Scandinavie) ont commencé à détecter une pollution radiologique importante avant qu'en soit connue l'origine. L'incident d'Algésiras (Espagne) en 1998, bien que d'ampleur moins dramatique, constitue également un exemple de pollution radiologique détectée mais d'origine inconnue. Plus récemment, à la fin de l'année 2011, de nombreux appareils de mesure en Europe centrale ont détecté une pollution à l'iode-131 dont l'origine n'était pas connue (et qui s'est révélée provenir d'un laboratoire de production de radioisotopes de Budapest en Hongrie). En plus de ces quelques cas réels, des campagnes expérimentales, par exemple ETEX I et ETEX II, ont été mises en place afin de tester et de valider dans un premier temps les modèles de dispersion ; par la suite ces campagnes ont été utilisées dans quelques études pour évaluer différentes

méthodes d'identification des sources [Bocquet, 2005c; Issartel, 2003; Krysta and Bocquet, 2007]. De son côté le CTBTO a déployé au niveau mondial un réseau de surveillance capable de détecter rapidement les conséquences d'un essai nucléaire et d'en localiser l'origine.

Si la localisation de la source est d'une importance évidente dans le cas où une intervention sur site est nécessaire, il s'agit d'une étape également importante dans le but d'estimer les rejets, car cela permet de diminuer fortement le nombre d'inconnues dans le système. Dans cette optique, plusieurs stratégies sont envisageables que l'on peut répartir en deux catégories : les méthodes non-paramétriques et les méthodes paramétriques.

les méthodes non-paramétriques

Les méthodes non-paramétriques s'inscrivent dans la continuité des méthodes présentées dans le chapitre 3. Le champ complet d'émission est discrétisé sur une grille spatio-temporelle de N cellules et le système réalise une estimation des N débits d'émission correspondants. L'utilisation des statistiques gaussiennes ou gaussiennes tronquées pour caractériser les distributions *a priori*, notamment de l'ébauche, conduit à la minimisation d'une fonction coût quadratique dont l'expression est rappelée dans le chapitre 4.2.1. Néanmoins dans les situations accidentelles où l'ébauche est très mal connue, l'utilisation de statistiques *a priori* non-gaussiennes peut permettre d'améliorer considérablement les résultats de l'inversion. Il est alors possible d'utiliser le caractère *certainement* localisé (dans le sens peu étendu spatialement) de la source comme information *a priori* sur l'ébauche [Bocquet, 2005b, c], par exemple en utilisant une distribution de Bernoulli. Cette méthode est présentée dans le chapitre 4.2.2.

les méthodes paramétriques

Malgré l'utilisation de distributions non-gaussiennes pour caractériser l'ébauche, les méthodes non-paramétriques peuvent ne pas être suffisamment contraintes et les solutions obtenues peuvent se révéler excessivement étendues dans le contexte accidentel. C'est particulièrement le cas si les observations disponibles se situent loin de la source. Une alternative consiste à contraindre formellement la source à être très peu étendue. Pour cela une paramétrisation de la source est réalisée, avec un nombre de paramètres le plus faible possible. Des méthodes stochastiques peuvent alors être utilisées pour obtenir les distributions *a posteriori* des paramètres et ainsi estimer la source. Au contraire des méthodes non-paramétriques, les méthodes paramétriques, présentées dans le chapitre 4.3, sont contraignantes pour le système dont l'estimation ne peut sortir du cadre de la paramétrisation choisie, même si celle-ci est incompatible avec les observations et/ou la situation réelle.

Application à l'expérience MUST : deux modèles de dispersion

Des méthodes paramétriques et non-paramétriques seront testées sur l'expérience MUST (Mock Urban Setting Test) dans le chapitre 4.5. Lors de cette expérience de dispersion à l'échelle locale (domaine d'environ $200 \times 200 \text{ m}^2$), des rejets de gaz traceurs ont été réalisés et des mesures sont disponibles. Un modèle de CFD et un modèle de panache gaussien seront utilisés pour modéliser la dispersion du traceur, et les méthodes d'inversion seront testées pour localiser la source.

Détermination du lieu d'un rejet parmi un ensemble de suspects identifiés

Dans le cas très particulier où le rejet de polluant ne peut être issu que de certains sites clairement identifiés, des méthodes très spécifiques peuvent également être proposées afin de déterminer le site responsable du rejet. Krysta and Bocquet [2007] ont proposé une méthode

empirique itérative, basée sur l'élimination progressive des sites les moins probables. Les méthodes proposées dans ce manuscrit reposent sur l'estimation de la vraisemblance des observations en supposant que les rejets sont tour à tour issus de chaque site suspecté et permettent de formaliser cette idée dans un cadre mathématique rigoureux. Quatre indicateurs bayésiens, différents par les hypothèses réalisées *a priori* sur le profil temporel du rejet, sont proposés et testés sur des expériences synthétiques à l'échelle européenne dans le chapitre 4.4, cette étude constituant en outre la seconde partie de Winiarek et al. [2011].

4.2 Méthodes non-paramétriques

4.2.1 Statistiques gaussiennes

Dans le cadre de statistiques gaussiennes pour décrire les statistiques *a priori* des erreurs d'ébauche et d'observation, les méthodes présentées dans le chapitre 3.3 restent applicables pour reconstruire un champ complet d'émission. Celui-ci est discrétisé sur une grille spatio-temporelle et le nombre N de débits d'émission inconnus constituant le vecteur $\boldsymbol{\sigma}$ est potentiellement très élevé alors que le nombre d'observations d est faible. L'estimation de la source passe par la minimisation de la fonction coût :

$$\mathcal{J}_q(\boldsymbol{\sigma}) = \frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) + \frac{1}{2} (\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)^T \mathbf{B}^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}^b), \quad (4.1)$$

où $\boldsymbol{\sigma}^b$ constitue une ébauche de la source (généralement nulle), \mathbf{R} et \mathbf{B} étant les matrices de covariance d'erreur d'observation et d'ébauche. \mathbf{H} est la matrice jacobienne du modèle et contient N colonnes et d lignes. Dans ce contexte, elle sera calculée ligne par ligne, à l'aide du modèle adjoint défini dans les chapitres 2.3.1 et 2.3.4. La minimisation de \mathcal{J}_q conduit à une solution analytique de la forme :

$$\boldsymbol{\sigma}^a = \boldsymbol{\sigma}^b + \mathbf{B}\mathbf{H}^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^b). \quad (4.2)$$

Cette formulation n'impose pas la positivité de chaque débit massique de la source. Pour cela, on peut minimiser, par des méthodes numériques, la fonction coût de l'équation 4.1 sous contrainte de positivité et ainsi obtenir une source positive.

4.2.2 Statistiques non-gaussiennes

Les statistiques gaussiennes ne sont pas nécessairement les plus adaptées à un contexte accidentel quand la localisation de la source n'est pas connue. En effet, dans ce cas le nombre d'inconnues est très grand par rapport au nombre d'observations et le terme d'ébauche est très incertain. Il peut donc être pertinent de se placer dans le cadre de statistiques non-gaussiennes pour caractériser l'ébauche [Bocquet, 2005b]. D'autres pdf *a priori* peuvent se révéler intéressantes dans ce contexte pour permettre de prendre en compte statistiquement les particularités de la source, telles que sa positivité et son étendue limitée par exemple.

Source *a priori* ponctuelle

Supposons qu'à l'échelle du domaine, la source est ponctuelle spatialement et instantanée, c'est-à-dire que la durée du rejet est plus faible que la résolution temporelle de la reconstruction. Nous supposons de plus que cette ébauche suit une loi de Bernoulli locale : dans ce cadre, il y a $N + 1$ sources potentielles : tout d'abord la source $\boldsymbol{\sigma} = \mathbf{0}$ qui a une probabilité d'occurrence valant $1 - \frac{1}{N} \sum_{i=1}^N \gamma_i$, puis N sources potentielles correspondant chacune à une émission

dans une des mailles de la grille de l'espace de contrôle. La probabilité d'émettre m_k dans la maille k est γ_k/N . Cette hypothèse définit complètement la pdf *a priori* de l'ébauche.

Des techniques utilisant de telles pdfs *a priori*, plus complexes que des gaussiennes, ont été développées notamment par Bocquet [2005b], et appliquées sur des situations accidentelles [Bocquet, 2005c; Krysta and Bocquet, 2007] ou expérimentales [Bocquet, 2005c; Krysta et al., 2008] réelles. Ces techniques s'appuient sur la méthode du maximum d'entropie sur la moyenne (MEM) [Jaynes, 1957a, b], qui permet grâce à la convexité des fonctions coût engendrées d'utiliser la relation primale-duale pour se placer dans l'espace des observations de dimension bien plus petite que l'espace des sources. Pour une ébauche ponctuelle définie précédemment, la fonction coût à minimiser s'écrit [Krysta and Bocquet, 2007] :

$$\mathcal{L}(\boldsymbol{\beta}) = \ln(Z_{\boldsymbol{\sigma}}(\boldsymbol{\beta})) + \frac{1}{2}\boldsymbol{\beta}^T \mathbf{R}\boldsymbol{\beta} - \boldsymbol{\beta}^T \boldsymbol{\mu}, \quad (4.3)$$

où $\boldsymbol{\beta} \in \mathbb{R}^d$, et $Z_{\boldsymbol{\sigma}}(\boldsymbol{\beta})$ est défini par :

$$Z_{\boldsymbol{\sigma}}(\boldsymbol{\beta}) = \frac{1}{N} \sum_{k=1}^N \{1 - \gamma_k + \gamma_k \exp(m_k [\boldsymbol{\beta}^T \mathbf{H}]_k)\}. \quad (4.4)$$

Une fois $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\beta})$ calculé, la pdf *a posteriori* $p(\boldsymbol{\sigma}|\boldsymbol{\mu})$ est connue et un estimateur $\boldsymbol{\sigma}^a$ de la source peut être obtenu, par exemple en prenant la moyenne. L'élément k_0 de cet estimateur s'écrit :

$$\sigma_{k_0}^a = m_{k_0} \frac{\gamma_{k_0} \exp\left(m_{k_0} [\hat{\boldsymbol{\beta}}^T \mathbf{H}]_{k_0}\right)}{\sum_{k=1}^N \{1 - \gamma_k + \gamma_k \exp\left(m_k [\hat{\boldsymbol{\beta}}^T \mathbf{H}]_k\right)\}}. \quad (4.5)$$

Dans la suite de ce manuscrit, nous nommerons cet algorithme *schéma de Krysta*. Il est à noter que l'estimateur $\boldsymbol{\sigma}^a$ n'appartient pas nécessairement à l'ensemble des sources potentielles définie par la pdf *a priori* (source ponctuelle qui émet m_k dans la maille k et 0 ailleurs), mais il appartient à l'enveloppe convexe de cet ensemble. Nous choisirons généralement $m_k = m$ et $\gamma_k = \gamma$, pour toutes les mailles k . γ est choisi suffisamment faible, et m et r seront estimés par itération de la L-curve.

Source *a priori* localisée

On entend par source localisée une source géographiquement ponctuelle mais dont le débit évolue au cours du temps. Il s'agit d'une généralisation du cas précédent. Pour chaque point d'émission potentiel, les émissions *a priori* sont décorrélées en temps et suivent un loi de Bernoulli (de paramètres (γ, m)) sur chaque intervalle de temps. Ces hypothèses définissent la pdf *a priori* de l'ébauche. Considérons que le nombre de points de la grille spatiale soit N_s et le nombre de pas de temps N_t , alors une maille k de la grille de l'espace de contrôle se décompose en (l, n) tels que $k = (l - 1) \times N_t + n$, où n est l'indice temporel et l l'indice spatial (et $N = N_s \times N_t$). Dans ce cas la fonction coût à minimiser est toujours donnée par l'équation 4.3 et $Z_{\boldsymbol{\sigma}}(\boldsymbol{\beta})$ devient [Krysta and Bocquet, 2007] :

$$Z_{\boldsymbol{\sigma}}(\boldsymbol{\beta}) = \sum_{l=1}^{N_s} \frac{1}{N_s} \prod_{n=1}^{N_t} \{1 - \gamma + \gamma \exp(m [\boldsymbol{\beta}^T \mathbf{H}]_k)\}. \quad (4.6)$$

L'élément du point de grille k_0 (site l_0 et instant n_0) de l'estimateur de la source reconstruite est :

$$\sigma_{k_0}^a = m \frac{\gamma \exp\left(m \left[\hat{\boldsymbol{\beta}}^T \mathbf{H}\right]_{k_0}\right)}{1 - \gamma + \gamma \exp\left(m \left[\hat{\boldsymbol{\beta}}^T \mathbf{H}\right]_{k_0}\right)} \times \frac{\prod_{n=1}^{N_t} \left\{1 - \gamma + \gamma \exp\left(m \left[\hat{\boldsymbol{\beta}}^T \mathbf{H}\right]_{(l_0-1) \times N_t + n}\right)\right\}}{Z_{\sigma}(\hat{\boldsymbol{\beta}})}, \quad (4.7)$$

avec $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\beta})$.

4.3 Méthodes paramétriques

Malgré l'utilisation de statistiques non-gaussiennes pour les erreurs d'ébauche, les méthodes précédentes restent de nature non-paramétrique : le système cherche à reconstruire un champ d'émission complet (en 4 dimensions dans les cas les moins favorables) et il peut ne pas être suffisamment contraint, notamment à cause du faible nombre d'observations, pour reconstruire des sources très localisées, ce qui est pourtant en général le cas en contexte accidentel. Une stratégie alternative consiste à paramétriser la source et, à l'aide d'hypothèses fortes sur sa nature, à réduire fortement le nombre de paramètres à estimer. Cette stratégie est très contraignante puisqu'elle n'autorise pas le système à estimer la source en dehors de l'ensemble des sources potentielles *a priori*. Dans le contexte de la dispersion atmosphérique, Keats et al. [2007] ont appliqué une méthode paramétrique pour localiser une source ponctuelle en utilisant des données issues de mesures en veine hydraulique. Delle Monache et al. [2008] ont tenté de localiser le lieu du rejet dans le contexte de l'incident d'Algésiras. Yee et al. [2008] ont proposé une généralisation de la méthode par rapport à ces études antérieures permettant de localiser plusieurs lieux de rejet éventuels. De telles méthodes sont également très utilisées dans d'autres domaines, notamment pour la localisation de sources de polluants, comme la pollution des sols ou des eaux souterraines par exemple [Hazart et al., 2014].

Les méthodes paramétriques proposées s'inscrivent dans un cadre bayésien et reposent sur l'estimation, à un coefficient multiplicatif près, de la pdf *a posteriori* $p(\sigma|\boldsymbol{\mu})$ définie par la formule de Bayes :

$$p(\sigma|\boldsymbol{\mu}) = \frac{p(\sigma)p(\boldsymbol{\mu}|\sigma)}{p(\boldsymbol{\mu})} \propto p(\boldsymbol{\mu}|\sigma)p(\boldsymbol{\mu}). \quad (4.8)$$

La vraisemblance $p(\boldsymbol{\mu}|\sigma)$ peut être vue comme la pdf des erreurs d'observation $p_{\epsilon}(\boldsymbol{\epsilon})$, où le vecteur $\boldsymbol{\epsilon}$ représente l'écart entre les observations $\boldsymbol{\mu}$ et les observations simulées par le modèle, notée $\boldsymbol{\mu}^{\text{simu}}$:

$$\boldsymbol{\epsilon} = \boldsymbol{\mu} - \boldsymbol{\mu}^{\text{simu}}. \quad (4.9)$$

L'opérateur d'observation permet de transformer le champ de concentration calculé par le modèle de dispersion en un observable, par l'intermédiaire d'une fonction d'échantillonnage, ce qui permet d'exprimer l'observation simulée i par la formule :

$$\mu_i^{\text{simu}} = \int_{\Omega} c(\mathbf{x}, t) \pi_i(\mathbf{x}, t) d\mathbf{x} dt, \quad (4.10)$$

où $\Omega = \mathcal{D} \times [0, T]$ représente le domaine spatio-temporelle de simulation et π_i est la fonction d'échantillonnage relative à l'observable i . c est le champ de concentration du polluant calculé

par le modèle direct par résolution de l'équation de dispersion

$$\frac{\partial c}{\partial t} + \operatorname{div}(\mathbf{u}c) = \operatorname{div}\left(\rho \mathbf{K} \nabla \left(\frac{c}{\rho}\right)\right) - \Lambda^s c - \Lambda^d c + \sigma. \quad (4.11)$$

Pour calculer μ^{simu} , il est donc possible d'utiliser le modèle numérique direct, qui résout l'équation 4.11, suivi de l'opérateur d'observation de l'équation 4.10. Une alternative consiste à utiliser la formulation adjointe décrite dans les chapitres 2.3.1 et 2.3.4. Dans ce cadre l'opérateur d'observation s'écrit :

$$\mu_i^{\text{simu}} = \int_{\Omega} c_i^*(\mathbf{x}, t) \sigma(\mathbf{x}, t) \, d\mathbf{x} \, dt, \quad (4.12)$$

où c_i^* est le champ de concentration du rétro-panache i , calculé par le modèle numérique adjoint en résolvant l'équation de dispersion

$$\frac{\partial c_i^*}{\partial(-t)} + \operatorname{div}(-\mathbf{u}c_i^*) = \operatorname{div}\left(\rho \mathbf{K} \nabla \left(\frac{c_i^*}{\rho}\right)\right) - \Lambda^s c_i^* - \Lambda^d c_i^* + \pi_i. \quad (4.13)$$

L'intérêt de la formulation adjointe se révèle si on souhaite estimer μ_i^{simu} pour un grand nombre de sources différentes. En effet la concentration c dépend de σ , au contraire de c_i^* . Ainsi à chaque nouvelle proposition de σ , il est nécessaire, dans le cas de la formulation directe, de procéder à une nouvelle simulation en utilisant le modèle direct, ce qui est très coûteux en temps de calcul. Au contraire, une fois le rétro-panache i calculé avec le modèle adjoint, l'estimation de μ_i^{simu} pour chaque nouvelle proposition de source nécessite seulement l'application de l'opérateur d'observation de l'équation 4.12.

Le principe des méthodes paramétriques consiste à représenter un champ complet d'émission σ par un vecteur Θ de paramètres caractérisant totalement la source. Dans ce cas et pour faire apparaître clairement la dépendance avec les paramètres, nous noterons

$$\mu^{\text{simu}} = \mathcal{H}(\Theta), \quad (4.14)$$

où \mathcal{H} est un opérateur incorporant le modèle numérique de dispersion et l'opérateur d'observation (formulation directe ou adjointe).

L'estimation de la pdf $p(\sigma|\mu)$ par l'équation 4.8 est donc équivalente à l'estimation d'une pdf $p(\Theta|\mu)$ définie par :

$$p(\Theta|\mu) \propto p(\mu|\Theta)p(\Theta). \quad (4.15)$$

Pour cela, il est nécessaire au préalable de définir les pdf *a priori* $p(\mu|\Theta)$ et $p(\Theta)$. Nous proposons de conserver une hypothèse de statistiques gaussiennes pour les erreurs d'observation :

$$p(\mu|\Theta) = \frac{\exp\left(-\frac{1}{2}(\mu - \mathcal{H}(\Theta))^T \mathbf{R}^{-1}(\mu - \mathcal{H}(\Theta))\right)}{\sqrt{(2\pi)^d |\mathbf{R}|}}, \quad (4.16)$$

\mathbf{R} représentant la matrice de covariance des erreurs d'observation et $|\mathbf{R}|$ son déterminant.

La définition de $p(\Theta)$ dépendra des hypothèses réalisées *a priori* sur la nature de la source, sa paramétrisation et les hypothèses *a priori* sur ces paramètres. Nous allons tout d'abord évoquer le cas d'une source supposée ponctuelle, avant de généraliser la méthode.

4.3.1 Cas d'une source ponctuelle

La paramétrisation la plus restrictive consiste à supposer que la source est ponctuelle de coordonnées (x_s, y_s, z_s) et de débit massique constant q_s entre les instants t^{on} et t^{off} . Dans ce cas $\Theta = (x_s, y_s, z_s, q_s, t^{\text{on}}, t^{\text{off}})$. On suppose de plus que les 6 paramètres de Θ sont indépendants, ce qui se traduit par :

$$p(\Theta|\mu) \propto p(\mu|\Theta) p(x_s) p(y_s) p(z_s) p(q_s) p(t^{\text{on}}) p(t^{\text{off}}), \quad (4.17)$$

et, de plus, chaque paramètre Θ_j est *a priori* distribué uniformément dans un intervalle $[\Theta_j^{\text{min}}, \Theta_j^{\text{max}}]$. Ainsi si on note $\mathcal{R}^1 = [\Theta_1^{\text{min}}, \Theta_1^{\text{max}}] \times \dots \times [\Theta_6^{\text{min}}, \Theta_6^{\text{max}}]$ alors l'équation 4.17 s'écrit :

$$p(\Theta|\mu) \propto \frac{\mathbb{I}_{\mathcal{R}^1}(\Theta)}{\sqrt{(2\pi)^d |\mathbf{R}|}} \exp\left(-\frac{1}{2}(\mu - \mathcal{H}(\Theta))^T \mathbf{R}^{-1}(\mu - \mathcal{H}(\Theta))\right), \quad (4.18)$$

où $\mathbb{I}_{\mathcal{R}^1}(\Theta) = 1$ si $\Theta \in \mathcal{R}^1$ et 0 sinon.

Dans le cas d'une source σ ponctuelle définie par le vecteur de paramètres Θ , l'estimation de $\mathcal{H}(\Theta)$ par l'équation 4.12 devient

$$[\mathcal{H}(\Theta)]_i = \mu_i^{\text{simu}} = q_s \int_{t^{\text{on}}}^{t^{\text{off}}} c_i^*(x_s, y_s, z_s, t) dt, \quad (4.19)$$

où c_i^* est le champ de concentration dans le rétro-panache défini par l'équation 4.13.

Un algorithme MCMC (Markov Chain Monte Carlo) permet de réaliser un échantillonnage de $p(\Theta|\mu)$ et d'obtenir un chaîne de Markov $\{\Theta^{(1)}, \dots, \Theta^{(N_e)}\}$ dont les éléments sont distribués suivant $p(\Theta|\mu)$. Il est alors possible d'obtenir une bonne estimation des probabilités marginales *a posteriori* de chaque paramètre de la source $p(x_s|\mu)$, $p(y_s|\mu)$, etc. [Keats et al., 2007; Delle Monache et al., 2008; Senocak et al., 2008].

4.3.2 Généralisation

La paramétrisation présentée dans le chapitre 4.3.1 contraint très fortement le système. Si la vraie source n'est pas conforme à cette paramétrisation, par exemple en étant composée de plusieurs sources ponctuelles, celui-ci peut avoir un comportement imprévisible. En utilisant notamment le concept de saut réversible dans les algorithmes de Markov introduit par Green [1995], il est néanmoins possible d'étendre la méthode dans un cadre plus général. On peut par exemple supposer que la source est composée de N_s sources ponctuelles définies dans le chapitre 4.3.1. Le vecteur de paramètres s'écrit alors :

$$\Theta = (N_s, x_{s1}, y_{s1}, z_{s1}, q_{s1}, t_1^{\text{on}}, t_1^{\text{off}}, \dots, x_{sN_s}, y_{sN_s}, z_{sN_s}, q_{sN_s}, t_{N_s}^{\text{on}}, t_{N_s}^{\text{off}}). \quad (4.20)$$

N_s est bien entendu non connu et doit être estimé par le système, mais on peut supposer qu'il est distribué par une loi *a priori* $p(N_s)$. De plus tous les paramètres de Θ sont supposés indépendants, et chacun des paramètres Θ_j pour $j > 1$ sont distribués uniformément sur $[\Theta_j^{\text{min}}, \Theta_j^{\text{max}}]$. L'équation 4.20 s'écrit ainsi :

$$p(\Theta|\mu) \propto \frac{p(N_s) \mathbb{I}_{\mathcal{R}^{N_s}}(\Theta)}{\sqrt{(2\pi)^d |\mathbf{R}|}} \exp\left(-\frac{1}{2}(\mu - \mathcal{H}(\Theta))^T \mathbf{R}^{-1}(\mu - \mathcal{H}(\Theta))\right), \quad (4.21)$$

où $\mathcal{R}^{N_s} = \mathbb{N} \times [\Theta_1^{\text{min}}, \Theta_1^{\text{max}}] \times \dots \times [\Theta_{6N_s}^{\text{min}}, \Theta_{6N_s}^{\text{max}}]$.

L'estimation de $\mathcal{H}(\Theta)$ peut s'écrire :

$$[\mathcal{H}(\Theta)]_i = \sum_{k=1}^{N_s} q_{sk} \int_{t_k^{\text{on}}}^{t_k^{\text{off}}} c_i^*(x_{sk}, y_{sk}, z_{sk}, t) dt. \quad (4.22)$$

Un algorithme MCMC avec saut réversible permet d'estimer le nombre N_s de sources ponctuelles le plus probable ainsi que les probabilités marginales *a posteriori* des paramètres associés [Yee et al., 2008].

4.4 Application I : Identifier le lieu d'un rejet radiologique

Dans le chapitre 3.7, nous avons proposé un algorithme semi-automatique pour reconstruire le terme source issu d'une installation nucléaire déjà identifiée. Il est malheureusement envisageable, et l'histoire l'a déjà prouvé, d'ignorer dans un premier temps l'origine de la pollution radiologique constatée par des appareils de mesure. Du fait du passif historique de certains de leurs voisins, les autorités finlandaises notamment, par l'intermédiaire de l'Institut finlandais de météorologie avec qui nous avons collaboré sur ce sujet, souhaitent pouvoir faire face à ce type de situation. L'objectif de ce chapitre est donc de proposer des algorithmes fiables et rapides qui permettent, dans le cas très particulier d'une pollution radiologique, d'identifier l'installation nucléaire responsable. Dans une optique d'opérationnalité de l'outil proposé, les méthodes académiques présentées dans les chapitres 4.2 et 4.3 sont exclues, même si elles ont pu montrer une certaine efficacité sur quelques cas tests. En lieu et place, nous proposons ici des tests statistiques qui, dans un cadre toujours bayésien, permettent d'identifier l'origine du rejet parmi un ensemble de sites suspectés. Cette étude constitue la seconde partie de Winiarek et al. [2011], publié dans Atmospheric Environment et visible en annexe D.

4.4.1 Principes et formalisme

Par rapport aux objectifs recherchés par les méthodes proposées au début de ce chapitre, le contexte ici est différent : il ne s'agit plus de localiser une ou plusieurs sources de polluant sur un domaine plus ou moins étendu, mais de déterminer parmi un ensemble d'installations suspectées celle qui est le plus probablement responsable. Les possibilités de situations de la source potentielle sont donc beaucoup moins nombreuses et il s'agit de tirer profit de cette particularité.

Les différents indicateurs statistiques que nous proposons reposent sur l'estimation de la vraisemblance des observations sachant que le rejet a lieu dans l'installation i . Dans le cas général, cette vraisemblance peut s'écrire :

$$p_i(\boldsymbol{\mu}) = p(\boldsymbol{\mu}|i) = \int p(\boldsymbol{\sigma}|i) p(\boldsymbol{\mu}|\boldsymbol{\sigma}, i) d\boldsymbol{\sigma}, \quad (4.23)$$

où $\boldsymbol{\sigma}$ représente un profil temporel de terme source. $p(\boldsymbol{\sigma}|i)$ représente la probabilité *a priori* du profil $\boldsymbol{\sigma}$ sachant que le rejet a lieu au niveau de l'installation i . Dans un premier temps, il ne semble pas pertinent de supposer que certains profils temporels sont plus ou moins probables selon l'installation d'origine et nous choisirons $p(\boldsymbol{\sigma}|i)$ indépendant de i . $p(\boldsymbol{\mu}|\boldsymbol{\sigma}, i)$ représente la vraisemblance des observations sachant que le rejet, de profil temporel $\boldsymbol{\sigma}$, a pour origine l'installation i .

Il est ainsi possible d'estimer, au fur et à mesure de l'obtention de nouvelles observations, l'évolution de l'indicateur p_i pour chaque installation suspectée. En normalisant l'ensemble des p_i par leur somme $\sum_j p_j$, chacun des indicateurs représente la probabilité que le site correspondant soit responsable du rejet radiologique, à condition que le rejet ait effectivement

pour origine un (et un seul) des sites suspectés. Si cette dernière hypothèse était violée, les indicateurs normalisés $p_i / (\sum_j p_j)$ n'auraient pas de sens, mais les valeurs brutes de tous les p_i resteraient anormalement faibles, révélant la violation d'hypothèse.

En ce qui concerne le choix de la distribution *a priori* $p(\boldsymbol{\mu}|\boldsymbol{\sigma}, i)$, nous proposons de conserver des statistiques gaussiennes. Dans ce cadre, la pdf s'écrit alors

$$p(\boldsymbol{\mu}|\boldsymbol{\sigma}, i) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}_i\boldsymbol{\sigma})^T \mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{H}_i\boldsymbol{\sigma})\right)}{\sqrt{(2\pi)^d |\mathbf{R}|}}, \quad (4.24)$$

où \mathbf{H}_i représente la matrice jacobienne du modèle correspondant à un rejet issu de l'installation i .

Cadre gaussien

Comme pour la distribution $p(\boldsymbol{\mu}|\boldsymbol{\sigma}, i)$, il est envisageable d'opter pour $p(\boldsymbol{\sigma}|i)$ pour une loi gaussienne, qui s'écrit alors

$$p(\boldsymbol{\sigma}|i) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)^T \mathbf{B}^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)\right)}{\sqrt{(2\pi)^N |\mathbf{B}|}}, \quad (4.25)$$

où $\boldsymbol{\sigma}^b$ représente une ébauche de profil temporel, généralement nulle comme expliqué dans le chapitre 3.7.3.

L'intégration de l'équation 4.23 conduit à une expression semblable à celle obtenue dans le chapitre 3.6.3 :

$$p_i(\boldsymbol{\mu}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}_i\boldsymbol{\sigma}^b)^T (\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T + \mathbf{R})^{-1}(\boldsymbol{\mu} - \mathbf{H}_i\boldsymbol{\sigma}^b)\right)}{\sqrt{(2\pi)^d |\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T + \mathbf{R}|}}. \quad (4.26)$$

Cadre non-gaussien

Le cadre gaussien du paragraphe précédent n'est pas très contraignant. Par exemple, il autorise chacun des débits indépendants qui composent le terme source à être négatif. Comme discuté dans le chapitre 3, l'utilisation de statistiques semi-gaussiennes permettraient certainement de contraindre davantage le système, mais au prix d'une augmentation importante du temps de calcul.

Par ailleurs, les centres opérationnels ont déjà construit une liste de termes sources possibles pour différents types d'incidents envisageables. Une alternative consiste ainsi à supposer que le terme source fait partie de cette liste. Néanmoins il est peu probable que l'amplitude du rejet réel corresponde à celui de cette ébauche notée $\boldsymbol{\sigma}^b$. L'hypothèse est donc plutôt que le vrai terme source est de la forme $\lambda\boldsymbol{\sigma}^b$, λ étant un paramètre positif distribué suivant une loi *a priori* $p(\lambda)$.

Formellement, la distribution *a priori* du terme source est donc donnée par

$$p(\boldsymbol{\sigma}|i) = \int_0^\infty \delta(\boldsymbol{\sigma} - \lambda\boldsymbol{\sigma}^b) p(\lambda) d\lambda. \quad (4.27)$$

Il s'ensuit que la vraisemblance des observations, donnée par l'équation 4.23 peut s'écrire

$$p_i(\boldsymbol{\mu}) = p(\boldsymbol{\mu}|i) = \int p(\boldsymbol{\sigma}|i) p(\boldsymbol{\mu}|\boldsymbol{\sigma}, i) d\boldsymbol{\sigma} = \int_0^\infty p(\lambda) p(\boldsymbol{\mu}|\lambda\boldsymbol{\sigma}^b, i) d\lambda. \quad (4.28)$$

Un premier choix envisagé pour $p(\lambda)$ est une loi Gamma

$$p(\lambda) = \frac{\lambda^{k-1} \exp(-\lambda/\theta)}{\theta^k \Gamma(k)}. \quad (4.29)$$

Dans le cas $k = 1$ (loi exponentielle), la vraisemblance s'écrit (voir Annexe D pour les détails du calcul) :

$$p_i(\boldsymbol{\mu}) = \sqrt{\frac{\pi}{2}} \frac{\exp\left(-\frac{1}{2}\boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}\right) \exp\left(\frac{(\boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}_i^b - \theta^{-1})^2}{2\boldsymbol{\mu}_i^{bT} \mathbf{R}^{-1} \boldsymbol{\mu}_i^b}\right)}{\theta \sqrt{(2\pi)^d |\mathbf{R}|} \sqrt{2\boldsymbol{\mu}_i^{bT} \mathbf{R}^{-1} \boldsymbol{\mu}_i^b}} \times \left[1 - \Phi\left(\frac{\theta^{-1} - \boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}_i^b}{\sqrt{2\boldsymbol{\mu}_i^{bT} \mathbf{R}^{-1} \boldsymbol{\mu}_i^b}}\right) \right], \quad (4.30)$$

où $\Phi(u) = \frac{2}{\sqrt{\pi}} \int_0^u e^{-x^2} dx$ est la fonction d'erreur et $\boldsymbol{\mu}_i^b = \mathbf{H}_i \boldsymbol{\sigma}^b$ est le jeu d'observations obtenu à partir de l'ébauche $\boldsymbol{\sigma}^b$ et du modèle considéré sans erreur.

La loi Gamma avec $k = 1$ favorise les facteurs d'échelle λ faibles. Une alternative consiste à choisir pour loi de distribution *a priori* de λ une loi Gamma avec $k = 2$ qui exclut les valeurs très faibles de λ . Dans ce cas la vraisemblance s'écrit :

$$p_i(\boldsymbol{\mu}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}\right)}{\theta^2 \sqrt{(2\pi)^d |\mathbf{R}|}} \frac{1}{\boldsymbol{\mu}_i^{bT} \mathbf{R}^{-1} \boldsymbol{\mu}_i^b} \times \left\{ 1 - \sqrt{\frac{\pi}{2}} \frac{\theta^{-1} - \boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}_i^b}{\sqrt{\boldsymbol{\mu}_i^{bT} \mathbf{R}^{-1} \boldsymbol{\mu}_i^b}} \left[1 - \Phi\left(\frac{\theta^{-1} - \boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}_i^b}{\sqrt{2\boldsymbol{\mu}_i^{bT} \mathbf{R}^{-1} \boldsymbol{\mu}_i^b}}\right) \right] \right\}. \quad (4.31)$$

Enfin une distribution semi-gaussienne peut également être envisagée

$$p(\lambda) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\lambda^2}{2\theta}\right), \quad (4.32)$$

la vraisemblance s'écrivant alors

$$p_i(\boldsymbol{\mu}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}\right) \exp\left(\frac{(\boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}_i^b)^2}{2(\theta^{-1} + \boldsymbol{\mu}_i^{bT} \mathbf{R}^{-1} \boldsymbol{\mu}_i^b)}\right)}{\sqrt{\theta} \sqrt{(2\pi)^d |\mathbf{R}|} \sqrt{\theta^{-1} + \boldsymbol{\mu}_i^{bT} \mathbf{R}^{-1} \boldsymbol{\mu}_i^b}} \times \left[1 + \Phi\left(\frac{\boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}_i^b}{\sqrt{2(\theta^{-1} + \boldsymbol{\mu}_i^{bT} \mathbf{R}^{-1} \boldsymbol{\mu}_i^b)}}\right) \right]. \quad (4.33)$$

4.4.2 Validation

Nous disposons donc de quatre indicateurs statistiques, donnés par les équations 4.26, 4.30, 4.31 et 4.33. Ces équations, en particulier dans les cas non-gaussiens, peuvent sembler compliquées mais sont en réalité simples à implémenter et remarquablement peu coûteuse en temps de calcul.

Nous avons procédé à différentes expériences synthétiques afin de valider la pertinence de ces indicateurs statistiques. Pour cela nous nous sommes à nouveau placé dans le cadre présenté dans la partie 3.7. La carte de situation est rappelée sur la figure 4.1. Nous nous sommes en particulier intéressés aux situations extrêmes possibles dans cette configuration : rejet au bord ou à l'extérieur de l'enveloppe convexe des capteurs en particulier, mais nous avons également procédé à une évaluation systématique, sur un grand nombre d'expériences, de ces indicateurs.

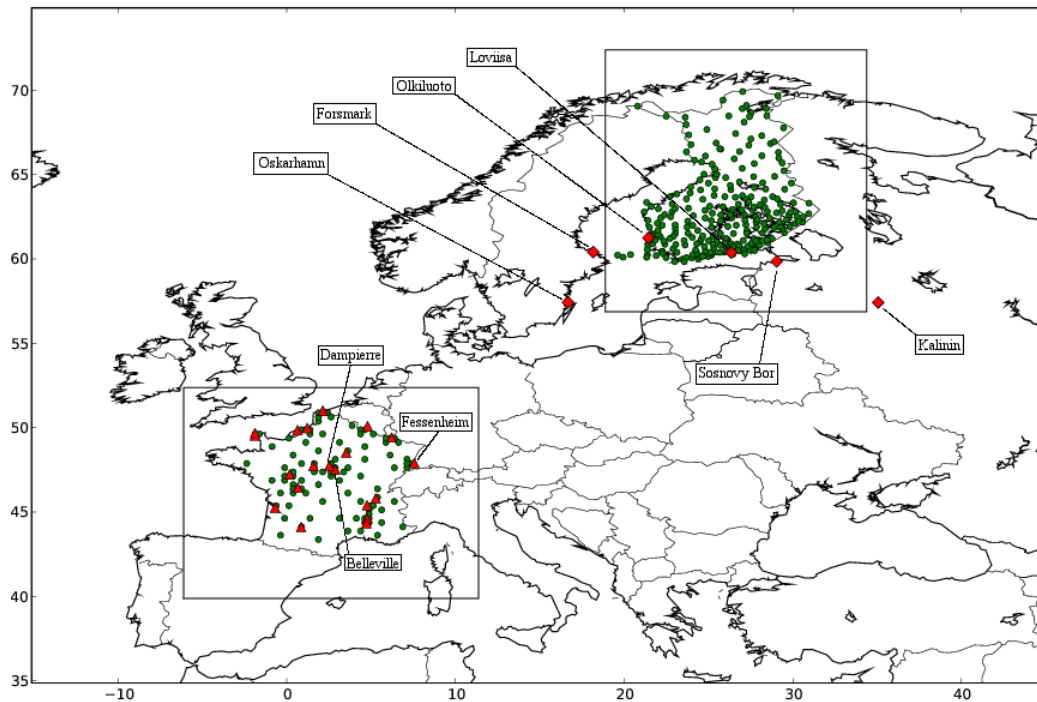


FIG. 4.1: Carte du domaine de simulation et des deux sous-domaines d’étude (France et Finlande). Les diamants rouges représentent les installations nucléaires surveillées par le système et les points verts les stations de mesure.

Illustration des performances des indicateurs bayésiens sur des cas critiques

Pour cette première expérience synthétique, le rejet fictif a lieu dans la centrale de Sosnovy Bor en Russie et a pour profil temporel celui représenté sur la figure 3.2 de la partie 3.7. Le réseau de mesures considéré est le réseau finlandais “uljas”. Les indicateurs statistiques sont calculés pour les centrales de Sosnovy Bor, Loviisa, Olkiluoto, Kalinin, Forsmark et Oskarhamn 1, 2, . . . , 7 heures après le début de l’accident.

Pour l’indicateur gaussien, dont les résultats sont reportés sur le panel (a) de la figure 4.2, nous avons défini $\mathbf{B} = m^2 \mathbf{I}_N$ avec $m = 10^9 \text{ Bq.s}^{-1}$. Pour les autres indicateurs, nous avons défini $\sigma^b = \sigma^t$ (le vrai profil temporel). Ces choix constituent des conditions “idéales” pour l’efficacité des indicateurs mais sont évidemment inaccessibles en conditions réelles. Les résultats des indicateurs non-gaussiens sont très proches de ceux de l’indicateur gaussien mais parviennent à identifier la centrale responsable environ 2h plus tôt.

Une seconde expérience synthétique, plus difficile, avec un panache moins observé par le réseau finlandais, consiste à choisir la centrale de Kalinin en Russie comme responsable d’un rejet de même profil temporel. Dans ce cas, l’indicateur gaussien échoue à identifier la centrale responsable, alors que les trois autres indicateurs y parviennent après quelques heures d’observation, comme illustré sur les panels (b), (c) et (d) de la figure 4.2. On distingue clairement sur ces figures les instants où de nouvelles observations permettent d’éliminer certaines centrales de la liste des suspects les plus probables. Les performances de ces trois indicateurs non-gaussiens sont comparables, même si l’indicateur issu d’une loi Gamma avec $k = 2$ et dont le mode correspond à $\lambda = 1$ est ici plus performant, notamment du fait du choix “idéal” de σ^b .

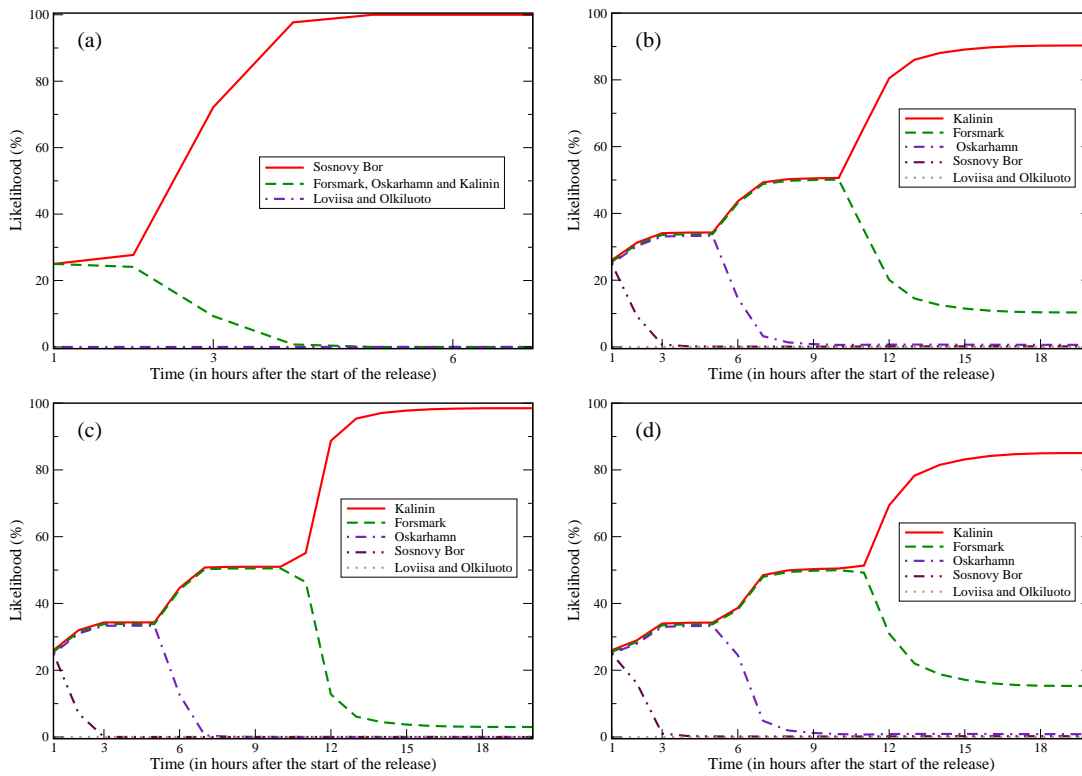


FIG. 4.2: Panel (a) : Indicateurs gaussiens dans le cas d'un accident survenu à Sosnovy Bor. Panels (b), (c), et (d) : Indicateurs non-gaussiens dans le cas de Kalinin. Le panel (b) correspond à la loi Gamma avec $k = 1$, le panel (c) à la loi Gamma avec $k = 2$, et le panel (d) à la loi semi-gaussienne.

Robustesse aux facteurs de masse *a priori*

Comme évoqué précédemment, en situation réelle, les paramètres *a priori* : m pour l'indicateur gaussien et σ^b pour les indicateurs non-gaussiens, ne seront pas connus. Il est donc nécessaire d'évaluer la robustesse des indicateurs à des paramètres dégradés.

Pour illustrer cela, nous avons réalisé une expérience synthétique en choisissant la centrale de Belleville en France comme lieu du rejet de radionucléides et en évaluant la valeur des indicateurs pour chacune des 20 installations nucléaires françaises. Pour l'indicateur gaussien, différentes valeurs de m ont été testées. La valeur de cet indicateur 6h après l'accident est reportée sur la figure 4.3 pour trois centrales particulières : Belleville, Dampierre et Fessenheim. Pour des valeurs de m faibles, le système n'est pas capable de différencier les centrales et chacune des 20 installations a une probabilité de 5%. De l'autre côté de la courbe, pour des grandes valeurs de m , le test indique la mauvaise centrale : Fessenheim. Cette centrale est au bord de l'enveloppe convexe du réseau d'observation (près de la frontière allemande) et un rejet très important est plus compatible lorsqu'il provient de la centrale de Fessenheim que de toute autre centrale située plus à l'intérieur, ce qui a notamment conduit à l'exclusion, à tort, de Belleville. On voit ici qu'il existe une plage de validité d'environ 4 ordres de grandeur sur laquelle l'indicateur sélectionne assez clairement la bonne centrale.

Dans le cas des indicateurs non-gaussiens, nous avons fait varier un paramètre λ^t défini par $\sigma^b = \lambda^t \sigma^t$. Pour des valeurs faibles de λ^t , un comportement identique que celui de l'indicateur gaussien avec m faible est observé : le système est incapable de distinguer les installations les unes des autres. Par contre, pour des valeurs élevées de λ^t , les indicateurs se sont montrés plus

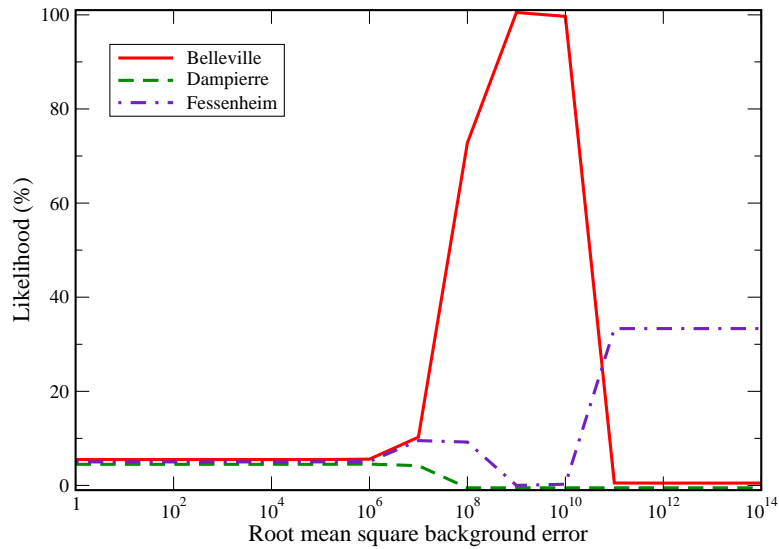


FIG. 4.3: Indicateur gaussien calculé 6h après un accident survenu à Belleville. Plusieurs valeurs du paramètre m (en abscisse) sont testées : $\log_{10}(m) = 0, 1, \dots, 14$.

robustes que l'indicateur gaussien et les tests ont échoué seulement pour des valeurs totalement irréalistes.

Performances globales des indicateurs bayésiens

Il est nécessaire d'évaluer les performances des indicateurs proposés non seulement sur quelques cas d'illustration mais également par une approche plus systématique, en prenant en compte plusieurs centrales responsables et des conditions météorologiques variées. Pour cela, la période du 25 novembre au 4 décembre 2007 est sélectionnée, caractérisée par des vents forts. Des conditions de vents forts devraient mener à des performances plus faibles comparées à une situation de vents plus faibles [Krysta and Bocquet, 2007; Abida and Bocquet, 2009].

À chaque heure, un accident est simulé sur chacune des 20 installations nucléaires françaises. Pour chaque ensemble de $p = 20$ accidents et à chaque heure qui suit, une matrice \mathbf{L} de dimension $p \times p$ est construite. L'élément $[\mathbf{L}]_{ij}$ est défini par la probabilité de l'installation i d'être responsable de l'accident survenu en j . L'évolution de la trace normalisée $\frac{1}{p} \text{tr}(\mathbf{L})$, qui indique la capacité de l'indicateur à identifier correctement l'installation responsable, est étudiée au fur et à mesure que de nouvelles observations sont disponibles.

L'évolution de la trace normalisée (sa moyenne sur tous les accidents simulés, la médiane de cet ensemble ainsi que les écart-types autour de la moyenne) dans le cas de l'indicateur gaussien est affichée sur la figure 4.4(a). Sur la figure 4.4(b) celle dans le cas de l'indicateur semi-gaussien, avec $\sigma^b = \sigma^t$ et enfin sur la figure 4.4(c) celle dans le cas de l'indicateur semi-gaussien avec σ^b choisie égale à une constante (situation fortement dégradée). Les performances moyennes sont impactées par quelques cas d'échec, mais les médianes montrent d'excellents résultats. Les résultats sont comparables pour les indicateurs gaussiens et non-gaussiens. Même avec une ébauche σ^b très éloignée de la vraie source, l'indicateur semi-gaussien parvient à obtenir des performances excellentes (médiane à 100% en 6h environ contre 3h dans les deux premiers cas).

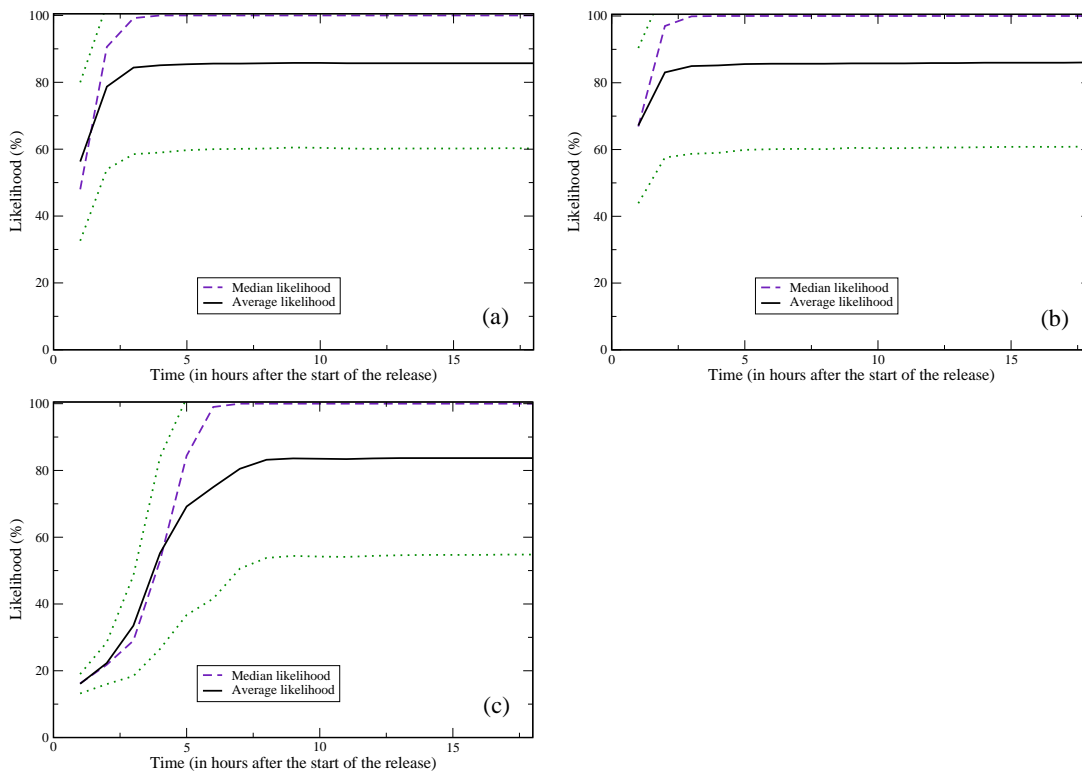


FIG. 4.4: Trace normalisée moyenne de la matrice L (ligne pleine) comparée avec la trace normalisée médiane (ligne tiretée). Les lignes pointillées indiquent les écarts types autour de la moyenne. Panel (a) : Dans le cas de l'indicateur gaussien. Panel (b) : Dans le cas de l'indicateur semi-gaussien, avec l'ébauche σ^b égale à la vraie source. Panel (c) : Dans le cas de l'indicateur semi-gaussien, avec l'ébauche σ^b constante.

4.5 Application II : MUST

4.5.1 Contexte - présentation de l'expérience

MUST (Mock Urban Setting Test) est une expérience de terrain à échelle quasi-réelle, organisée par la Defense Threat Reduction Agency (DTRA) en 2001 et dont l'objectif est d'aider au développement et à la validation des modèles numériques. L'expérience et les analyses des résultats sont décrites en détails dans Bilotft [2001] et Yee and Bilotft [2004].

La canopée urbaine est représentée par un quadrillage de containers, disposés en 12 lignes et 10 colonnes sur un terrain de l'armée dans le désert de l'Utah aux États-Unis (coordonnées : $40^{\circ}12.606' N$, $113^{\circ}10.635' W$). Les dimensions des containers sont de 2.54 m de haut, 12.2 m de long et 2.42 m de large, l'espacement entre les lignes de 12.9 m et entre les colonnes de 7.9 m. La surface totale ainsi formée est d'environ $200 \times 200 \text{ m}^2$. Différentes vues de l'expérience sont visibles sur la figure 4.5.

L'expérience consiste en 63 rejets d'un gaz non réactif, le propylène (C_3H_6) dont la masse volumique est de 1.75 kg.m^{-3} pour une température de $20^{\circ}C$, pour différentes conditions de vent (angle, vitesse, etc.) et différentes positions pour les rejets (à l'intérieur ou à l'extérieur de la canopée, à différentes hauteurs). Les rejets ont été effectués dans des conditions de stratification allant de stable à neutre à l'extérieur de la canopée. Chaque expérience de rejet dure 15 minutes, période pendant laquelle des mesures de concentration sont disponibles. Ces mesures sont réalisées par 40 détecteurs à photo-ionisation (PIDs) placés sur quatre lignes horizontales



(a)



(b)

FIG. 4.5: Photographies de l'expérience MUST (Mock Urban Setting Test). Source : Milliez [2006]

à une hauteur de 1.6 m, ainsi que par 8 PIDs supplémentaires installés sur un mât de mesures à différentes altitudes approximativement entre 1 m et 16 m. Cette configuration est représentée sur la figure 4.6.

4.5.2 Cas d'étude sélectionnés

Yee and Biltoft [2004] ont sélectionné dans 21 séries temporelles de 15 min des périodes de 200 s considérées comme quasi-stationnaires. Ces 21 périodes de 200 s sont supérieures au temps de trajet du traceur sur le domaine, tout en permettant de considérer les conditions météo-

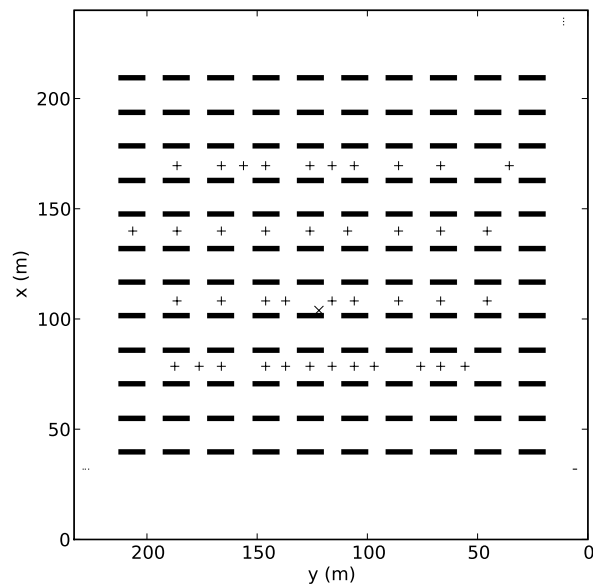


FIG. 4.6: Représentation de la scène de l'expérience avec la localisation des appareils de mesure de concentration (PIDs). Les '+' représentent les PIDs unitaires situés à 1.6 m. La 'x' représente le mât sur lequel sont situés 8 PIDs supplémentaires à 1.1 m, 2.1 m, 4.0 m, 6.0 m, 7.6 m, 9.7 m, 12.5 m et 16.3 m de hauteur.

rologiques comme stationnaires (la variabilité du vent étant trop importante sur des périodes plus longues) et autorisant ainsi des comparaisons avec des simulations numériques stationnaires. Parmi ces périodes, nous avons sélectionné 14 cas, déjà étudiés par Milliez [2006] et Milliez and Carissimo [2007] avec MERCURE (ancienne version de Code_Saturne que nous allons utiliser).

Dans le tableau 4.1 sont indiqués pour chacun des cas étudiés, l'heure à laquelle commence le test, l'angle d'incidence α_4 et la vitesse S_4 du vent en amont de la canopée et à la hauteur $z = 4$ m, l'énergie turbulente k_4 à la hauteur $z = 4$ m au niveau du mât central (recalculée par Milliez [2006]), le débit Q_s du rejet, la position de la source repérée sur la figure 4.7 ainsi que la hauteur z_s du rejet.

TAB. 4.1: Présentation des 14 cas sélectionnés : α_4 et S_4 sont respectivement la vitesse et la direction d'incidence (par rapport à l'axe des x) du vent moyen, au niveau situé à 4 m sur un mât en amont de la canopée, k_4 est l'énergie cinétique turbulente à une hauteur de 4 m au niveau du mât central. Q est le débit du rejet, "Localisation de la source" est la position de la source repérée sur la figure 4.7 et z_s sa hauteur.

Cas	Début de rejet (heure locale)	α_4 (deg)	S_4 (m.s ⁻¹)	k_4 (m ² .s ⁻²)	Q_s (L.min ⁻¹)	Localisation de la source	z_s (m)
2640138	0145	17	2.35	0.258	175	19	0.15
2640246	0251	30	2.01	0.292	200	21	0.15
2671852	1852	-49	3.06	0.379	200	25	0.15
2672213	2213	30	2.68	0.366	200	22	1.8
2681829	1830	-41	7.93	4.263	225	29	1.8
2681849	1849	-41	7.26	2.776	225	29	0.15
2682256	2256	-42	5.02	1.540	225	32	0.15
2682320	2321	-39	4.55	1.032	225	26	2.6
2682353	2354	-47	4.49	1.056	225	33	5.2
2692054	2055	39	3.34	0.483	225	37	1.3
2692131	2132	39	4.00	0.677	225	35	1.3
2692157	2158	43	2.98	0.510	225	36	2.6
2692223	2224	26	2.63	0.468	225	34	1.3
2692250	2251	36	3.38	0.577	225	37	1.3

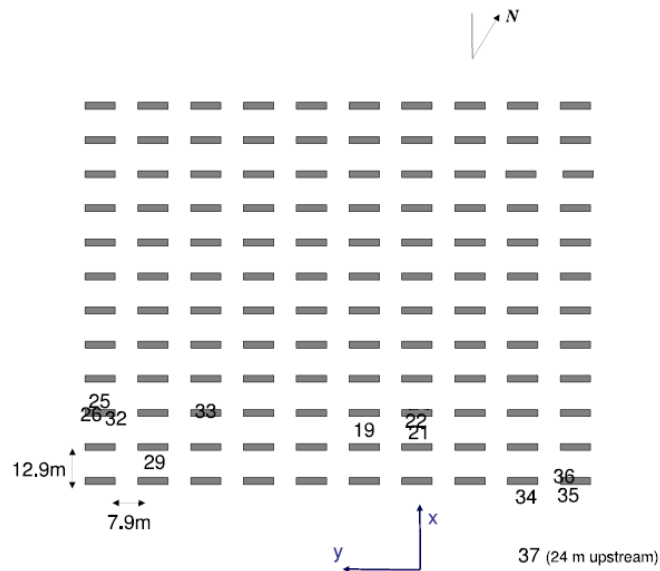


FIG. 4.7: Représentation de la scène de l'expérience avec la localisation des sources utilisées dans les cas sélectionnés. D'après Milliez [2006]

4.5.3 Modèles de dispersion

Le premier modèle de dispersion envisagé est un modèle simple de panache gaussien (présenté dans le chapitre 2.3.3). Ce modèle présente l'avantage de faible coût de calcul, la contrepartie étant une précision moindre surtout près des obstacles, qui ne sont pas pris en compte.

Le second modèle utilisé est Code_Saturne, le code de mécanique des fluides (CFD) développé par EDF R&D et dont le module atmosphérique est développé par le CEREAA. Le domaine de simulation maillé est représenté sur la figure 4.8. Le maillage, proposé par Milliez [2006], est un maillage non-structuré constitué d'environ 800 000 mailles hexaédriques. La résolution horizontale varie de 4 m (loin des obstacles) à 0.6 m (à proximité des containers). La résolution verticale varie de 0.3 m près du sol à 4 m vers le haut du domaine située à une hauteur de 35 m.

Modélisation de l'écoulement par Code_Saturne

Les variables caractérisant l'écoulement (champ de vent, viscosité turbulente) sont calculées par résolution des équations 2.51, 2.52, 2.53, 2.55, 2.56 et 2.57 détaillées dans le chapitre 2.3.4.

Les conditions initiales et aux limites aux frontières supérieures et latérales rendent compte de l'état de référence grande échelle (en amont des perturbations dues aux obstacles).

Pour la frontière supérieure, nous utilisons des conditions de symétrie (flux nul imposé pour toutes les grandeurs moyennes).

Les conditions aux limites latérales dépendent de la direction du vent :

- si la vitesse est sortante, les valeurs sont imposées par l'advection pour toutes les grandeurs transportées. La vitesse est ensuite corrigée pour assurer la conservation du débit.
- si la vitesse est entrante, des conditions de Dirichlet sont utilisées : les valeurs sont fixées pour toutes les grandeurs transportées.

Pour modéliser les profils de vitesse, d'énergie cinétique turbulente et de dissipation de l'écoulement amont, nous utilisons les méthodes de Milliez and Carissimo [2007].

Modélisation de la dispersion dans Code_Saturne

Une fois calculés, les champs qui décrivent l'écoulement sont stockés et vont servir comme données d'entrée au modèle pour simuler la dispersion.

Pour modéliser la dispersion du panache avec le modèle direct, ils sont utilisés sans modification pour résoudre l'équation de dispersion directe (en stationnaire) :

$$u_j \frac{\partial c}{\partial x_j} = \frac{\partial}{\partial x_j} \left[\left(D_m + \frac{\nu_t}{Sc_t} \right) \frac{\partial c}{\partial x_j} \right] + \sigma, \quad (4.34)$$

où D_m est le coefficient de diffusion moléculaire du traceur, ν_t est la viscosité turbulente calculé par Code_Saturne, Sc_t est le nombre de Schmidt turbulent et σ le terme source.

Pour modéliser la dispersion du rétro-panache issu de l'appareil de mesure i , le champ de vent est inversé (changement de signe) alors que les autres champs, décrivant la turbulence, sont conservés, et le modèle calcule la concentration dans le rétro-panache en résolvant l'équation

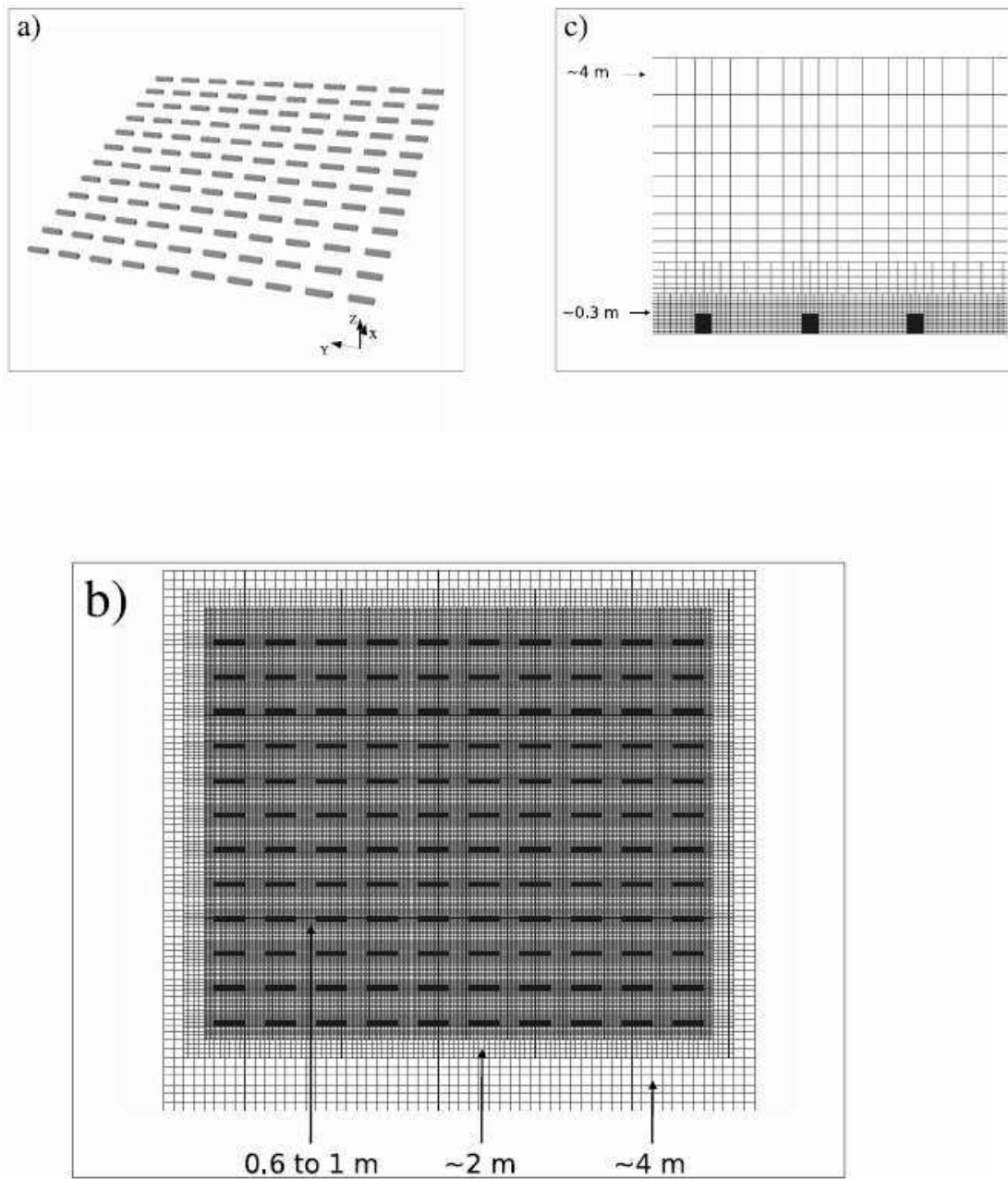


FIG. 4.8: Domaine de simulation et son maillage. (a) Domaine tridimensionnel (b) Coupe horizontale (c) Coupe verticale. Source : Milliez [2006]

de dispersion adjointe :

$$-u_j \frac{\partial c_i^*}{\partial x_j} = \frac{\partial}{\partial x_j} \left[\left(D_m + \frac{\nu_t}{Sc_t} \right) \frac{\partial c_i^*}{\partial x_j} \right] + \pi_i. \quad (4.35)$$

Cette procédure nécessite une phase préliminaire de vérification de la dualité, car le modèle adjoint utilisé est seulement une approximation du modèle adjoint de Code_Saturne. En effet le modèle (numérique) Code_Saturne utilisé pour résoudre l'équation 4.35 n'est pas le modèle adjoint de Code_Saturne utilisé pour résoudre l'équation 4.34, quand bien même les modèles *continus* décrits par ces deux équations sont bien adjoints l'un de l'autre. L'explication vient de

l'existence d'un certain nombre de phénomènes numériques, comme l'utilisation de limiteurs de flux dans les schémas numériques de résolution des équations différentielles. Le test de dualité est réalisé en simulant la dispersion de plusieurs traceurs et en vérifiant la suffisante proximité entre des observations estimées par le modèle direct et par le modèle adjoint. Les résultats de ce test agglomérés pour les 14 cas considérés sont visibles sur la figure 4.9 : un coefficient de corrélation de 0.9991 est obtenu, pour une droite de coefficient directeur 1.001 et d'ordonnée à l'origine de -0.0005, ce qui montre que le modèle adjoint utilisé est une bonne approximation du modèle adjoint exact.

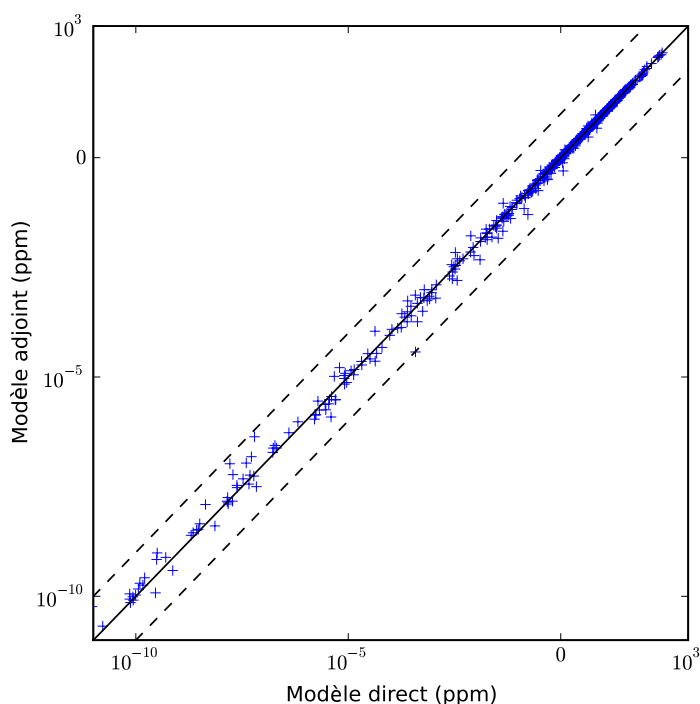


FIG. 4.9: Test de dualité : Comparaison entre des observations simulées par le modèle direct Code_Saturne (en abscisses) et les mêmes observations calculées par le modèle adjoint approché (en ordonnées).

4.5.4 Reconstruction d'une source ponctuelle à partir des données réelles

Performances du modèle direct Code_Saturne

Avant de tenter la reconstruction de source par modélisation inverse, il convient d'estimer les performances du modèle de dispersion. Pour cela, les 14 cas sont simulés en donnant au modèle toutes les informations sur la vraie source. Les performances sont évaluées à l'aide d'indicateurs statistiques classiques adaptés à la dispersion atmosphérique et en particulier aux situations accidentelles. Pour certains d'entre eux des valeurs critiques, définissant un modèle *acceptable*, ont été proposées par Chang and Hanna [2004]. On comparera ainsi les moyennes des observations (notées $\bar{\mu}$) avec les moyennes des observations simulées par le modèle (notées $\bar{\mu}^{\text{simu}}$), ainsi que le coefficient de corrélation de Pearson (noté R) de ces deux ensembles. La

racine carrée de l'erreur quadratique moyenne (RMSE) est définie par :

$$\text{RMSE} = \sqrt{\frac{1}{d} \sum_{i=1}^d (\mu_i - \mu_i^{\text{simu}})^2}. \quad (4.36)$$

D'autres indicateurs sont généralement utilisés en contexte accidentel. Il s'agit notamment du biais moyen géométrique (MG) :

$$\text{MG} = \exp\left(\overline{\ln(\boldsymbol{\mu}) - \ln(\boldsymbol{\mu}^{\text{simu}})}\right), \quad (4.37)$$

de la variance géométrique (VG) :

$$\text{VG} = \exp\left(\frac{1}{d} \sum_{i=1}^d (\ln(\mu_i) - \ln(\mu_i^{\text{simu}}))^2\right), \quad (4.38)$$

de l'indicateurs d'accord FAC2, compris entre 0 et 1 et défini par :

$$\text{FAC2} = \frac{1}{d} \text{Card}\left(\left\{\mu_i^{\text{simu}} / \frac{1}{2}\mu_i \leq \mu_i^{\text{simu}} \leq 2\mu_i\right\}\right), \quad (4.39)$$

de la figure de mérite, également comprise entre 0 et 1 et définie par :

$$\text{FM} = \frac{\sum_{i=1}^d \min(\mu_i, \mu_i^{\text{simu}})}{\sum_{i=1}^d \max(\mu_i, \mu_i^{\text{simu}})}. \quad (4.40)$$

De plus, afin de pouvoir se comparer à d'autres études, notamment Milliez and Carissimo [2007], l'erreur quadratique normalisée (NMSE) sera également évaluée :

$$\text{NMSE} = \frac{\text{RMSE}^2}{\overline{\boldsymbol{\mu}\boldsymbol{\mu}^{\text{simu}}}}, \quad (4.41)$$

ainsi que le biais fractionnel (FB) :

$$\text{FB} = 2 \frac{\overline{\boldsymbol{\mu} - \boldsymbol{\mu}^{\text{simu}}}}{\overline{\boldsymbol{\mu} + \boldsymbol{\mu}^{\text{simu}}}}, \quad (4.42)$$

bien que ces indicateurs aient une tendance naturelle à privilégier les modèles qui sur-estiment les concentrations et doivent être manipulés avec attention.

Un modèle *parfait* obtiendrait des valeurs de RMSE, NMSE et FB nulles et de R, MG, VG, FAC2 et FM égales à 1. Chang and Hanna [2004] ont proposé les valeurs limites suivantes pour un modèle "acceptable" :

$$\text{NMSE} < 4, -0.3 < \text{FB} < 0.3, 0.7 < \text{MG} < 1.3, \text{VG} < 1.6, \text{FAC2} > 0.5. \quad (4.43)$$

Les indicateurs relatifs aux 14 cas simulés avec Code_Saturne sont reportés dans le tableau 4.2. Les valeurs ne répondant pas aux critères préconisés par Chang and Hanna [2004] sont indiquées en rouge. Les résultats obtenus sont très comparables à ceux obtenus par Milliez and Carissimo [2007] avec un modèle très proches. Dans l'ensemble, Code_Saturne présente de bonnes performances (NMSE, FB et FAC2 notamment). Quelques cas présentent néanmoins des indicateurs plus faibles ; il sera intéressant de comparer la qualité de la reconstruction de sources dans ces cas-là et dans ceux présentant de très bonnes performances directes.

Dans le tableau 4.3 sont indiquées les valeurs moyennes des indicateurs sur les 14 cas simulés avec Code_Saturne et avec le modèle de panache gaussien. Il est évident que Code_Saturne, qui prend en compte l'influence des obstacles sur l'écoulement, obtient de meilleurs scores. On peut s'interroger sur les conséquences de ces écarts de performance dans la modélisation inverse, et dans la précision de la localisation de la source en utilisant Code_Saturne et le modèle de panache gaussien.

TAB. 4.2: Scores statistiques des simulations directes par le modèle Code_Saturne sur les 14 cas d'étude. Pour les cinq dernières colonnes, les valeurs dépassant les critères de Chang and Hanna [2004] sont indiquées en rouge.

Cas	$\bar{\mu}$ (ppm)	$\bar{\mu}^{\text{simu}}$	R (ppm)	RMSE	FM (ppm)	NMSE	FB	MG	VG	FAC2 (%)
2640138	3.26	4.15	0.87	4.60	0.53	1.56	-0.24	1.38	8.91	51.1
2640246	3.26	3.15	0.89	2.50	0.63	0.61	0.03	1.10	6.55	63.8
2671852	2.43	3.14	0.97	2.44	0.67	0.78	-0.26	0.92	1.54	74.5
2672213	2.70	2.96	0.91	3.17	0.64	1.26	-0.09	1.47	4.36	83.0
2681829	0.77	0.69	0.95	0.29	0.76	0.15	0.12	1.28	1.29	80.9
2681849	0.76	0.68	0.85	0.51	0.67	0.50	0.11	1.32	1.66	78.7
2682256	1.55	3.03	0.97	5.49	0.46	6.40	-0.64	0.84	2.92	66.0
2682320	1.06	1.55	0.91	2.36	0.54	3.40	-0.38	1.0	1.80	66.0
2682353	1.07	1.33	0.91	0.95	0.69	0.64	-0.21	1.01	1.31	93.6
2692054	1.20	1.29	0.98	0.34	0.84	0.07	-0.07	1.07	1.33	91.5
2692131	0.94	1.13	0.96	0.61	0.76	0.35	-0.18	1.03	1.60	87.2
2692157	1.24	1.28	0.97	0.42	0.81	0.11	-0.03	1.00	1.41	91.5
2692223	1.13	1.20	0.95	1.02	0.67	0.76	-0.06	1.51	1.91	66.0
2692250	1.24	1.72	0.92	1.08	0.62	0.55	-0.33	1.02	2.33	74.5

TAB. 4.3: Scores statistiques moyens des 14 simulations directes par le modèle Code_Saturne et par le modèle de panache gaussien comparés aux critères préconisés par Chang and Hanna [2004]

	Code_Saturne	modèle de panache gaussien	valeurs limites [Chang and Hanna, 2004]
$\bar{\mu}$ (ppm)	1.62	1.62	
$\bar{\mu}^{\text{simu}}$ (ppm)	1.95	2.70	
R	0.93	0.59	
RMSE (ppm)	1.84	4.58	
FM	0.66	0.39	
NMSE	1.23	4.24	NMSE < 4
FB	-0.16	-0.33	-0.3 < FB < 0.3
MG	1.14	0.72	0.7 < MG < 1.3
VG	2.78	40.41	VG < 1.6
FAC2	0.76	0.45	FAC2 > 0.5

Inversion à l'aide de méthodes non-paramétriques

Le champ d'émission à reconstruire est discrétisé sur une grille en 3 dimensions (les situations étant stationnaires), de résolution horizontale 5m et de 1m de résolution verticale.

La première méthode testée est la méthode "classique" des moindres carrés présentée dans le chapitre 4.2.1. Elle s'appuie sur une hypothèse de semi-gaussianité des erreurs d'ébauche (la source obtenue est donc positive). Les variances des erreurs *a priori* sont estimées par une méthode de L-curve (méthode présentée dans le chapitre 3.6.2). Deux exemples de champs d'émission obtenus sont présentés sur la figure 4.10 (pour les cas 2640138 et 2682256 modélisés avec Code_Saturne). Comme nous l'avions anticipé, ces champs sont relativement étendus, ce qui est un inconvénient pour la reconstruction de sources accidentelles, généralement limitée spatialement. Il est à noter que le débit total retrouvé, c'est-à-dire intégré sur tout le domaine, est relativement proche du débit de la vraie source : $132 \text{ L}\cdot\text{min}^{-1}$ pour le cas 2640138 (soit 75% du vrai débit : $175 \text{ L}\cdot\text{min}^{-1}$) et $154 \text{ L}\cdot\text{min}^{-1}$ pour le cas 2682256 (68% du vrai débit : $225 \text{ L}\cdot\text{min}^{-1}$). Par contre, le débit maximal sur un point de grille est faible (respectivement $6 \text{ L}\cdot\text{min}^{-1}$ et $7 \text{ L}\cdot\text{min}^{-1}$, soit environ 3% du vrai débit pour les deux cas).

La seconde méthode non-paramétrique testée est le schéma de Krysta, présenté dans le chapitre 4.2.2. Ce schéma s'appuie sur des erreurs d'ébauche *a priori* non-gaussiennes. Les résultats obtenus sur les deux mêmes cas que précédemment sont présentés sur la figure 4.11. Comme attendu, les champs d'émission obtenus sont beaucoup moins étendus et la source n'est constituée que de quelques points d'émission. Malheureusement, ces points d'émission sont systématiquement localisés près des capteurs reportant les mesures les plus élevées, ce qui ne permet pas de localiser une source ponctuelle loin des capteurs. Ce comportement a notamment été mis en évidence par Ashbaugh et al. [1985] et Stohl [1998] avec un modèle lagrangien et Issartel [2003] avec un modèle eulérien, et expliqué par Bocquet [2005a] comme étant directement lié à la nature diffusive de la dynamique des écoulements atmosphériques. Au niveau des débits estimés, le débit total est plus faible par rapport à la méthode précédente (respectivement $70 \text{ L}\cdot\text{min}^{-1}$ et $84 \text{ L}\cdot\text{min}^{-1}$) ; à l'opposé les débits maximaux estimés sont plus élevés (respectivement $11 \text{ L}\cdot\text{min}^{-1}$ et $20 \text{ L}\cdot\text{min}^{-1}$).

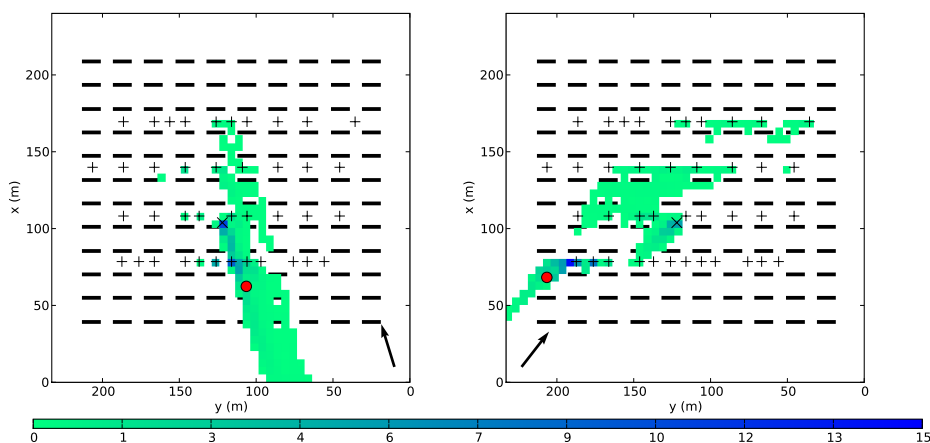


FIG. 4.10: Champ d'émission reconstruit (intégré sur la verticale et en $\text{L}\cdot\text{min}^{-1}$) par la méthode des moindres carrés (erreurs d'ébauche semi-gaussiennes) pour les cas 2640138 (à gauche) et 2682256 (à droite). Les disques rouges représentent la vraie source.

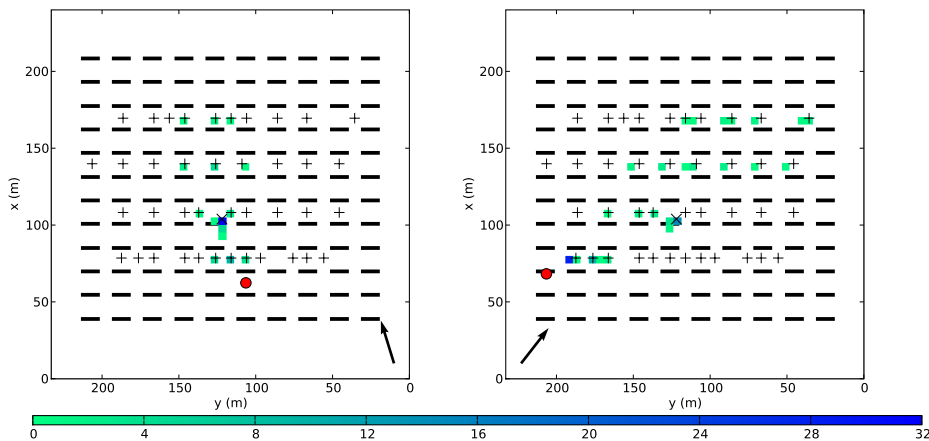


FIG. 4.11: . Champ d'émission reconstruit (intégré sur la verticale et en $L \cdot \text{min}^{-1}$) par la méthode de Krysta (erreurs d'ébauche non-gaussiennes) pour les cas 2640138 (à gauche) et 2682256 (à droite). Les disques rouges représentent la vraie source.

Inversion à l'aide d'une méthode paramétrique

La méthode paramétrique utilisée est la méthode présentée dans le chapitre 4.3.1. La source est supposée ponctuelle et, comme les situations modélisées sont stationnaires, le vecteur de paramètres à estimer se réduit à un vecteur de 4 paramètres : les trois coordonnées spatiales et le débit d'émission, soit $\Theta = (x_s, y_s, z_s, q_s)^T$. Un algorithme de Monte Carlo par chaîne de Markov (MCMC) est implémenté et permet d'obtenir un échantillonnage proche de la distribution *a posteriori* du vecteur de paramètres (voir l'annexe B pour de plus amples détails sur l'algorithme). Pour chacun des 14 cas étudiés, deux chaînes de Markov sont construites : la première avec le modèle de dispersion par panache gaussien et la seconde avec Code_Saturne. Puis l'estimateur moyen de la source est obtenue en prenant l'espérance des paramètres dans chacune des deux chaînes.

Pour évaluer la précision de cet estimateur, la distance entre celui-ci et la vraie source est calculée pour chacun des cas et pour les deux modèles de dispersion considérés. En moyenne sur les 14 cas, l'écart à la vraie source est de 22.6 m pour Code_Saturne, avec une distance minimale de 3.4 m et maximale de 41.7 m, et de 18.7 m pour le modèle de panache gaussien, avec une distance minimale de 9.7 m et maximale de 36.0 m. Les performances entre les deux types de modèles de dispersion sont assez comparables, ce qui peut paraître surprenant étant données les performances des deux modèles en simulation directe. Mais cette précision comparable cache en réalité un comportement assez différent qui se révèle si on considère l'écart à la vraie source dans l'axe du vent mesuré, noté \tilde{x} , et l'écart à la vraie source dans la direction orthogonale au vent mesuré, noté \tilde{y} . Ces valeurs sont reportées pour les 14 cas dans le tableau 4.4.

En ce qui concerne l'écart dans l'axe du vent, l'estimateur obtenu avec Code_Saturne présente des valeurs assez élevées, négatives dans 12 des 14 cas, c'est-à-dire que l'estimateur se trouve en amont de la vraie source. L'estimateur obtenu avec le modèle de panache gaussien présente des valeurs plus faibles. L'explication pourrait venir de la modélisation de la diffusion turbulente dans Code_Saturne, considérée isotrope, ce qui pourrait amener à sous-estimer la diffusion horizontale. Pour compenser cela, le système a tendance à reculer la source, afin que le nuage de polluant impacte les capteurs adéquats.

À l'opposé, l'estimateur obtenu avec Code_Saturne présente une très bonne précision dans

TAB. 4.4: Distances entre les estimateurs et la vraie source et comparaison des débits d'émission estimés, en utilisant Code_Saturne (CS) et un modèle de panache gaussien (PG).

Cas	Distance à la vraie source		Déviation dans l'axe du vent		Déviation orthogonale au vent		Débit d'émission		
	D^{CS} (m)	D^{PG} (m)	\tilde{x}^{CS} (m)	\tilde{x}^{PG} (m)	\tilde{y}^{CS} (m)	\tilde{y}^{PG} (m)	Q^{true} (L.min ⁻¹)	Q^{CS} (L.min ⁻¹)	Q^{PG} (L.min ⁻¹)
2640138	28.6	12.6	-30.2	-11.2	-0.4	5.5	175.0	149.7	147.3
2640246	22.1	18.1	-22.2	-11.1	1.6	14.1	200.0	277.3	292.1
2671852	17.5	16.0	-17.5	-12.4	-6.0	-10.8	200.0	213.0	312.78
2672213	39.7	36.0	-38.9	-35.9	-9.0	2.0	200.0	331.0	253.9
2681829	31.6	25.6	-31.7	-13.3	-3.0	-21.7	225.0	374.4	468.6
2681849	25.6	25.4	-25.1	-11.6	-2.2	-22.4	225.0	348.1	429.4
2682256	7.2	15.0	-6.9	-10.2	-2.0	-10.7	225.0	210.5	360.9
2682320	25.8	9.7	-25.8	-0.9	-0.2	-7.9	225.0	255.2	322.6
2682353	41.7	17.1	-41.9	-13.7	3.8	-8.7	225.0	286.9	275.1
2692054	2.8	35.24	2.3	33.4	0.5	5.7	225.0	206.3	231.9
2692131	21.5	13.8	-21.6	-0.8	1.4	11.2	225.0	227.2	283.5
2692157	13.8	9.9	-13.0	8.3	4.4	4.1	225.0	257.8	163.9
2692223	30.9	7.8	-30.0	-2.8	7.2	0.3	225.0	259.0	160.0
2692250	3.4	19.0	0.4	11.1	-3.3	14.8	225.0	153.5	343.9

la direction orthogonale au vent, alors que celui obtenu avec le modèle de panache gaussien présente un écart systématique vers *l'intérieur* du vent. Cet écart systématique s'explique par le fait que le modèle de panache gaussien ne prend pas en compte l'influence des containers sur l'écoulement et en particulier la déviation du nuage de polluant, mise en avant notamment dans Milliez and Carissimo [2007] et illustrée sur la figure 4.12.

Les débits d'émission retrouvés sont également comparés avec ceux de la vraie source (tableau 4.4). On note de très bonnes performances générales, en particulier des estimateurs issus de la modélisation avec Code_Saturne. On peut remarquer également qu'il ne semble pas y avoir de lien évident entre la performance du modèle direct, évaluée au début du chapitre, et la précision de l'estimateur. Ainsi le modèle de panache gaussien, qui obtenait en simulation directe des performances moins bonnes que Code_Saturne, conduit à des précisions comparables dans la localisation des sources. De même, avec Code_Saturne, le cas 2682256 présentait des performances directes relativement faibles (par rapport aux autres cas) alors que la source est ici retrouvée à 7 m près. Au contraire, pour le cas 2681829, le modèle direct Code_Saturne présentait d'excellents scores alors que la source est ici localisée avec une erreur d'environ 32 m.

Les hauteurs de source z_s vraies et estimées sont reportées dans le tableau 4.5. En moyenne sur les 14 cas, l'écart vertical en valeur absolue entre la vraie source et l'estimateur reconstruit est de 1.44m pour Code_Saturne et 4.27m pour le modèle de panache gaussien. Les perturbations de l'écoulement par les containers, qui augmentent le mélange près du sol, ne sont pas

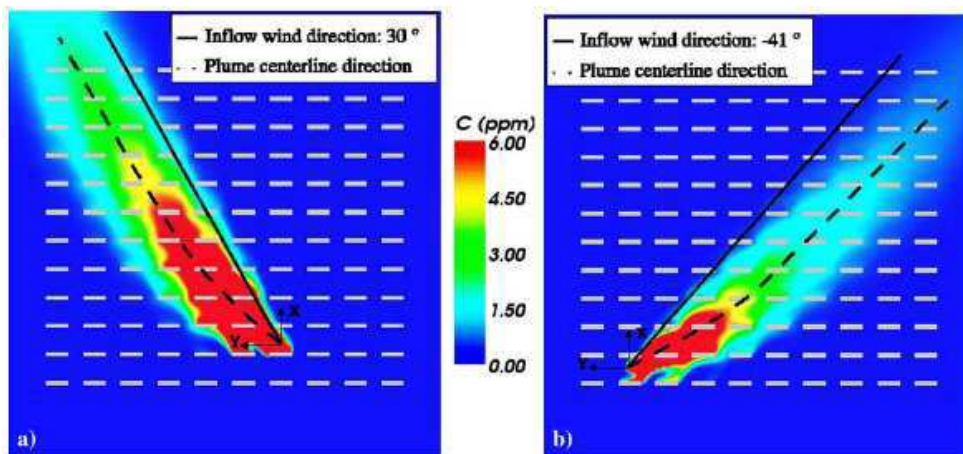


FIG. 4.12: Illustration de la déviation du nuage de polluant par les containers dans les cas 2640246 et 2681829. La ligne pleine représente l'axe du vent mesuré, la ligne tiretée l'axe du nuage de polluant modélisé avec Code_Saturne. Source : Milliez and Carissimo [2007]

TAB. 4.5: Comparaison entre les hauteurs des vraies sources et estimées en utilisant Code_Saturne (CS) et un modèle de panache gaussien (PG).

Cas	z_s^{true}	z_s^{CS}	z_s^{PG}
40138	0.15	1.88	2.36
40246	0.15	1.66	1.91
71852	0.15	1.94	3.82
72213	1.8	1.22	1.81
81829	1.8	2.65	4.51
81849	0.15	2.35	3.60
82256	0.15	1.8	2.07
82320	2.6	4.03	8.14
82353	5.2	6.88	10.73
92054	1.3	2.94	11.05
92131	1.3	3.19	9.20
92157	2.6	2.85	6.17
92223	1.3	3.60	8.80
92250	1.3	2.07	5.61

modélisées par le modèle de panache gaussien. Ce manque est alors compensé par l'algorithme par la sur-élévation de la source estimée.

Sur les panels de gauche des figures 4.13 et 4.14 sont représentées la vraie source et les estimateurs moyens des sources reconstruites, projetés sur le plan horizontal, pour les cas 2640138 et 2682256 (les autres cas sont visibles en annexe C). L'écart systématique dans la direction orthogonale au vent de l'estimateur issu du modèle de panache gaussien est clairement visible, ainsi que l'écart de l'estimateur issu de la modélisation avec Code_Saturne en amont de la vraie source.

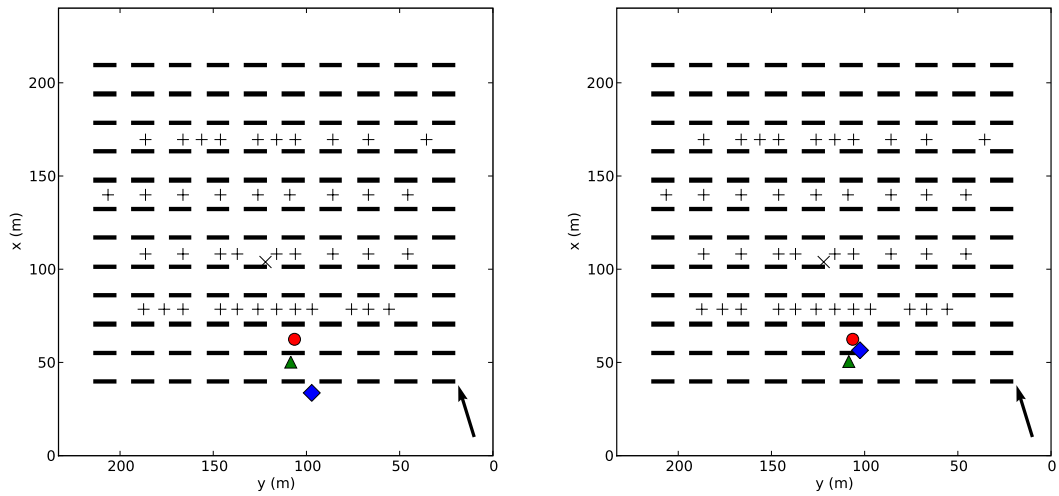


FIG. 4.13: Localisation horizontale de la vraie source et des sources reconstruites par méthode paramétrique pour le cas 2640138. Disques rouges : vraie source. Triangles verts : Modèle de panache gaussien. Diamants bleus : Code_Saturne. La flèche représente la direction du vent en amont des containers. Sur la figure de gauche, l'estimateur de la source reconstruite est l'espérance de la source dans la chaîne de Markov. Sur la figure de droite, l'estimateur est le mode des distributions marginales *a posteriori*.

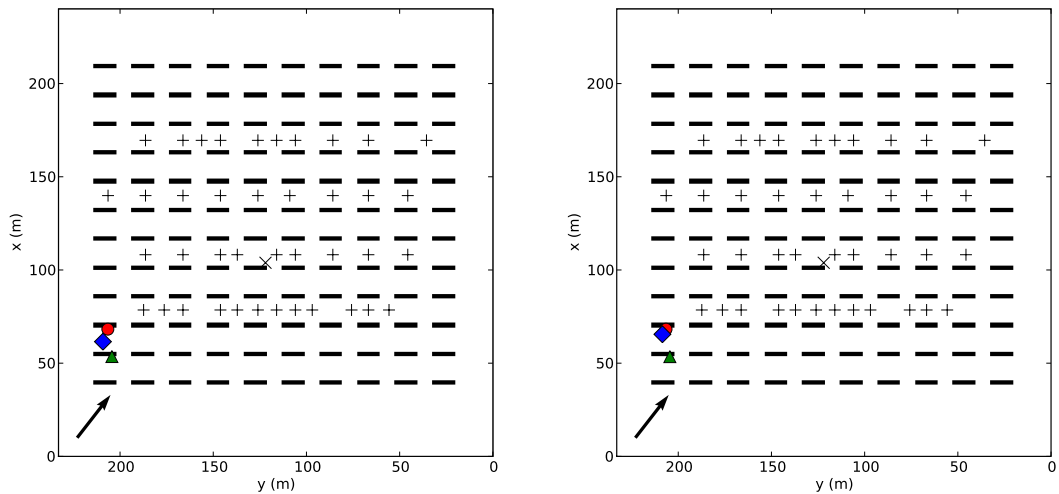


FIG. 4.14: Localisation horizontale de la vraie source et des sources reconstruites par méthode paramétrique pour le cas 2682256. Disques rouges : vraie source. Triangles verts : Modèle de panache gaussien. Diamants bleus : Code_Saturne. La flèche représente la direction du vent en amont des containers. Sur la figure de gauche, l'estimateur de la source reconstruite est l'espérance de la source dans la chaîne de Markov. Sur la figure de droite, l'estimateur est le mode des distributions marginales *a posteriori*.

L'intérêt d'un algorithme MCMC par rapport à d'autres méthodes d'optimisation, telles que le *recuit simulé* par exemple, est qu'il permet d'obtenir non seulement un estimateur du vecteur des paramètres, mais également les statistiques complètes de ce vecteur estimées par la chaîne de Markov. Il est par exemple possible de calculer pour chacun des paramètres composant la source la fonction de densité de probabilité marginale correspondante. Sur la figure 4.15 sont représentées les fonctions de densité de probabilité des quatre paramètres x_s, y_s, z_s, q_s et des deux écarts à la vraie source : dans l'axe du vent \tilde{x} et dans la direction orthogonale au vent \tilde{y} , pour le cas 2640138.

Dans tous les cas (cf. annexe C), il est intéressant de noter que la distribution de l'écart dans la direction orthogonale au vent est toujours plus fine (quelques mètres) que celle de l'écart dans l'axe du vent (jusqu'à plusieurs dizaines de mètres). Ce type de performance est en accord avec d'autres études de méthodes paramétriques [Keats et al., 2007; Yee et al., 2008]. Dans le cas du modèle de panache gaussien, les fonctions de densité de probabilité sont assez proches de gaussiennes et l'estimateur utilisé rend assez bien compte des statistiques des paramètres. Par contre, dans le cas de Code_Saturne, les distributions *a posteriori* peuvent être fortement asymétriques, multi-modales et/ou avec des queues de distributions importantes. C'est par exemple le cas sur la figure présentée pour le cas 2640138. Si la distribution de l'écart dans la direction orthogonale au vent est assez précise, celle de l'écart dans l'axe du vent semble bi-modale (témoignant de la présence des containers) et assez large en amont de la vraie source. De plus la moyenne est plus éloignée de la vraie source que le mode le plus important.

Nous avons alors considéré un second estimateur pour la source reconstruite : celui qui maximise les densités de probabilité *a posteriori*. Les résultats pour les cas 2640138 et 2682256 sont visibles sur les panels de droite des figures 4.13 et 4.14. Pour les simulations avec le modèle de panache gaussien, les écarts entre ces deux estimateurs sont faibles. Pour les simulations avec Code_Saturne, sur l'ensemble des 14 cas, la distance moyenne à la vraie source est inchangée, les performances meilleures sur certains cas (comme le 2640138 et le 2682256 présentés sur les figures citées ci-dessus) étant compensées par des dégradations sur d'autres cas (par exemple le 2692054 et 2692250 visibles dans l'annexe C).

En conclusion, si l'obtention d'un estimateur reste un objectif de toute méthode de localisation de sources, il semble important de considérer les fonctions de densité de probabilité pour avoir une vision globale de la reconstruction. Les algorithmes MCMC permettent d'avoir accès aux statistiques *a posteriori*, par exemple par l'intermédiaire des pdfs de la figure 4.15. On peut également tracer sur le plan horizontal, en plus des estimateurs considérés, des lignes d'iso-densité qui permettent de rendre compte des possibilités de localisation de la source et des incertitudes associées. Ce type de représentation est proposé sur les figures 4.16 et 4.17. On distingue le caractère non-symétrique des distributions dans le cas de Code_Saturne, conséquence d'un écoulement non-trivial perturbé par les containers. Si les distances à la vraie source sont comparables pour les inversions utilisant Code_Saturne ou le modèle de panache gaussien, on peut remarquer que la vraie source est généralement à l'intérieur de iso-densités issues des inversions avec Code_Saturne, ce qui n'est pas le cas avec le modèle de panache gaussien.

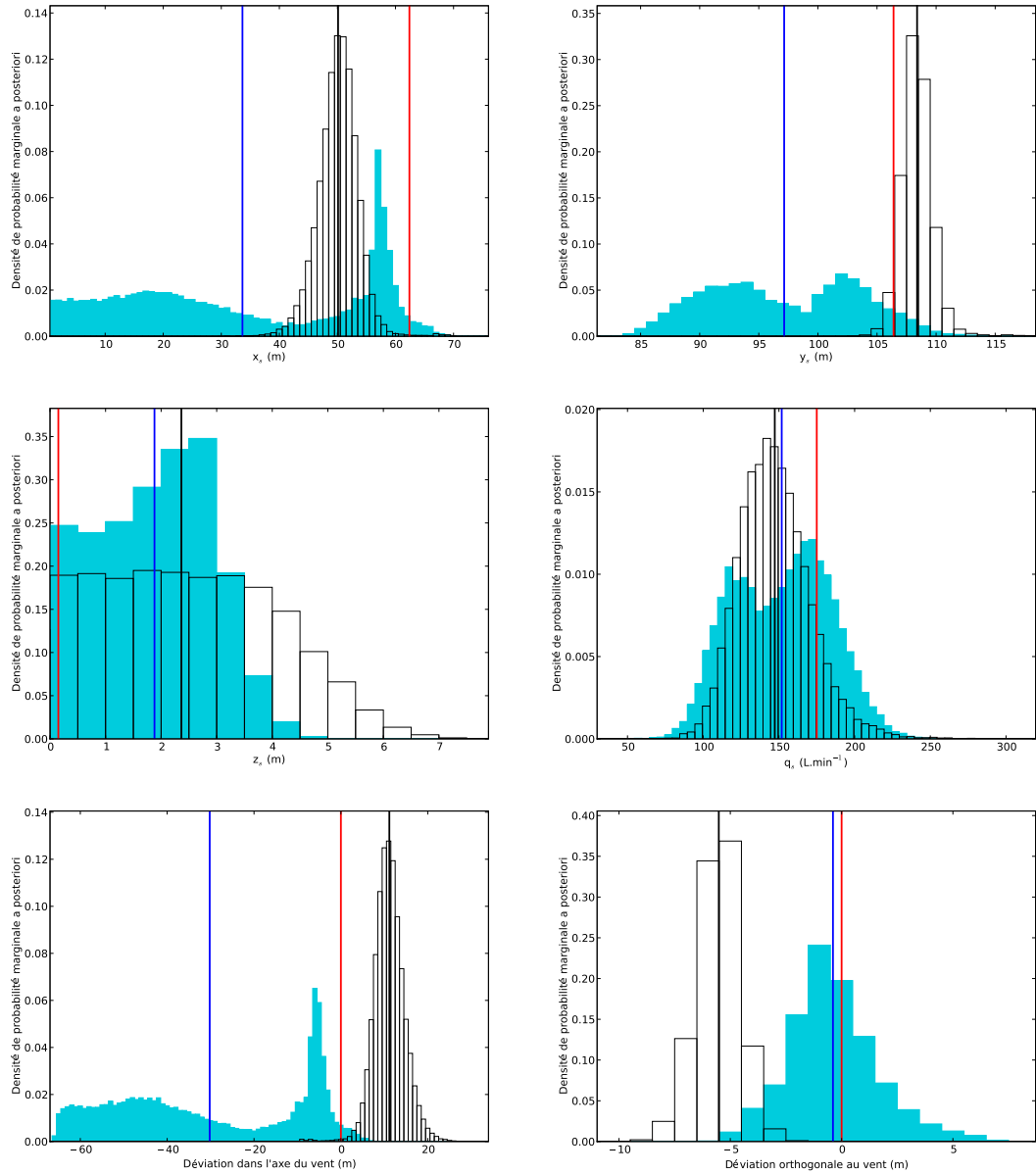


FIG. 4.15: Densité de probabilité des paramètres de la source reconstruite pour le cas 2640138. Dans chaque panel, le trait vertical rouge représente la valeur vraie du paramètre, le trait bleu la moyenne suivant la distribution bleue issue de la chaîne de Markov avec Code_Saturne et le trait noir la moyenne de la distribution noire issue de la chaîne de Markov avec le modèle de panache gaussien.

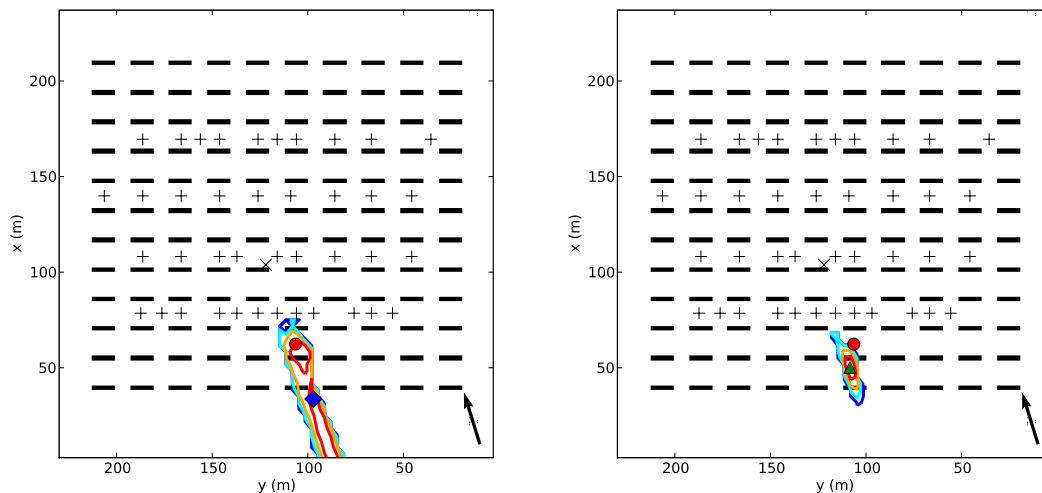


FIG. 4.16: Contours de densité des sources reconstruites dans la chaîne de Markov pour le cas 2640138. Figure de gauche : avec Code_Saturne (diamant bleu : estimateur moyen). Figure de droite : avec le modèle de panache gaussien (triangle vert : estimateur moyen). Disques rouges : vraie source. La flèche représente la direction du vent en amont des containers. Entre deux couleurs successives (rouge, orange, cyan, bleu), la densité de probabilité est divisée par 10.

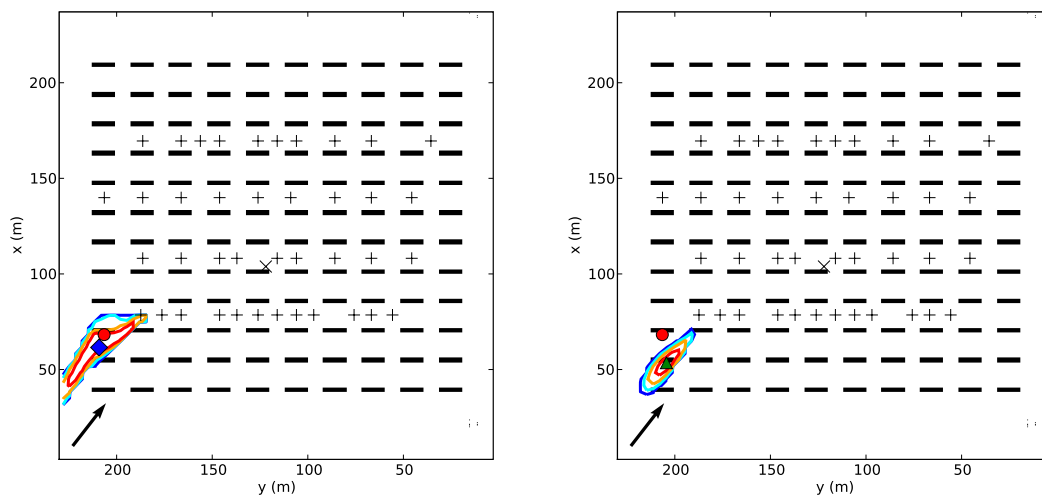


FIG. 4.17: Contours de densité des sources reconstruites dans la chaîne de Markov pour le cas 2682256. Figure de gauche : avec Code_Saturne (diamant bleu : estimateur moyen). Figure de droite : avec le modèle de panache gaussien (triangle vert : estimateur moyen). Disques rouges : vraie source. La flèche représente la direction du vent en amont des containers. Entre deux couleurs successives (rouge, orange, cyan, bleu), la densité de probabilité est divisée par 10.

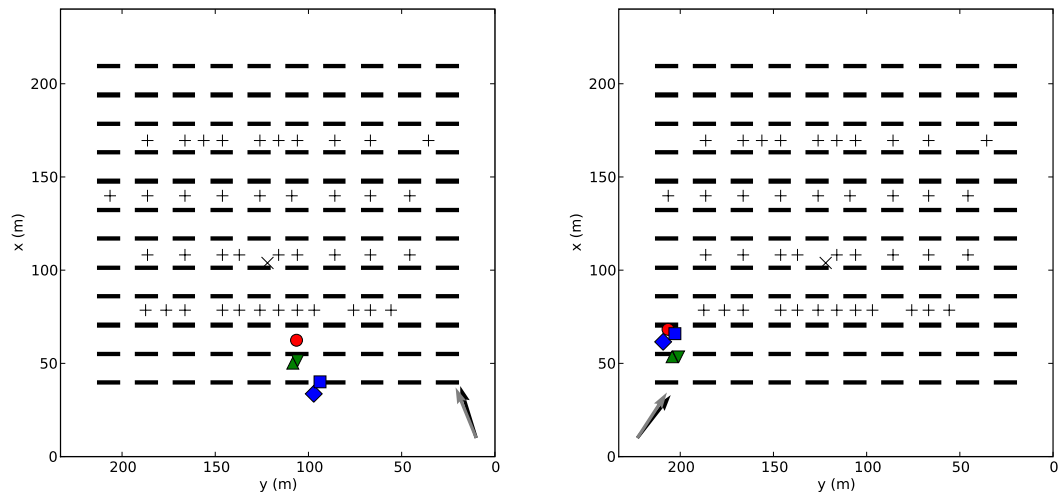


FIG. 4.18: Localisation horizontale de la vraie source et des estimateurs moyens des sources reconstruites par méthode paramétrique pour les cas 2640138 (à gauche) et 2682256 (à droite), avec et sans rotation de $+5^\circ$ du vent amont. Disques rouges : vraie source. Triangles verts : Modèle de panache gaussien sans rotation du vent. Triangles renversés verts : Modèle de panache gaussien avec rotation du vent amont (distance entre triangle et triangle renversé : 2.9 m sur la figure de gauche et 3.5 m sur la figure de droite). Diamants bleus : Code_Saturne sans rotation du vent amont. Carrés bleus : Code_Saturne avec rotation du vent amont (distance entre diamant et carré : 7.3 m sur la figure de gauche et 7.7 m sur la figure de droite). La flèche noire représente la direction du vent mesuré en amont des containers, La flèche grise le vent après rotation de $+5^\circ$.

Influence du vent mesuré en amont

La précision de la mesure de vent en amont des containers constitue une des principales sources d'erreurs possibles dans le système. Pour évaluer la sensibilité de l'inversion à ce paramètre d'entrée, nous avons à nouveau réalisé les reconstructions de source en ayant au préalable fait pivoter le vent amont de $\pm 5^\circ$. L'objectif n'est pas d'améliorer la reconstruction de la source, mais d'étudier, pour chacun des 14 cas et pour les deux modèles de dispersion, l'écart entre les sources estimées avec ces trois valeurs de vent (vent mesuré, rotation de $+5^\circ$ et rotation de -5°).

Pour le modèle de panache gaussien, la distance moyenne entre les sources reconstruites avec vent modifié (dans l'un ou l'autre sens) et les sources reconstruites avec le vent mesuré est de 5.8 m, avec un écart maximal de 10.1 m. Pour le modèle Code_Saturne, l'écart moyen est de 8.3 m et maximal de 10.9 m. Comparées aux écarts trouvés dans le paragraphe précédent entre la reconstruction et la vraie source, de l'ordre de 20 m, ces écarts ne sont pas négligeables et montrent que l'inversion est sensible à la précision du vent mesuré en entrée du domaine d'étude. Les résultats de ces tests pour les cas 2640138 et 2682256, avec rotation du vent amont de $+5^\circ$, sont reportés sur la figure 4.18.

Chapitre 5

Conclusion et perspectives

5.1 Conclusion

Dans le contexte d'une pollution atmosphérique accidentelle ou terroriste, des modèles numériques sont utilisés pour déterminer les zones contaminées. La précision de ces modèles dépend notamment de la précision des données d'entrée, parmi lesquelles le terme source est particulièrement incertain dans ce contexte. Dans cette thèse, différentes méthodes de modélisation inverse ont été proposées pour reconstruire une source accidentelle de polluant atmosphérique. Dans une première partie, différents aspects de la dispersion atmosphérique ont été présentés, depuis les processus mis en jeu jusqu'à leur modélisation, de l'échelle urbaine à l'échelle inter-continentale. Dans une deuxième partie, le cadre général bayésien de la modélisation inverse a été présenté et des méthodes ont été proposées pour l'estimation de sources. Dans une troisième partie, des méthodes spécifiques ont été proposées pour localiser la source de polluant, dans le cas où le lieu du rejet n'est pas connu. Plusieurs études d'application, à différentes échelles et utilisant différents modèles, ont de plus été proposées dans ce manuscrit.

5.1.1 La dispersion atmosphérique de polluant : des processus à la modélisation numérique

Le chapitre 2 a été consacré à la dispersion atmosphérique de polluant dans le cadre accidentel. Dans un premier temps, les processus physiques régissant cette dispersion ont été présentés :

- le transport sous l'influence des circulations d'air. Pour prendre en compte la turbulence de la couche limite atmosphérique, ce transport est décomposé en un terme d'advection par le champ moyen des vitesses de vent (à l'échelle du modèle) et en un terme de diffusion qui rend compte des effets de la turbulence aux échelles plus petites.
- les transformations physico-chimiques. Ces réactions peuvent être complexes (chimie des aérosols, multiphasique, ...). Aux échelles de temps considérées et étant donnés les polluants mis en jeu, et afin d'éviter l'utilisation de modèles trop coûteux, seule une éventuelle filiation radioactive est prise en compte.
- Les processus de perte par dépôts au sol : dépôt sec et lessivage par les précipitations. Dans la mesure du possible et pour les mêmes raisons, des paramétrisations simples de ces processus sont recherchées.

Ces processus sont formalisés par une équation d'advection-diffusion-chimie décrivant l'évolution du champ de concentration du polluant considéré. Suivant l'échelle de modélisation, plusieurs modèles sont alors envisageables et les plus utilisés en contexte accidentel, ainsi que leur formulation adjointe, ont été présentés dans ce chapitre.

Dans ce travail de thèse, trois modèles particuliers ont été utilisés :

- à l'échelle inter-continentale et méso-échelle, un modèle eulérien POLAIR3D, développé par le CEREAs au sein de la plateforme de modélisation de qualité de l'air POLYPHEMUS. Ce modèle, déterministe, utilise une discrétisation du domaine spatio-temporel de simulation pour calculer l'évolution de la concentration en polluant. Des paramétrisations adaptées au contexte accidentel et à une utilisation intensive des modèles ont été sélectionnées pour modéliser les différents processus mis en jeu, par exemple les dépôts.
- à l'échelle locale, un modèle de panache gaussien, simple à implémenter de par sa formule analytique et très rapide d'exécution, a été utilisé. La contrepartie est une mauvaise prise en compte de l'influence des obstacles sur l'écoulement de l'air et la turbulence mise en jeu.
- enfin à l'échelle locale également, un modèle de dynamique des fluides, Code_Saturne, développé par le CEREAs et EDF R&D, permet de prendre en compte explicitement la présence des obstacles et de calculer les caractéristiques d'un écoulement complexe, au prix d'un coût de calcul nettement supérieur.

5.1.2 Estimation de source accidentelle par modélisation inverse

Les modèles numériques présentés dans le chapitre 2 et permettant de simuler la dispersion de polluants sont incertains. Ces incertitudes peuvent être liées aux paramétrisations du modèle, par exemple des processus de dépôt, ou aux données d'entrée, telles que les champs météorologiques. En situation accidentelle, la principale source d'incertitude est la mauvaise connaissance du terme source de polluant. Afin d'estimer le terme source, des techniques de modélisation inverse, utilisant à la fois modèle numérique et observations, peuvent se révéler très efficaces. Les méthodes proposées dans le chapitre 3, mais aussi dans le chapitre 4, se placent toutes dans un cadre bayésien, et réalisent ainsi un équilibre objectif entre une information *a priori* et de nouvelles informations issues des observations et du modèle numérique, permettant *in fine* de diminuer l'incertitude associée au terme source.

L'information *a priori* est exprimée par le biais d'hypothèses sur les statistiques des erreurs *a priori* dans le système : erreurs d'observation et erreurs d'ébauche. Nous nous sommes tout d'abord placés dans un cadre statistique relativement classique dans lequel les erreurs *a priori* suivent des statistiques gaussiennes. Ce choix conduit à estimer le terme source par minimisation d'une fonction coût quadratique. En contexte accidentel où le nombre d'observations est généralement faible et ne contraint pas suffisamment le système, cet estimateur, appelé BLUE (Best Linear Unbiased Estimator), peut contenir des termes négatifs. Pour forcer la positivité de la source reconstruite, nous avons donc proposé l'utilisation de statistiques gaussiennes tronquées pour les erreurs d'ébauche. Cette formulation conduit à minimiser la même fonction coût que précédemment, mais sous contrainte de positivité.

La simplicité des fonctions coûts engendrées, associée à la linéarité des modèles numériques de dispersion en situation accidentelle, permettent d'envisager l'implémentation de telles méthodes dans une optique opérationnelle. Dans le chapitre 3.7, nous avons ainsi proposé et évalué, sur des expériences synthétiques à l'échelle européenne, un algorithme semi-automatique

de surveillance d'un parc d'installations nucléaires. Cet algorithme permet de reconstruire le profil temporel d'un rejet de polluant en cas d'accident sur une installation déjà identifiée et de simuler la dispersion du panache radioactif, et ainsi les zones contaminées, aussi bien en analyse qu'en prévision à l'aide des CTM eulériens POLAIR3D, développé par le CEREAA, et SILAM, développé par l'Institut finlandais de météorologie (FMI). Ce travail a été publié en 2011 dans *Atmospheric Environment* [Winiarek et al., 2011].

Les modèles statistiques utilisés pour caractériser les erreurs *a priori* dépendent de paramètres, appelés hyper-paramètres dans ce manuscrit, qu'il convient d'estimer rigoureusement car le résultat de l'inversion y est particulièrement sensible. Nous avons proposé trois méthodes pour réaliser cette estimation : la première méthode repose sur la technique de la L-curve, les deuxième et troisième méthodes reposent sur la maximisation de la vraisemblance des observations. Parmi ces trois méthodes, la plus rapide, appelée ici schéma de Desrozières est seulement une approximation, mais elle conduit à des résultats très comparables aux méthodes exactes quand le nombre d'observations augmente, ce qui est très encourageant dans une optique d'utilisation opérationnelle. De plus, l'estimation des erreurs *a priori* permet aussi d'estimer, par simulation de Monte Carlo, l'incertitude *a posteriori* du terme source reconstruit.

Nous avons appliqué, dans les chapitres 3.8 et 3.9, ces méthodes pour estimer les termes sources de césium-137 et d'iode-131 consécutifs à l'accident de la centrale de Fukushima Daiichi en mars 2011. Dans un premier temps (chapitre 3.8), seules des mesures de concentration d'activité dans l'air ont été utilisées pour l'inversion. Des rejets totaux de 12 PBq, avec une incertitude comprise entre 40 et 100%, ont été estimés pour ^{137}Cs , ce qui est comparable avec d'autres études, utilisant des méthodes et des données différentes. Pour ^{131}I , les rejets totaux estimés sont compris entre 190 et 380 PBq, avec une incertitude comprise entre 40 et 45%. Ce travail a été publié dans *JGR Atmospheres* [Winiarek et al., 2012]. Dans un second temps (chapitre 3.9), différents types de données (concentration d'activité dans l'air, dépôts) ont été utilisés dans la même inversion pour estimer le terme source de ^{137}Cs . L'équilibre entre les différents types de données est réalisé par les méthodes d'estimation des erreurs *a priori*. Les rejets totaux ont alors été estimés entre 12 et 19 PBq, suivant les données et la méthode utilisées, avec une incertitude réduite à environ 15 – 20%. Cette étude a également été publiée dans *Atmospheric Environment* [Winiarek et al., 2014].

5.1.3 Localisation de sources accidentelles

En contexte accidentel ou terroriste, le lieu même du rejet de polluant peut ne pas être connu. Des exemples historiques existent (Tchernobyl, Algésiras) et montrent la nécessité de proposer des méthodes permettant de localiser la source. Différentes méthodes sont présentées dans le chapitre 4. Ces méthodes diffèrent par l'information *a priori* apportée au système sur le terme source, notamment son étendue spatiale souvent limitée.

Dans le contexte déjà évoqué de la surveillance d'un parc d'installations nucléaires, nous avons proposé dans le chapitre 4.4 des indicateurs statistiques bayésiens capables d'identifier, parmi un ensemble de sites suspectés, le responsable d'une pollution radiologique détectée. Nous avons évalué ces indicateurs sur des expériences synthétiques à l'échelle européenne et cette étude constitue la seconde partie de Winiarek et al. [2011].

Dans le chapitre 4.5, nous avons utilisé les données issues de la campagne expérimentale MUST (Mock Urban Setting Test) pour tester les méthodes de localisation de sources. Pour cette étude, à l'échelle locale, deux modèles ont été utilisés : un modèle simple de panache gaussien et un modèle de CFD, Code_Saturne, prenant en compte l'influence des obstacles sur l'écoulement. Pour chacun des 14 cas stationnaires étudiés, 48 capteurs ont permis d'observer un panache de gaz traceur consécutif à un rejet ponctuel. L'objectif est de localiser la source de ce rejet, c'est-à-dire d'estimer ces trois coordonnées spatiales ainsi que son débit d'émission.

Les résultats des différentes méthodes d'inversion, en utilisant l'un ou l'autre modèle de dispersion, ont été analysés et discutés dans ce chapitre. Les méthodes paramétriques, en utilisant Code_Saturne ou le modèle de panache gaussien, ont permis de localiser la source avec une erreur de 20m en moyenne (soit 10% de la taille caractéristique du domaine de simulation). Un décalage systématique dans la direction orthogonale au vent a également été mise en évidence avec le modèle de panache gaussien, du fait de la non prise en compte par celui-ci de la déviation du panache par les obstacles. Il est en outre possible, par des méthodes stochastiques, d'obtenir les statistiques *a posteriori* estimées de la source et ainsi d'estimer l'incertitude de cette localisation.

5.2 Perspectives

5.2.1 Temps de calcul et cadre opérationnel

Pour envisager la mise en place des méthodes proposées dans un cadre opérationnel, il est nécessaire d'obtenir des temps de calcul raisonnables. Les algorithmes d'inversion proposés dans ce manuscrit sont en général peu coûteux et l'étape la plus coûteuse se révèle être le calcul par le modèle de dispersion de la matrice jacobienne ou des rétro-panaches nécessaires.

En ce qui concerne la surveillance d'installations nucléaires à l'échelle régionale ou continentale, une solution envisageable est de calculer la matrice jacobienne en temps réel, en tâche de fond. La mise à jour de la matrice est réalisée au rythme de l'acquisition de nouvelles observations par l'utilisation du modèle de dispersion sur une durée courte. Au fur et à mesure, les observations les plus anciennes et les lignes de la matrice jacobienne correspondantes sont éliminées.

Pour les simulations à l'échelle locale en utilisant un modèle de CFD, il peut être envisageable d'établir en amont une base de données de rétro-panaches correspondant à différentes conditions météorologiques (vent et stabilité). En cas de besoin, les rétro-panaches nécessaires seront estimés par interpolations à partir de cette base de données [Vendel, 2010], ce qui évite l'utilisation, coûteuse, du modèle de CFD en situation d'urgence.

5.2.2 Amélioration des modèles de dispersion

Les résultats de la modélisation inverse sont sensibles aux performances des modèles de dispersion. Ainsi, une des voies d'amélioration des méthodes de modélisation inverse consiste à améliorer les modèles, en gardant si possible des temps de calcul relativement faibles.

À l'échelle régionale et continentale, la précision des champs météorologiques (notamment de la représentation de la turbulence dans la couche limite atmosphérique) et la précision de la modélisation des processus de perte semblent être particulièrement importantes. Pour ce deuxième point, la précision des champs de précipitations, calculées par un modèle météorologique, doit dans un premier temps être améliorée. Si des observations de précipitations au sol sont disponibles, des méthodes d'assimilation de données sont envisageables pour corriger les champs calculés par le modèle. Dans un second temps, les paramétrisations de ces processus dans le modèle de dispersion doivent également être améliorées. Pour cela, des méthodes d'estimation de paramètres par modélisation inverse sont envisageables. Bocquet [2012] a proposé un algorithme 4D-Var pour estimer certains paramètres (coefficient de diffusion verticale, vitesse de dépôt sec) et le terme source pour la modélisation de l'épisode de pollution consécutif à l'accident de Tchernobyl.

À l'échelle locale, l'influence de la valeur du vent amont fourni au modèle en tant que donnée d'entrée est importante : lors de notre étude sur l'expérience MUST, un décalage de 5° en direction a conduit en moyenne à un écart de 10m dans la localisation de la source. On

peut alors envisager des méthodes d'inversion estimant dans le même temps le terme source et la valeur du vent amont. De telles méthodes avec un modèle gaussien ne présentent pas de surcoût, étant donnée la simplicité du modèle, et ont déjà été utilisées [Krysta et al., 2006; Kovalets et al., 2008; Senocak et al., 2008]; avec un modèle de CFD, l'utilisation du modèle tangent linéaire peut être envisagée [Kovalets et al., 2011], mais aussi l'utilisation d'une base de données de rétro-panaches présentée dans le chapitre 5.2.1.

5.2.3 Assimilation des débits de dose γ

Plus le nombre d'observations est important et plus les méthodes proposées sont robustes. Dans cette optique, et dans le cadre de l'estimation de termes sources consécutifs à un accident sur une installation nucléaire, il peut être intéressant de disposer de méthodes capables d'utiliser les mesures de débits de dose γ . En effet, ces données bénéficient d'une distribution spatiale et temporelle très dense et sont donc disponibles en très grand nombre. Elles sont néanmoins difficiles à utiliser dans une inversion car elles rendent compte d'une part de l'influence de plusieurs radionucléides, et d'autre part de l'influence des polluants dans l'air et déjà déposés au sol. Saunier et al. [2013] ont néanmoins proposé une méthode complète pour estimer le terme source de plusieurs radionucléides en utilisant ces données sur le cas de Fukushima. Ils ont notamment montré qu'une dizaine d'espèces pouvaient être prises en compte, mais que l'algorithme pouvait avoir quelques difficultés à équilibrer l'influence de chacun des radionucléides.

L'idée est alors de proposer une méthode capable d'utiliser simultanément tous les types de données disponibles : concentrations d'activité dans l'air de plusieurs espèces, dépôts au sol et débits de doses. L'équilibre entre toutes ces différentes données peut être réalisé avec les méthodes présentées dans cette thèse, mais de nouvelles difficultés sont probables : du fait de la distribution dense des observations disponibles, les corrélations dans les erreurs d'observation ne pourraient plus être négligées, nécessitant l'utilisation de méthodes plus complexes telles que la sélection optimale d'observations, la génération de super-observations, ou l'utilisation d'un modèle de corrélation dans les matrices de covariance des erreurs; de plus par nature les mesures de débits de doses sont très influencées par les dépôts et une meilleure prise en compte des erreurs de modèles pourraient être nécessaires, en plus d'une amélioration de la modélisation des dépôts.

5.2.4 Généralisation pour l'estimation d'émissions non-accidentelles

Les techniques de modélisation inverse sont utilisées depuis de nombreuses années pour estimer les émissions, et plus généralement les flux (émission et puits), de certains polluants en contexte non-accidentel. C'est le cas par exemple du dioxyde de carbone ou du méthane [Bousquet et al., 1999]. Les techniques d'estimation des erreurs présentées dans ce manuscrit peuvent être pour certaines généralisées à ces situations. Une différence importante provient du fait qu'il n'est généralement pas possible de négliger les corrélations dans les erreurs, en particulier pour l'ébauche, et que les matrices de covariance d'erreurs mises en jeu ne sont donc pas diagonales, entraînant notamment des temps de calcul plus longs. De telles méthodes ont été par exemple appliquées pour l'estimation des flux de CO₂ [Wu et al., 2013] ou pour les émissions de composés organiques volatiles [Koohkan et al., 2013] et il serait intéressant de prolonger ces travaux pour proposer des méthodes générales capables de résoudre ces problèmes de grande taille.

5.2.5 Localisation de sources multiples

Il serait intéressant d'évaluer les méthodes de localisation de sources proposées dans le chapitre 4 dans le cas de lieux de rejets multiples. La comparaison entre les méthodes paramétriques et non-paramétriques ainsi que la capacité des méthodes paramétriques proposées (notamment MCMC avec saut réversible) à retrouver le nombre exact de sources et leur localisation doivent être étudiées. Dans le cadre de l'expérience MUST, cette évaluation doit passer par des expériences synthétiques, aucune expérience de rejets simultanés issus de plusieurs points n'ayant été mise en œuvre.

Bibliographie

- Abida, R., Bocquet, M., 2009. Targeting of observations for accidental atmospheric release monitoring. *Atmos. Env.* 43, 6312–6327. 40, 53, 97
- Anderson, J. L., 2007. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A* 59, 210–224. 41
- Ashbaugh, L., Malm, W., Sadeh, W., 1985. A residence time probability analysis of sulfur concentrations at Grand Canyon National Park. *Atmos. Env.* 19, 1263–1270. 107
- Baklanov, B., Sørensen, J., 2001. Parameterisation of radionuclides deposition in atmospheric long-range transport modelling. *Phys. Chem. Earth (B)* 26 (10), 787–799. 52
- Belge, M., Kilmer, M., Miller, E., 2002. Efficient determination of multiple regularization parameters in a generalized l-curve framework. *Inverse Problems* 18, 1161–1183. 48
- Biltoft, C., 2001. Customer report for mock urban setting test. Tech. rep., DPG Document No. WDTC-FR-01-121, West Desert Test Center, U.S. Army Dugway Proving Ground, Dugway, Utah, 58pp. 98
- Bocquet, M., 2005a. Grid resolution dependence in the reconstruction of an atmospheric tracer source. *Nonlinear Processes in Geophys.* 12, 219–234. 107
- Bocquet, M., 2005b. Reconstruction of an atmospheric tracer source using the principle of maximum entropy. I : Theory. *Q. J. Roy. Meteor. Soc.* 131, 2191–2208. 86, 87, 88
- Bocquet, M., 2005c. Reconstruction of an atmospheric tracer source using the principle of maximum entropy. II : Applications. *Q. J. Roy. Meteor. Soc.* 131, 2209–2224. 86, 88
- Bocquet, M., 2012. Parameter field estimation for atmospheric dispersion : Application to the Chernobyl accident using 4D-Var. *Q. J. Roy. Meteor. Soc.* 138, 664–681. 29, 75, 120
- Bornstein, R., 1987. Urban barrier effects on mesoscale and synoptic systems. Preprint 9, Conference on Mesoscale Processes, Amer. Met. Soc., 21–26, August. 21
- Bousquet, P., Ciais, P., Peylin, P., Ramonet, M., Monfray, P., 1999. Inverse modeling of annual atmospheric CO₂ sources and sinks 1. method and control inversion. *J. Geophys. Res.* 104, 26161–26178. 121
- Brandt, J., Christensen, J., Frohn, M., 2002. Modelling transport and deposition of caesium and iodine from chernobyl accident using the dream model. *Atmos. Chem. Phys.* 2, 397–417. 60
- Byrd, R. H., Lu, P., Nocedal, J., 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing* 16, 1190–1208. 44, 53

- Chang, J., Hanna, S., 2004. Air quality model performance evaluation. *Meteorol. Atmos. Phys.* 87, 167–196. 104, 105, 106
- Chapnik, B., Desroziers, G., Rabier, F., Talagrand, O., 2004. Properties and first application of an error-statistics tuning method in variational assimilation. *Q. J. Roy. Meteor. Soc.* 130, 2253–2275. 41
- Chapnik, B., Desroziers, G., Rabier, F., Talagrand, O., 2006. Diagnosis and tuning of observational error in a quasi-operational data assimilation setting. *Q. J. Roy. Meteor. Soc.* 132, 543–565. 41, 50
- Chazette, P., Bocquet, M., Royer, P., Winiarek, V., Raut, J., Labazuy, P., Gouhier, M., Lardier, M., Cariou, J., 2012. Eyjafjallajökull ash concentrations derived from both lidar and modeling. *J. Geophys. Res.* 117, D00U14. 23
- Chino, M., Nakayama, H., Nagai, H., Terada, H., Katata, G., Yamazawa, H., 2011. Preliminary estimation of release amounts of I131 and Cs137 accidentally discharged from the Fukushima Daiichi nuclear power plant into the atmosphere. *Journal of Nuclear Science and Technology* 48, 1129–1134. 58, 68, 70, 78
- Davoine, X., Bocquet, M., 2007. Inverse modelling-based reconstruction of the Chernobyl source term available for long-range transport. *Atmos. Chem. Phys.* 7, 1549–1564. 29, 41, 47, 53
- De Boor, C., 2001. *A practical guide to Splines*. Springer-Verlag, New-York. 62
- Dee, D. P., 1995. On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.* 123, 1128–1145. 46, 48
- Delle Monache, L., Lundquist, J. K., Kosovic, B., Johannesson, G., Dyer, K. M., Aines, R. D., Chow, F. K., Belles, R. D., Hanley, W. G., Larsen, S. C., Loosmore, G. A., Nitao, J. J., Sugiyama, G. A., Vogt, P. J., 2008. Bayesian inference and Markov chain Monte Carlo sampling to reconstruct a contaminant source on a continental scale. *Journal of Applied Meteorology and Climatology* 47, 2600–2613. 40, 89, 91
- Desroziers, G., Ivanov, S., 2001. Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Q. J. Roy. Meteor. Soc.* 127, 1433–1452. 49
- Doury, A., 1976. Une méthode de calcul pratique et générale pour la prévision numérique des pollutions véhiculées par l'atmosphère. Tech. rep., Rapport technique 4280, CEA, France. 33
- Elbern, H., Strunk, A., Schmidt, H., Talagrand, O., 2007. Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmos. Chem. Phys.* 7, 3749–3769. 41
- Enting, I., 2002. *Inverse Problems in Atmospheric Constituent Transport*. Atmospheric and Space Science Series, Cambridge. 40, 68
- Estournel, C., Bosc, E., Bocquet, M., Ulses, C., Marsaleix, P., Winiarek, V., Osvath, I., Nguyen, C., Duhaut, T., Lyard, F., Michaud, H., Auclair, F., 2012. Assessment of the amount of cesium-137 released to the Pacific Ocean after the Fukushima accident and analysis of its dispersion in the Japanese coastal waters. *J. Geophys. Res.* 117, C11014. 29, 75

- Galperin, M., 2000. The approaches to correct computation of airborne pollution advection In : Problems of Ecological Monitoring and Ecosystem Modelling. XVII. St. Petersburg, Gidrometeoizdat. 52
- Garratt, J., 1992. The Atmospheric Boundary Layer. Cambridge Atmospheric and Space Science Series - Cambridge University Press. 20, 34, 36
- Green, P., 1995. Reversible-jump markov chain monte carlo computation and bayesian model determination. *Biometrika* 82, 711–732. 91
- Hajivassiliou, V., McFadden, D., Ruud, P., 1996. Simulation of multivariate normal rectangle probabilities and their derivatives - Theoretical and computational results. *Journal of Econometrics* 72, 85–134. 49
- Hansen, P. C., 1992. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review* 34, 561–580. 41, 46, 47
- Hastings, W., 1970. Monte carlo sampling methods using markov chains and their applications. *Biometrika* 57(1), 97–109. 135
- Hazart, A., Giovannelli, J., Dubost, S., Chatellier, L., 2014. Inverse transport problem of estimating point-like source using a Bayesian parametric method with MCMC. *Signal Processing* 96, 346–361. 89
- Issartel, J., 2003. Rebuilding sources of linear tracers after atmospheric concentration measurements. *Atmos. Chem. Phys.* 3, 2111–2125. 86, 107
- Jaynes, E., 1957a. Information statistics and statistical mechanics. *Phys. Rev.* 106, 620. 88
- Jaynes, E., 1957b. Information statistics and statistical mechanics II. *Phys. Rev.* 108, 171. 88
- Katata, G., Terada, H., Nagai, H., Chino, M., 2012. Numerical reconstruction of high dose rate zones due to the Fukushima Dai-ichi Nuclear Power Plant accident. *J. Environ. Radioactivity* 111, 2–12. 76, 79
- Keats, A., Yee, E., Lien, F., 2007. Bayesian inference for source determination with applications to a complex urban environment. *Atmos. Env.* 41, 465–479. 89, 91, 112
- Kolmogorov, A., 1991. The local structure of turbulence in incompressible viscous-fluid for very large reynolds numbers. *Proc. R. Soc. A* 434, 9–13. 22
- Koohkan, R., Bocquet, M., Roustan, Y., Kim, Y., Seigneur, C., 2013. Estimation of volatile organic compound emissions for europe using data assimilation. *Atmos. Chem. Phys.* 13, 5887–5905. 121
- Korsakissok, I., 2009. Changements d'échelle en modélisation de la qualité de l'air et estimation des incertitudes associées. Thèse de doctorat de l'Université Paris-Est. 33
- Korsakissok, I., Mathieu, A., Didier, D., 2013. Atmospheric dispersion and ground deposition induced by the Fukushima Nuclear Power Plant accident : A local-scale simulation and sensitivity study. *Atmos. Env.* 70, 267–279. 29
- Kovalets, I., Andronopoulos, S., Venetsanos, A., Bartzis, J., 2011. Identification of strength and location of stationary point source of atmospheric pollutant in urban conditions using computational fluid dynamics model. *Math. Comput. Simul.* 82(2), 244–257. 121

- Kovalets, I., Tsiouri, V., Andronopoulos, S., Bartzis, J., 2008. Improvement of source and wind field input of atmospheric dispersion model by assimilation of concentration measurements : methods and applications in idealized settings. *Appl. Math. Model.* 33, 3511–3521. 121
- Krysta, M., Bocquet, M., 2007. Source reconstruction of an accidental radionuclide release at European scale. *Q. J. Roy. Meteor. Soc.* 133, 529–544. 86, 88, 97
- Krysta, M., Bocquet, M., Brandt, J., 2008. Probing ETEX-II data set with inverse modelling. *Atmos. Chem. Phys.* 8, 3963–3971. 41, 47, 88
- Krysta, M., Bocquet, M., Sportisse, B., Quélo, D., 2006. Data assimilation for short-range dispersion of radionuclides : An application to wind tunnel data. *Atmos. Env.* 40, 7269–7279. 121
- Launder, B., Spalding, D., 1974. The numerical computation of turbulent flow. *Comp. Math. Appl. Mech. Eng.* 3, 369–389. 11, 36, 37
- Li, H., Kalnay, E., Miyoshi, T., 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Q. J. Roy. Meteor. Soc.* 135, 523–533. 41
- Liu, Z.-Q., Rabier, F., 2002. The interaction between model resolution, observation resolution and observation density in data assimilation : A one-dimensional study. *Q. J. Roy. Meteor. Soc.* 128, 1367–1386. 81
- Louis, J., 1979. A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteor.* 17, 197–202. 28, 36, 52, 60, 74
- Lumley, J., Panofsky, H., 1964. *The structure of atmospheric turbulence.* Wiley-Interscience. 22
- Maryon, R. H., Smith, F. B., Conway, B. J., Goddard, D. M., 1991. The U.K. nuclear accident model. *Progress in Nuclear Energy* 26, 85–104. 29, 75
- Mathieu, A., Korsakissok, I., Quélo, D., Groëll, J., Tombette, M., Didier, D., Quentric, E., Saunier, O., Benoit, J.-P., Isnard, O., 2012. Atmospheric dispersion and deposition of radionuclides from the Fukushima Daiichi nuclear power plant accident. *Elements* 8, 195–200. 58, 68, 70, 78, 79
- Ménard, R., Cohn, S. E., Chang, L.-P., Lyster, P. M., 2000. Assimilation of stratospheric chemical tracer observations using a kalman filter. Part I : Formulation. *Mon. Wea. Rev.* 128, 2654–2671. 41, 47
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E., 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087. 135
- Michalak, A. M., Bruhwiler, L., Tans, P. P., 2004. A geostatistical approach to surface flux estimation of atmospheric trace gases. *J. Geophys. Res.* 109, D14109. 41
- Milliez, M., 2006. *Modélisation micro-météorologique en milieu urbain : dispersion des polluants et prise en compte des effets radiatifs.* Thèse de doctorat de L'École Nationale des Ponts et Chaussées. 35, 99, 100, 101, 102, 103
- Milliez, M., Carissimo, B., 2007. Numerical simulations of pollutant dispersion in an idealized urban area, for different meteorological conditions. *Boundary Layer Meteorol.* 122, 321–341. 100, 102, 105, 109, 110

- Mitchell, H. L., Houtekamer, P. L., 1999. An adaptive ensemble Kalman filter. *Mon. Wea. Rev.* 128, 416–433. 41
- Oke, T., 1987. *Boundary layer climates*. second edition, Routledge, London-New York. 21
- Panofsky, H., Dutton, J., 1984. *Atmospheric Turbulence*. Wiley and Sons. 25
- Pasquill, F., 1961. The estimation of the dispersion of windborne material. *Meteorol. Mag.* 90(1063), 33–49. 33
- Perkins, R., Soulhac, L., Mejean, P., Rios, I., 2002. Modélisation de la dispersion des émissions atmosphériques d'un site industriel – vers un guide de l'utilisateur. 1ère partie : état de l'art. Tech. rep., Rapport technique, Laboratoire de Mécanique des Fluides et d'Acoustique. UMR CNRS 5509 – UCBL1 – ECL. 21
- Pudykiewicz, J., 1989. Simulation of the chernobyl dispersion with a 3-d hemispheric tracer model. *Tellus B* 41B, 391–412. 29, 60, 76
- Pudykiewicz, J., 1998. Application of adjoint tracer transport equations for evaluating source parameters. *Atmos. Env.* 32, 3039–3050. 30
- Quélo, D., Krysta, M., Bocquet, M., Isnard, O., Minier, Y., Sportisse, B., 2007. Validation of the Polyphemus platform on the ETEX, Chernobyl and Algeciras cases. *Atmos. Env.* 41, 5300–5315. 29, 51
- Roustan, Y., Bocquet, M., 2006. Inverse modeling for Mercury over Europe. *Atmos. Chem. Phys.* 6, 3085–3098. 30
- Saïde, P., Bocquet, M., Osses, A., Gallardo, L., 2011. Constraining surface emissions of air pollutants using inverse modeling : method intercomparison and a new two-step multiscale approach. *Tellus B* 63, 360–370. 41
- Saunier, O., Bocquet, M., Mathieu, A., Isnard, O., 2009. Model reduction via principal component truncation for the optimal design of atmospheric monitoring networks. *Atmos. Env.* 43, 4940–4950. 50
- Saunier, O., Mathieu, A., Didier, D., Tombette, M., Quélo, D., Winiarek, V., Bocquet, M., 2013. An inverse modeling method to assess the source term of the Fukushima nuclear power plant accident using gamma dose rate observations. *Atmos. Chem. Phys.* 13, 11403–11421. 58, 68, 70, 78, 121
- Schwinger, J., Elbern, H., 2010. Chemical state estimation for the middle atmosphere by four-dimensional variational data assimilation : A posteriori validation of error statistics in observation space. *J. Geophys. Res.* 115, D18307. 41
- Seinfeld, J., Pandis, S., 1998. *Atmospheric chemistry and physics : from air pollution to climate change*. Wiley-Interscience. 23
- Senocak, I., Hengartner, N., Short, M., Brent Daniel, W., 2008. Stochastic event reconstruction of atmospheric contaminant dispersion using Bayesian inference. *Atmos. Env.* 42, 7718–7727. 91, 121
- Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., Huang, S.-Y., Wang, W., Powers, J. G., 2008. A description of the advanced research wrf version 3. Tech. rep., NCAR. 74

- Sofiev, M., 2000. A model for the evaluation of long-term airborne pollution transport at regional and continental scales. *Atmos. Env.* 34, 2481–2493. 52
- Sofiev, M., 2002. Extended resistance analogy for construction of the vertical diffusion scheme for dispersion models. *J. Geophys. Res.* 107 (D12), 4159. 52
- Sofiev, M., Siljamo, P., Valkama, I., Ilvonen, M., Kukkonen, J., 2006. A dispersion modelling system silam and its evaluation against etex data. *Atmos. Env.* 40, 674–685. 52
- Stauffer, D. R., Seaman, N. L., 1994. On multi-scale four-dimensional data assimilation. *J. Appl. Meteor.* 33, 416–434. 74
- Stohl, A., 1998. Computation, accuracy and applications of trajectories - a review and bibliography. *Atmos. Env.* 32, 947–966. 107
- Stohl, A., Forster, C., Frank, A., Seibert, P., Wotawa, G., 2005. Technical note : The Lagrangian particle dispersion model FLEXPART version 6.2. *Atmos. Chem. Phys.* 5, 2461–2474. 31, 32
- Stohl, A., Seibert, P., Wotawa, G., Arnold, D., Burkhardt, J. F., Eckhardt, S., Vargas, A., Yasunari, T. J., 2012. Xenon-133 and caesium-137 releases into the atmosphere from the Fukushima Dai-ichi nuclear power plant : determination of the source term, atmospheric dispersion. *Atmos. Chem. Phys.* 12, 2313–2343. 41, 58, 68, 78, 79
- Stull, R., 1988. *An introduction to Boundary Layer Meteorology*. Kluwer Academic Publishers, Dordrecht, The Netherlands. 20, 34
- Tarantola, A., 2005. *Inverse problem theory and methods for model parameter estimation*. SIAM. 47
- Terada, H., Katata, G., Chino, M., Nagai, H., 2012. Atmospheric discharge and dispersion of radionuclides during the Fukushima Dai-ichi Nuclear Power Plant accident. Part ii : verification of the source term and analysis of regional-scale atmospheric dispersion. *J. Environ. Radioactivity* 112, 141–154. 68, 70, 78, 79
- Tikhonov, A., Arsenin, V., 1977. *Solutions of Ill Posed Problems*. Winston, Washington DC. 41, 43, 131
- Troen, I., Mahrt, L., 1986. A simple model of the atmospheric boundary layer : sensitivity to surface evaporation. *Boundary-Layer Meteor.* 37, 129–148. 28
- Vendel, F., 2010. *Modélisation de la dispersion atmosphérique en présence d'obstacles complexes : application à l'étude de sites industriels*. Thèse de doctorat de l'École Centrale de Lyon. 120
- Viollet, P., 1988. On the numerical modeling of stratified flows. *J. Dronkers and W. Leussen Eds, Springer-Verlag, 257-277*. 37
- Wesely, M., 1989. Parametrization of surface resistances to gaseous dry deposition in regional-scale numerical models. *Atmos. Env.* 23, 1293–1304. 29, 52
- Winiarek, V., Bocquet, M., Duhanyan, N., Roustan, Y., Saunier, O., Mathieu, A., 2014. Estimation of the caesium-137 source term from the Fukushima daiichi nuclear power plant using a consistent joint assimilation of air concentration and deposition observations. *Atmos. Env.* 82, 268–279. 59, 72, 83, 119

- Winiarek, V., Bocquet, M., Saunier, O., Mathieu, A., 2012. Estimation of errors in the inverse modeling of accidental release of atmospheric pollutant : Application to the reconstruction of the cesium-137 and iodine-131 source terms from the Fukushima Daiichi power plant. *J. Geophys. Res.* 117, D05122. 53, 59, 63, 78, 119
- Winiarek, V., Vira, J., Bocquet, M., Sofiev, M., Saunier, O., 2011. Towards the operational estimation of a radiological plume using data assimilation after a radiological accidental atmospheric release. *Atmos. Env.* 45, 2944–2955. 50, 87, 92, 119
- Wu, L., Bocquet, M., Chevallier, F., Lauvaux, T., Davis, K., 2013. Hyperparameter estimation for uncertainty quantification in mesoscale carbon dioxide inversions. *Tellus B* 65, 20894. 121
- Yee, E., Bilotft, C., 2004. Concentration fluctuation measurements un a plume dispersing through a regular array of obstacles. *Boundary-Layer Meteorol.* 111, 363–415. 98, 99
- Yee, E., Lien, F.-S., Keats, A., D'Amours, R., 2008. Bayesian inversion of concentration data : Source reconstruction in the adjoint representation of atmospheric diffusion. *Journal of Wind Engineering and Industrial Aerodynamics* 96, 1805–1816. 40, 89, 92, 112
- Zhang, L., Gao, S., Padro, J., Barrie, L., 2001. A size-segregated dry deposition scheme for an atmospheric aerosol module. *Atmos. Env.* 35, 549–560. 29, 52

Annexe A

Calcul de la vraisemblance marginale des observations

La vraisemblance marginale des observations s'exprime en fonction des pdf des erreurs d'observation et des erreurs d'ébauche :

$$p(\boldsymbol{\mu}|\boldsymbol{\theta}) = \int p_{\epsilon}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) p_b(\boldsymbol{\sigma}) d\boldsymbol{\sigma}. \quad (\text{A.1})$$

Statistiques gaussiennes

Dans le cadre de statistiques gaussiennes, la pdf des erreurs d'observation s'écrit :

$$p_{\epsilon}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})}}{\sqrt{(2\pi)^d |\mathbf{R}|}}, \quad (\text{A.2})$$

et la pdf des erreurs d'ébauche :

$$p_b(\boldsymbol{\sigma}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)^T \mathbf{B}^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)}}{\sqrt{(2\pi)^N |\mathbf{B}|}}. \quad (\text{A.3})$$

Ainsi l'équation A.1 s'écrit

$$\begin{aligned} p(\boldsymbol{\mu}|\boldsymbol{\theta}) &= \int \frac{e^{-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) - \frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)^T \mathbf{B}^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}^b)}}{\sqrt{(2\pi)^{N+d} |\mathbf{R}| |\mathbf{B}|}} d\boldsymbol{\sigma} \\ &= \frac{1}{\sqrt{(2\pi)^{N+d} |\mathbf{R}| |\mathbf{B}|}} \int e^{-\mathcal{J}_q(\boldsymbol{\sigma})} d\boldsymbol{\sigma}. \end{aligned} \quad (\text{A.4})$$

Or la fonction coût est une forme quadratique qui admet un unique minimum $\boldsymbol{\sigma}^{\text{BLUE}}$ [Tikhonov and Arsenin, 1977] et peut s'écrire

$$\mathcal{J}_q(\boldsymbol{\sigma}) = \mathcal{J}_q(\boldsymbol{\sigma}^{\text{BLUE}}) + \frac{1}{2} (\boldsymbol{\sigma} - \boldsymbol{\sigma}^{\text{BLUE}})^T \mathbf{P}_{\text{BLUE}}^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}^{\text{BLUE}}). \quad (\text{A.5})$$

avec

$$\begin{aligned} \boldsymbol{\sigma}^{\text{BLUE}} &= \boldsymbol{\sigma}^b + \mathbf{B}\mathbf{H}^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^b) \\ \mathbf{P}_{\text{BLUE}} &= \mathbf{B} - \mathbf{B}\mathbf{H}^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} \mathbf{H}\mathbf{B} = (\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1})^{-1}. \end{aligned} \quad (\text{A.6})$$

L'équation A.5 permet d'écrire la vraisemblance marginale des observations sous la forme :

$$p(\boldsymbol{\mu}|\boldsymbol{\theta}) = \frac{e^{-\mathcal{J}_q(\boldsymbol{\sigma}^{\text{BLUE}})}}{\sqrt{(2\pi)^{N+d} |\mathbf{R}||\mathbf{B}|}} \int e^{-\frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^{\text{BLUE}})^T \mathbf{P}_{\text{BLUE}}^{-1}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^{\text{BLUE}})} d\boldsymbol{\sigma}. \quad (\text{A.7})$$

Nous allons estimer d'un côté $\mathcal{J}_q(\boldsymbol{\sigma}^{\text{BLUE}})$ et de l'autre le terme intégral. $\mathcal{J}_q(\boldsymbol{\sigma}^{\text{BLUE}})$, par définition, s'écrit

$$\begin{aligned} \mathcal{J}_q(\boldsymbol{\sigma}^{\text{BLUE}}) &= \frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^{\text{BLUE}})^T \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^{\text{BLUE}}) \\ &\quad + \frac{1}{2} (\boldsymbol{\sigma}^{\text{BLUE}} - \boldsymbol{\sigma}^b)^T \mathbf{B}^{-1} (\boldsymbol{\sigma}^{\text{BLUE}} - \boldsymbol{\sigma}^b). \end{aligned} \quad (\text{A.8})$$

En remplaçant $\boldsymbol{\sigma}^{\text{BLUE}}$ dans cette équation par son expression des équations A.6, nous obtenons après simplification :

$$\mathcal{J}_q(\boldsymbol{\sigma}^{\text{BLUE}}) = \frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^b)^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}^b). \quad (\text{A.9})$$

Le terme intégral est, à une constante multiplicative près, l'intégrale d'une distribution Normale multivariée dont la valeur est obtenue en utilisant l'équation A.6 :

$$\begin{aligned} \int e^{-\frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^{\text{BLUE}})^T \mathbf{P}_{\text{BLUE}}^{-1}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^{\text{BLUE}})} d\boldsymbol{\sigma} &= \sqrt{(2\pi)^N |\mathbf{P}_{\text{BLUE}}|} \\ &= \sqrt{\frac{(2\pi)^N}{|\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1}|}}. \end{aligned} \quad (\text{A.10})$$

Nous avons donc

$$p(\boldsymbol{\mu}|\boldsymbol{\theta}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}^b)^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} (\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}^b)}}{\sqrt{(2\pi)^d |\mathbf{R}||\mathbf{B}||\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1}|}}. \quad (\text{A.11})$$

En utilisant le lemme du déterminant :

$$|\mathbf{R}||\mathbf{B}||\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1}| = |\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}|, \quad (\text{A.12})$$

nous obtenons finalement l'expression de la vraisemblance marginale des observations dans le cadre de statistiques gaussiennes :

$$p(\boldsymbol{\mu}|\boldsymbol{\theta}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}^b)^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} (\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}^b)}}{\sqrt{(2\pi)^d |\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}|}}. \quad (\text{A.13})$$

Statistiques gaussiennes tronquées

Dans le cas où des statistiques gaussiennes tronquées ont été choisies pour décrire les erreurs d'ébauche, la pdf des erreurs d'observation n'est pas modifiée et la pdf des erreurs d'ébauche s'écrit :

$$p_b(\boldsymbol{\sigma}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^b)^T \mathbf{B}^{-1}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^b)}}{\Phi(\boldsymbol{\sigma}^b, \mathbf{B}, \mathbf{0}) \sqrt{(2\pi)^N |\mathbf{B}|}} \mathbb{I}_{\boldsymbol{\sigma} \geq \mathbf{0}}, \quad (\text{A.14})$$

où $\mathbb{I}_{\boldsymbol{\sigma} \geq \mathbf{0}}$ vaut 1 si tous les termes $\sigma_i \geq 0$ ($i = 1, \dots, N$), sinon sa valeur est 0. $\Phi(\boldsymbol{\sigma}^b, \mathbf{B}, \mathbf{0})$ est la fonction de distribution cumulative complémentaire de $\mathcal{N}(\boldsymbol{\sigma}^b, \mathbf{B})$ en $\mathbf{0}$, soit :

$$\Phi(\boldsymbol{\sigma}^b, \mathbf{B}, \mathbf{0}) = \int_{\boldsymbol{\sigma} \geq \mathbf{0}} \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^b)^T \mathbf{B}^{-1}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^b)}}{\sqrt{(2\pi)^N |\mathbf{B}|}} d\boldsymbol{\sigma}. \quad (\text{A.15})$$

Ainsi l'équation A.1 s'écrit :

$$\begin{aligned}
 p(\boldsymbol{\mu}|\boldsymbol{\theta}) &= \int \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}) - \frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^b)^T \mathbf{B}^{-1}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^b)}}{\Phi(\boldsymbol{\sigma}^b, \mathbf{B}, \mathbf{0}) \sqrt{(2\pi)^{N+d} |\mathbf{R}||\mathbf{B}|}} \mathbb{I}_{\boldsymbol{\sigma} \geq \mathbf{0}} d\boldsymbol{\sigma} \\
 &= \frac{1}{\Phi(\boldsymbol{\sigma}^b, \mathbf{B}, \mathbf{0}) \sqrt{(2\pi)^{N+d} |\mathbf{R}||\mathbf{B}|}} \int e^{-\mathcal{J}_q(\boldsymbol{\sigma})} \mathbb{I}_{\boldsymbol{\sigma} \geq \mathbf{0}} d\boldsymbol{\sigma}. \tag{A.16}
 \end{aligned}$$

Les équations A.5 et A.9 permettent alors d'écrire :

$$\begin{aligned}
 p(\boldsymbol{\mu}|\boldsymbol{\theta}) &= \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}^b)^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}^b)}}{\Phi(\boldsymbol{\sigma}^b, \mathbf{B}, \mathbf{0}) \sqrt{(2\pi)^{d+N} |\mathbf{R}||\mathbf{B}|}} \\
 &\quad \times \int e^{-\frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^{\text{BLUE}})^T \mathbf{P}_{\text{BLUE}}^{-1}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^{\text{BLUE}})} \mathbb{I}_{\boldsymbol{\sigma} \geq \mathbf{0}} d\boldsymbol{\sigma}. \tag{A.17}
 \end{aligned}$$

Il n'est maintenant plus possible d'obtenir une formule simple pour l'intégrale qui est, à une constante multiplicative près, l'intégrale d'une distribution Normale multivariée tronquée. Il est possible de faire apparaître cette constante en manipulant les déterminants de matrice pour finalement obtenir l'expression de la vraisemblance marginale des observations dans le cadre de statistiques gaussiennes tronquées :

$$\begin{aligned}
 p(\boldsymbol{\mu}|\boldsymbol{\theta}) &= \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}^b)^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}^b)}}{\Phi(\boldsymbol{\sigma}^b, \mathbf{B}, \mathbf{0}) \sqrt{(2\pi)^d |\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}|}} \\
 &\quad \times \int \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^{\text{BLUE}})^T \mathbf{P}_{\text{BLUE}}^{-1}(\boldsymbol{\sigma}-\boldsymbol{\sigma}^{\text{BLUE}})}}{\sqrt{(2\pi)^N |\mathbf{P}_{\text{BLUE}}|}} \mathbb{I}_{\boldsymbol{\sigma} \geq \mathbf{0}} d\boldsymbol{\sigma}. \tag{A.18}
 \end{aligned}$$

Annexe B

Présentation de l'algorithme MCMC

Dans le cas où la source à reconstruire est supposée ponctuelle et en situation stationnaire, le vecteur Θ des paramètres se réduit à (x_s, y_s, z_s, q_s) , les coordonnées spatiales de la source et son débit d'émission. À un facteur multiplicatif près, la pdf *a posteriori* de Θ est donnée par :

$$p(\Theta|\mu) \propto \frac{\mathbb{I}_{\mathcal{R}^1}(\Theta)}{\sqrt{(2\pi)^d |\mathbf{R}|}} \exp\left(-\frac{1}{2}(\mu - \mathcal{H}(\Theta))^T \mathbf{R}^{-1}(\mu - \mathcal{H}(\Theta))\right), \quad (\text{B.1})$$

où $\mathbb{I}_{\mathcal{R}^1}(\Theta)$ vaut 1 si la source est dans le domaine, et 0 sinon.

Un algorithme de Monte Carlo par chaîne de Markov (MCMC) permet d'obtenir un échantillonnage de $p(\Theta|\mu)$, c'est-à-dire une chaîne $\{\Theta^{(0)}, \dots, \Theta^{(N_e)}\}$ dont les éléments sont distribués suivant $p(\Theta|\mu)$, en 3 étapes [Metropolis et al., 1953; Hastings, 1970] :

1. Initialisation de la chaîne de Markov

Un premier vecteur de paramètres $\Theta^{(0)} = (x_s^{(0)}, y_s^{(0)}, z_s^{(0)}, q_s^{(0)})^T$ est proposé dans le domaine de simulation et $p(\Theta^{(0)}|\mu)$ est calculée d'après l'équation B.1. Si le modèle de dispersion est Code_Saturne, la valeur de l'observation i calculée par le modèle est donnée par $[\mathcal{H}(\Theta^{(0)})]_i = q_s^{(0)} c_i^*(x_s^{(0)}, y_s^{(0)}, z_s^{(0)})$, les d rétro-panaches $(c_i^*)_{i \leq d}$ ayant été préalablement calculées. Avec le modèle de panache gaussien, la valeur de l'observation i calculée par le modèle est calculée par la formule analytique 2.35 du chapitre 2.3.3.

2. Itération dans la chaîne de Markov

Supposons que la chaîne de Markov contienne $n + 1$ éléments $\{\Theta^{(0)}, \dots, \Theta^{(n)}\}$. Un nouveau vecteur Θ^* est alors proposé dans le domaine, suivant une loi de proposition qui peut être par exemple une distribution normale d'espérance $\Theta^{(n)}$. $p(\Theta^*|\mu)$ est ensuite calculée ; le nouvel élément $\Theta^{(n+1)}$ de la chaîne de Markov est Θ^* avec une probabilité égale à $\min\left(1, \frac{p(\Theta^*|\mu)}{p(\Theta^{(n)}|\mu)}\right)$, sinon $\Theta^{(n+1)} = \Theta^{(n)}$.

Cette étape est itérée un grand nombre de fois, 10^6 fois dans notre étude MUST, ce qui, dans le cas du modèle Code_Saturne, est rendu possible par l'utilisation du modèle ad-joint.

3. Élimination des éléments de la chaîne avant convergence

Les premiers éléments de la chaîne (environ 10% de la chaîne totale comme on peut le voir sur la figure B.1) ne sont pas distribués suivant la pdf cible. Cette phase, appelée

burn-in, est nécessaire à l'algorithme pour localiser la zone de l'espace des solutions qui contient de l'information (où $p(\Theta|\mu)$ est élevée). Ces premiers éléments sont donc retirés de la chaîne de Markov.

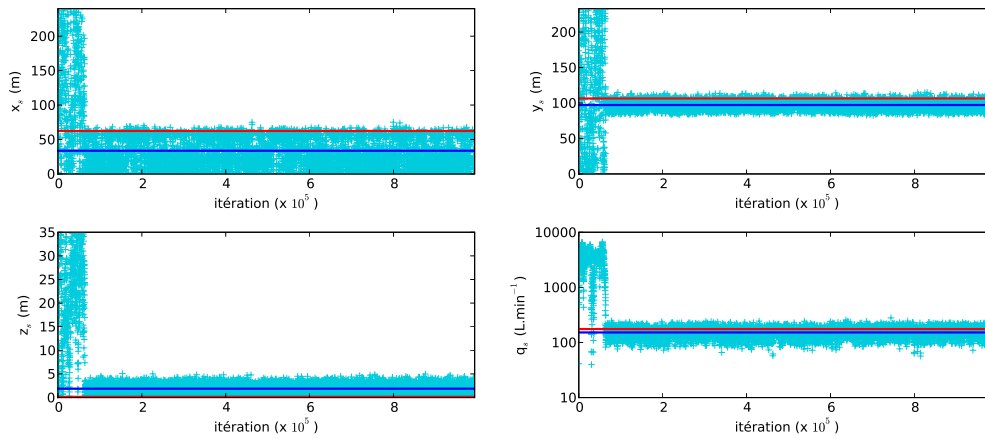


FIG. B.1: Convergence d'une chaîne de Markov calculée pour le cas 2640138 avec Code_Saturne. Les lignes rouges représentent les vraies valeurs des paramètres. Les lignes bleues représentent les moyennes des paramètres dans la chaîne, après élimination des éléments du *burn-in*.

Annexe C

Ensemble des résultats sur l'expérience MUST

Localisation des estimateurs sur le plan horizontal

Figures C.1 à C.14. Sur chaque figure, le disque rouge représente la vraie source, le diamant bleu l'estimateur issu de la simulation avec Code_Saturne et le triangle vert l'estimateur issu de la simulation avec le modèle de panache gaussien. Sur les figures de gauche, l'estimateur utilisé est l'espérance de sources dans la chaîne de Markov. Sur les figures de droite, l'estimateur utilisé est celui qui maximise les densité de probabilité *a posteriori*.

Densités de probabilités *a posteriori* des paramètres de la source

Figures C.15 à C.28. Sur chaque panel de chaque figure, le trait rouge représente la valeur vraie du paramètre, le trait bleu la moyenne de la distribution bleue issue de la simulation avec Code_Saturne et le trait noir la moyenne de la distribution noire issue de la simulation avec le modèle de panache gaussien.

Lignes d'iso-densité de probabilité *a posteriori* de localisation horizontale des sources

Figures C.29 à C.42. Sur chaque figure, le panel de gauche représente l'estimateur moyen de la source reconstruite avec Code_Saturne (diamant bleu), ainsi que les lignes d'iso-densité de probabilité *a posteriori*. Le panel de droite représente l'estimateur moyen de la source reconstruite avec un modèle de panache gaussien (triangle vert) et les lignes d'iso-densité de probabilité *a posteriori*. Sur les deux panels, le disque rouge représente la vraie source.

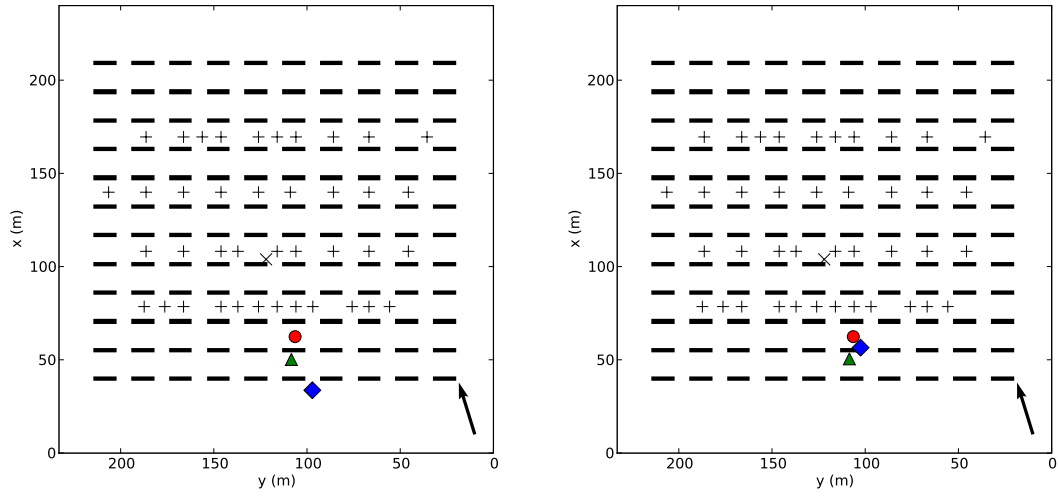


FIG. C.1: Cas 2640138.

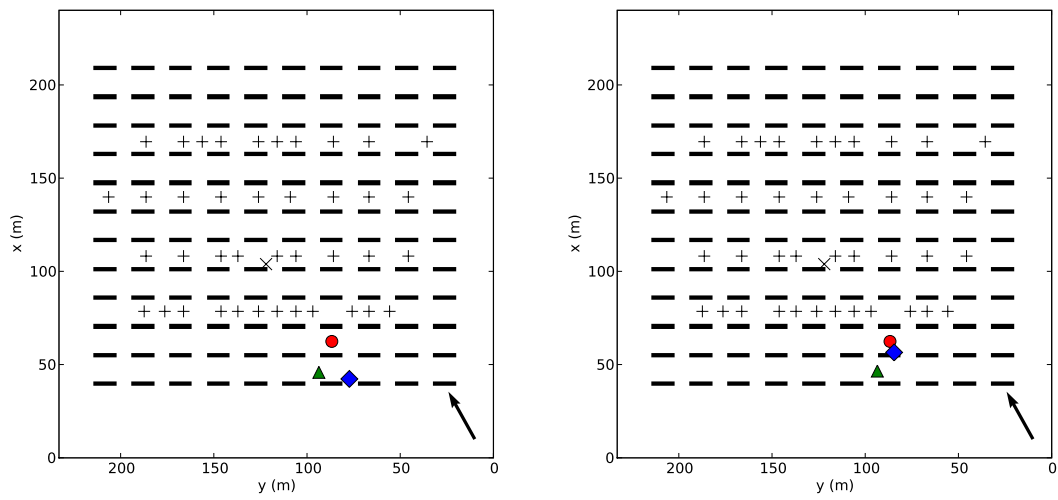


FIG. C.2: Cas 2640246.

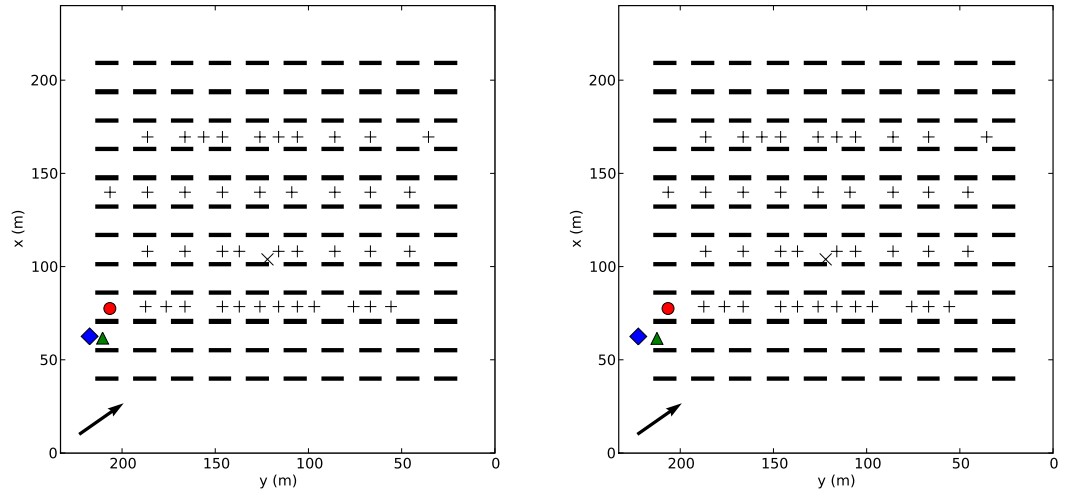


FIG. C.3: Cas 2671852.

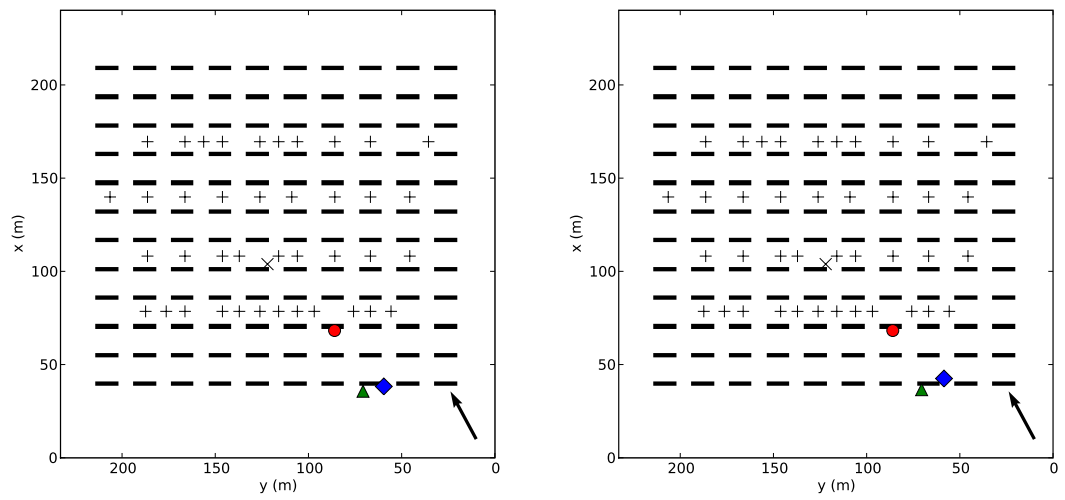


FIG. C.4: Cas 2672213.

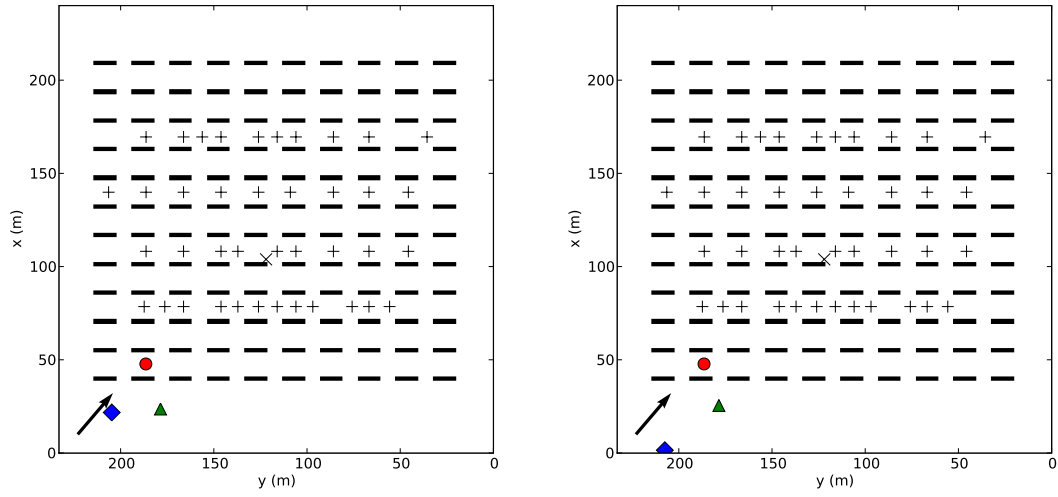


FIG. C.5: Cas 2681829.

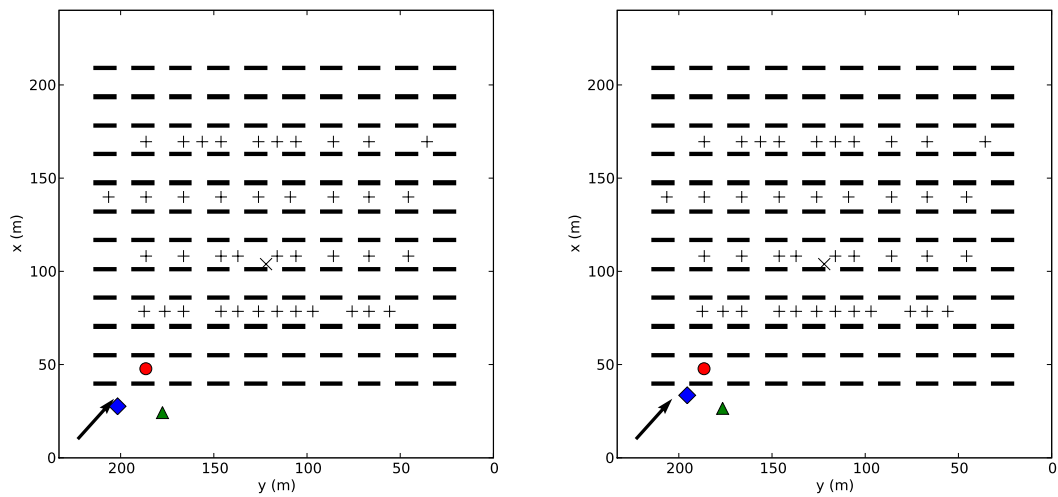


FIG. C.6: Cas 2681849.

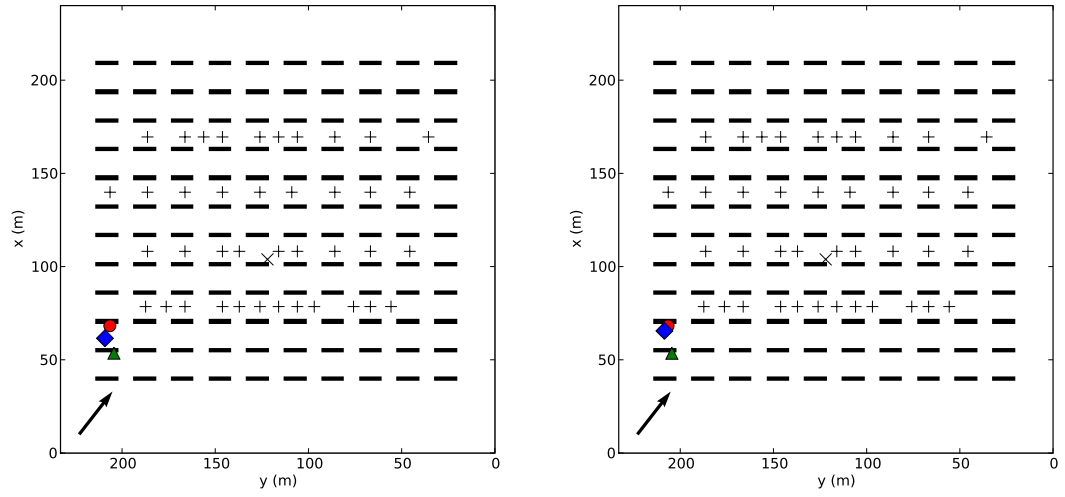


FIG. C.7: Cas 2682256.

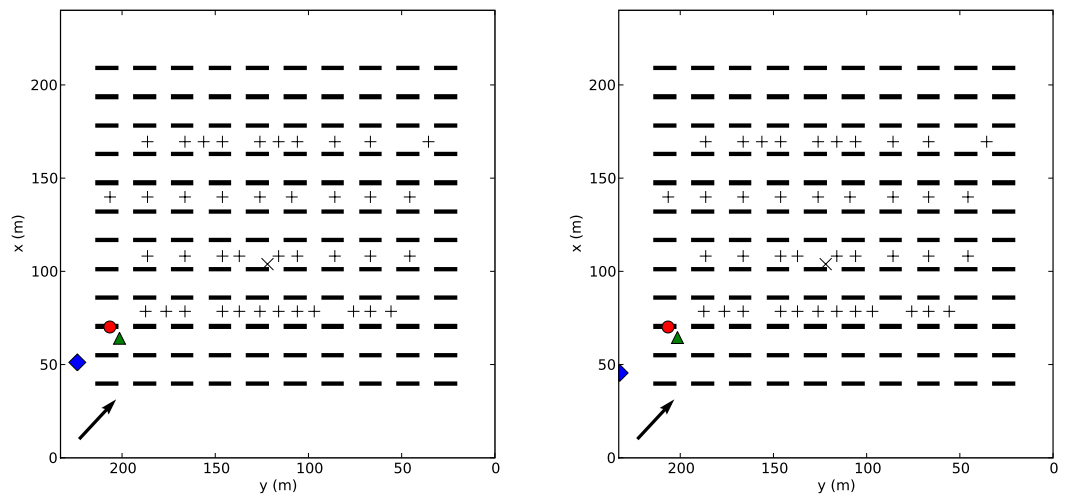


FIG. C.8: Cas 2682320.

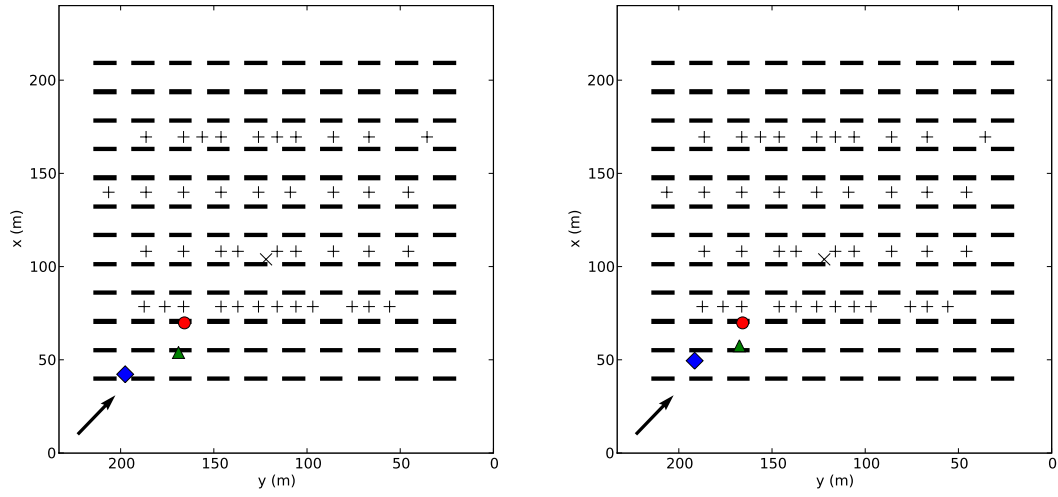


FIG. C.9: Cas 2682353.

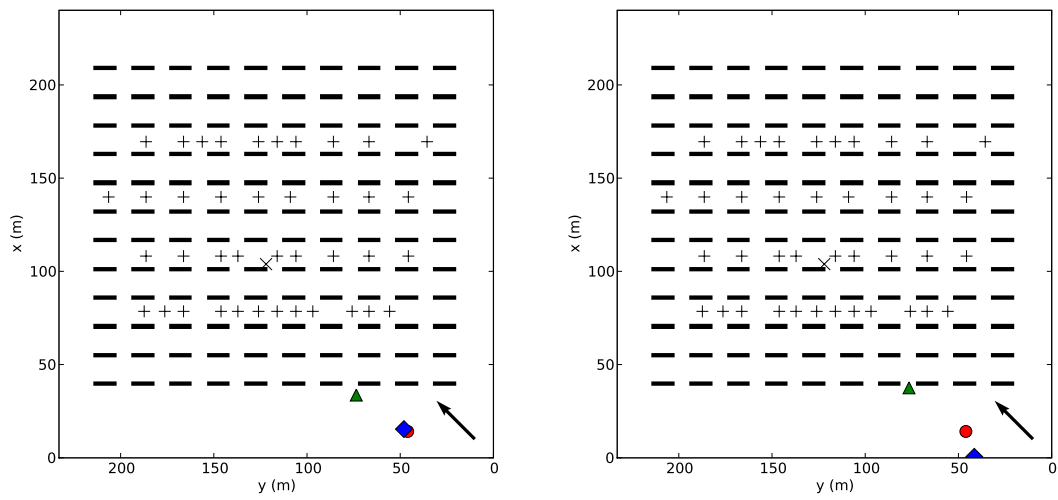


FIG. C.10: Cas 2692054.

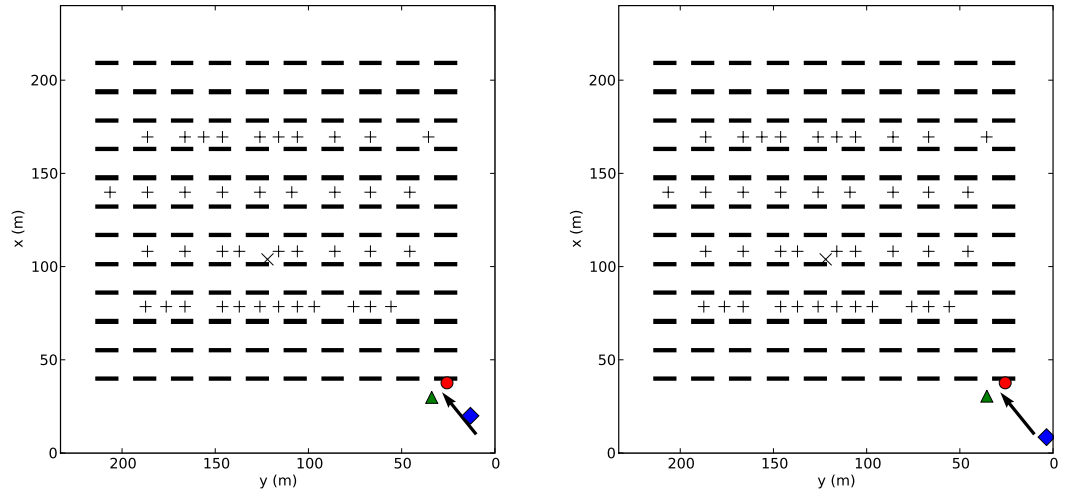


FIG. C.11: Cas 2692131.

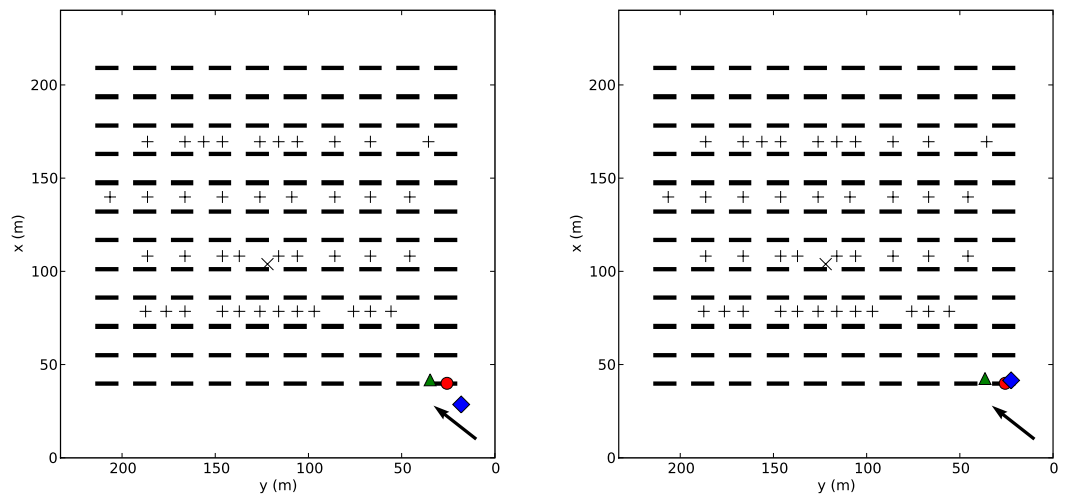


FIG. C.12: Cas 2692157.

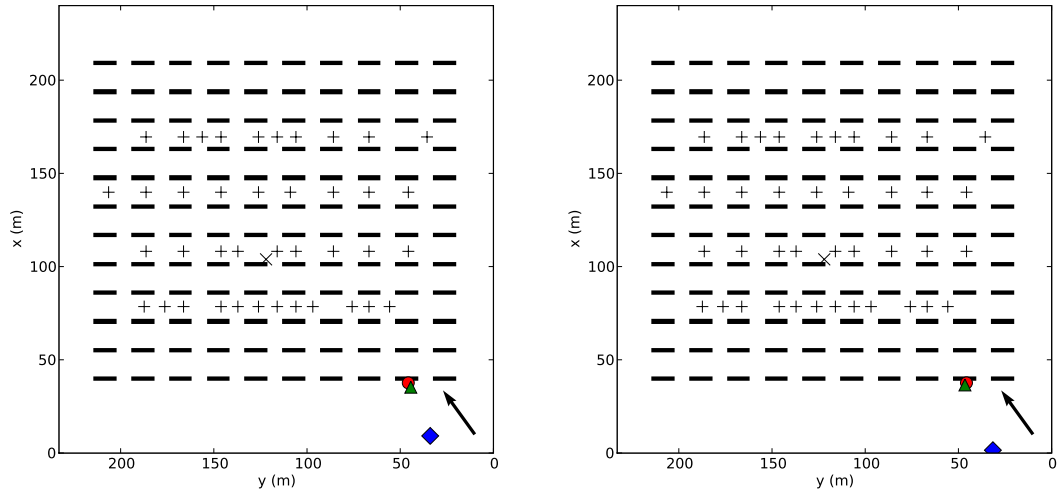


FIG. C.13: Cas 2692223.

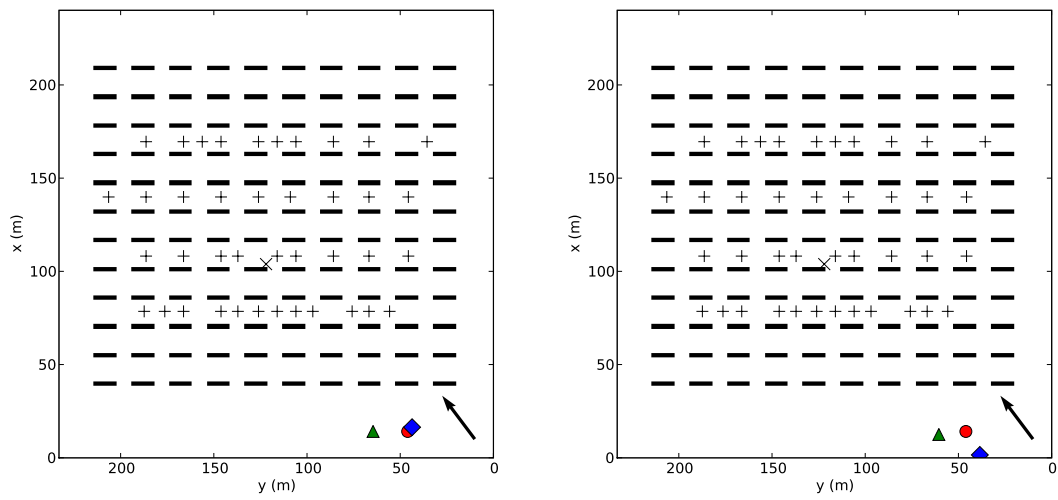


FIG. C.14: Cas 2692250.

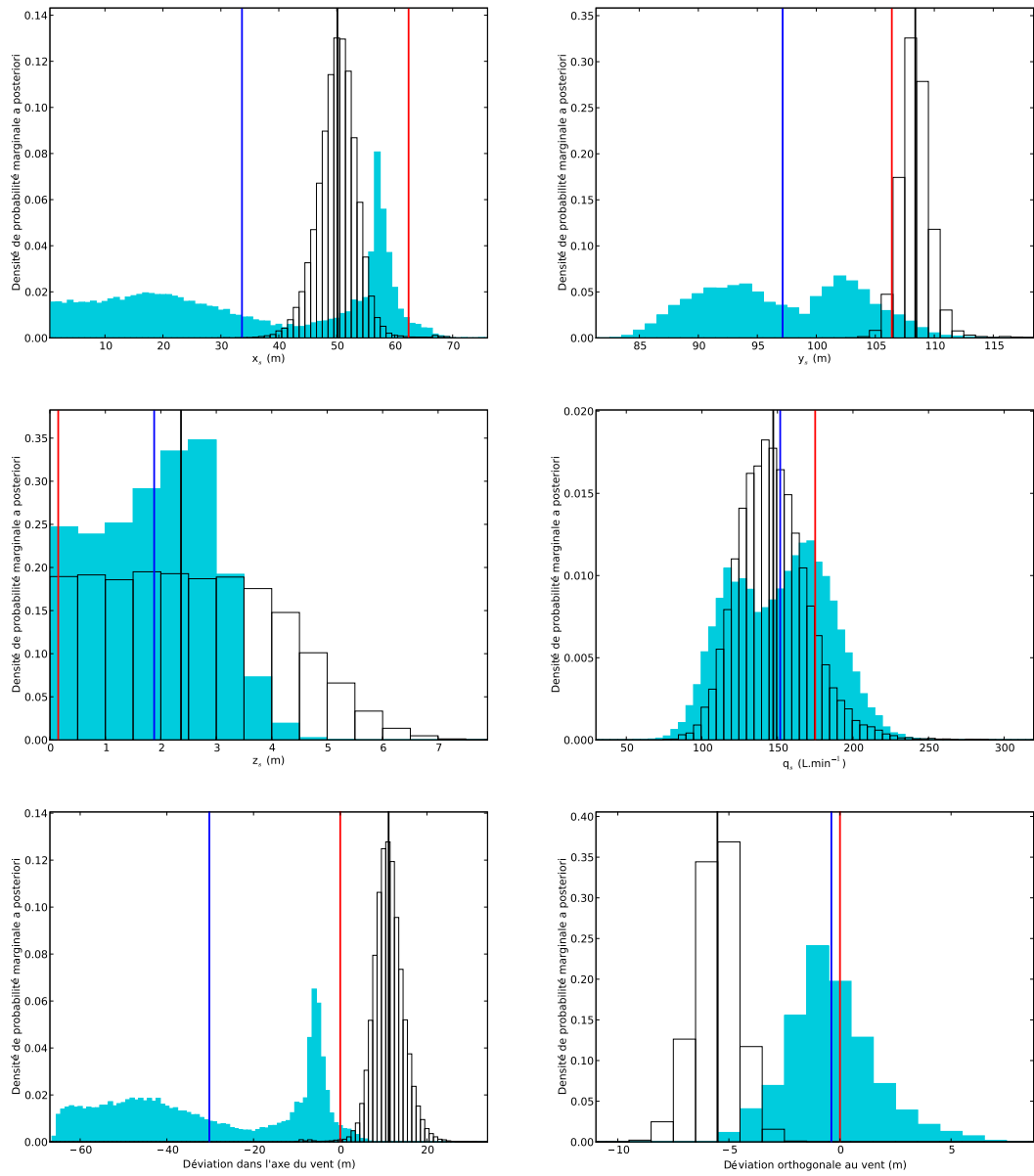


FIG. C.15: Cas 2640138.

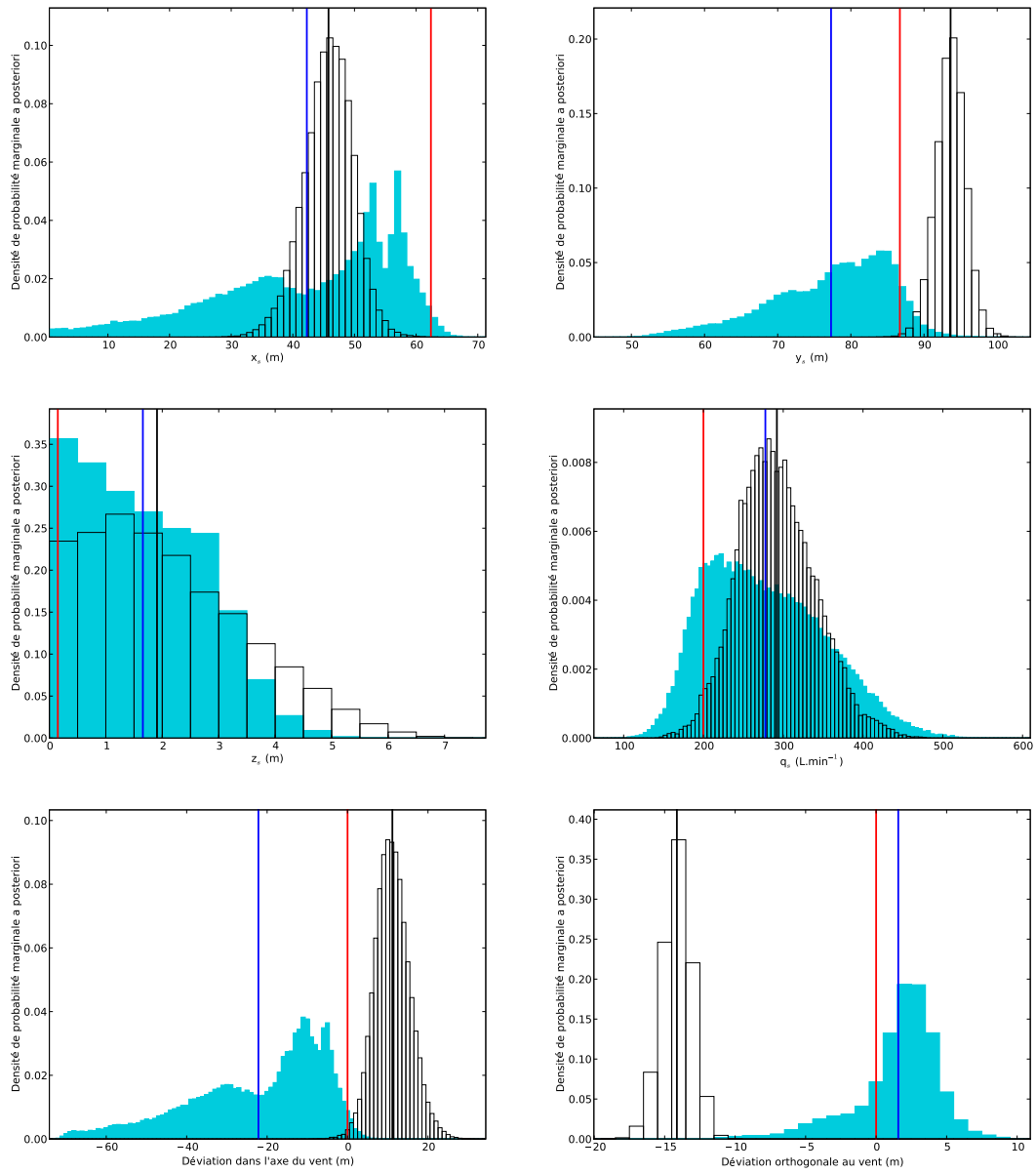


FIG. C.16: Cas 2640246.

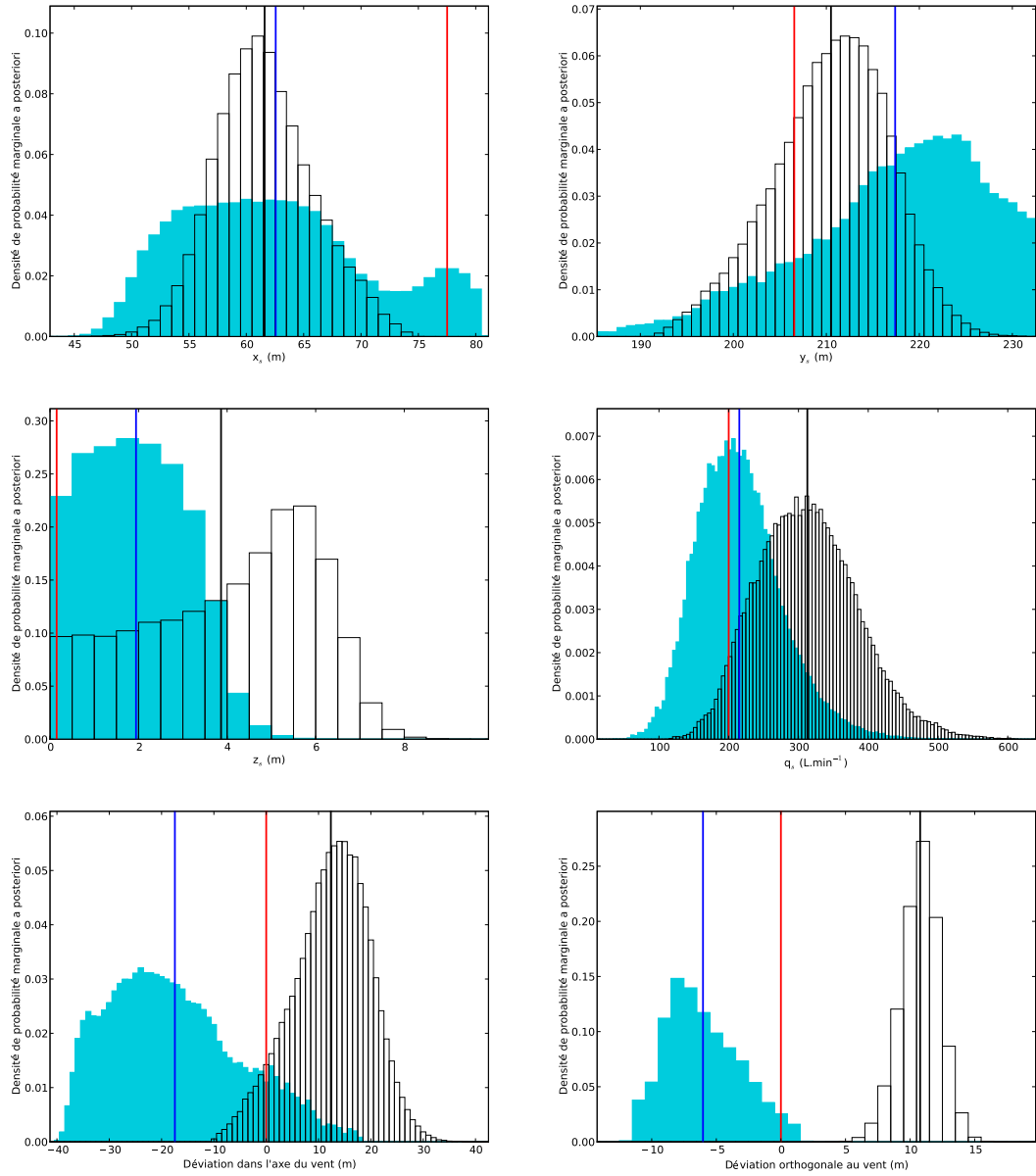


FIG. C.17: Cas 2671852.

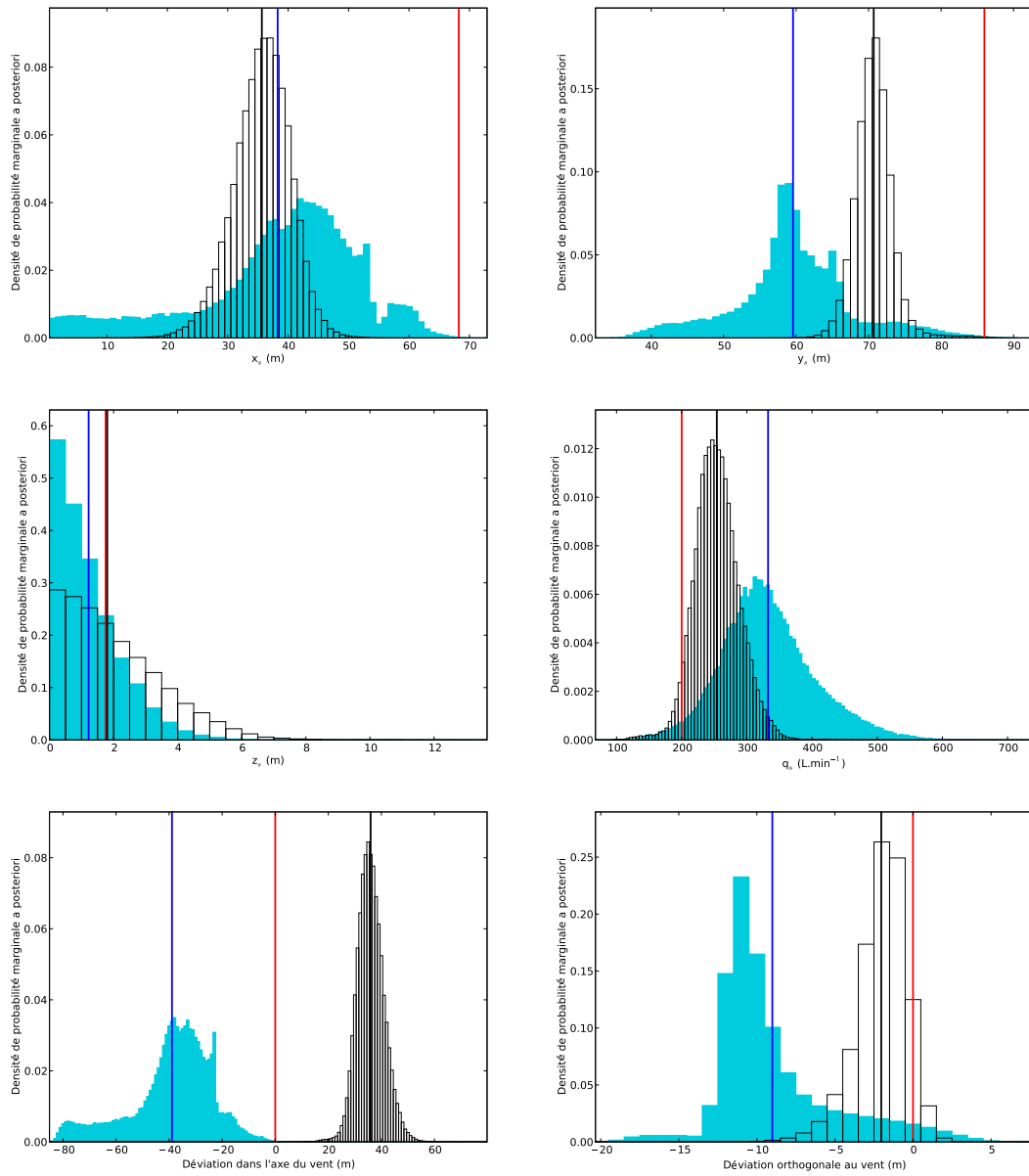


FIG. C.18: Cas 2672213.

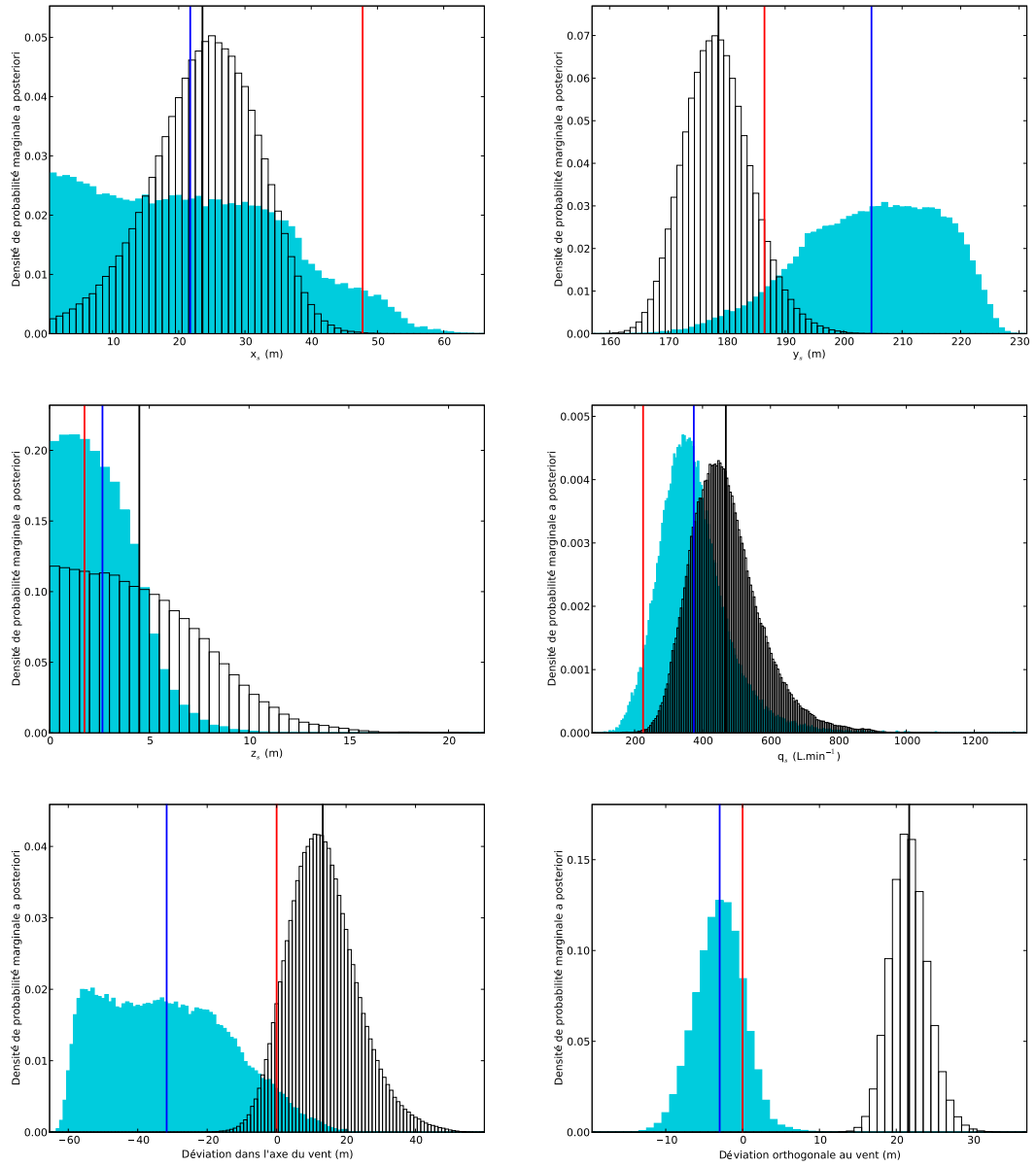


FIG. C.19: Cas 2681829.

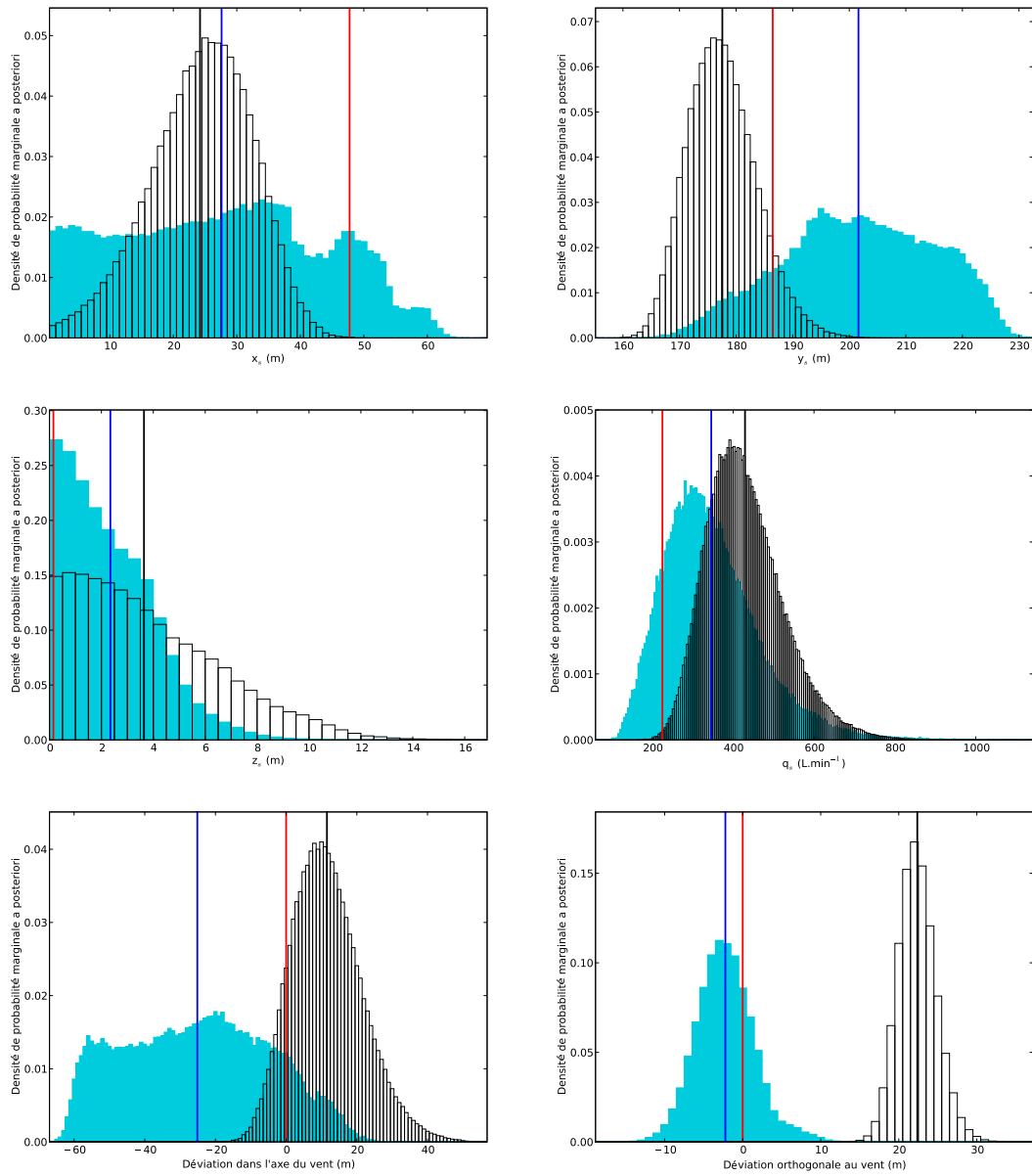


FIG. C.20: Cas 2681849.

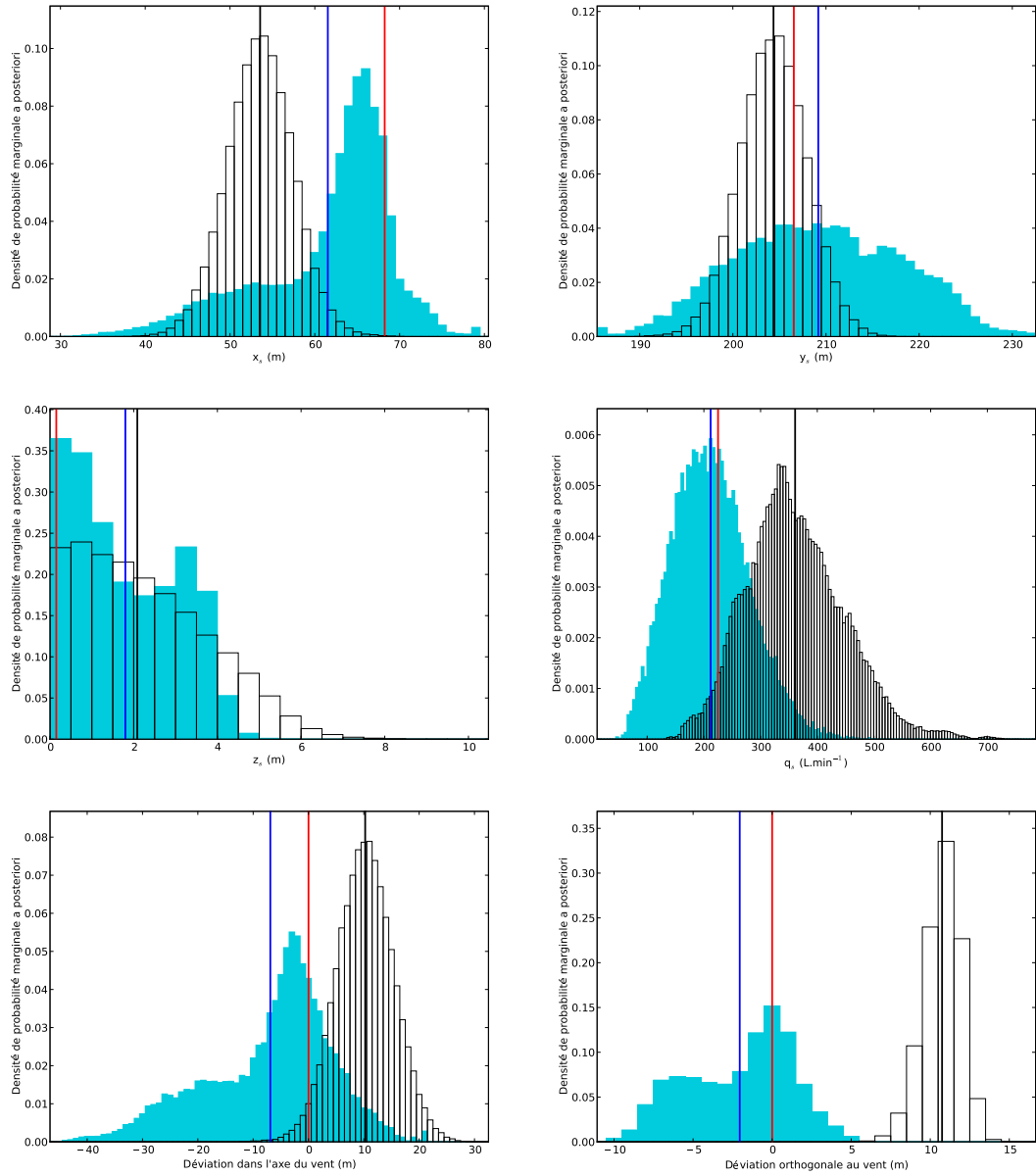


FIG. C.21: Cas 2682256.

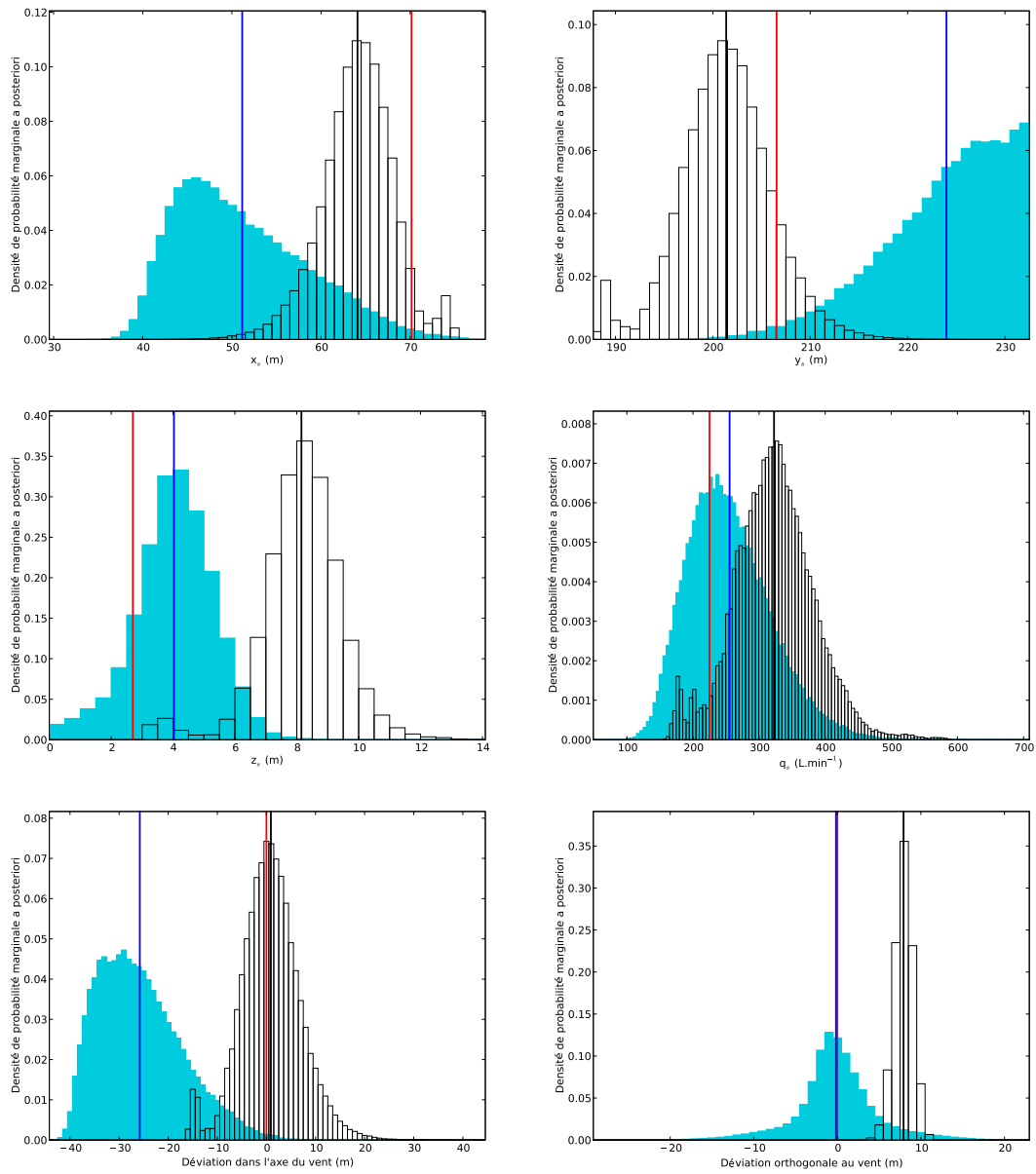


FIG. C.22: Cas 2682320.

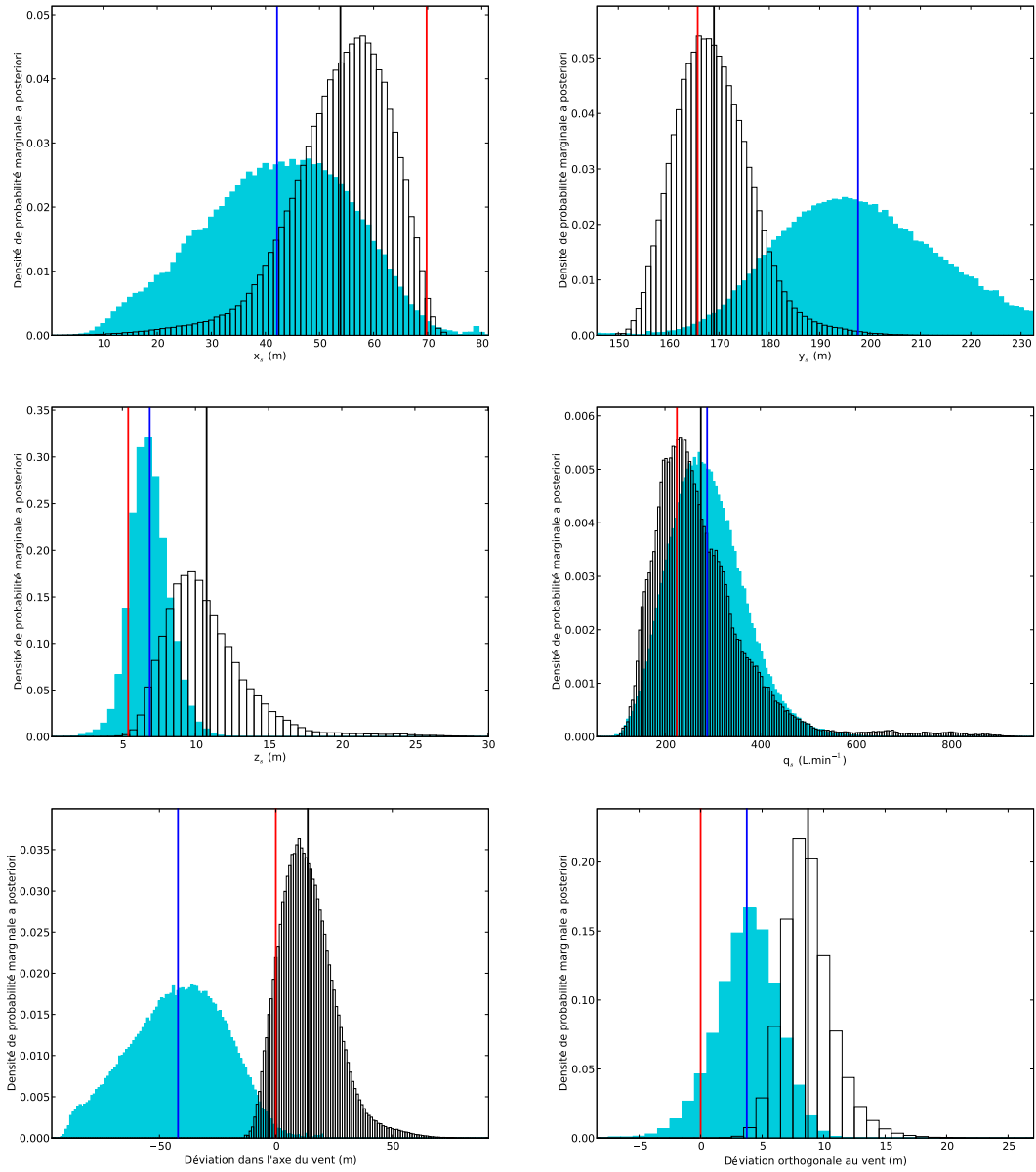


FIG. C.23: Cas 2682353.

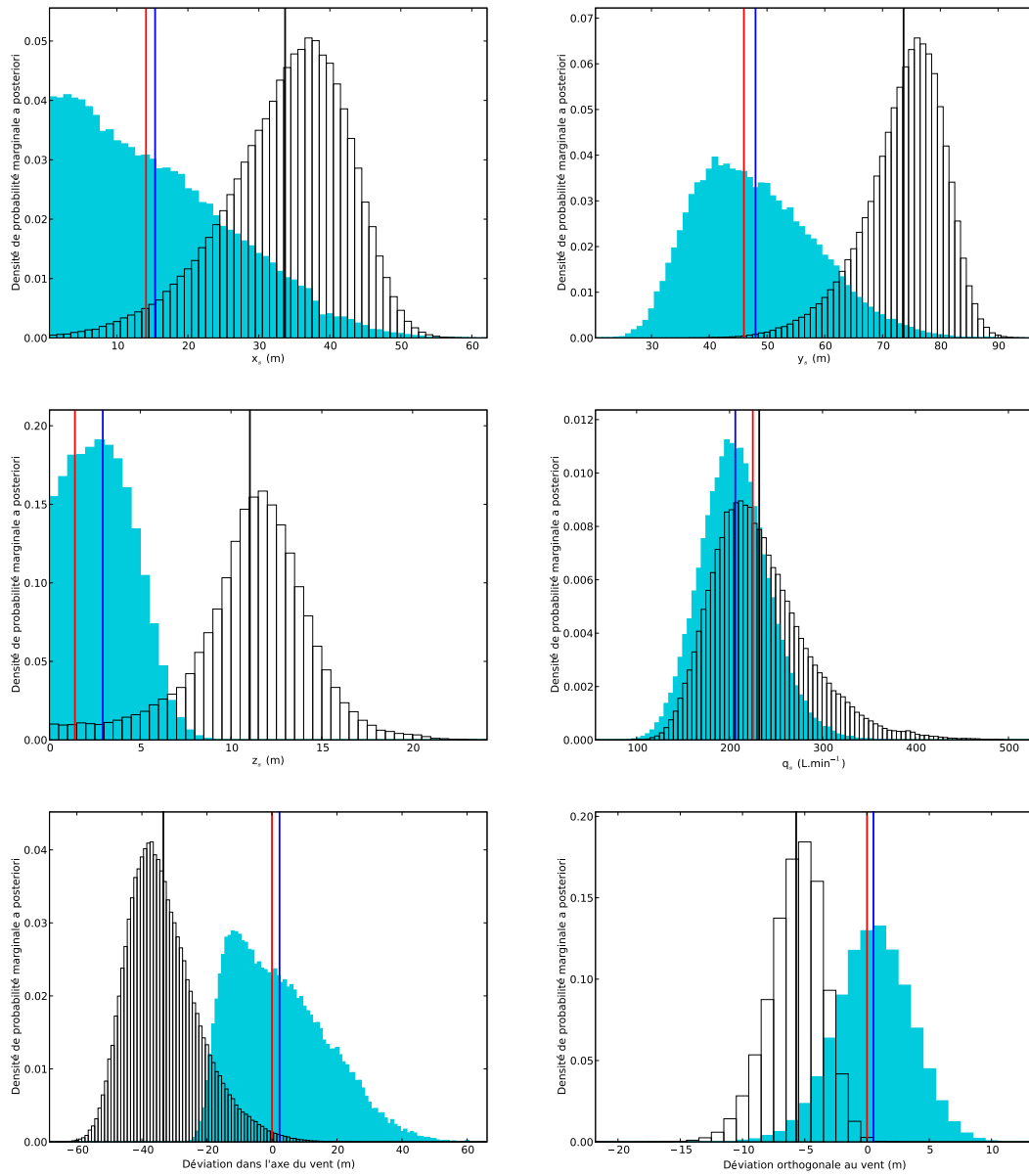


FIG. C.24: Cas 2692054.

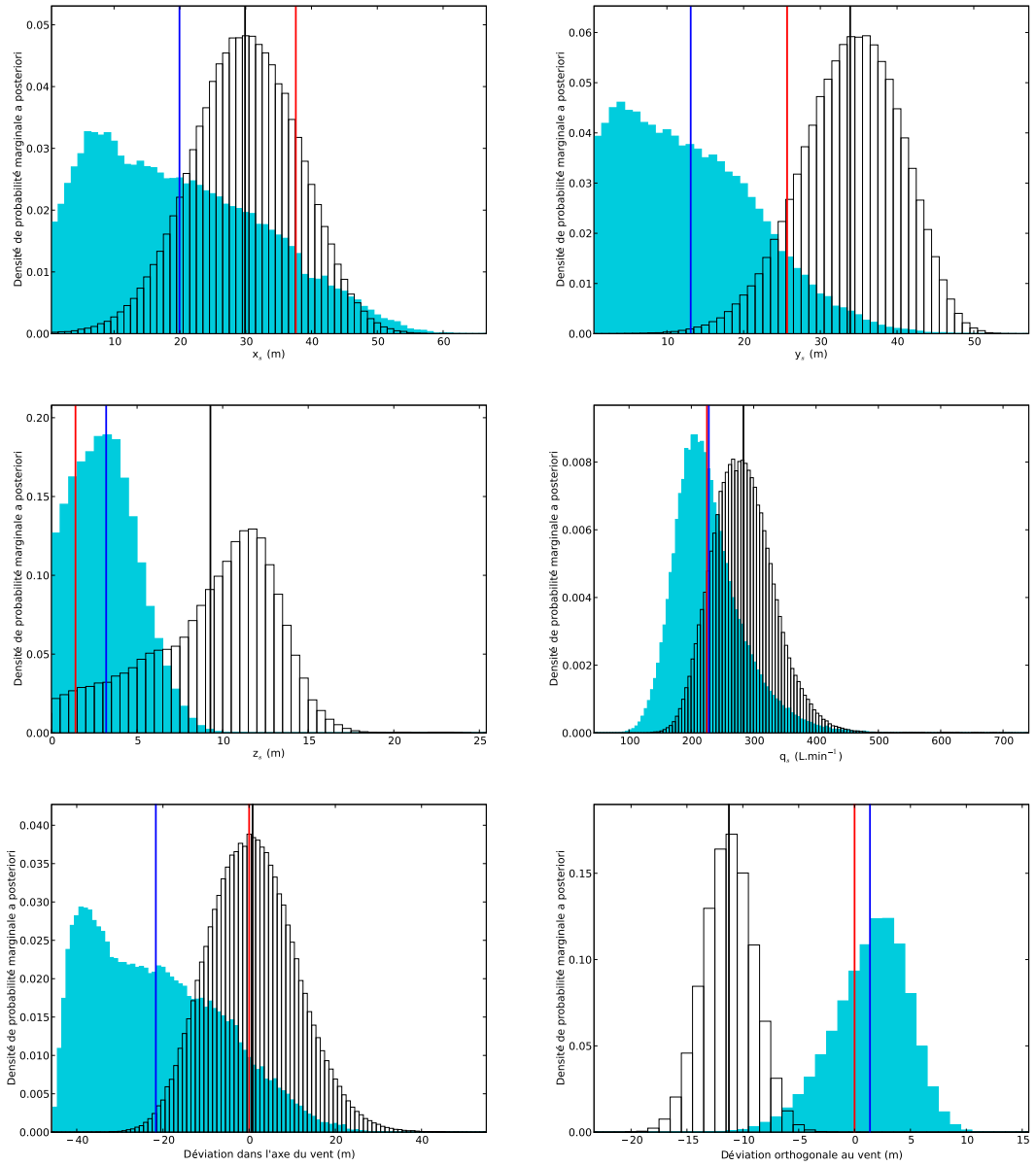


FIG. C.25: Cas 2692131.

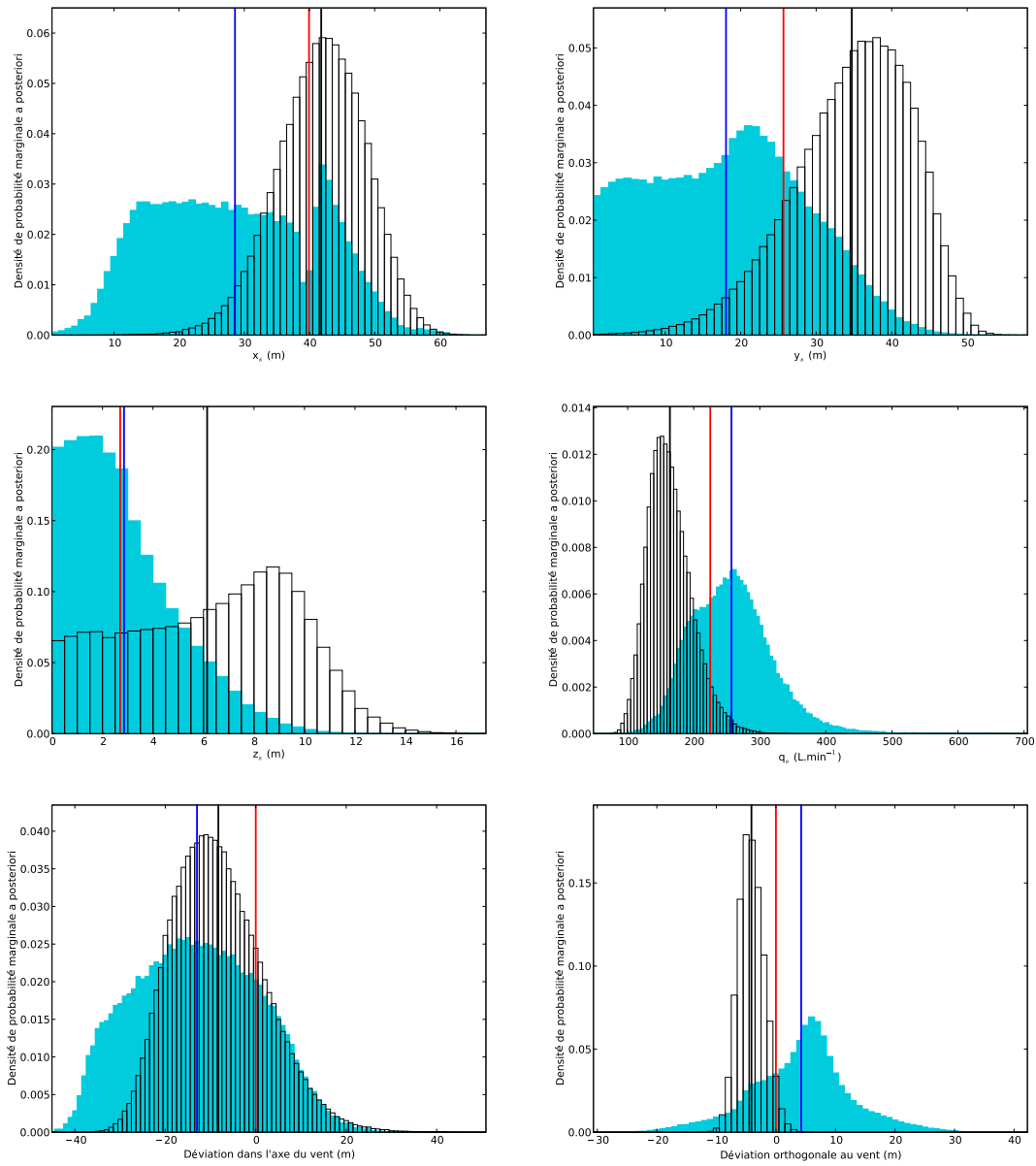


FIG. C.26: Cas 2692157.

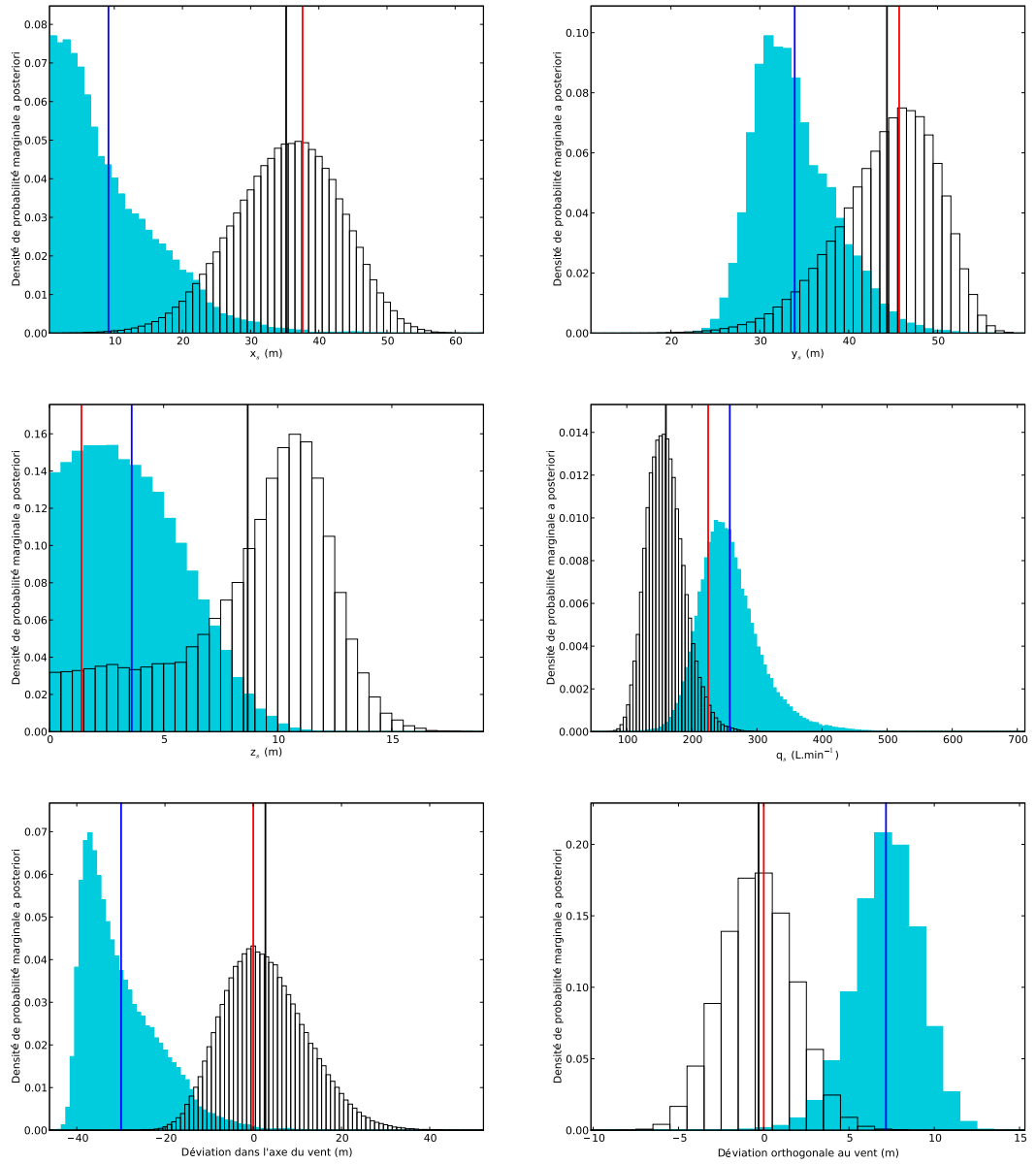


FIG. C.27: Cas 2692223.

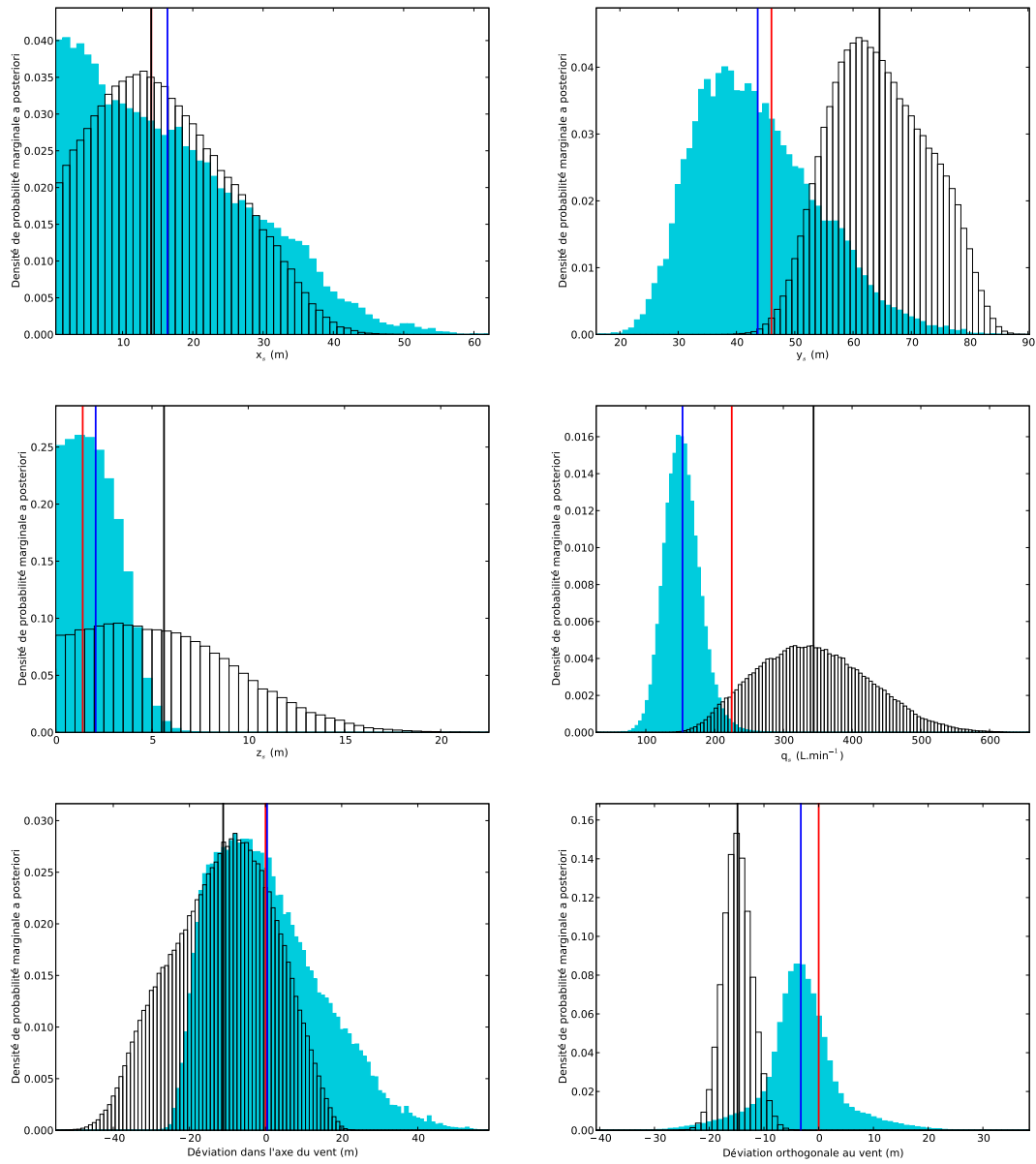


FIG. C.28: Cas 2692250.

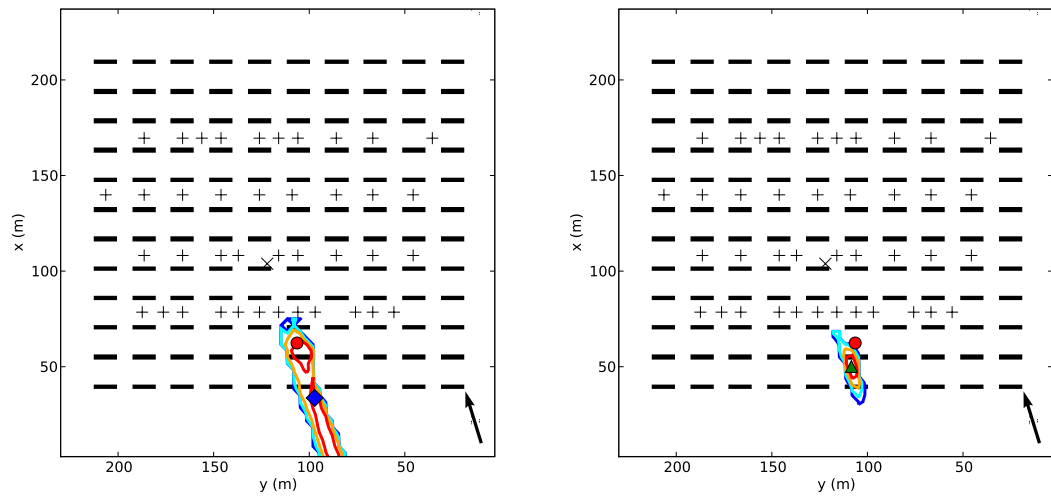


FIG. C.29: Cas 2640138.

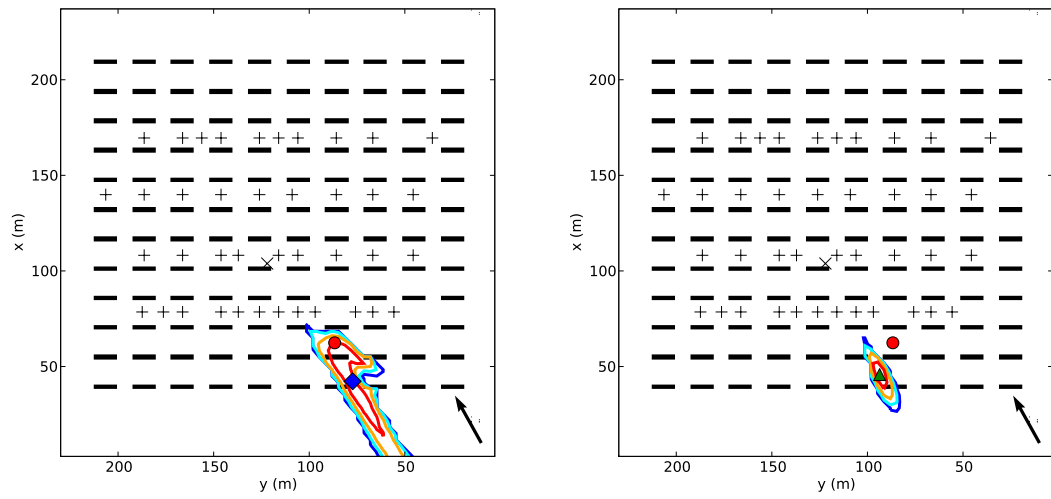


FIG. C.30: Cas 2640246.

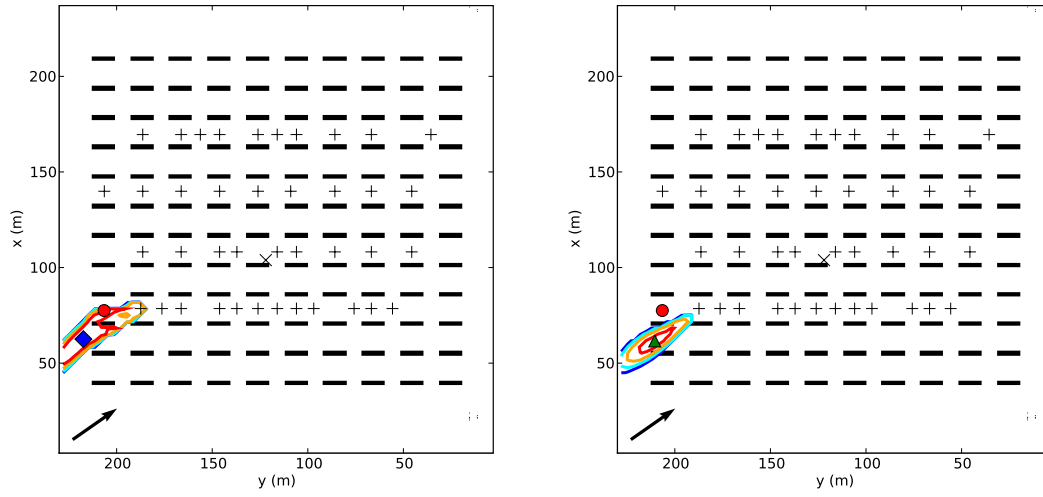


FIG. C.31: Cas 2671852.

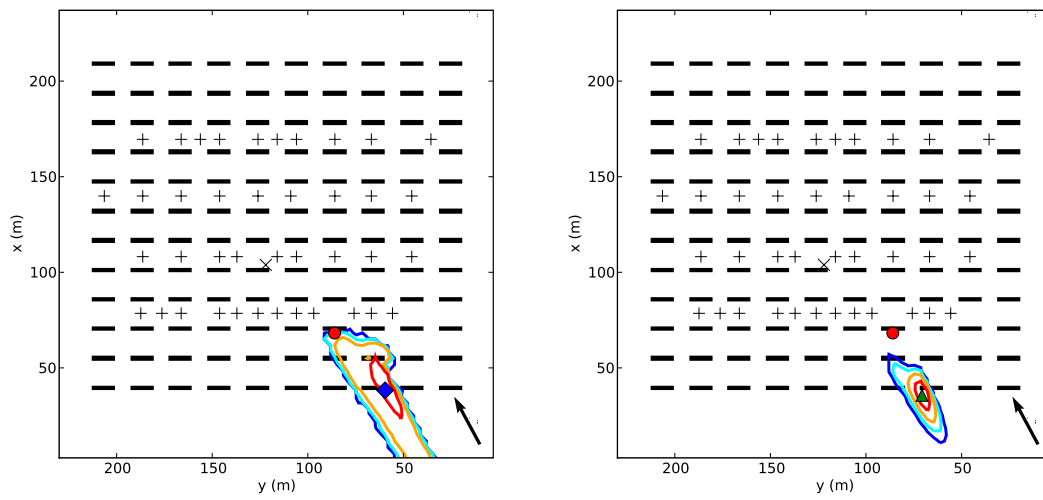


FIG. C.32: Cas 2672213.

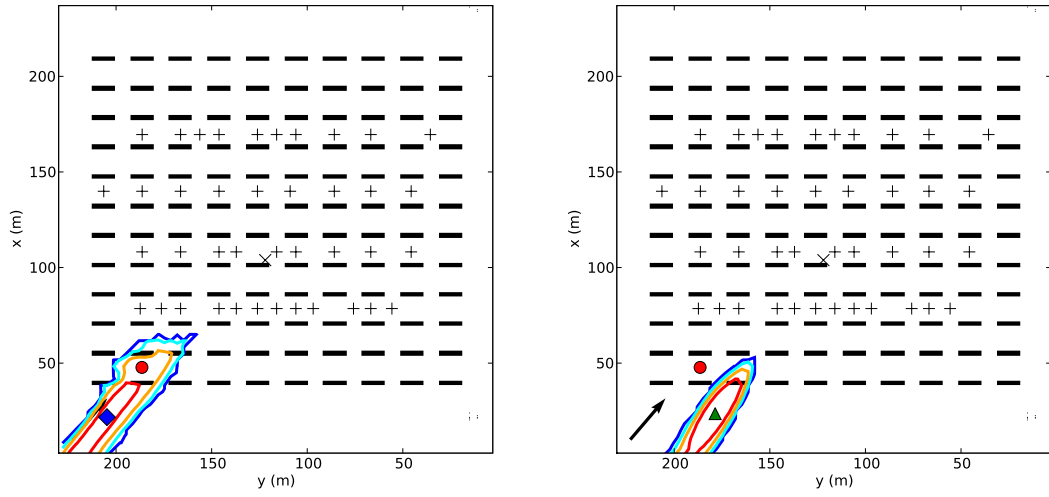


FIG. C.33: Cas 2681829.

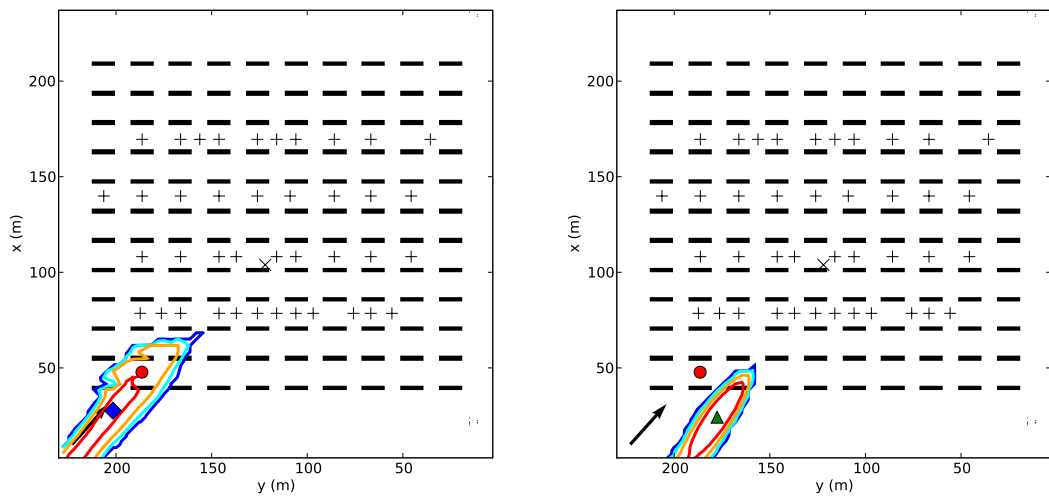


FIG. C.34: Cas 2681849.

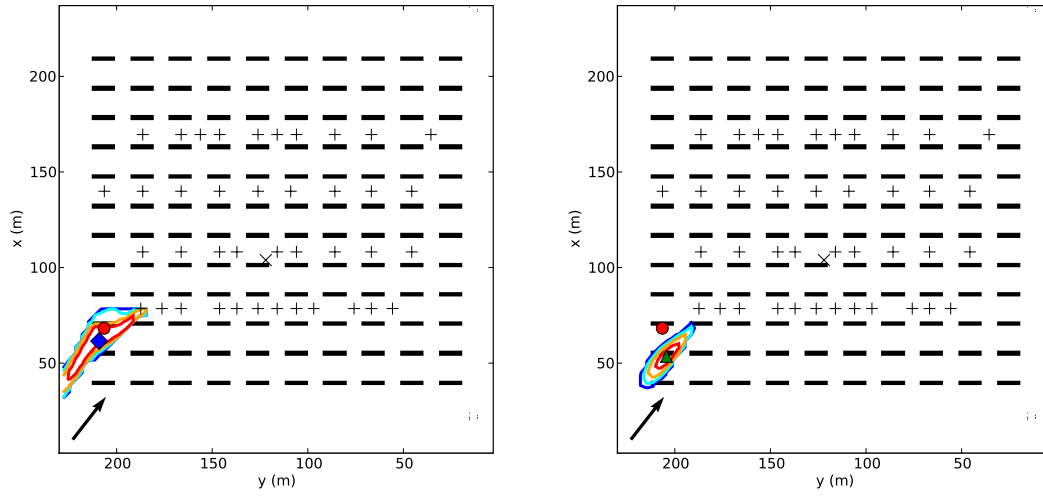


FIG. C.35: Cas 2682256.

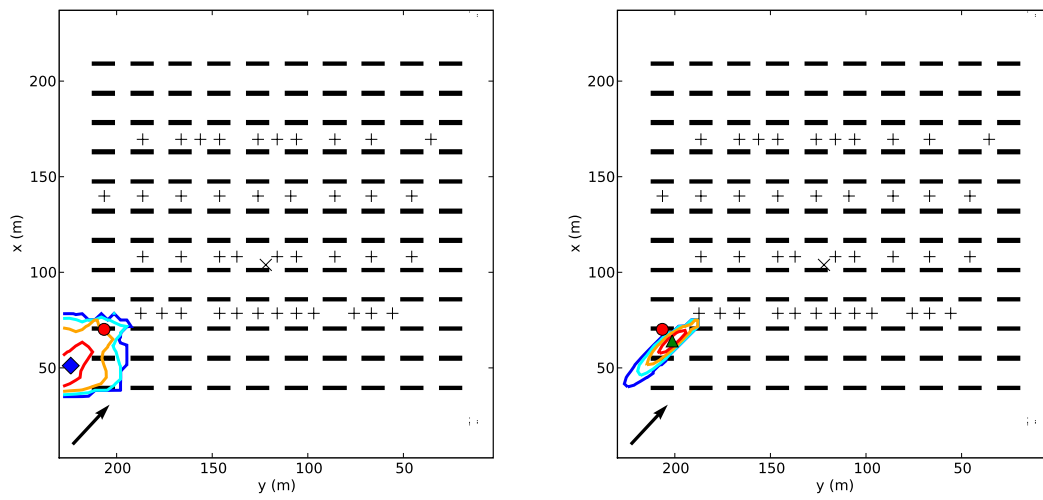


FIG. C.36: Cas 2682320.

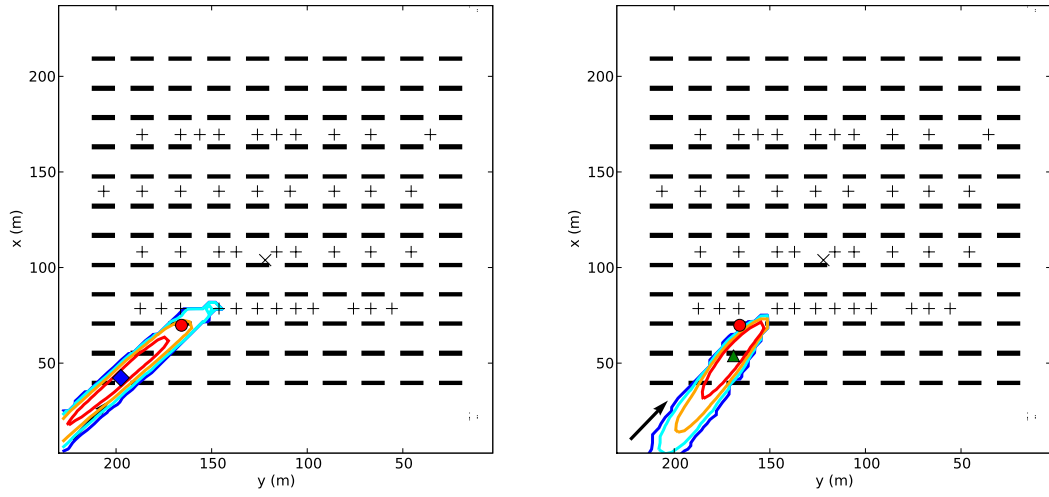


FIG. C.37: Cas 2682353.

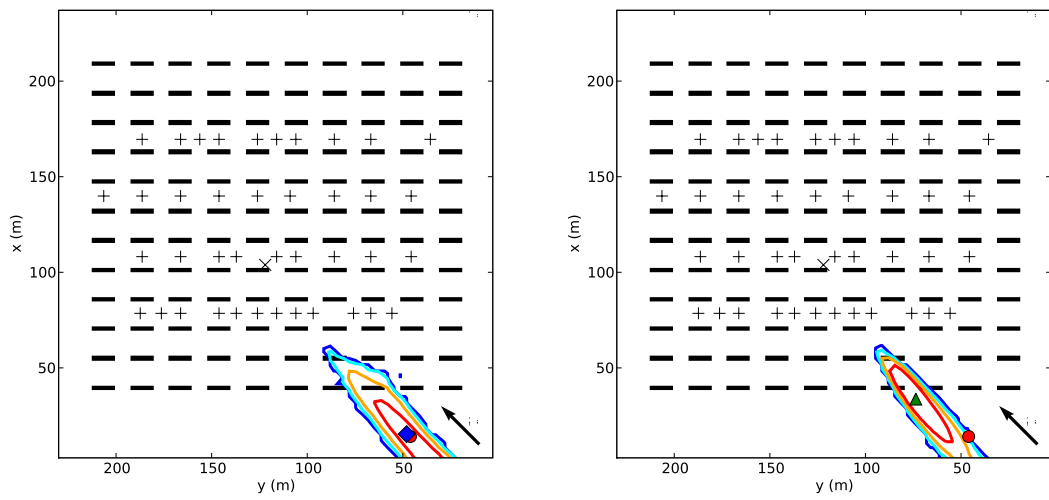


FIG. C.38: Cas 2692054.

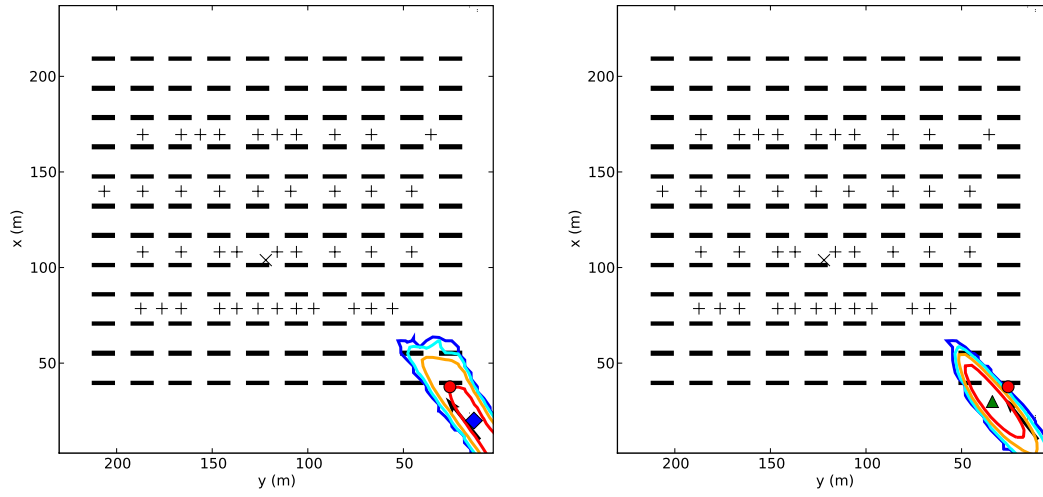


FIG. C.39: Cas 2692131.

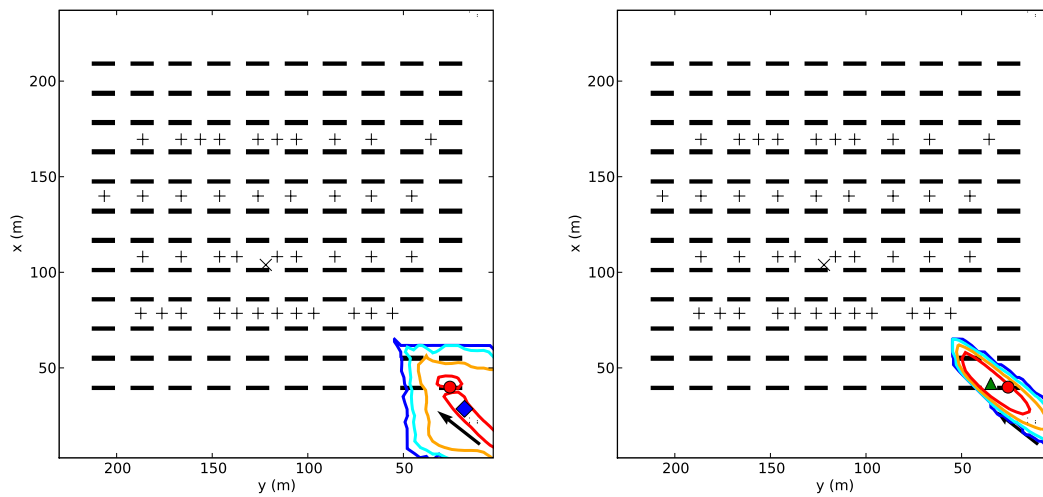


FIG. C.40: Cas 2692157.

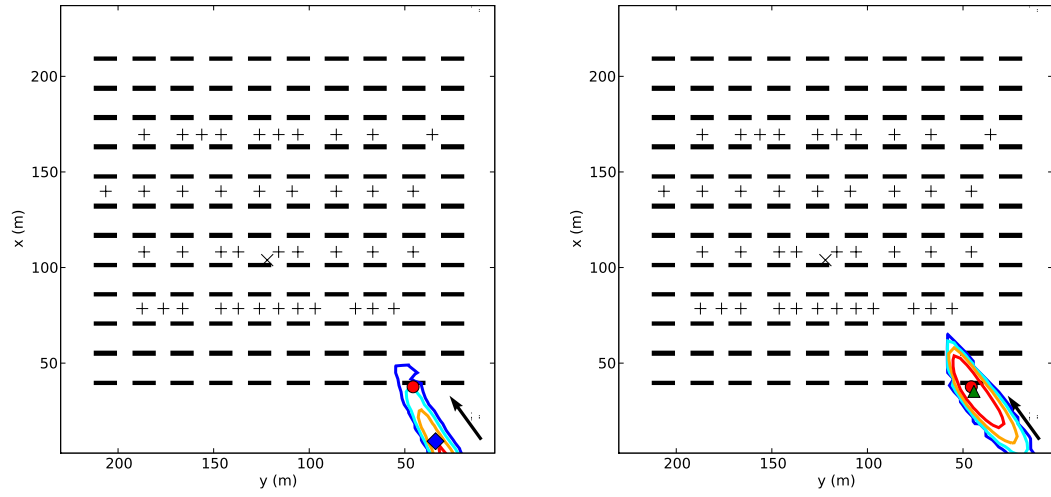


FIG. C.41: Cas 2692223.

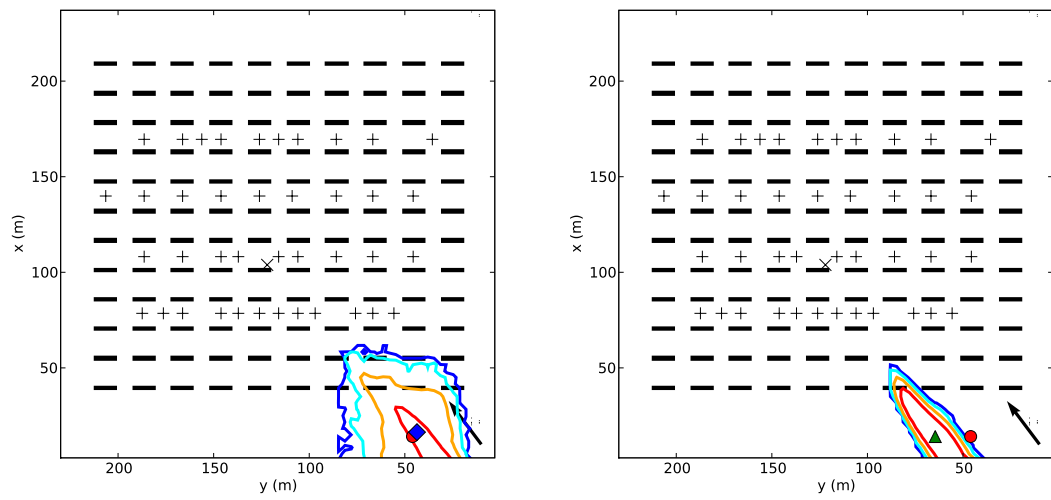


FIG. C.42: Cas 2692250.

Annexe D

Publication I

Article publié en 2011 dans Atmospheric Environment :

Winiarek, V., Vira, J., Bocquet, M., Sofiev, M., Saunier, O., 2011. Towards the operational estimation of a radiological plume using data assimilation after a radiological accidental atmospheric release. *Atmos. Env.* **45**, 2944-2955.

Dans cet article, un algorithme semi-automatique de surveillance d'un parc d'installations nucléaires est proposé. Cet algorithme permet, dans le cas d'un incident sur une installation préalablement identifiée et en utilisant les concentrations d'activité dans l'air mesurées par un réseau de mesures, de reconstruire en quasi temps réel le terme source de radionucléides, et de prévoir la dispersion du panache de polluant. Dans une seconde partie, des indicateurs statistiques bayésiens sont proposés pour déterminer, parmi un ensemble de sites potentiels, l'installation responsable d'un rejet dont les conséquences sont mesurées par un réseau de mesures mais dont l'origine est inconnue. Des expériences synthétiques sont utilisées pour évaluer les performances de l'algorithme semi-automatique et des indicateurs statistiques d'identification.

Towards the operational estimation of a radiological plume using data assimilation after a radiological accidental atmospheric release

Victor Winiarek^a, Julius Vira^b, Marc Bocquet^{a,c}, Mikhail Sofiev^b, Olivier Saunier^{d,a}

^aUniversité Paris-Est, CEREAs, joint laboratory École des Ponts ParisTech and EDF R&D, Champs-sur-Marne, France

^bFinnish Meteorological Institute, Helsinki, Finland

^cINRIA, Paris Rocquencourt research centre, France

^dInstitute of Radiation Protection and Nuclear Safety, BP 17, 92262, Fontenay-aux-Roses, France

Abstract

In the event of an accidental atmospheric release of radionuclides from a nuclear power plant, accurate real-time forecasting of the activity concentrations of radionuclides is required by the decision makers for the preparation of adequate countermeasures. The accuracy of the forecast plume is highly dependent on the source term estimation. On several academic test cases, including real data, inverse modelling and data assimilation techniques were proven to help in the assessment of the source term.

In this paper, a semi-automatic method is proposed for the sequential reconstruction of the plume, by implementing a sequential data assimilation algorithm based on inverse modelling, with a care to develop realistic methods for operational risk agencies. The performance of the assimilation scheme has been assessed through the intercomparison between French and Finnish frameworks. Two dispersion models have been used: POLAIR3D and SILAM developed in two different research centres. Different release locations, as well as different meteorological situations are tested. The existing and newly planned surveillance networks are used and realistically large multiplicative observational errors are assumed. The inverse modelling scheme accounts for strong error bias encountered with such errors. The efficiency of the data assimilation system is tested via statistical indicators. For France and Finland, the average performance of the data assimilation system is strong. However there are outlying situations where the inversion fails because of a too poor observability. In addition, in the case where the power plant responsible for the accidental release is not known, robust statistical tools are developed and tested to discriminate candidate release sites.

This article has been published in Atmospheric Environment with the reference: Winiarek, V., Vira, J., Bocquet, M., Sofiev, M., Saunier, O., 2011. Towards the operational estimation of a radiological plume using data assimilation after a radiological accidental atmospheric release. *Atmos. Env.* **45**. 2944-2955.

Keywords: Data assimilation, Atmospheric dispersion, Radionuclides, Emergency response

1. Introduction

1.1. Context

Emergency centres that are in charge of real-time estimation of the impact of an accidental release of pollutant, are getting interested in testing and implementing data assimilation techniques for source term estimation as well as for plume dispersion forecast.

In the case of the IRSN, the French nuclear safety agency, a simplistic Gaussian plume model was used so

far as an operational forecast model. An Eulerian modelling platform, *ldX* based on the POLYPHEMUS platform (Quélo et al., 2007), is being implemented and should serve as their reference forecast tool.

Since the beginning of this century, academic researchers have been developing and testing, sometimes with real data, inverse modelling techniques for accidental release of pollutant, using Gaussian, Eulerian and Lagrangian models. With increasing confidence, they were able to demonstrate, at least at an academic level, the potential of such techniques to operational centres. In numerical weather forecasting, data assim-

Email address: victor.winiarek@cerea.enpc.fr (Victor Winiarek)

ilation has been implemented operationally very early: the European Centre for Medium-range Weather Forecast demonstrated that mathematical techniques, such as 4D-Var, could be used operationally with success on an even more complex system.

That is why an agency such as the IRSN is getting interested in the potential of data assimilation for their operations. This paper addresses with synthetic examples what a data assimilation system built on hypotheses as simple as possible, but without compromising the efficiency of the system (essentially the forecast skill), could achieve.

1.2. *State of the art*

Firstly, we would like to review the current status of inverse modelling and data assimilation of atmospheric tracer from an accidental origin, at continental scale. All the methods cited below are Bayesian in nature, and that is why they should not fundamentally yield different results.

Many of published inverse modelling techniques in atmospheric tracer source inversion, are of non-parametric nature: they attempt to retrieve large spatial fields of variables, rather than a few selected parameters. One typically looks for a source field, usually multi-dimensional, typically one dimension for time and two dimensions for the ground. If one knows the location of the source, then a one-dimensional time rate profile is usually sought. Robertson and Langner (1997); Seibert (2001) were among the first to attempt such an inversion using real data of the European Tracer Experiment (Nodop et al., 1998) (ETEX). Issartel and Baverel (2003); Issartel (2003) used projection methods (that are equivalent to a 4D-Var approach in a linear model context) and demonstrated the interest of a simplex method on synthetic data case inspired from ETEx. Bocquet (2005a) suggested to use a maximum entropy on the mean inference in order to implement additional constraints on the source such as positivity and boundedness. It was used successfully on the real ETEx data (Bocquet, 2005b, 2007; Krysta et al., 2008), on the Algeciras dispersion incident (Krysta and Bocquet, 2007) and on the Chernobyl accident (Davoine and Bocquet, 2007).

An advantage of these non-parametrical methods is the robustness that often stems from the convexity of the underlying cost functions. Another advantage is their immediate generalisation to more complex problems, such as multiple sources or source with a significant duration. They also share common grounds with other atmospheric chemistry problems such as ozone precursors source inversion in air pollution modelling

or greenhouse gas fluxes inversion. Being deterministic methods, their aim is to provide an estimate of the source, but second-order sensitivity analysis is also possible (Bocquet, 2008).

In the case of the inversion of a point-wise accidental atmospheric source, one could use parametrical methods instead (relying on a few parameters), by looking for the location, time and rate of the source. For instance, Yee et al. (2008) simply compute the marginal posterior probability density function (pdf) on a few selected parameters on the ETEx experiment. Delle Monache et al. (2008) apply Markov Chain Monte Carlo techniques for a stochastic estimation of the Algeciras source location. The major advantage of these methods is the limited number of parameters that are retrieved. Owing to the small number of parameters, the full pdf (or marginal pdf) of these parameters can be retrieved.

Assimilation of observations for the forecast of accidental plume dispersion has been investigated in Politis and Robertson (2004); Bocquet (2007) on the ETEx-I case, and on a hypothetical regional case around a power plant in Abida and Bocquet (2009). It was shown that the plume can efficiently be controlled because of the non-chaoticity of dispersion dynamics, provided the source is well retrieved.

1.3. *Objectives and outline*

The objective of this paper is to assess the performance of a simple inverse modelling and data assimilation system. In particular one is interested in not only the average performance of such a tool, but also in cases where it may fail, which is of utmost importance toward an implementation in operational centres. Different situations (sources, monitoring network, dispersion tools) of two different countries, France and Finland, are explored.

The setup of the data assimilation experiments, along with the models that are employed are described in Section 2. An efficient data assimilation system, common to the two national contexts, is presented in the same section. Although the methodology is simple, it differs from simpler least-square schemes, because it is assumed that the errors, being multiplicative, are source dependent.

Within the framework of Section 2, a sensitivity analysis is carried out in Section 3, varying the context: France or Finland, the meteorology for a fixed typical power plant, the power plants, the meteorological conditions, the monitoring network, etc. The data assimilation methodology assumes here that the origin of the radiological release, one of the power plants, is

known. This should be realistic as the management of the nuclear power plant at risk should immediately raise an alert (to the national authority, possibly European Union, or the International Atomic Energy Agency). However, history has shown that this is not always the case. Besides, the accident may occur on a location abroad. It is therefore necessary to complement the operational data assimilation system with either a more complex system that does not assume that the location of the source is known, or with statistical tests that would assess the probability of a site to be responsible for the release. Since this paper is aimed at easily implementable tools, the focus is on this second option. Bayesian statistical tools are developed and tested in Section 4.

The conclusions and perspectives are given in Section 5.

2. Methods and setup

2.1. France and Finland setup

As far as France is concerned, all the civil nuclear facilities are monitored, that is 19 nuclear power plants and a nuclear fuel reprocessing plant located at La Hague. The monitoring network that is used is the one that has been proposed by Saunier et al. (2009), called DESCARTES then, and renamed recently OPERA-AIR ($\alpha = 1$ configuration as described in Saunier et al. (2009)). It is made of 100 stations, that measure activity concentrations over a range of $10^{-6} - 10^9$ Bq.m⁻³.

As for Finland, it was chosen to monitor 6 sites. Two of them are located in Finland (one of them, Olkiluoto, potentially stands for two power plants as a third generation European pressurised reactor is under construction there). Two others are in Sweden and the last two ones are in Russia. The monitoring network is the actual ‘‘Uljas’’ Finnish ambient rate dose monitoring network of 255 stations. The power plant locations as well as the monitoring networks are represented in Fig. 1.

As a simplification, we have hypothesised that the stations of the Finnish and the French networks are measuring activity concentrations instead of ambient dose. This assumption simplifies the method since the assimilation of ambient dose would require a spatial model integration, and would require to properly take into account contamination effects of dose instruments.

The Finnish power plants are located on the shores, and the Swedish and Russian power plants are quite far from the Finnish monitoring network. This situation can lead to difficulties in the inverse modelling process. This is an important difference with the French setup where

the monitored power plants cover the country land, even though not uniformly, and they are therefore well surrounded by the measuring stations.

2.2. Dispersion modelling with POLAIR3D and SILAM

The first numerical chemistry-transport model used in this study is POLAIR3D, the Eulerian model of the POLYPHEMUS platform. As far as radionuclides are concerned, it has been validated on the European Tracer Experiment, on the Algeciras incident and on the Chernobyl accident (Qu  lo et al. (2007)).

After the gamma dose instruments, most of the current radionuclides monitoring devices are particle filters. That is why caesium-137, which disperses mostly as fine particulate matter, was chosen as the species of interest. The model integrates the concentration field c of ¹³⁷Cs, following the transport equation

$$\frac{\partial c}{\partial t} + \text{div}(\mathbf{u}c) = \text{div}\left(\rho \mathbf{K} \nabla \left(\frac{c}{\rho}\right)\right) - \Lambda^s c - \Lambda^d c + \sigma \quad (1)$$

where Λ^s is the wet scavenging rate, Λ^d represents the radioactive decay and σ is the point-wise source for ¹³⁷Cs. \mathbf{K} is the matrix of turbulent diffusion, diagonal in practice. The vertical component is given by K_z , computed with Louis parametrisation (Louis, 1979). The horizontal component K_H is taken null. Except for the boundary condition on the ground, all boundary conditions are taken null. The boundary condition on the ground is

$$-K_z \nabla c \cdot \mathbf{n} = -v^{\text{dep}} c \quad (2)$$

where \mathbf{n} is the unitary vector upward oriented, and v^{dep} is the dry deposition velocity of ¹³⁷Cs.

The advection is implemented thanks to a third-order direct space-time scheme, with a Koren-Sweby flux limiter function. Because of the sharp gradients found, it is important that such a limiter be used. The diffusion scheme is integrated through an implicit second-order Rosenbrock scheme, with a three-point spatial scheme, and directional splitting.

Caesium-137 is modelled as a passive gaseous tracer with radioactive decay. Its half-life is 30 years. Its dry deposition is modelled by using a simple scheme with a constant deposition velocity: $v^{\text{dep}} = 0.5$ cm s⁻¹. As far as the wet scavenging is concerned, the parametrisation used in this study is Belot of the form $\Lambda^s = ap_0^b$, with $a = 8 \times 10^{-5}$ and $b = 0.8$. p_0 is the rain intensity, in mm/h (Baklanov and S  rensen, 2001).

The SILAM dispersion model includes both Eulerian and Lagrangian dynamical cores. A description of the model and the Lagrangian core is given by Sofiev et al.

(2006). This work uses the more recent Eulerian core, where horizontal and vertical advection is computed using the scheme by Galperin (2000). Vertical diffusion is discretised as described in Sofiev (2002) with the diffusion coefficient K_z parametrised following Genikhovich et al. (2004).

Dry deposition is computed using the standard resistance analogy approach of Wesely (1989), with resistances for aerosols parametrised after Zhang et al. (2001). The wet deposition parametrisation follows Sofiev (2000) for gases and Sofiev et al. (2006) for aerosols.

Both transport models are then integrated in the European domain with a spatial resolution of $0.25^\circ \times 0.25^\circ$. The number of grid points in the domain simulation is 240×160 for both models. The POLAIR3D model is configured with 9 vertical levels ranging from 0 to 3300 m, while 7 levels ranging up to 6300 m are used with SILAM. Results are then saved respectively on French and Finnish subdomains. On these subdomains, defined as shown in Fig. 1, the indicators of performance concerning the ability of the system to forecast the radioactive plume will be calculated.

For this study, three periods of simulation are selected. The first one is the whole 2007 year, used to evaluate the performance of the system for one specific power plant, varying the meteorological conditions. Additionally, in order to test the average performance of the system on every power plant, two shorter periods (one week long) have been chosen. The first one is the week starting 2 December 2007, selected for its strong wind conditions. The second one is the week starting 8 June 2007, characterised by its weak wind speeds. The meteorological fields used in this study are the fields calculated by the operational model from the European Centre for Medium-range Weather Forecasts (ECMWF). They have a resolution of $0.25^\circ \times 0.25^\circ$ and are available every 3 hours.

2.3. Inverse modelling methodology

The observation equation

$$\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\sigma} + \boldsymbol{\epsilon}, \quad (3)$$

encodes the source-receptor relation: $\boldsymbol{\mu} \in \mathbb{R}^d$ is the vector of activity concentration measurements; $\boldsymbol{\sigma} \in \mathbb{R}^N$ is the vector that results from the discretisation of the source term; $\boldsymbol{\epsilon} \in \mathbb{R}^d$ is a vector that represents errors in the system (instrumental, representativity or model errors); $\mathbf{H} \in \mathbb{R}^{d \times N}$ is an operator, the Jacobian matrix, that not only stands for the observation operator but also the full dispersion model.

The Jacobian matrix can be computed column by column using the forward numerical model when the number of source parameters is smaller than the number of measurements. A column corresponds to the response of the system (a set of measurements in \mathbb{R}^d) to an elementary source, typically a point-wise instantaneous source with unit emission rate (read Abida and Bocquet (2009) for more details). Alternatively, it can be computed row by row when the number of source parameters is greater than the number of measurements. In this latter case, it is necessary to use the adjoint model since each row correspond to a solution of the adjoint of the numerical model, which can be seen as a drawback since deriving the adjoint can sometimes be a technological challenge. Equation (3) sets the inverse problem when one looks for the vector components of $\boldsymbol{\sigma}$.

In order to compute each element of Eq. (3) with reliable, simple and efficient tools, with a view to operational constraints, we have made the following choices:

1. It is assumed that the source location is known and is one of the nuclear power plants that are monitored. This is consistent with the fact that in France and in Finland, the power plant management must report instantly any incident to the authorities.
2. Because the source location is known, we aim at retrieving the source rate profile. For a time resolution of one hour, this may represent of the order of a few hundreds of parameters to solve for (for instance 168 for a week of reconstruction).
3. This number of unknown parameters is small compared to the number of activity measurements that can be routinely obtained from the French or the Finnish monitoring networks. Thus, when posing the inverse problem, \mathbf{H} should be computed column by column, using the forward model. This scheme avoids the derivation of the adjoint.
4. The computation of the columns or impulses, can be performed routinely. If the time resolution is set to one hour, operational centre would hourly compute one impulse for each one of the monitored power plants. It would be stored to build up \mathbf{H} . Note that these simulations can be trivially parallelised.
5. As a consequence, an up-to-date Jacobian matrix \mathbf{H} should always be available.

The synthetic measurements $\boldsymbol{\mu}$ that are generated for the experiments contains errors $\boldsymbol{\epsilon}^t$ (called true errors) that are lognormally distributed:

$$\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\sigma}^t + \boldsymbol{\epsilon}^t \quad (4)$$

with

$$\epsilon_i^t = (e_i - 1) [\mathbf{H}\boldsymbol{\sigma}^t]_i \quad (5)$$

$\ln(e_i)$ being distributed according to $0.5 \times \mathcal{N}(0, 1)$, the Gaussian distribution of mean 0 and standard deviation 0.5. This implies that there is no bias in the median of ϵ^t . However the mean of ϵ^t is biased. As far as amplitude is concerned, these errors are significant and realistic for an atmospheric tracer dispersion event. From this point, the true errors are unknown and even the distribution of the true errors is unknown. One has only access to the (noisy) observations $\boldsymbol{\mu}$. Nevertheless prior hypotheses can be made on the errors.

If the prior hypotheses on the errors used in the data assimilation scheme would follow the true errors (that have been generated with the lognormal distribution), then one should solve the inversion using a cost function that accounts for the prior hypothesis such as in Abida and Bocquet (2009).

But realistic errors do not have to follow a lognormal distribution law, and it would be unwise to suppose a priori that the errors are lognormal, even though the actual distribution might be similar. Here, we assume a priori that the errors are Gaussian, following a normal distribution

$$p(\boldsymbol{\epsilon}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{\epsilon}^T \mathbf{R}^{-1} \boldsymbol{\epsilon}\right)}{\sqrt{(2\pi)^d |\mathbf{R}|}}, \quad (6)$$

with error covariance matrix \mathbf{R} (and $|\mathbf{R}|$ its determinant). From this point, elementary Bayes inference leads to the following cost function :

$$\mathcal{L}(\boldsymbol{\sigma}) = \frac{1}{2} \ln |\mathbf{R}| + \frac{1}{2} (\mathbf{H}\boldsymbol{\sigma} - \boldsymbol{\mu})^T \mathbf{R}^{-1} (\mathbf{H}\boldsymbol{\sigma} - \boldsymbol{\mu}), \quad (7)$$

where the first term, seemingly constant, will be at use shortly. At first, we are not assuming any regularisation of the inversion. Indeed, because $N \ll d$, no background term seems necessary, though this hypothesis will be discussed in section 3.

Then, we assume a priori that the errors are Gaussian but of multiplicative nature: \mathbf{R} is diagonal and the variances $[\mathbf{R}]_{ii}$ depend on the measurements μ_i . Assuming that the variances are multiplicatively related to the *true* (errorless) measurements $\mu_i^t = [\mathbf{H}\boldsymbol{\sigma}^t]_i$ would be accurate but unfortunately an inversion crime, since one has only access to the *noisy* observations. Thus, a second idea is to relate the variances to the (noisy) observations μ_i . But, we have proved heuristically (see Appendix B) and numerically checked that, on the assumption that the true errors are lognormal, this would lead to a large underestimation of the source by a factor of $\exp(-\frac{3}{2}\chi)$,

where χ is the variance parameter of the lognormal distribution. In an accidental context, one must avoid such an undershoot.

As a consequence, it is assumed that the error attached to the measurement μ_i is proportional to the corresponding analysed measurement $[\mathbf{H}\boldsymbol{\sigma}]_i$, which is the best approximation of μ_i that can be known: $\epsilon_i \sim r[\mathbf{H}\boldsymbol{\sigma}]_i \mathcal{N}(0, 1)$, with r chosen to be $r = 50\%$. This also accounts for the fact that prior error distribution can only ideally match the true distribution. With this formulation, one obtains the following cost function:

$$\mathcal{L}(\boldsymbol{\sigma}) = \sum_{i=1}^d \left(\ln([\mathbf{H}\boldsymbol{\sigma}]_i) + \frac{1}{2r^2} \frac{([\mathbf{H}\boldsymbol{\sigma}]_i - \mu_i)^2}{[\mathbf{H}\boldsymbol{\sigma}]_i^2} \right). \quad (8)$$

The retrieval of the source term $\boldsymbol{\sigma}$ is carried out through the minimisation of the cost function Eq. (8), using the L-BFGS-B limited-memory quasi-Newton minimiser (Liu and Nocedal (1989)). More generally the online estimation of error covariances for atmospheric problems has been discussed by Dee (1995).

2.4. Data assimilation methodology

In an emergency situation, and at current time t_n the sequential data assimilation system counts three main steps :

1. *Preprocessing step*: All available data in the interval $[t_n - \Delta t_a, t_n]$ are collected (where Δt_a is the assimilation interval). In parallel, the elementary solutions corresponding to this time interval are computed using the forward chemistry-transport model. The system needs to compute $\Delta t_a / \Delta t_{source}$ elementary solutions and store them (where Δt_{source} is the reconstructed source resolution).
2. *Analysis step*: The updated \mathbf{H} matrix is computed column-by-column using all elementary solutions from t_0 to t_n . The measurement vector $\boldsymbol{\mu}_n$ containing all the observations from t_0 to t_n is also built up. Then an estimation of the source term $\boldsymbol{\sigma}_n$ is obtained by minimisation of the cost function Eq. (8).
3. *Forecast step*: A forecast of Δt_f hours from t_n to $t_n + \Delta t_f$ is performed using the chemistry-transport model from t_0 and $t_n + \Delta t_f$. The source term used in this simulation is $\boldsymbol{\sigma}_n$ between t_0 to t_n and an estimation of the source rate between current time t_n and $t_n + \Delta t_f$. We chose to implement the persistence assumption, which means that the source rate beyond current time is taken equal to the last retrieved component of $\boldsymbol{\sigma}_n$. It is usually safer because for a decreasing rate profile, it will overestimate the source term rather than underestimate it.

This can be invalidated by the presence of a second peak in the source term as we shall see.

This simple procedure has been tested successfully by Abida and Bocquet (2009), with a view to the targeting of observations, but at a regional scale rather than at continental scale. More details can be found there.

2.5. The source term

The source term that has been selected, provided by the IRSN, corresponds to a core-meltdown without hull breach. There is a second peak 24 hours after the beginning of the accident due to an intentional release to alleviate the pressure inside the confinement hull. The source term is displayed in Fig. 2, among further results.

3. Sensitivity analysis of the data assimilation system

To compare the reference and the retrieved field (the source rate profile or the concentration field of the plume), we use several statistical indicators. For the plume, a specific figure of merit indicator fm is defined at some time τ as follows:

$$fm = \frac{\sum_{h \in S} \min([\bar{c}]_h, [c^t]_h)}{\sum_{h \in S} \max([\bar{c}]_h, [c^t]_h)}, \quad (9)$$

where S denotes the set of spatial grid cells index, c^t is the true concentration field at time τ , and \bar{c} represents the reconstructed plume field at the same time. It has been used in a similar context but in the ETEX context by Bocquet (2007).

The root mean square error (rmse) compares the source term reference and the estimate of the source at some time τ and is given by:

$$rmse = \sqrt{\frac{1}{N} \sum_{n=1}^N ([\sigma^t]_n - [\bar{\sigma}]_n)^2}, \quad (10)$$

where σ^t , $\bar{\sigma}$ are respectively the source reference and the estimated source at time τ . A Pearson correlation coefficient is also used:

$$\rho = \frac{\sum_{n=1}^N [\bar{\sigma} - \langle \bar{\sigma} \rangle]_n [\sigma^t - \langle \sigma^t \rangle]_n}{\sqrt{(\sum_{n=1}^N [\bar{\sigma} - \langle \bar{\sigma} \rangle]_n^2) (\sum_{n=1}^N [\sigma^t - \langle \sigma^t \rangle]_n^2)}}, \quad (11)$$

where $\langle \rangle$ denotes the average over time.

3.1. Time-averaged performance for one power plant

To demonstrate the performance of the data assimilation system, the focus is first on the Belleville nuclear power plant which is central in France. Over one year of meteorology (2007), an accident was considered every hour. These 8760 accidents required as many seven-day simulations with POLAIR3D. For each accident, the source term was estimated by inverse modelling using all available observations (during one week, i.e. 168 hours), and the methodology described earlier. Averaging over all accidents, the mean reconstructed source is reported in Fig. 2, along with the true source. The dotted curves are the standard deviation lower and upper limits around the mean solution. The main estimation errors occur in the vicinity of the peaks. Yet it was checked that these errors are attenuated by the use of the objective function Eq. (8) that accounts for multiplicative errors.

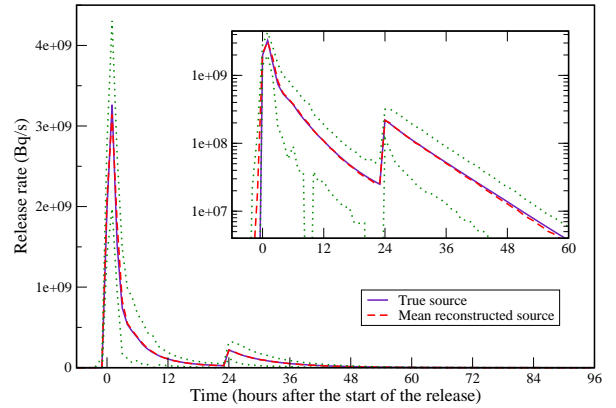


Figure 2: Mean reconstructed source (dashed line) as compared to the true source (full line) using all available observations. The upper and lower curves indicate the standard deviation range of the reconstructed sources. In insert, the same curves are represented using a logarithmic scale.

In an emergency situation, the observations are assimilated sequentially and an estimate of the source term is produced every hour. This is different from the previous a posteriori estimation which relies on all observations. Figure 3 reports the average sequential rmse on the source reconstruction. No impact of the second peak is visible on the curves because they are mainly driven by the errors on the first peak. The Pearson correlation coefficient is also reported. Obviously, the latter is not very stringent, and one should rather rely on rmse.

As for the plume forecast estimation, the figure of merit of the concentration field 3 hours after the assimilation time is plotted in Fig. 4. The impact of the uncer-

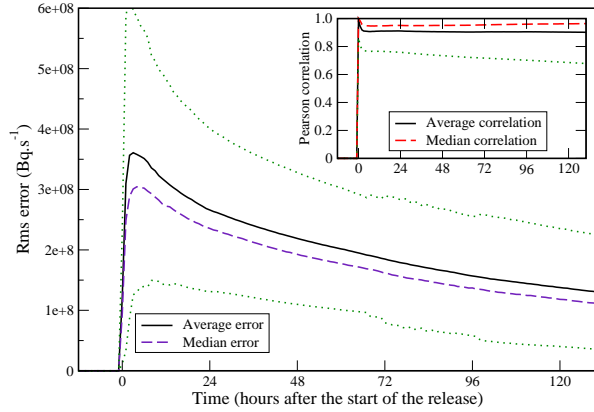


Figure 3: Mean (full line) and median (dashed line) root mean square error, assimilating observations sequentially. The upper and lower curves (dotted lines) indicate the standard deviation range of the average rmse. In inset, the Pearson correlation coefficient is displayed with a full line for the mean value and a dashed line for the median value. The mean correlation plus/minus the standard deviations are represented by the dotted curves.

tainty related to the second peak is patent with a drop of performance over four hours.

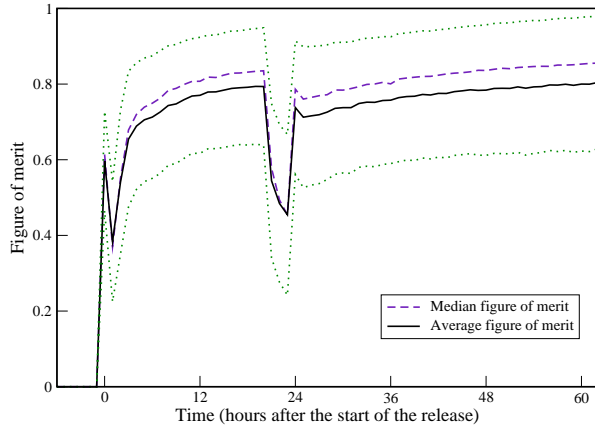


Figure 4: Mean (full line) and median (dashed line) figure of merit. The upper and lower curves (dotted lines) indicate the standard deviation range of the average figure of merit.

3.2. Global average performance

In the next step, the performance of the data assimilation system is evaluated through a more global systematic test. For the French and the Finnish context, all the power plants (20 in the French context and 6 in the Finnish one) is tested, over two different periods of 168 accidents (one accident every hour during one week).

The first period is one week in June 2007, chosen for its weak wind situation. The second period is one week in December 2007, with a strong wind situation. Like in Section 3.1, averaged performance are investigated for the source reconstruction and the forecast plume. For both periods, the results are similar to those obtained for the Belleville power plant over a whole year (same order of magnitude and evolution for the RMSE, correlations and figures of merit). The median solution is always better than the mean one, and the performance of the system are higher in the French context than in the Finnish one. This indicates that the system is mostly efficient, except in some situations where it completely fails (around 5% of the situations). These failure situations are associated to power plants located on borders or totally outside the monitoring network convex envelop.

3.3. Usefulness of a source term regularisation

In some cases, especially in the Finnish context, with monitored power plants near the shores or abroad, it was found that the inversion (through the minimisation of the cost function) may fail. This happens when the monitoring network has little overlap with the plume. It is then necessary to apply a regularisation, to use background information in the terminology of data assimilation. Some prior knowledge on the source term is hypothesised. For instance, simple prior statistics on the source is given by the Gaussian multivariate distribution

$$p(\sigma) = \frac{\exp\left(-\frac{1}{2}\sigma^T \mathbf{B}^{-1} \sigma\right)}{\sqrt{(2\pi)^N |\mathbf{B}|}}, \quad (12)$$

where \mathbf{B} is the background error covariance matrix, which will be taken diagonal in this context. The diagonal elements $[\mathbf{B}]_{ii}$ represent the a priori variances of the source, characterising some uncertainty on the source

$$[\mathbf{B}]_{ij} = E[\sigma_i \sigma_j], \quad (13)$$

for all $i, j = 1, \dots, N$.

The solution of the inverse problem is then the minimum $\bar{\sigma}$ of the cost function

$$\mathcal{L}(\sigma) = \frac{1}{2} \ln |\mathbf{R}| + \frac{1}{2} (\mu - \mathbf{H}\sigma)^T \mathbf{R}^{-1} (\mu - \mathbf{H}\sigma) + \frac{1}{2} \sigma^T \mathbf{B}^{-1} \sigma. \quad (14)$$

\mathbf{B} can simply be taken proportional to the identity $\mathbf{B} =$

$m^2 \mathbf{I}$ and then the cost function can be written as :

$$\mathcal{L}(\boldsymbol{\sigma}) = \sum_{i=1}^d \left(\ln([\mathbf{H}\boldsymbol{\sigma}]_i) + \frac{1}{2r^2} \frac{([\mathbf{H}\boldsymbol{\sigma}]_i - \mu_i)^2}{[\mathbf{H}\boldsymbol{\sigma}]_i^2} \right) + \frac{1}{2} \sum_{i=1}^N \frac{\sigma_i^2}{m^2}. \quad (15)$$

The value of scale parameter m matters, because it determines the scale of the fluctuations in the source that should be considered relevant or not. A very rough estimation is given by the maximum rate of the source. It depends on the type of accident one is monitoring and a lower one should be chosen for a less critical accident, such as a breakdown of a vapour generating tube, with a preventive release into the atmosphere of overpressure.

In our case one has $m \simeq 10^9$ Bq/s. A too large value of m would not render the regularisation operative for the critical cases (nevertheless it would not render the inversion impossible for the cases where regularisation is not necessary) whereas a too low value might smooth out relevant part of the source with high rate values that should have been reconstructed otherwise.

The scalar m could be determined objectively using hyper-parameter estimation techniques (e.g. Davoine and Bocquet (2007); Krysta et al. (2008) in a similar context). But its use in operational systems remains risky, since the result of this estimation is very sensitive to the context and is also computationally demanding.

As an example let us consider the nuclear power plant of Kalinin (in Russia) in a strong wind case. Without regularisation the inversion fails, whereas it is quite successful when a background term is added, as reported in Fig. 5. In this case, both observations generated by POLAIR3D and SILAM are employed. The two retrievals are consistent.

4. Bayesian tests for the discrimination of sources

Because one cannot always trust the nuclear facility management to report an accident immediately and because one cannot guarantee that an international alert will be raised without delay, one needs a diagnostic tool that does not assume that the location of the accident is known. Since we aim at an easily implementable tool, more academic methods that seek for the location are excluded, whatever successful they could have been on a few test-cases. Instead, we have developed Bayesian tests to help decide which nuclear site could be at the origin of the radiological plume.

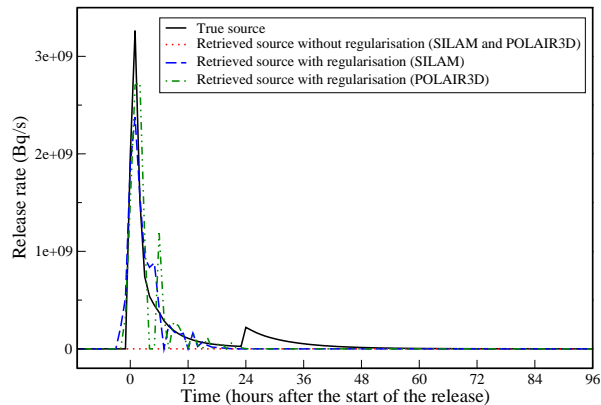


Figure 5: Impact of the regularisation in the extreme case of an inversion at Kalinin. The full line represents the true source. The retrieved source without regularisation using observations generated from POLAIR3D or SILAM are null. The retrieved sources obtained with a regularised inversion are plotted with a dashed line (SILAM) and a dashed-dotted line (POLAIR3D).

4.1. Bayesian test with Gaussian assumptions

As a first example, it is assumed that the prior statistics on the source follows the Gaussian distribution Eq. (12). This may be considered as a rough approximation since the source should not be allowed to have negative values. Yet, this simple assumption is often used in atmospheric dispersion inverse modelling with reasonable outcomes (Issartel and Baverel, 2003), although non-Gaussian assumptions are preferable (Bocquet, 2008).

The source-receptor relationship is given again by Eq. (3). Taking into account this mild prior information on the source, the likelihood of the measurements is

$$p(\boldsymbol{\mu}) = \int p(\boldsymbol{\sigma}) p(\boldsymbol{\mu}|\boldsymbol{\sigma}) d\boldsymbol{\sigma}, \quad (16)$$

where $d\boldsymbol{\sigma}$ represent the integration element over all σ_k control variables. The conditional density of the dataset knowing the source $p(\boldsymbol{\mu}|\boldsymbol{\sigma})$, is easily given by the errors

$$p(\boldsymbol{\mu}|\boldsymbol{\sigma}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})\right)}{\sqrt{(2\pi)^d |\mathbf{R}|}}, \quad (17)$$

where $\mathbf{R} \in \mathbb{R}^{d \times d}$ is the observation error covariance matrix. To simplify the integration, the error attached to measurement μ_i is no longer assumed proportional to the corresponding analysed measurement, but to the measurement itself: $\epsilon_i \sim r\mu_i \mathcal{N}(0, 1)$. It was explained earlier that it may lead to a severe under-estimation of the source if for instance the real errors follow a lognormal distribution. But source estimation is not the point

here, and this approximation is secondary for these tests. The Gaussian integration in Eq. (16) on σ then leads to (see Appendix A for details)

$$p(\boldsymbol{\mu}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{\mu}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}\boldsymbol{\mu}\right)}{\sqrt{(2\pi)^d|\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}|}}. \quad (18)$$

The likelihood $p(\boldsymbol{\mu})$ represents the probability of observing $\boldsymbol{\mu}$ assuming that the statistical distribution of the source follows Eq. (12). If one knows the location of the release, say nuclear site i , then it is straightforward to obtain the likelihood of the dataset $\boldsymbol{\mu}$ provided the source prior statistics are Gaussian, and that the source is located at site i . The Jacobian matrix \mathbf{H} can be replaced by that Jacobian matrix of site i (a submatrix of \mathbf{H}).

$$p_i(\boldsymbol{\mu}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{\mu}^T(\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T + \mathbf{R})^{-1}\boldsymbol{\mu}\right)}{\sqrt{(2\pi)^d|\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T + \mathbf{R}|}}. \quad (19)$$

$\mathcal{L} = \frac{1}{2}\boldsymbol{\mu}^T(\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T + \mathbf{R})^{-1}\boldsymbol{\mu}$, in the exponential of Eq. (19) has been shown by Bocquet (2008) to be an objective measure of the gain of information by the observations $\boldsymbol{\mu}$ with respect to the hypothesis that the source is at site i . Intuitively, the bigger this information, the smaller the likelihood of the measurements to be explained by a source at site i . It is straightforward to check that this is indeed the case on Eq. (19).

As a consequence, it is possible to use the set of p_i , with i running on all the suspected release sites. The sum $\sum_i p_i$, does not need to be 1, but the p_i can be normalised by their sum. For practical purposes, since the inversion of the matrix $\mathbf{H}_i\mathbf{B}\mathbf{H}_i^T + \mathbf{R}$ in $\mathbb{R}^{d \times d}$ can be expensive, the Sherman-Morrison-Woodbury transform can be used to invert the matrix $\mathbf{H}_i^T\mathbf{R}^{-1}\mathbf{H}_i + \mathbf{B}^{-1}$ in $\mathbb{R}^{N \times N}$ ($N \ll d$) instead, so as to speed up the computation of each $p_i(\boldsymbol{\mu})$. Then the p_i would represent normalised indices assessing the likelihood of each site, on the condition that the release took place on one of the predetermined sites.

When the latter assumption is violated, the normalised $p_i/(\sum_j p_j)$ are meaningless but the absolute low values of all p_i should reveal this violation.

4.2. Bayesian tests knowing the source term shape

The Gaussian prior Eq. (12) is not very constraining. For instance, each one of the source independent scalar rates could either be positive or negative. On the other hand, operational centres have built a list of possible source terms for several types of accidents. As a prior

piece of information, one could assume that the source term (say σ_b) belongs to this list. However it is unlikely that the magnitude of the real event would match the magnitude of this prior, even though the shape of the source rates is similar. Thus, one could rather assume that the source is of the form $\lambda\sigma_b$, with λ a positive parameter distributed according to some given law $p(\lambda)$.

The prior distribution of the source is then formally given by

$$p(\sigma) = \int_0^\infty \delta(\sigma - \lambda\sigma_b)p(\lambda) d\lambda. \quad (20)$$

As a result the observations likelihood reads

$$p(\boldsymbol{\mu}) = \int p(\sigma)p(\boldsymbol{\mu}|\sigma) d\sigma = \int_0^\infty p(\lambda)p(\boldsymbol{\mu}|\lambda\sigma_b) d\lambda. \quad (21)$$

For instance, λ could follow a gamma distribution

$$p(\lambda) = \frac{\lambda^{k-1}e^{-\lambda/\theta}}{\theta^k\Gamma(k)}. \quad (22)$$

In the case ($k = 1$, exponential law), the likelihood reads (see Appendix C for details)

$$p(\boldsymbol{\mu}) = \sqrt{\frac{\pi}{2}} \frac{e^{-\frac{1}{2}\boldsymbol{\mu}^T\mathbf{R}^{-1}\boldsymbol{\mu}}}{\theta\sqrt{(2\pi)^d|\mathbf{R}|}} \frac{e^{\frac{(\boldsymbol{\mu}^T\mathbf{R}^{-1}\boldsymbol{\mu}_b - \theta^{-1})^2}{2\boldsymbol{\mu}_b^T\mathbf{R}^{-1}\boldsymbol{\mu}_b}}}{\sqrt{2\boldsymbol{\mu}_b^T\mathbf{R}^{-1}\boldsymbol{\mu}_b}} \times \left[1 - \Phi\left(\frac{\theta^{-1} - \boldsymbol{\mu}^T\mathbf{R}^{-1}\boldsymbol{\mu}_b}{\sqrt{2\boldsymbol{\mu}_b^T\mathbf{R}^{-1}\boldsymbol{\mu}_b}}\right) \right], \quad (23)$$

where $\Phi(u) = \frac{2}{\sqrt{\pi}} \int_0^u e^{-x^2} dx$ is the error function, and $\boldsymbol{\mu}_b = \mathbf{H}\sigma_b$ is the errorless set of observations that would be obtained from the first guess source term σ_b . This case favours low scale factor λ . Setting $\theta = 1$ implies that the average source term is σ_b .

A more appropriate prior distribution of λ could be a gamma distribution of index $k = 2$ which excludes too low values for λ , and reaches a maximum at θ (the mean being 2θ). The likelihood reads

$$p(\boldsymbol{\mu}) = \frac{e^{-\frac{1}{2}\boldsymbol{\mu}^T\mathbf{R}^{-1}\boldsymbol{\mu}}}{\theta^2\sqrt{(2\pi)^d|\mathbf{R}|}} \frac{1}{\boldsymbol{\mu}_b^T\mathbf{R}^{-1}\boldsymbol{\mu}_b} \times \left\{ 1 - \sqrt{\frac{\pi}{2}} \frac{\theta^{-1} - \boldsymbol{\mu}^T\mathbf{R}^{-1}\boldsymbol{\mu}_b}{\sqrt{\boldsymbol{\mu}_b^T\mathbf{R}^{-1}\boldsymbol{\mu}_b}} \left[1 - \Phi\left(\frac{\theta^{-1} - \boldsymbol{\mu}^T\mathbf{R}^{-1}\boldsymbol{\mu}_b}{\sqrt{2\boldsymbol{\mu}_b^T\mathbf{R}^{-1}\boldsymbol{\mu}_b}}\right) \right] \right\}. \quad (24)$$

In order for the average source term to be σ_b , one could choose in this case $\theta = 1/2$. But $\theta = 1$ is preferable

because it would set the mode of the distribution to $\lambda = 1$ and hence to the source term σ_b .

Lastly, a semi-Gaussian distribution of density $p(\lambda) = \sqrt{\frac{2}{\pi}} e^{-\frac{\lambda^2}{2\theta}}$, would yield

$$p(\boldsymbol{\mu}) = \frac{e^{-\frac{1}{2}\boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}}}{\sqrt{\theta} \sqrt{(2\pi)^d |\mathbf{R}|}} \frac{e^{-\frac{(\boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}_b)^2}{2(\theta^{-1} + \boldsymbol{\mu}_b^T \mathbf{R}^{-1} \boldsymbol{\mu}_b)}}}{\sqrt{\theta^{-1} + \boldsymbol{\mu}_b^T \mathbf{R}^{-1} \boldsymbol{\mu}_b}} \times \left[1 + \Phi \left(\frac{\boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}_b}{\sqrt{2(\theta^{-1} + \boldsymbol{\mu}_b^T \mathbf{R}^{-1} \boldsymbol{\mu}_b)}} \right) \right]. \quad (25)$$

One may investigate the skills of these Bayesian tests where the shape of σ_b is the one of the true source but with the wrong scale $\sigma_b = \lambda_t \sigma_t$, since it is quite difficult to guess a priori the right scale. In that case defining $\|\boldsymbol{\mu}\|_{\mathbf{R}^{-1}}^2 = \boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}$, $p(\boldsymbol{\mu})$ only depends on the norm $\|\boldsymbol{\mu}_b\|_{\mathbf{R}^{-1}}$, and on the true scale parameter λ_t .

4.3. Validation

The following validation of the tests has been carried out for both POLAIR3D and SILAM.

4.3.1. Performance with time

At first we choose $m = 10^9$ Bq/s for the Gaussian test and $\lambda_t = 1$ for the other (non-Gaussian) Bayesian tests, which corresponds to optimal conditions for these tests. We implement these new Bayesian tests on the case of Sosnovy Bor. The (synthetic) true accident originates from this plant and the likelihood of Sosnovy Bor, Loviisa, Olkiluoto, Kalinin, Oskarshamn, and Forsmark to be the source is computed at 1, 2, ..., 7 hours after the beginning of the accident. The results for the Gaussian test are displayed in panel (a) of Fig. 6. The results for the other Bayesian tests look very similar but they manage to discriminate the correct release site faster than the Gaussian test (typically 2 hours faster).

A much more tangential case, with a poorly observed plume, is given by a fictitious accident occurring at Kalinin. In that specific example, the Gaussian Bayesian test fails to discriminate Kalinin from the other sites mentioned above. However the three other non-Gaussian tests managed to do so gradually, as reported in panels (b), (c) and (d) of Fig. 6. They almost perform equally well, with an understandable asset for the tests that have knowledge of the rate profile, up to a scalar factor.

4.3.2. Changing the prior fluctuation scale

As mentioned earlier, the prior scale of the source should be chosen wisely depending on the expected type of accident, and the order of magnitude of the source term. We have studied the performance of these Bayesian tests when the scales m or λ_t are varied and deviate from the true scale.

One typical example is reported on Fig. (7), when Belleville is the source, and the 19 other French sites are tested. The curves showing the likelihood for three sites with prominent features (Belleville, Dampierre and Fessenheim) are shown 6 hours after the beginning of the accident. The Gaussian test is used. For very low values of m , the test cannot discriminate the nuclear power plants. For very high values of m , the test picks up the

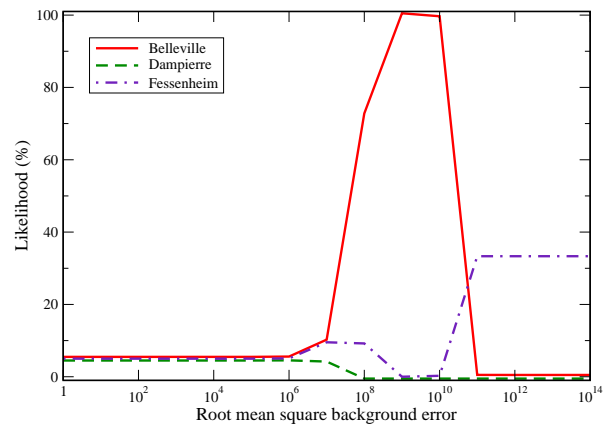


Figure 7: Gaussian Bayesian test for an accident in Belleville 6 hours after the start of the release. The root mean square of the background error which is the scale m of the fluctuation of the hypothesised source, is varied and the test is performed for $\log_{10}(m) = 0, 1, \dots, 14$.

wrong site: Fessenheim. Indeed, this site is farther away (on the French-German border) and a very large source is compatible with observations then, whereas the true source magnitude is several orders of magnitude different from the expected fluctuations thus leading to the exclusion of Belleville.

Thus, in this case, there is a range of validity for the test of four orders of magnitude.

In the case of the non-Gaussian tests, the behaviour for low values of λ_t is the same, but the performance for high values of λ_t are widely more robust, and the tests fail only for unrealistic high values of λ_t .

4.3.3. Global performance of the Bayesian tests

These indicators need to be tested on all potential sources, so as to estimate their global performance. In order to do so, the strong wind context, that is from 25

November 2007 to 25 December 2007, is considered. Every hour an accident is considered on each one of the 20 nuclear sites in France. For each batch of $p = 20$ accidents, a $p \times p$ matrix \mathbf{L} is built for all subsequent times after the start of the accidents. The entry $[\mathbf{L}]_{ij}$ is defined by the likelihood of site i in the accident from site j . Only the normalised trace, $\frac{1}{p}\text{Tr}(\mathbf{L})$ is studied. It tells how well the correct site is identified, on average. A trace is computed for each time lag from the start of the release: 1h, 2h, Each one of them is averaged over the whole strong wind 10-day period. According to Krysta and Bocquet (2007); Abida and Bocquet (2009), we expect the performance to be weaker compared to a context with more stationary winds.

First, the Gaussian Bayesian test is used, with a \mathbf{B} matrix proportional to the identity. Its proportionality coefficient is taken to be $m^2 = M^2/\tau$, where M is the total mass expected to be released, and τ is the number of one-hour time-steps after the start of the release. The non-Gaussian Bayesian tests have also been extensively tested. First the prior σ_b is taken equal to the true source, which represents the most favourable case. Then these tests were carried out with priors σ_b whose shapes were significantly different from the true source profile (for example taken as a monotonic or even as a constant term). The results are reported in Fig. 8(a) for the Gaussian tests, Fig. 8(b) for the semi-Gaussian tests, and Fig. 8(c) for the semi-Gaussian test with a wrong first guess. The average performances are impacted by outliers, but the skills of the median are excellent. These outliers are due to a poor observability of an event, that the swift dispersion of radionuclides by strong advection makes worse. The results are comparable for the Gaussian test and the non-Gaussian ones. Even when the prior σ_b is different from the true source, the skill of the median still reaches almost 100%. However, it does so later, typically 5 hours later. These results confirm the robustness of these non-Gaussian Bayesian tests.

5. Conclusion

In this paper, a semi-automatic sequential data assimilation system has been developed. In the event of an accidental atmospheric release of radionuclides from a nuclear power plant, the objective of such a system is mainly to accurately forecast the radioactive plume. This ability is strongly constrained by the knowledge of the source field and inverse modelling techniques have been used in this aim. The main concern of this paper was to use algorithms simple enough to be implemented in an operational context, but still state-of-the-art.

It was assumed that the source location is known soon enough. Consequently the Jacobian matrix can be computed efficiently using the forward dispersion numerical model, without resorting to the adjoint, and trivial parallelism can be used. Then the inversion of the source-receptor equation is performed with Gaussian assumptions for errors statistics distributions. Yet it is assumed that the variance of these errors are proportional to the activity concentration value at the observation site. At first, no regularisation and no background term has been used, though this is contemplated later.

To evaluate the performance of the system, synthetic experiments have been generated, varying the meteorological situations, the models, the monitoring network, the release locations, etc., based on the cases of France and Finland. The reconstruction statistical indicators show good results in average. The source is swiftly and well estimated and the forecast is accurate, even though the assumption of persistence for the source first guess can lead to a temporary (from 1 to 6 hours depending on the forecast time) deterioration in the forecast quality if the true source is not monotonic. Cases where the system failed (about 5% in our context) have been identified and alternate solutions have been proposed. The critical situations are all due to a lack of observability and concern mainly the power plants located on the periphery of the monitoring networks (near the frontiers, on the shores or abroad).

Aside from accounting for multiplicative errors, a second methodological novelty has been introduced. Bayesian statistical indicators have been developed and tested, when the origin of the release is not known or not soon enough. These tests determine the likelihood of a plant to be responsible for the release, knowing the observations and some prior ideas on the source term. They differ from each other by the assumptions made on the prior statistics on the source. The first indicator assumes a Gaussian distribution for these statistics. The results are very good in average, over a large range of value for the parameters in the prior. Yet, this indicator may fail in critical situations, for example for a release site far from the monitoring network. A second type of indicators have been proposed, assuming that the temporal profile of the source is known but not its exact magnitude. Then the results are excellent even when the prior temporal profile does not match the true one. The guilty power plant is quickly identified even when its location is far from the monitoring network.

Additional tests have been performed that are not reported here because they have not made systematic. For instance, what is result of the inversion for an uncharted source location?

In this paper, it was assumed that the monitoring network were measuring directly activity concentrations in the air whereas current networks are mostly reporting gamma dose. Assimilating the gamma dose measurements using the proper integrating observation operator may amplify the potential errors, and this would need to be taken into account in a future operational system.

Acknowledgements. The authors thank R. Abida for sharing his expertise and interesting discussions. This paper is a contribution to the INSU/LEFE-ASSIM project 2007-21: "Application of advanced data assimilation techniques to the dispersion of accidental release of pollutants in case of emergency".

Appendix A. Bayesian test with Gaussian assumption

If the prior statistics of the source follows the Gaussian distribution Eq. (12), i.e.

$$p(\boldsymbol{\sigma}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{\sigma}^T \mathbf{B}^{-1} \boldsymbol{\sigma}\right)}{\sqrt{(2\pi)^N |\mathbf{B}|}}, \quad (\text{A.1})$$

and since the conditional density of the dataset knowing the source is given by the errors (Eq. (17))

$$p(\boldsymbol{\mu}|\boldsymbol{\sigma}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})\right)}{\sqrt{(2\pi)^d |\mathbf{R}|}}, \quad (\text{A.2})$$

the likelihood of the measurements is given by Eq. (16)

$$p(\boldsymbol{\mu}) = \int p(\boldsymbol{\sigma}) p(\boldsymbol{\mu}|\boldsymbol{\sigma}) d\boldsymbol{\sigma} \quad (\text{A.3})$$

Direct integration of Eq. (A.3) gives

$$p(\boldsymbol{\mu}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}\right)}{\sqrt{(2\pi)^d |\mathbf{B}||\mathbf{R}|}} \times \frac{\exp\left(\frac{1}{2}\boldsymbol{\mu}^T \mathbf{R}^{-1} \mathbf{H} \left(\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1}\right)^{-1} \mathbf{H}^T \mathbf{R}^{-1} \boldsymbol{\mu}\right)}{\sqrt{|\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1}|}} \quad (\text{A.4})$$

The Sherman-Morrison-Woodbury transform, applied to the numerator of the last member of the right hand side term in Eq. (A.4), gives

$$\begin{aligned} (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} &= \mathbf{R}^{-1} - \\ &\mathbf{R}^{-1} \mathbf{H} \left(\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1}\right)^{-1} \mathbf{H}^T \mathbf{R}^{-1} \end{aligned} \quad (\text{A.5})$$

and the determinant matrix lemma for the denominator gives

$$|\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}| = |\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} + \mathbf{B}^{-1}| |\mathbf{B}||\mathbf{R}| \quad (\text{A.6})$$

Thus Eq. (A.4) becomes

$$p(\boldsymbol{\mu}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{\mu}^T \left(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}\right)^{-1} \boldsymbol{\mu}\right)}{\sqrt{(2\pi)^d |\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}|}} \quad (\text{A.7})$$

The Eq. (A.4) and Eq. (A.7) are strictly equivalent. One can use either one equation or the other to retrieve the likelihood of the set of measurements $\boldsymbol{\mu}$ provided that the source is located at site i $p_i(\boldsymbol{\mu})$ by replacing all \mathbf{H} by its submatrix \mathbf{H}_i associated with site i . The interpretation seems more convenient with Eq. (A.7) since one retrieves in the exponential the term $\mathcal{L} = \frac{1}{2}\boldsymbol{\mu}^T \left(\mathbf{H}_i \mathbf{B} \mathbf{H}_i^T + \mathbf{R}\right)^{-1} \boldsymbol{\mu}$, which has been shown by Bocquet (2008) to be an objective measure of the gain of information by the observations $\boldsymbol{\mu}$ with respect to the hypothesis that the source is at site i . Nevertheless, for practical purposes, when $N \ll d$, it is more convenient to implement Eq. (A.4), as the size of the matrix to invert (which is the critical process in the implementation) is much smaller.

Appendix B. Source bias without online error estimation

We assume that the true errors are lognormal. However, unaware of the exact distribution of the errors, the operator assumes normal observational errors that are proportional to the measurements. The observational error covariance matrix \mathbf{R} is diagonal with diagonal element $[\mathbf{R}]_{ii} = r^2 \mu_i^2$, as explained in Section 2.3. Like for the whole paper, it is assumed that the number of observation is significantly greater than the number of emission rates to retrieve. Then the solution of the inverse problem Eq. (3) is

$$\bar{\boldsymbol{\sigma}} = \left(\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{R}^{-1} \boldsymbol{\mu}, \quad (\text{B.1})$$

Each matrix element of the denominator is of the form

$$[\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}]_{kl} = \sum_{i=1}^d r^{-2} \mu_i^{-2} [\mathbf{H}]_{ik} [\mathbf{H}]_{il}. \quad (\text{B.2})$$

By the definition of the synthetic experiment, the observation μ_i reads $x_i \mu_i^t$, where μ_i^t is the true observation, which cannot be accessed directly, and x_i is a log-normal random draw. A heuristic argument is that with

many observations for a limited rate variables, this entry averages out on the log-normal errors:

$$\begin{aligned} [\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}]_{kl} &\simeq \text{E}_x \left[[\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}]_{kl} \right] \\ &\simeq \sum_{i=1}^d \text{E} \left[x_i^{-2} \right] (r \mu_i')^{-2} [\mathbf{H}]_{ik} [\mathbf{H}]_{jl}. \end{aligned} \quad (\text{B.3})$$

And, similarly for $\mathbf{H}^T \mathbf{R}^{-1} \boldsymbol{\mu}$, one obtains

$$\begin{aligned} [\mathbf{H}^T \mathbf{R}^{-1} \boldsymbol{\mu}]_k &\simeq \text{E}_x \left[[\mathbf{H}^T \mathbf{R}^{-1} \boldsymbol{\mu}]_k \right] \\ &\simeq \sum_{i=1}^d \text{E} \left[x_i^{-1} \right] r^{-2} (\mu_i')^{-1} [\mathbf{H}]_{ik}. \end{aligned} \quad (\text{B.4})$$

With a lognormal law of standard deviation parameter σ , one has $\text{E} \left[x_i^{-2} \right] \propto \exp(2\sigma^2)$, and $\text{E} \left[x_i^{-1} \right] \propto \exp(\sigma^2/2)$. As a consequence, the retrieved source is

$$\bar{\boldsymbol{\sigma}} \simeq e^{-\frac{3}{2}\sigma^2} (\mathbf{H}^T \mathbf{R}_t^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{R}_t^{-1} \boldsymbol{\mu}^t, \quad (\text{B.5})$$

where \mathbf{R}_t is the unbiased observation error covariance matrix that would have been used if one knew the true measurements. This explains the bias of $e^{-\frac{3}{2}\sigma^2}$ claimed in the text. Even though the argument is heuristic, this value has been confirmed by numerical experiments using the data assimilation system presented in this paper.

Appendix C. Bayesian tests knowing the source term shape

If we assume that the source term is of the form $\lambda \boldsymbol{\sigma}_b$ with $\boldsymbol{\sigma}_b$ being a known source shape and λ a positive parameter distributed according to some given law $p(\lambda)$, the likelihood of the dataset of measurements $\boldsymbol{\mu}$ is given by Eq. (21)

$$p(\boldsymbol{\mu}) = \int_0^{+\infty} p(\lambda) p(\boldsymbol{\mu} | \lambda \boldsymbol{\sigma}_b) d\lambda \quad (\text{C.1})$$

If λ follows a gamma distribution $p(\lambda) = \frac{\lambda^{k-1} e^{-\lambda/\theta}}{\theta^k \Gamma(k)}$ with $k = 1$ (in other words an exponential law), then this equation becomes

$$\begin{aligned} p(\boldsymbol{\mu}) &= \int_0^{+\infty} \frac{\exp\left(-\frac{\lambda}{\theta}\right)}{\theta} \times \\ &\frac{\exp\left(-\frac{1}{2}(\lambda \mathbf{H} \boldsymbol{\sigma}_b - \boldsymbol{\mu})^T \mathbf{R}^{-1} (\lambda \mathbf{H} \boldsymbol{\sigma}_b - \boldsymbol{\mu})\right)}{\sqrt{(2\pi)^d |\mathbf{R}|}} d\lambda \end{aligned} \quad (\text{C.2})$$

The first step of the integration consists in isolating the constant terms (i.e. not depending on λ), which leads,

if we name $\boldsymbol{\mu}_b = \mathbf{H} \boldsymbol{\sigma}_b$ the errorless set of observations that would be obtained from the first guess source term $\boldsymbol{\sigma}_b$, to

$$\begin{aligned} p(\boldsymbol{\mu}) &= \frac{e^{-\frac{1}{2} \boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}}}{\theta \sqrt{(2\pi)^d |\mathbf{R}|}} \times \\ &\int_0^{+\infty} \exp\left(-\frac{1}{2} \boldsymbol{\mu}_b^T \mathbf{R}^{-1} \boldsymbol{\mu}_b \lambda^2 + (\boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}_b - \theta^{-1}) \lambda\right) d\lambda \end{aligned} \quad (\text{C.3})$$

Gaussian integration over $[0, +\infty[$ is given by

$$\int_0^{+\infty} e^{-ax^2+bx} dx = \sqrt{\frac{\pi}{2}} \frac{e^{\frac{b^2}{4a}}}{\sqrt{2a}} \left[1 - \Phi\left(\frac{-b}{2\sqrt{a}}\right) \right] \quad (\text{C.4})$$

where $\Phi(u) = \frac{2}{\sqrt{\pi}} \int_0^u e^{-x^2} dx$ is the error function. In our case, $a = \frac{1}{2} \boldsymbol{\mu}_b^T \mathbf{R}^{-1} \boldsymbol{\mu}_b$ and $b = \boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}_b - \theta^{-1}$, and it leads to Eq. (23), i.e.

$$\begin{aligned} p(\boldsymbol{\mu}) &= \sqrt{\frac{\pi}{2}} \frac{e^{-\frac{1}{2} \boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}}}{\theta \sqrt{(2\pi)^d |\mathbf{R}|}} \frac{e^{\frac{(\boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}_b - \theta^{-1})^2}{2 \boldsymbol{\mu}_b^T \mathbf{R}^{-1} \boldsymbol{\mu}_b}}}{\sqrt{2 \boldsymbol{\mu}_b^T \mathbf{R}^{-1} \boldsymbol{\mu}_b}} \times \\ &\left[1 - \Phi\left(\frac{\theta^{-1} - \boldsymbol{\mu}^T \mathbf{R}^{-1} \boldsymbol{\mu}_b}{\sqrt{2 \boldsymbol{\mu}_b^T \mathbf{R}^{-1} \boldsymbol{\mu}_b}}\right) \right], \end{aligned} \quad (\text{C.5})$$

The same calculation is done for λ following a Gamma law with $k = 2$ (with a further integration by parts) and for λ following a semi-Gaussian distribution to obtain respectively Eq. (24) and Eq. (25).

References

- Abida, R., Bocquet, M., 2009. Targeting of observations for accidental atmospheric release monitoring. *Atmos. Environ.* 43, 6312–6327.
- Baklanov, B., Sørensen, J. H., 2001. Parameterisation of radionuclides deposition in atmospheric long-range transport modelling. *Phys.Chem.Earth(B)* 26 (10), 787–799.
- Bocquet, M., 2005a. Reconstruction of an atmospheric tracer source using the principle of maximum entropy. I: Theory. *Q. J. Roy. Meteor. Soc.* 131, 2191–2208.
- Bocquet, M., 2005b. Reconstruction of an atmospheric tracer source using the principle of maximum entropy. II: Applications. *Q. J. Roy. Meteor. Soc.* 131, 2209–2223.
- Bocquet, M., 2007. High resolution reconstruction of a tracer dispersion event. *Q. J. Roy. Meteor. Soc.* 133, 1013–1026.
- Bocquet, M., 2008. Inverse modelling of atmospheric tracers: Non-Gaussian methods and second-order sensitivity analysis. *Nonlin. Processes Geophys.* 15, 127–143.
- Davoine, X., Bocquet, M., 2007. Inverse modelling-based reconstruction of the Chernobyl source term available for long-range transport. *Atmos. Chem. Phys.* 7, 1549–1564.
- Dee, D., 1995. On-line estimation of error covariance parameters for atmospheric data assimilation. *Monthly weather review* 123, 1128–1144.

- Delle Monache, L., Lundquist, J. K., Kosovic, B., Johannesson, G., Dyer, K. M., Aines, R. D., Chow, F. K., Belles, R. D., Hanley, W. G., Larsen, S. C., Loosmore, G. A., Nitao, J. J., Sugiyama, G. A., Vogt, P. J., 2008. Bayesian inference and markov chain monte carlo sampling to reconstruct a contaminant source on a continental scale. *Journal of Applied Meteorology and Climatology* 47, 2600–2613.
- Galperin, M. V., 2000. The approaches to correct computation of airborne pollution advection. In: *Problems of Ecological Monitoring and Ecosystem Modelling. XVII*, St. Petersburg, Gidrometeoizdat, 2000. pp. 54–68.
- Genikhovich, E., Sofiev, M., Gracheva, I., 2004. Interactions of meteorological and dispersion models at different scales. Springer, pp. 158–166.
- Issartel, J.-P., 2003. Rebuilding sources of linear tracers after atmospheric concentration measurements. *Atmos. Chem. Phys.* 3, 2111–2125.
- Issartel, J.-P., Baverel, J., 2003. Inverse transport for the verification of the comprehensive nuclear test ban treaty. *Atmos. Chem. Phys.* 3, 475–486.
- Krysta, M., Bocquet, M., 2007. Source reconstruction of an accidental radionuclide release at European scale. *Q. J. Roy. Meteor. Soc.* 133, 529–544.
- Krysta, M., Bocquet, M., Brandt, J., 2008. Probing ETEX-II data set with inverse modelling. *Atmos. Chem. Phys.* 8, 3963–3971.
- Liu, D., Nocedal, J., 1989. On the limited memory method for large scale optimization. *Math. Program. B.* 45(3), 503–528.
- Louis, J., 1979. A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteor.* 17, 197–202.
- Nodop, K., Connolly, R., Girardi, F., 1998. The field campaigns of the European Tracer Experiment (ETEX): Overview and results. *Atmos. Environ.* 32, 4095–4108.
- Politis, K., Robertson, L., 2004. Bayesian updating of atmospheric dispersion after a nuclear accident. *Appl. Statist.* 53, 583–600.
- Quélo, D., Krysta, M., Bocquet, M., Isnard, O., Minier, Y., Sportisse, B., 2007. Validation of the Polyphemus platform on the ETEX, Chernobyl and Algeiras cases. *Atmos. Env.* 41, 5300–5315.
- Robertson, L., Langner, J., 1997. Variational assimilation of ETEX-I data. In: Nodop, K. (Ed.), *Proceedings of the ETEX Symposium on long-range atmospheric transport, model verification and emergency response*.
- Saunier, O., Bocquet, M., Matthieu, A., Isnard, O., 2009. Model reduction via principal component truncation for the optimal design of atmospheric monitoring networks. *Atmos. Env.* 43, 4940–4950.
- Seibert, P., 2001. Inverse modelling with a lagrangian particle dispersion model: application to point releases over limited time intervals. In: *Air Pollution Modeling and its Application XIV*. Plenum Press.
- Sofiev, M., 2000. A model for the evaluation of long-term airborne pollution transport at regional and continental scales. *Atmos. Env.* 34, 2481–2493.
- Sofiev, M., 2002. Extended resistance analogy for construction of the vertical diffusion scheme for dispersion models. *J. Geophys. Res.* 107(D12), 4159.
- Sofiev, M., Siljamo, P., Valkama, I., Ilvonen, M., Kukkonen, J., 2006. A dispersion modelling system silam and its evaluation against etex data. *Atmos. Env.* 40, 674–685.
- Wesely, M. L., 1989. Parametrization of surface resistances to gaseous dry deposition in regional-scale numerical models. *Atmos. Env.* 23, 1293–1304.
- Yee, E., Lien, F.-S., Keats, A., D'Amours, R., 2008. Bayesian inversion of concentration data: Source reconstruction in the adjoint representation of atmospheric diffusion. *Journal of Wind Engineering and Industrial Aerodynamics* 96, 1805–1816.
- Zhang, L., Gao, S., Padro, J., Barrie, L., 2001. A size-segregated dry deposition scheme for an atmospheric aerosol module. *Atmos. Env.* 35, 549–560.

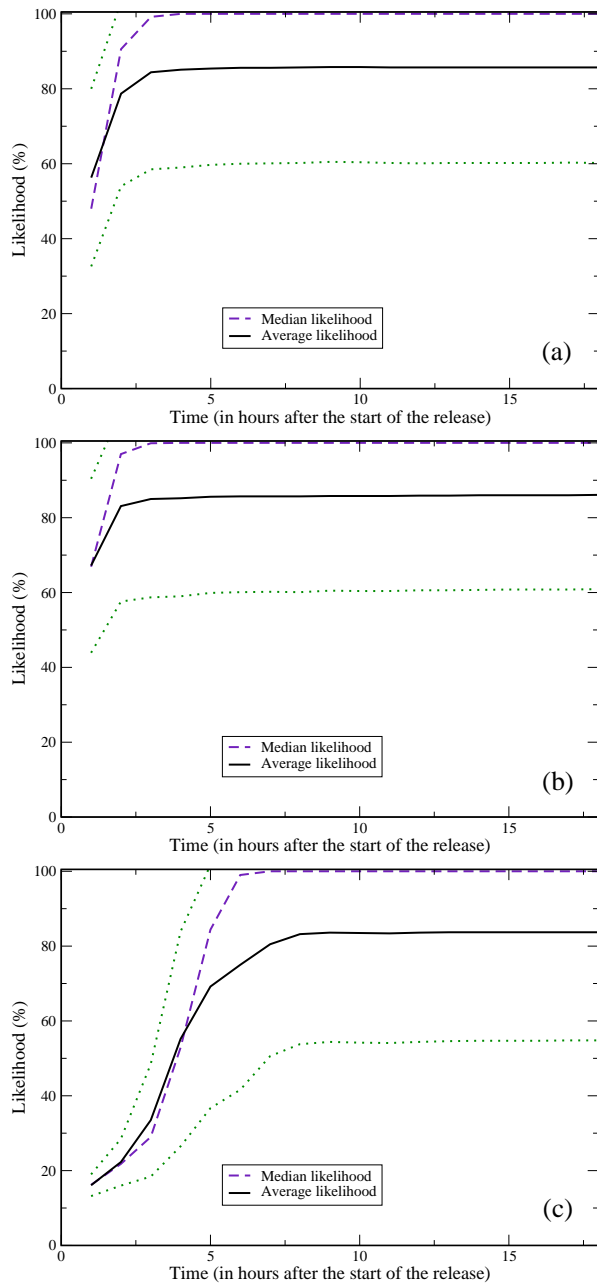


Figure 8: Mean normalised trace of the likelihood matrix L (full line) as compared to the median normalised trace (dashed line). The upper and lower curves indicate the standard deviation range of the normalised trace. Panel (a): in the case of the Gaussian Bayesian test. Panel (b): in the semi-Gaussian case of the non-Gaussian Bayesian test, with the prior σ_b equal to the true source term. Panel (c): in the semi-Gaussian case of the non-Gaussian Bayesian test, with the prior σ_b taken as a constant term, equal to the mean of the true source.

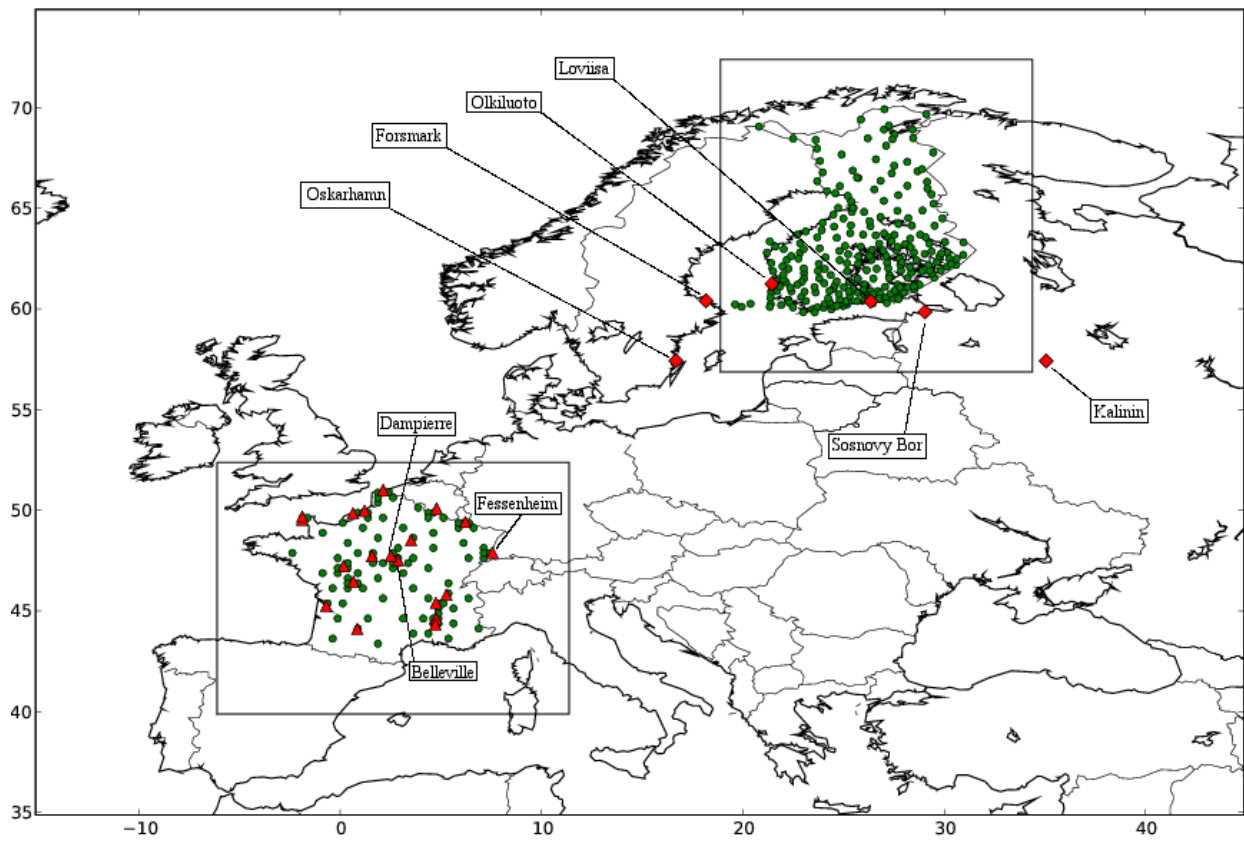


Figure 1: Representation of the domain of study and of the two subdomains centred on France and Finland. The black dots represent the monitoring stations: the OPERA-AIR network which is virtual in the French case (optimal network to be built), or the Finnish monitoring network. The locations of the civil nuclear sites are indicated by triangles and rhombuses respectively.

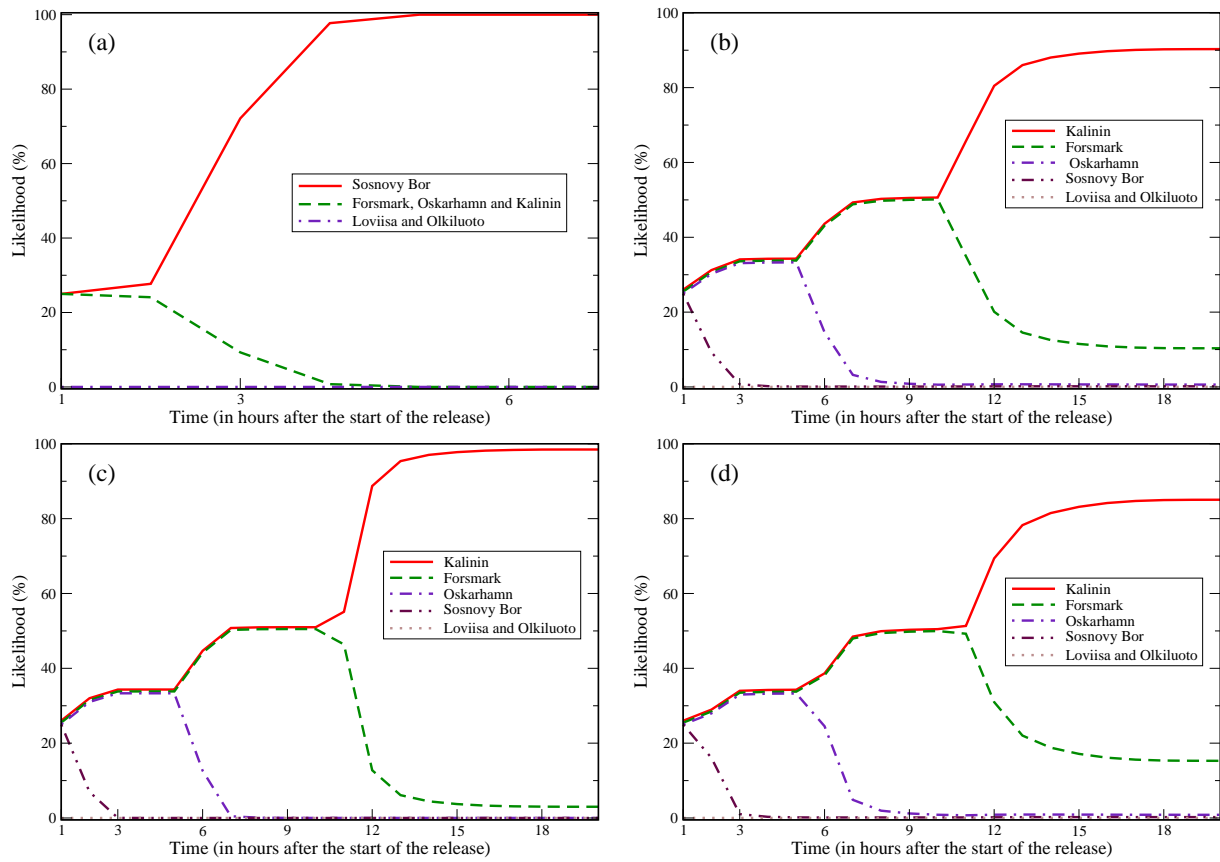


Figure 6: Panel (a): Gaussian Bayesian tests on the Sosnovy Bor case. Panel (b), (c), and (d): non-Gaussian Bayesian tests on the Kalinin case. Panel (b) corresponds to the exponential, $k = 1$ case. Panel (c) corresponds to the exponential, $k = 2$ case. Panel (d) corresponds to the semi-Gaussian case.

Annexe E

Publication II

Article publié en 2012 dans Journal of Geophysics Research Atmospheres :
Winiarek, V., Bocquet, M., Saunier, O., Mathieu, A., 2012. Estimation of errors in the inverse modeling of accidental release of atmospheric pollutant : Application to the reconstruction of the Cesium-137 and iodine-131 source terms from the Fukushima Daiichi power plant. *J. Geophys. Res.* **117**, D05122.

Dans cet article, les termes sources de césium-137 et d'iode-131 consécutifs à l'accident de Fukushima en mars 2011 sont estimés par modélisation inverse, en utilisant les mesures de concentrations d'activité dans l'air mesurés au Japon et sur un large domaine couvrant l'Océan Pacifique Nord et une partie de l'Amérique du Nord. Nous avons mis en évidence la dépendance des inversions aux statistiques des erreurs *a priori* et avons donc proposé différentes méthodes pour estimer les paramètres caractérisant ces erreurs. Les estimations obtenues, bien que représentant seulement une limite inférieure des vrais rejets du fait des conditions météorologiques particulières, sont comparables à d'autres études utilisant des méthodes et des données différentes.

Estimation of Errors in the Inverse Modeling of Accidental Release of Atmospheric Pollutant: Application to the Reconstruction of the Cesium-137 and Iodine-131 Source Terms from the Fukushima Daiichi Power Plant

Victor Winiarek,^{1,2} Marc Bocquet,^{1,2} Olivier Saunier,³ Anne Mathieu.³

Abstract. A major difficulty when inverting the source term of an atmospheric tracer dispersion problem is the estimation of the prior errors: those of the atmospheric transport model, those ascribed to the representativity of the measurements, the instrumental errors, and those attached to the prior knowledge on the variables one seeks to retrieve. In the case of an accidental release of pollutant, the reconstructed source is sensitive to these assumptions. This sensitivity makes the quality of the retrieval dependent on the methods used to model and estimate the prior errors of the inverse modeling scheme.

We propose to use an estimation method for the errors' amplitude based on the maximum likelihood principle. Under semi-Gaussian assumptions, it takes into account, without approximation, the positivity assumption on the source. We apply the method to the estimation of the Fukushima Daiichi source term using activity concentrations in the air. The results are compared to an L-curve estimation technique, and to Desroziers's scheme. The total reconstructed activities significantly depend on the chosen method. Because of the poor observability of the Fukushima Daiichi emissions, these methods provide lower-bounds for cesium-137 and iodine-131 reconstructed activities. These lower-bound estimates, 1.2×10^{16} Bq for cesium-137, with an estimated standard deviation range of 15 – 20%, and $1.9 - 3.8 \times 10^{17}$ Bq for iodine-131, with an estimated standard deviation range of 5 – 10%, are of the same order of magnitude as those provided by the Japanese Nuclear and Industrial Safety Agency, and about 5 to 10 times less than the Chernobyl atmospheric releases.

1. Introduction

1.1. Estimation of Errors in the Inverse Modeling of Accidental Release

Like any inverse modeling problem, the reconstruction of the source term of an accidental release of atmospheric tracer depends on the balance between the information load in the observation set, and the number of source parameters to retrieve, as well as the nature of the source-receptor relationship between them, provided by a numerical model of atmospheric transport (ATM). The source-receptor relationship reads

$$\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\sigma} + \boldsymbol{\epsilon}, \quad (1)$$

where $\boldsymbol{\mu}$ in \mathbb{R}^d is the measurement vector, $\boldsymbol{\sigma}$ in \mathbb{R}^N is the source vector, and \mathbf{H} is the Jacobian matrix of the transport model which is linear in this context for gaseous and particulate matter. It is clear that \mathbf{H} also incorporates the observation operator. The vector $\boldsymbol{\epsilon}$ in \mathbb{R}^d , called the observation error in this article, represents the instrumental errors, as well as the representativeness errors. By Eq. (1), it accounts for model errors as well. For the transport of tracer at regional to global scale, and in particular long-range radionuclides, it is well known that the Jacobian matrix is ill-conditioned because of the dispersive nature of large-scale

transport [Enting, 2002]. As a result, the source-receptor relationship Eq. (1) constitutes an ill-posed inverse problem.

Additionally, in an accidental context, when remote sensing means are essentially unavailable such as in the case of radionuclides, the number of observations is limited.

One solution to alleviate the lack of constraints is to parameterize the source with a very limited number of variables. This corresponds to the so-called parametric methods. *De facto*, this regularizes the inverse problem, and even allows to compute posterior probability density function for the parameters through stochastic sampling techniques [Delle Monache et al., 2008; Yee et al., 2008]. However, it is highly constraining: if the true source does not abide to this parametric model, the inverse may fail, have multiple minima or be meaningless.

Another route which has been followed for the retrieval of tracer sources, is of non-parametric nature: one seeks to retrieve a source/emission field. It must ultimately be discretized, which leads to a set of variables but this set can be large, and even significantly larger than the number of observations.

The latter approach can rely on the set of techniques developed by the field of geophysical data assimilation. It is very flexible since the source is not parameterized a priori but it has its own constraints. Firstly, under simple Gaussian statistical assumptions on the observation errors, it essentially relies in the minimization of

$$\mathcal{L}(\boldsymbol{\sigma}) = \frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}), \quad (2)$$

where \mathbf{R} is the observation error covariance matrix: $\mathbf{R} = \mathbf{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]$, where $\boldsymbol{\epsilon}$ has been defined by Eq. (1). One simple choice is to set \mathbf{R} to $r^2 \mathbf{I}_d$, assuming no correlation in the

¹Université Paris-Est, CERE, joint laboratory École des Ponts ParisTech - EDF R&D, Marne la Vallée, France.

²INRIA, Paris-Rocquencourt research center, France

³Institute of Radiation Protection and Nuclear Safety, BP 17, 92262, Fontenay-aux-Roses, France

observation errors. With a large number of source parameters, and/or an ill-conditioned physical model representation (\mathbf{H}), the minimization of Eq. (2) offers infinitely many solutions. To regularize the inverse problem, one usually adds a Tikhonov term to the cost function

$$\mathcal{L}(\boldsymbol{\sigma}) = \frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) + \frac{1}{2} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_b)^T \mathbf{B}^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_b), \quad (3)$$

which formally guarantees the existence of a unique solution, at the cost of introducing additional parameters: \mathbf{B} is the background error covariance matrix, and $\boldsymbol{\sigma}_b$ is the first guess for the source (known a priori).

Unfortunately, as opposed to the forecasting context of data assimilation in meteorology, the background information specified by $\boldsymbol{\sigma}_b$ and \mathbf{B} is uncertain. In atmospheric chemistry data assimilation that focuses on the emission retrieval, the first guess $\boldsymbol{\sigma}_b$ is usually provided by an inventory built up from survey (anthropogenic sources), or from a vegetation model (biogenic sinks and sources). As a result, \mathbf{B} can be roughly estimated from these bottom-up methodologies. Correlations in the prior source errors (related to the off-diagonal entries of \mathbf{B}) are usually used to constrain the inverse modelling problem which is a smooth way of aggregating variables. More fundamentally they may correspond to correlations in the errors when building the inventory. Recent examples using these concepts for greenhouses gas flux inversions are given by Villani et al. [2010]; Wu et al. [2011] and references therein, or, in air quality emission inverse modeling by Elbern et al. [2007] and references therein.

In the context of an accidental release, the status of $\boldsymbol{\sigma}_b$ and \mathbf{B} significantly differ, because it is difficult to form a precise idea on the source a priori. The first guess $\boldsymbol{\sigma}_b$ is often set to zero. One reason is that most of the emission rates are likely to be zero. Other reasons are discussed in Bocquet [2005]; Davoine and Bocquet [2007]. Alternatively one can contemplate using a first guess built from a model of the accident. However, an estimation of the errors made by this model needed to define \mathbf{B} is even more uncertain than the models used to build up inventories of non-accidental emission fields. Besides, one must be aware that such a prior $\boldsymbol{\sigma}_b$ is usually calibrated by observations which could lead to a multiple and uncontrolled use of observations (often referred to as an *inversion crime* in inverse modeling). In addition, the use of $\boldsymbol{\sigma}_b$ built from a model of the accident will lead to a non-independent estimation of the source term. That is why we favor the choice $\boldsymbol{\sigma}_b = \mathbf{0}$ in this article.

Still in the context of an accidental release, the presence of off-diagonal terms in \mathbf{B} , which corresponds to correlation in the estimated errors, would imply some persistence in the emission rates and the way the model makes errors about them. For accidental release this persistence is often negligible. For instance in the case of the Fukushima Daiichi accident, the gamma dose measurements at the nuclear power plant site reveals that the source term must certainly be composed of uncorrelated events.

That is why we opt for the simple uncorrelated statistical model: $\mathbf{B} = m^2 \mathbf{I}_N$. However the choice of scalar m remains crucial. In the case of Chernobyl source term reconstruction, Davoine and Bocquet [2007] have shown that the total reconstructed activity is highly dependent on the choice of r and m . This required the use of an hyper-parameter (r and m) estimation technique. In Davoine and Bocquet [2007]; Krysta et al. [2008]; Saide et al. [2011], the L-curve technique of Hansen [1992] is used to either estimate the ratio r/m or both parameters. In the context of the Chernobyl inverse modeling, Bocquet [2011] has shown that even with a number of source parameters three times inferior to the number

of observations, the poor observability of some of the source variables still leads to an ill-conditioned problem with a too high sensitivity of the results to the hyper-parameters, so that an hyper-parameter estimation technique is required.

Recently, methodological developments of data assimilation in meteorology have focussed on the estimation of such hyper-parameters. They are either based on a maximum likelihood estimation technique, or on a cross-validation technique: Mitchell and Houtekamer [1999]; Chapnik et al. [2004, 2006]; Anderson [2007]; Li et al. [2009] to cite just a few. These techniques allow online estimation of the errors. They can lead to significant improvement of the forecasting skills. These techniques have also made their way to atmospheric chemistry inverse modelling, using for instance the χ^2 criterion [Ménard et al., 2000; Elbern et al., 2007] or the maximum likelihood principle [Michalak et al., 2004].

Our inverse modeling point of view, where the background/regularization is poorly known, is different but the mathematical approach remains relevant to our problem. In this article, we would like to adapt these hyper-parameter estimation techniques to the demanding case of the reconstruction of the Fukushima Daiichi source term.

1.2. Fukushima Daiichi Accident

On March 11, 05:46 UTC, a magnitude 9.0 (M_w) undersea megathrust earthquake occurred about 70 km east of the Oshika Peninsula (Japan) in the Pacific Ocean. About one hour later, an extremely destructive tsunami hit the Pacific coastline of Japan's northern islands and resulted in the loss of thousands of lives and devastated entire towns.

These two combined disasters caused the automatic shut-down of eleven nuclear reactors in Japan, located on four nuclear power plants (Fukushima Daiichi, Fukushima Daini, Onagawa and Tokai power plants). The reactor cooling process was then operated through diesel backup power systems. Nevertheless, the damages caused by the earthquake and the tsunami rendered the backup cooling systems inoperative in the Fukushima Daiichi power plant and the reactors began to overheat.

The Fukushima Daiichi power plant has six nuclear reactors. At the time of the earthquake, reactor 4 had been de-fuelled and reactors 5 and 6 were in a cold shut-down for planned maintenance.

In the hours and days that followed, reactors 1, 2 and 3 experienced at least partial meltdown and hydrogen explosions that destroyed the upper cladding of the buildings housing reactors. Multiple fires broke out at reactor 4. Additionally fuel rods stored in pools in each reactor building began to overheat as water levels in the pools dropped.

All these events caused a massive discharge of radioactive materials in the atmosphere, especially during the first two weeks, which has to be quantified in order to estimate the sanitary and environmental impact of the accident.

1.3. Objectives and Outline

Our objective is two-fold. The final objective is to estimate the Fukushima Daiichi radionuclide source term available for long-range transport with a health impact, using radionuclides activity concentrations in the air. Because of the small number of publicly released observations, we need to extend the current available methodologies to estimate the errors in order to take into account all the information available, such as the positivity of the source term. This is the intermediary objective.

In Section 2, we briefly recall what the current available methodologies for the inverse modeling of accidental release of pollutant are. The amplitude of the statistics of the errors that is required by these algorithms was shown to be crucial for a proper estimation. Hence, the techniques to estimate these errors, are briefly recalled: L-curve, generalized cross-validation, and maximum likelihood. We propose

a generalization of the maximum likelihood scheme, that rigorously takes into account the positivity of the source. How to estimate the posterior errors, attached to the estimated sources, is also explained.

In Section 3, we apply the new method, as well as former ones, to the estimation of the atmospheric release of cesium-137 and iodine-131 from the Fukushima Daiichi plant. We demonstrate the poor observability of the release by the activity measurements, and provide lower bounds for the source term. From the methodological stand-point, the former schemes are shown to lead to significantly different estimates, and several pitfalls. The posterior uncertainty of the retrieved source is computed. Several sensitivity tests are performed. In particular we discuss the interest and impact of using a non-zero first guess.

Results are summarized and conclusions are given in Section 4.

2. Methodology: Modeling Errors

2.1. Observation Errors

Throughout this article, the observation errors, as defined by Eq. (1), which represent model, representativeness and instrumental errors altogether, are assumed Gaussian, with probability density function (pdf):

$$p(\boldsymbol{\epsilon}) = \frac{e^{-\frac{1}{2}\boldsymbol{\epsilon}^T \mathbf{R}^{-1} \boldsymbol{\epsilon}}}{\sqrt{(2\pi)^d |\mathbf{R}|}}, \quad (4)$$

where $|\mathbf{R}|$ is the determinant of the error covariance matrix \mathbf{R} . Quite naturally for errors of activity concentrations at different site, \mathbf{R} is assumed diagonal. (However, in theory, one should include small correlations induced by model error.) Besides, as mentioned in the introduction, it is assumed they have the same variance, so that $\mathbf{R} = r^2 \mathbf{I}_d$ (a property called homoscedasticity). More sophisticated approaches are possible [Abida and Bocquet, 2009; Winiarek et al., 2011]. However the latter necessarily rely on more sophisticated models for the statistics of the observation errors. With a small set of observations, they could prove less robust and too contrived.

2.2. Gaussian Modeling of Background Errors

As mentioned in the introduction we assume $\boldsymbol{\sigma}_b = \mathbf{0}$. Yet, for the sake of generality, we shall assume in this section that $\boldsymbol{\sigma}_b$ can be different. As a first step, we shall assume that the background errors are Gaussian:

$$p(\boldsymbol{\sigma}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_b)^T \mathbf{B}^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_b)}}{\sqrt{(2\pi)^N |\mathbf{B}|}}. \quad (5)$$

Because the background errors on the source are assumed uncorrelated, as discussed in the introduction, \mathbf{B} is supposed to be diagonal. For the sake of simplicity, we shall assume these errors are homoscedastic: $\mathbf{B} = m^2 \mathbf{I}_N$.

Bayes' rule helps to formulate the inference, after the acquisition of the measurement vector $\boldsymbol{\mu}$:

$$\begin{aligned} p(\boldsymbol{\sigma}|\boldsymbol{\mu}) &= \frac{p(\boldsymbol{\mu}|\boldsymbol{\sigma})p(\boldsymbol{\sigma})}{p(\boldsymbol{\mu})} \\ &\propto \exp \left\{ -\frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) \right. \\ &\quad \left. - \frac{1}{2} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_b)^T \mathbf{B}^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_b) \right\}. \end{aligned} \quad (6)$$

To obtain the maximum a posteriori estimate, one should maximize $\ln p(\boldsymbol{\sigma}|\boldsymbol{\mu})$, which is equivalent to minimizing cost function Eq. (3). In this case, the result, known as the Best

Linear Unbiased Estimator, is

$$\boldsymbol{\sigma}_a = \boldsymbol{\sigma}_b + \mathbf{B}\mathbf{H}^T (\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_b). \quad (7)$$

The analysis of the posterior error leads to the analysis error covariance matrix:

$$\mathbf{P}_a = \mathbf{B} - \mathbf{B}\mathbf{H}^T (\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1} \mathbf{H}\mathbf{B}. \quad (8)$$

2.3. Semi-Gaussian Modeling of Background Errors

Enforcing the positivity of the source in inverse modeling is known to provide valuable information to the data assimilation system (see Bocquet et al. [2010] for a review). To do so, we can assume that the background errors are semi-Gaussian. The pdf of this normalized truncated normal distribution reads

$$\begin{cases} \text{if } \boldsymbol{\sigma} \geq \mathbf{0} & p(\boldsymbol{\sigma}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_b)^T \mathbf{B}^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_b)}}{\sqrt{(\pi/2)^N |\mathbf{B}|}} \\ \text{otherwise} & p(\boldsymbol{\sigma}) = 0. \end{cases} \quad (9)$$

Using Bayes' rule Eq. (5) we observe that the inference is unchanged: one still has to minimize cost function Eq. (3). However the positivity of $\boldsymbol{\sigma}$ should be enforced during the numerical minimization, and there is no analytical solution to this problem.

As a consequence, Eq. (8) is only an approximation to the posterior errors statistics, which are non-Gaussian. In this non-Gaussian statistical context, a proper way to sample the posterior error pdf is to use a Monte-Carlo second-order sensitivity analysis. To compute sample k , one perturbs the observations ($\boldsymbol{\mu}$) and the first guess ($\boldsymbol{\sigma}_b = \mathbf{0}$), with noises that respectively follow the observational and background error prior statistics, which leads to an analysis $\boldsymbol{\sigma}_a^{(k)}$ for the sample. Statistics can then be inferred from the empirical statistics of numerous draws. This will be used later.

2.4. Estimation of the Hyper-parameters

In the Gaussian case and with the particular forms given to \mathbf{R} and \mathbf{B} , it should be noted that the analysis increment $\boldsymbol{\sigma}_a - \boldsymbol{\sigma}_b$ only depends on the ratio r/m , while \mathbf{P}_a depends on both. A prior estimation of r is possible but certainly very difficult since, in particular, it requires a fine knowledge of model errors. A prior estimation of m is almost impossible, since it would require a large database of similar accidents, which fortunately does not exist. In the semi-Gaussian case, the analysis state and the analysis error covariance matrix depend on both r and m .

Hence, to estimate r/m , or both r and m , one should rely on a posteriori techniques. Several techniques are available and are well described in recent books [Vogel, 2002; Hansen, 2010]. For instance, the general cross-validation method (Wahba [1990] and references therein) is based on the idea that, with a proper estimation of the hyper-parameters using a subset of the observations, the rest of the observations should be consistently accounted for. Another available method is the L-curve method, which is a semi-empirical technique that determines the right balance between overfitting to the data, and over-smoothing by the regularization [Hansen, 1992]. The turning point is indicated by the corner formed by the plot of $\ln \|\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_a\|$ against $\ln \|\boldsymbol{\sigma}_a - \boldsymbol{\sigma}_b\|$. It has been used in the context of accidental atmospheric release inverse modeling by Davoine and Bocquet [2007]; Krysta et al. [2008]; Saide et al. [2011]; Bocquet [2011]. In the ETEX-I and Chernobyl cases, it was shown that only the use of such techniques was able to deliver an estimate

of the source terms consistent with the officially reported emissions.

Another approach is based on the maximum likelihood. As mentioned in the introduction, in the field of data assimilation an increasing number of methodologies for the estimation of parameters of the prior error statistics are explicitly or implicitly based on this paradigm. One seminal article on the topic in data assimilation is Dee [1995]. In our case, $p(\boldsymbol{\mu})$, the likelihood of the observation set, obtained as a normalization factor

$$p(\boldsymbol{\mu}) = \int d\boldsymbol{\sigma} p(\boldsymbol{\mu}|\boldsymbol{\sigma})p(\boldsymbol{\sigma}), \quad (10)$$

is actually an hidden function of r and m . We should write instead $p(\boldsymbol{\mu}|r, m)$. If they exist, the values of r and m that maximize $p(\boldsymbol{\mu}|r, m)$ are considered the most likely values of the parameters consistent with $\boldsymbol{\mu}$. One can seek for a numerical scheme that localize these values, for example the Desroziers' scheme, or we can screen the likelihood values for a large range of (r, m) parameters.

In the light of this discussion and early experiments, we have decided to use three hyper-parameter estimation techniques. Each one of the three approaches is tested in the Gaussian case and semi-Gaussian case.

2.4.1. First Method: L-curve and χ^2

The first approach makes use of the L-curve coupled to a χ^2 diagnosis. The reason to use the L-curve technique is that it is also valid in the non-Gaussian context since it does not assume Gaussian statistics for the errors. However this determines only one degree of freedom of the error statistics. In order to obtain the second one, we would like to tune the general level of errors in the system. If we note $J_o(\boldsymbol{\sigma}_a)$ the degree of freedom for the noise defined by

$$J_o(\boldsymbol{\sigma}_a) = \frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_a)^T \mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_a) \quad (11)$$

and $J_b(\boldsymbol{\sigma}_a)$ the degree of freedom for the signal defined by

$$J_b(\boldsymbol{\sigma}_a) = \frac{1}{2}(\boldsymbol{\sigma}_a - \boldsymbol{\sigma}_b)^T \mathbf{B}^{-1}(\boldsymbol{\sigma}_a - \boldsymbol{\sigma}_b) \quad (12)$$

then the quantity $2(J_o(\boldsymbol{\sigma}_a) + J_b(\boldsymbol{\sigma}_a))$ should have the statistics of a χ^2 if the prior errors are Gaussian, and as shown by Ménard et al. [2000], it should equal the number of observations when the prior statistics matches the genuine ones:

$$J_o(\boldsymbol{\sigma}_a) + J_b(\boldsymbol{\sigma}_a) = \frac{d}{2}. \quad (13)$$

where d represents the number of observations. Note that in the semi-Gaussian case, the use of the χ^2 diagnosis is necessarily approximate. The algorithm is as follows: Take a range of values for r discretized into a set of values r_j . For any r_j perform the L-curve analysis to get the optimal $m_j = m(r_j)$. Then, pick the right value of r_j , which satisfies the χ^2 diagnosis for r_j and m_j .

2.4.2. Second Method: Desroziers' Scheme

Under Gaussian assumptions, integral Eq. (10) can be computed analytically:

$$p(\boldsymbol{\mu}|r, m) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_b)^T (\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_b)}}{\sqrt{(2\pi)^d |\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T|}}. \quad (14)$$

Our approach consists in trying to numerically maximize the likelihood. The log-likelihood can be written:

$$\begin{aligned} \ln p(\boldsymbol{\mu}|r, m) &= -\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_b)^T (\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_b) \\ &\quad - \frac{1}{2} \ln |\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T| + C \\ &= -\frac{1}{2} \text{tr} (\mathbf{S}_{r,m}^{-1} \bar{\mathbf{S}}) - \frac{1}{2} \ln |\mathbf{S}_{r,m}| + C, \end{aligned} \quad (15)$$

where $\mathbf{S}_{r,m} = \mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T$ is the prior statistics of the innovation, $\bar{\mathbf{S}} = (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_b)(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_b)^T$ is the empirical statistics of the innovation, and C is an irrelevant constant. The optimization of the log-likelihood with respect to a parameter denoted λ , leads to the zero gradient condition

$$0 = \nabla_\lambda \ln p(\boldsymbol{\mu}|r, m) = \frac{1}{2} \text{tr} [(\mathbf{S}_{r,m}^{-1} \bar{\mathbf{S}} - \mathbf{I}_d) \mathbf{S}_{r,m}^{-1} \nabla_\lambda \mathbf{S}_{r,m}]. \quad (16)$$

Choosing $\lambda = r$ and $\lambda = m$, one obtains two scalar equations for the likely hyper-parameters. These equations can be put in the form

$$r^2 = \frac{2J'_o(\boldsymbol{\sigma}_a)}{\text{tr}(\mathbf{I}_d - \mathbf{H}\mathbf{K})}, \quad m^2 = \frac{2J'_b(\boldsymbol{\sigma}_a)}{\text{tr}(\mathbf{K}\mathbf{H})}, \quad (17)$$

where

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T (\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1}, \quad (18)$$

$$J'_o(\boldsymbol{\sigma}) = \frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}), \quad (19)$$

$$J'_b(\boldsymbol{\sigma}) = \frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_b)^T (\boldsymbol{\sigma} - \boldsymbol{\sigma}_b). \quad (20)$$

Desroziers and Ivanov [2001] have shown that these two equations can be used as an iterative system which converges to a fixed point which could be a maximum of the likelihood [Chapnik et al., 2004]. They have shown that they could successfully be used in an operational meteorological model [Chapnik et al., 2006].

Let us first remark that this scheme is consistent for Gaussian statistics, not necessarily for non-Gaussian ones. Secondly this scheme, that we shall call Desroziers' scheme later on, converges to one local maximum, whereas there could be several maxima for the likelihood. The context in which we shall use this system is different from Chapnik et al. [2004], because their operator \mathbf{H} is the observation operator, while ours represents the ATM model as well. As a consequence, $\mathbf{H}\mathbf{B}\mathbf{H}^T$ partly populates off-diagonal terms of the innovation statistics $\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T$ so that the sensitivity of the likelihood with respect to r or m should be sufficient to discriminate the two hyper-parameters.

2.4.3. Third Method: Unapproximated Maximum Likelihood Values Screening

The third method consists in estimating the exact value of the likelihood for a large range of (r, m) parameters.

In the Gaussian Case, we use the exact formula Eq. (14) to estimate the likelihood. With two hyper-parameters and computer power, it is possible to perform a value screening of the likelihood, if the number of observations is limited.

In the Semi-Gaussian case, the Gaussian likelihood Eq. (14) is not valid anymore. We will numerically compute integral Eq. (10) in the semi-Gaussian case:

$$\begin{aligned} p(\boldsymbol{\mu}|r, m) &= \frac{1}{\sqrt{(2\pi)^d |\mathbf{R}| (\pi/2)^N |\mathbf{B}|}} \\ &\quad \times \int_{\boldsymbol{\sigma} \geq 0} d\boldsymbol{\sigma} e^{-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) - \frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_b)^T \mathbf{B}^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_b)} \\ &= \frac{e^{-\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_b)^T (\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_b)}}{\sqrt{(2\pi)^d |\mathbf{R}| (\pi/2)^N |\mathbf{B}|}} \\ &\quad \times \int_{\boldsymbol{\sigma} \geq 0} d\boldsymbol{\sigma} e^{-\frac{1}{2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_a)^T \mathbf{P}_a^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_a)}, \end{aligned} \quad (21)$$

where σ_a has been defined by Eq. (7) and \mathbf{P}_a by Eq. (8). A Monte-Carlo estimation of the integral term will be used. To do so, Eq. (21) can be written as:

$$p(\boldsymbol{\mu}|r, m) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}_b)^T(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}_b)}}{\sqrt{(2\pi)^d|\mathbf{H}\mathbf{B}\mathbf{H}^T+\mathbf{R}|}} \times 2^N \int_{\boldsymbol{\sigma} \geq 0} \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_a)^T\mathbf{P}_a^{-1}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_a)}}{\sqrt{(2\pi)^N|\mathbf{P}_a|}} d\boldsymbol{\sigma}, \quad (22)$$

which can be seen as the product of the regular ‘‘Gaussian-case’’ term by a correction taking into account the positivity of the source term. To estimate the value of the integral, we used the GHK simulator method from Hajivassiliou et al. [1996]. Our implementation is detailed in Appendix A.

In total, six cases will be tested depending on whether the background statistics is assumed Gaussian or semi-Gaussian, and depending on the choice of the approaches (L-curve and χ^2 , Desroziers’ scheme, and exact value screening of the likelihood).

3. Application to the Fukushima Daiichi Accident

3.1. Observations

If all the radionuclides measurements of the International Monitoring System (IMS) of the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) have still not been released publicly, some activity concentrations measurements are available. In the United States, they come from the Radnet monitoring network which is supported by the Environmental Protection Agency (EPA) and from the CTBTO air monitoring stations (USP70 in California, USP77, USP78, USP79 and USP80 on islands in the Pacific Ocean), whose data have been publicly released by the United States National Data Center (USNDC). In Canada, the Health ministry has also released some monitoring measurement data. The Philippine Nuclear Research Institute (PNRI) has displayed the measurements from the CTBTO air monitoring station PHP52 on its website. In Japan, the High Energy Accelerator Research Organization (KEK) and the National Institute for Environmental Studies (NIES), located in Tsukuba, and the Metropolitan Industrial Technology Research Institute (MITRI), in Tokyo, have performed some measurements. In Fukushima prefecture, the Ministry of Education, Culture, Sport, Science and Technology (MEXT) has performed a large number of measurements. Because the Eulerian atmospheric transport model in this study is used at a spatial resolution of 0.25° , we have excluded measurements within a radius of 30 km around the power plant. Finally, the Center for the Promotion of Disarmament and Non-proliferation (CPDNP) has released the

measurements from the JPP38 CTBTO air monitoring station. Altogether, 267 measurements of cesium-137 activity concentration (104 in Japan and 163 outside) and 428 measurements of iodine-131 activity concentration (233 in Japan and 195 outside) are assimilated in this study (see Tab. 1 and Fig. 1). The iodine measurements should be understood as a lower bound, since only the aerosol form may be collected in particle filters.

3.2. Atmospheric Transport Model

The simulations of the dispersion of radionuclides from Daiichi nuclear power plant have been performed with the chemistry-transport model POLAIR3D, the Eulerian model of the POLYPHEMUS platform. As far as radionuclides are concerned, it has been validated on the European Tracer Experiment, on the Algeiras incident and on the Chernobyl accident (Qu  lo et al. [2007]).

The model integrates the concentration field c of either ^{137}Cs or ^{131}I , following the transport equation

$$\frac{\partial c}{\partial t} + \text{div}(\mathbf{u}c) = \text{div}\left(\rho\mathbf{K}\nabla\left(\frac{c}{\rho}\right)\right) - \Lambda^s c - \Lambda^d c + \sigma \quad (23)$$

where Λ^s is the wet scavenging rate, Λ^d represents the radioactive decay and σ is the point-wise source for the radionuclide. \mathbf{K} is the matrix of turbulent diffusion, diagonal in practice. The vertical component is given by K_z , computed with Louis parameterization [Louis, 1979]. The horizontal component K_H is taken constant (equal to $25000 \text{ m}^2 \text{ s}^{-1}$ for a spatial resolution of 0.25°). Except for the boundary condition on the ground, all boundary conditions are taken null. The boundary condition on the ground is

$$-K_z \nabla c \cdot \mathbf{n} = -v^{\text{dep}} c \quad (24)$$

where \mathbf{n} is the unitary vector upward oriented, and v^{dep} is the dry deposition velocity of ^{137}Cs or ^{131}I .

The advection is implemented thanks to a third-order direct space-time scheme, with a Koren-Sweby flux limiter function. Because of the sharp gradients found, it is important that such a limiter be used. The diffusion scheme is integrated through an implicit second-order Rosenbrock scheme, with a three-point spatial scheme, and directional splitting.

Cesium-137 and iodine-131 are modeled as passive gaseous tracers with radioactive decay. Their half-lives are respectively 11000 and 8.04 days. Dry deposition is modeled by using a simple scheme with a constant deposition velocity: $v^{\text{dep}} = 0.2 \text{ cm s}^{-1}$ for ^{137}Cs and $v^{\text{dep}} = 0.5 \text{ cm s}^{-1}$ for ^{131}I . As far as the wet scavenging is concerned, the parameterization used in this study is of the form :

$$\begin{cases} \Lambda^s = 0 & \text{if } RH < RH_t \\ \Lambda^s = 3.5 \times 10^{-5} \left(\frac{RH - RH_t}{RH_s - RH_t} \right) & \text{otherwise,} \end{cases} \quad (25)$$

where RH stands for the relative humidity, RH_t being a threshold value ($= 80\%$) and RH_s is the saturation value ($= 100\%$) [Pudykiewicz, 1989; Brandt et al., 2002].

The domain of simulation is an extended domain covering Japan, the north-Pacific ocean and an important part of North-American continent, with a spatial resolution of $0.25^\circ \times 0.25^\circ$. The number of grid points in the domain simulation is 652×256 . The POLAIR3D model is configured with 15 vertical levels ranging from 0 to 8000 m.

This study attempts to reconstruct the source term from March 11 to March 26 (which represents 384 one-hour timesteps). However the simulations run over a longer period (from March 11 to April 5) in order to exploit the information content of later observations. A total of 384 direct simulations are performed to fill the Jacobian matrix

Table 1. Source of the measurements of activity concentration for cesium-137 and iodine-131 used in the study.

Source	Cs137	I131
KEK-NIES - Tsukuba	10	10
MITRI - Tokyo	58	160
JPP38 CTBTO station	20	9
MEXT - Fukushima Prefecture	16	54
Total Japan	104	233
USNDC - CTBTO - USA	77	82
RadNet - EPA - USA	44	69
Health Ministry Of Canada	29	31
PNRI - PHP52 CTBTO station	13	13
Total Outside	163	195
Total	267	428

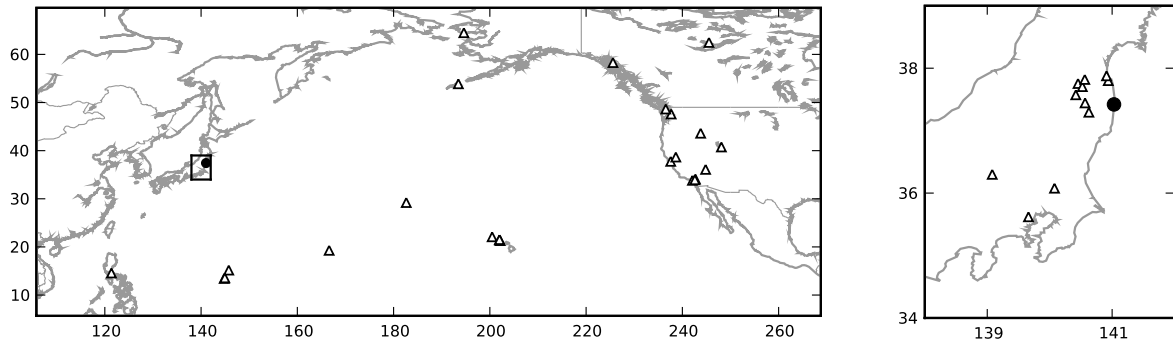


Figure 1. Map of the simulation domain used in POLAIRD3D. The triangles stand for the observations locations. The circle stands for the Fukushima Daiichi power plant. The right panel zooms in over the Japan and the observations sites available over the country.

H column by column, and no adjoint model is needed in this process (see Winiarek et al. [2011]). The meteorological fields used in this study are the fields computed by the operational model from the European Centre for Medium-range Weather Forecasts (ECMWF). They have a resolution of $0.25^\circ \times 0.25^\circ$ and are available every 3 hours.

3.3. Gaussian Case and Cesium-137

Observations collected in Japan and those collected elsewhere are of different nature. First, the distances between the Daiichi power plant and the monitoring stations are so different that they induce a large gap in the order of magnitude of the observation values. All observations in Japan are in the range $[10^{-1}, 10^2] \text{ Bq m}^{-3}$ whereas most observations outside Japan are in the range $[10^{-6}, 10^{-3}] \text{ Bq m}^{-3}$. Impacts of the model errors (meteorological fields and loss process parameterizations) accumulate along this long-distance transport path. Additionally, most of observations outside Japan are mean measurements over a long period (typically 24 hours), whereas in Japan measurements are typically representative of shorter periods (one hour or less). For these reasons, we have chosen to perform first the inversions using only the Japanese observations, and then using all the available observations to study the impact of those differences.

3.3.1. Estimation of Parameters and Total Released Activity

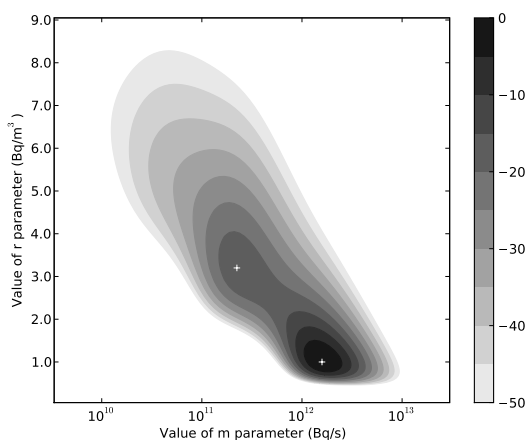


Figure 2. Density plot of the log-likelihood under Gaussian assumptions. The global maximum log-likelihood value has been subtracted to all log-likelihood values, so that the maximum value is set to 0.

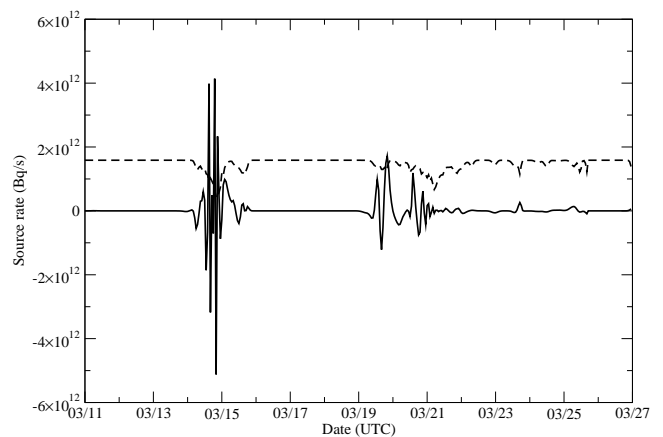


Figure 3. Reconstructed temporal profile of ^{137}Cs under Gaussian assumptions with the maximum likelihood method (equivalent to Desroziers's scheme), using the observations of Japan. The dashed line represents the posterior uncertainty on the reconstructed source (diagonal terms of matrix \mathbf{P}_a).

Under Gaussian assumptions for both observation and background errors statistics, the three methods are comparable. Estimates for r , m and the total released activity are reported in Tab. 2.

In particular we have verified that under Gaussian assumptions, Desroziers's scheme exactly converges to the values (r, m) given by the maximum likelihood method. The value screening of the likelihood (see Fig. 2) allows to plot the marginal posterior distribution function for both parameters r and m (see Fig. 5 and Fig. 8) on which two local maxima appear. Given our initial couple of parameters, Desroziers's scheme converges towards the absolute maximum location. (With a different initial condition in another basin of attraction, it could have converged to the other.)

The reconstructed total mass of ^{137}Cs released in the atmosphere lies between $2.8 \times 10^{16} \text{ Bq}$ and $2.1 \times 10^{17} \text{ Bq}$, which is consistent, though slightly higher, with the first estimates released by the Nuclear Safety Commission of Japan ($1.2 \times 10^{16} \text{ Bq}$ at first, then $1.3 \times 10^{16} \text{ Bq}$ in Chino et al. [2011]).

3.3.2. Temporal Profile and Uncertainty Reduction

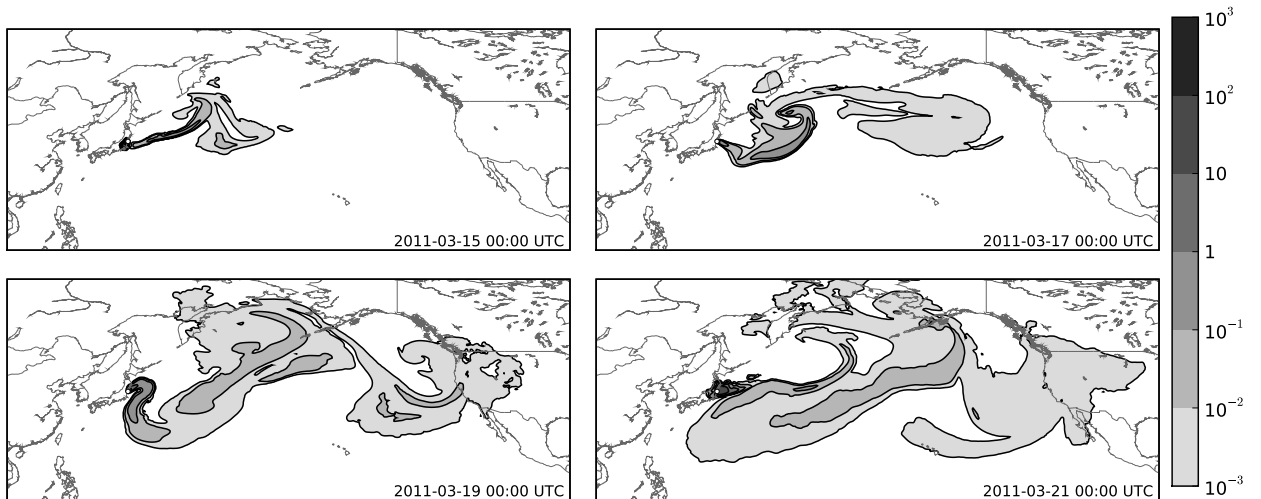
Most of the time, the meteorological conditions have transported the radionuclides toward the Pacific Ocean, sparing populations, but reducing the observability of the

Table 2. Estimation of parameters and corresponding reconstructed released activity with the different methods in Gaussian case for cesium-137 source reconstruction

parameter	method	with observations in Japan (104)	with all observations (267)
r (Bq m^{-3})	$\chi^2 + \text{L-curve}$	1.05	1.03
	Desroziers's scheme	}1.0	0.56
	Maximum likelihood		
m (Bq s^{-1})	$\chi^2 + \text{L-curve}$	1.5×10^{12}	1.5×10^{12}
	Desroziers's scheme	} 1.5×10^{12}	2.2×10^{12}
	Maximum likelihood		
Released activity (Bq)	$\chi^2 + \text{L-curve}$	2.8×10^{16}	6.9×10^{16}
	Desroziers's scheme	} 2.8×10^{16}	2.1×10^{17}
	Maximum likelihood		

Table 3. Estimation of parameters and corresponding reconstructed released activity with the different methods in semi-Gaussian case for cesium-137 source reconstruction

parameter	method	with observations in Japan (104)	with all observations (267)
r (Bq m^{-3})	$\chi^2 + \text{L-curve}$	4.55	2.88
	Desroziers's scheme	5.41	2.96
	Maximum likelihood	3.25	1.7
m (Bq s^{-1})	$\chi^2 + \text{L-curve}$	3.2×10^{11}	2.0×10^{11}
	Desroziers's scheme	5.3×10^{10}	1.3×10^{11}
	Maximum likelihood	2.0×10^{11}	3.5×10^{11}
Released activity (Bq)	$\chi^2 + \text{L-curve}$	1.2×10^{16}	1.3×10^{16}
	Desroziers's scheme	3.3×10^{15}	1.0×10^{16}
	Maximum likelihood	1.2×10^{16}	1.9×10^{16}

**Figure 4.** Map (in Bq m^{-3}) of simulated cesium-137 plume dispersion at four different dates after the accident start. The source term is the one reconstructed with all available data and the maximum likelihood method.

plume. As a consequence, the inversion methods are only able to reconstruct some time windows when attempting to retrieve the temporal profile of the source term. The profile of the reconstructed source using the set of observations over Japan is displayed in Fig. 3. In particular three periods are well-observed: the first one lies approximatively from March 14 to March 16, the second one from March 19 to March 21 and the last one from March 24 to March 26, when winds were blowing in the south-west direction, where the monitoring stations of Tokyo, Tsukuba and JPP38 are located.

Because of the Gaussian assumptions, the source parameters are allowed to take negative values. From Fig. 3, it is clear that the number of observations is not sufficient to constrain the solution to be positive for most parameters. A larger number of observations would have probably helped in that respect as shown in Winiarek et al. [2011].

Consequently, in the absence of a large number of observations, non-Gaussian statistical assumptions that constrain the source to be positive are paramount.

In Fig. 3, the uncertainty (standard deviation) of the retrieved source parameter is also plotted. It has been computed using the diagonal parameters of \mathbf{P}_a , which is an approximation since it neglects the correlations between parameters.

3.4. Semi-Gaussian Case and Cesium-137

3.4.1. Estimation of Parameters and Total Released Activity

Under semi-Gaussian assumptions for the background errors statistics, the three methods are expected to yield more distinct results. These results are displayed in Tab. 3. The χ^2 diagnosis, as well as the Desroziers' scheme, are based on Gaussian assumptions and this is why the first two methods

Table 4. Estimation of parameters and corresponding reconstructed released activity with the different methods in semi-Gaussian case for iodine-131 source reconstruction

parameter	method	with observations in Japan (233)	with all observations (428)
r (Bq m ⁻³)	χ^2 + L-curve	14.0	10.5
	Desroziers's scheme	18.4	10.6
	Maximum likelihood	5.60	4.01
m (Bq s ⁻¹)	χ^2 + L-curve	2.7×10^{12}	2.0×10^{12}
	Desroziers's scheme	2.0×10^{11}	1.9×10^{12}
	Maximum likelihood	5.6×10^{12}	7.1×10^{12}
Released activity (Bq)	χ^2 + L-curve	1.9×10^{17}	2.0×10^{17}
	Desroziers's scheme	1.6×10^{16}	1.9×10^{17}
	Maximum likelihood	3.8×10^{17}	7.0×10^{17}

are to be considered approximations in this case. Nevertheless, results obtained with these two methods are comparable to those given by the value screening of the likelihood (which is unapproximated), especially when more observations are assimilated. The posterior marginal pdf of r and m parameters are also displayed in Fig. (5) and Fig. (8). In the situation where all available data are used, the total released activity of ¹³⁷Cs released is estimated to be between 1.0×10^{16} Bq and 1.9×10^{16} Bq. Focussing on the most reliable methods (χ^2 +L-curve and maximum likelihood), and only assimilating observations over Japan to avoid large model errors, we obtain an estimation of 1.2×10^{16} Bq.

3.4.2. Temporal Profile and Uncertainty Reduction

As indicated in Section 2.3, we have performed a second-order Monte-Carlo analysis to estimate the posterior analysis error instead of using Eq. (8), which is only consistent (though approximate) with Gaussian statistics. In Fig. 9 the reconstructed source using the set of observations over Japan, is displayed as well as its standard deviation computed from this Monte-Carlo analysis. The same three periods of observability as in the Gaussian case are visible, but may be slightly enlarged. Even if the available data (activity concentration data as used in this study, but also deposition data and gamma dose rate within the NPP site) clearly indicate releases between March 19 and March 21, it may be surprising to retrieve such a high peak on March 20, as an emergency cooling process has been engaged using Tokyo Fire Department equipment since approximately March 18, as debated in the discussion manuscript of Stohl et al. [2011] and its subsequent comments. However, it can be

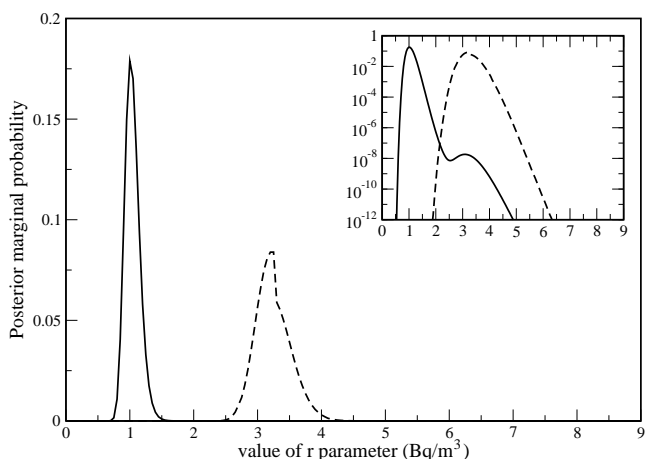


Figure 5. Marginal pdf of parameter r obtained with the maximum likelihood method under Gaussian assumptions (full line) or semi-Gaussian assumptions (dashed line) with the Japanese data of Cesium-137 activity concentrations. The inset displays the same pdfs in logarithmic scale.

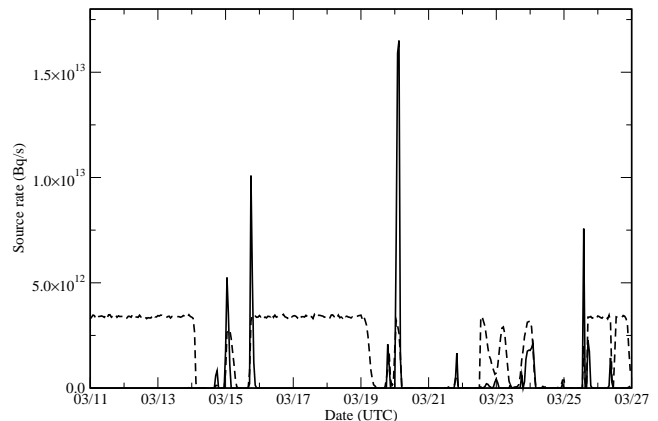


Figure 6. Reconstructed temporal profile of ¹³¹I under semi-Gaussian assumptions with maximum likelihood method, using observations over Japan. The dashed line represents the standard deviation given by the Monte-Carlo analysis.

seen on Fig. 9 that a high uncertainty, estimated by our Monte-Carlo analysis, is attached to the magnitude of this peak. Moreover, from the end of this peak to March 22, the retrieved emissions are weak and this is supported by a low uncertainty.

The plume concentration field that is simulated using the source term obtained from all available observations and the maximum likelihood method is displayed in Fig. 4. The activity concentration fields of cesium-137 are displayed at four different dates from March 15 to March 21. In Fig. 10 the comparison between simulated concentrations and observations is plotted at the MITRI station in Tokyo. The agreement between the simulation and the observations is good, especially at peak times, in spite of a time offset for the last peak (between 2 and 4 hours).

3.5. Application to Iodine-131

In the case of iodine-131, we only consider the non-Gaussian case since it was shown to have a significant advantage over the Gaussian methodology for cesium-137.

The results are reported in Tab. 4. With all the available observations, the total reconstructed released activity lies between 1.9×10^{17} Bq and 7.0×10^{17} Bq. This estimate is of the same order of magnitude as the NSC first official estimates, although higher (1.5×10^{17} Bq in Chino et al. [2011]).

Focussing on the most reliable methods (χ^2 +L-curve and maximum likelihood), and only assimilating observations over Japan to avoid large model errors, we obtain estimates of 1.9×10^{17} Bq and 3.8×10^{17} Bq respectively.

Similarly to the non-Gaussian case for cesium-137, the reconstruction profile when assimilating observations over

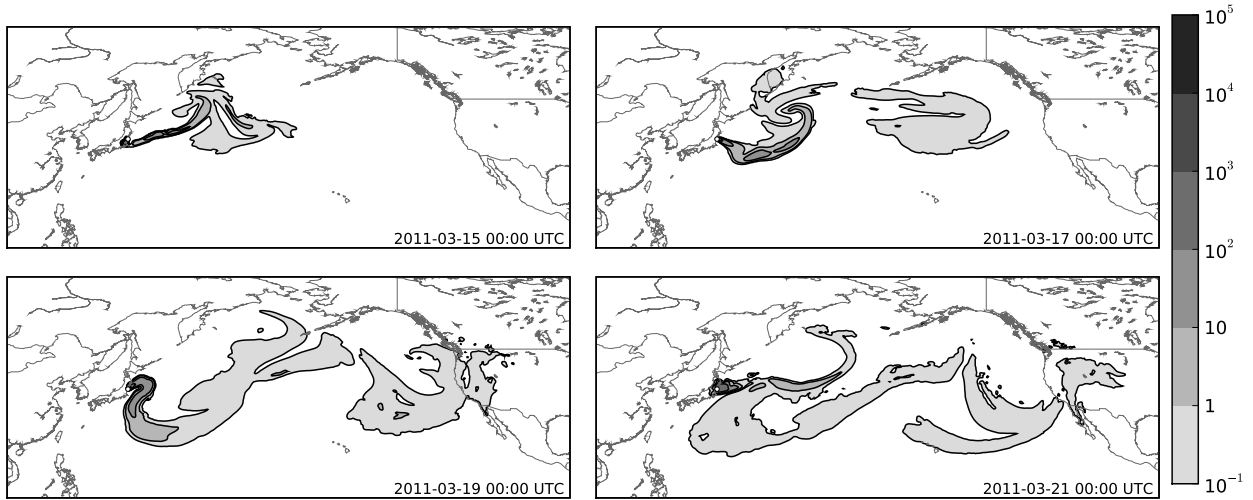


Figure 7. Map (in Bq m^{-3}) of simulated iodine-131 plume dispersion at four different dates after the accident start: The source term is the one reconstructed with all available data and the maximum likelihood method.

Japan is reported in Fig. 6. An estimation of the uncertainty of the retrieval obtained from Monte Carlo analysis is also plotted. Compared to cesium-137, there is a better observability from March 19 until March 26 without interruption. The reduction of uncertainty during this period is explained by a larger amount of available data, especially in Tokyo, MITRI station.

As for cesium-137, a cartoon of the dispersion plume between March 15 and March 21 is displayed in Fig. 7. The simulation is based on the iodine-131 source term estimated using all observations. In Fig. 11 the comparison between simulated concentrations and observations is plotted at the MITRI station in Tokyo. The agreement between the simulation and the observations is good, especially at peak times.

3.6. Uncertainty and Sensitivity Analysis

In this section, several key issues pertaining to the uncertainty of the prior assumptions, and of the retrievals are discussed.

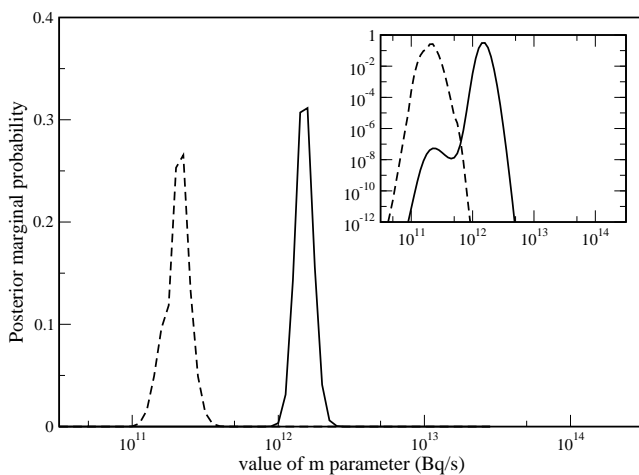


Figure 8. Marginal pdf of parameter m obtained with the maximum likelihood method under Gaussian assumptions (full line) or semi-Gaussian assumptions (dashed line) with the Japanese data of Cesium-137 activity concentrations. The inset displays the same pdfs in logarithmic scale.

3.6.1. Quantitative Results and Uncertainties

The quantitative results are summarized in Tab. 5: the range of estimates using all observations and all methods, and the robust range of estimates, using Japanese observations and (more robust) non-Gaussian methods.

In addition, the estimated standard deviations for the total retrieved activities are computed and reported in the table. It provides an estimation of the uncertainty attached to the robust result. Two standard deviations are computed. The first one results from the perturbation of the observations only. Therefore it is a measure of the uncertainty attached to the retrieved mass. A second one results from the perturbation of both the observations and the background. It is a measure of the expected total uncertainty that includes the releases that could not be observed. Obviously it is very speculative since it relies on the estimation of the background parameter m for the whole period but which has been determined with observations that only shed light on a few days in the period.

In Fig. 12, the temporal profiles of the ^{137}Cs source term retrieved using only the observations in Japan on one hand,

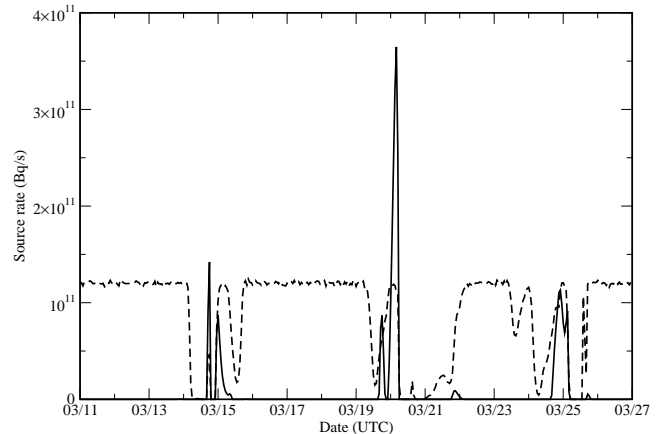


Figure 9. Reconstructed temporal profile of ^{137}Cs under semi-Gaussian assumptions with the maximum likelihood method, using the observations over Japan. The dashed line represents the standard deviation yielded by the Monte-Carlo analysis.

Table 5. Summary of the quantitative results and attached uncertainty.

Species	Released activity (Bq) all observations	Released activity (Bq) observations over Japan (robust results)	std. deviation with pert. observations	std. deviation with pert. obs. and background
cesium-137	$1.0 \times 10^{16} - 1.9 \times 10^{16}$	1.2×10^{16}	15% - 20%	60% - 100%
iodine-131	$1.9 \times 10^{17} - 7.0 \times 10^{17}$	$1.9 \times 10^{17} - 3.8 \times 10^{17}$	5% - 10%	40% - 45%

and all available observations on the other hand, are plotted. The use of all observations does not introduce new peaks in the profile, and the order of magnitude of the total released activity is unchanged. Nevertheless, significant emissions rates are retrieved during several periods (from March 11 to March 14, from March 16 to March 19 and on March 26), although at a much lower magnitude compared to the peaks

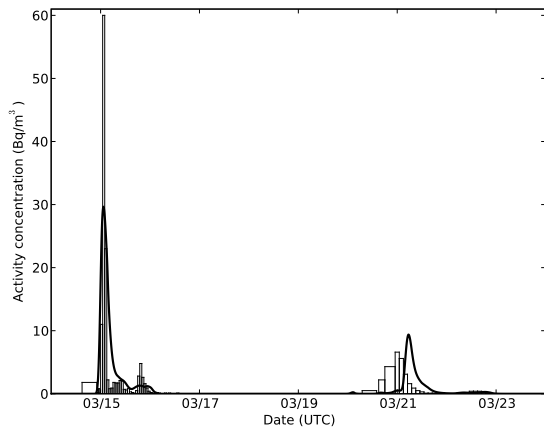


Figure 10. Comparison between simulated concentrations of cesium-137 and measurements data in Tokyo station (MITRI). The full line stands for the simulated concentrations using the source term provided by the maximum likelihood method. The histogram represents the observations. The width of the bars represents the duration of the measurements.

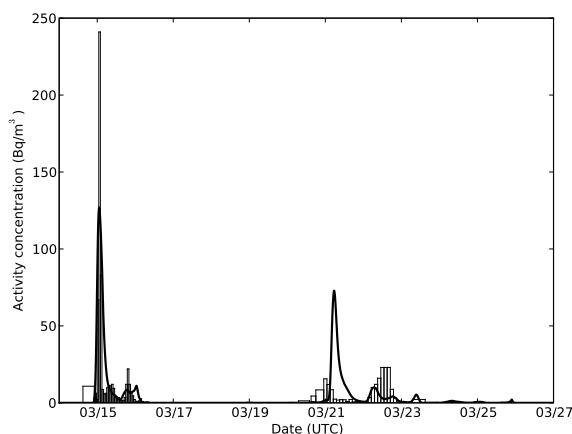


Figure 11. Comparison between simulated concentrations of iodine-131 and measurements data in Tokyo station (MITRI). The full line stands for the simulated concentrations using the source term provided by the maximum likelihood method. The histogram represents the observations. The width of the bars represents the duration of the measurements.

of March 15 and March 20. The smoother shape of the distant observations contribution is due to the diffusive nature of atmospheric transport over such long distances, in addition to their longer duration. The total retrieved activity of this distant observations contribution only represents about 1% of the total retrieved activity.

3.6.2. Background Uncertainty

As argued in the introduction, following [Bocquet, 2005; Davoine and Bocquet, 2007] and differently from standard practice in greenhouse gas flux inversions and atmospheric chemistry emission inversion where an inventory is often available, we have chosen a first guess (background term)

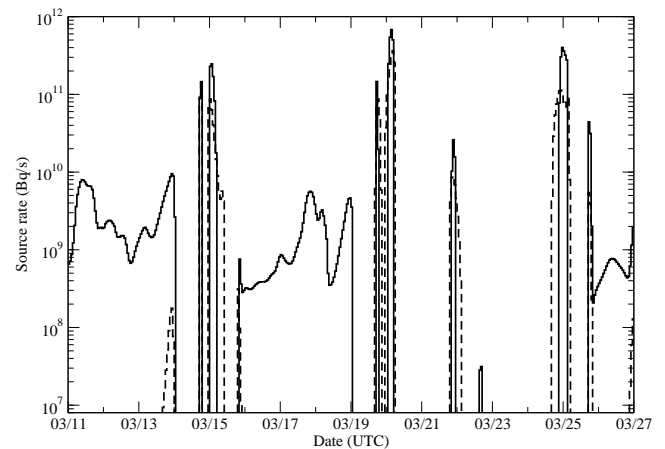


Figure 12. Reconstructed temporal profiles of ^{137}Cs under semi-Gaussian assumptions with the maximum likelihood method, using observations over Japan (dashed line) and all observations (full line).

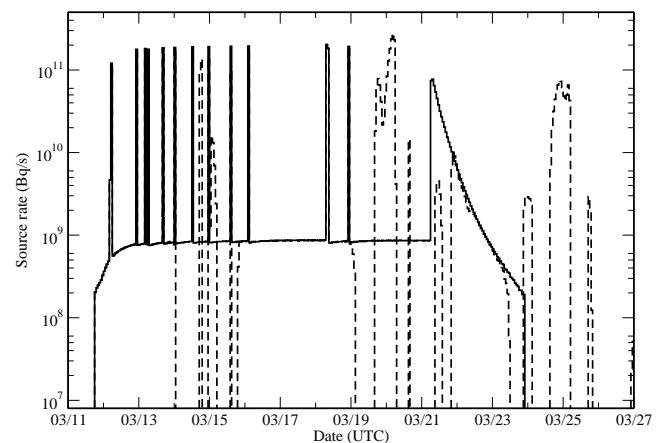


Figure 13. Reconstructed temporal profile of ^{137}Cs under semi-Gaussian assumptions with the maximum likelihood method, with the use of a non-null first guess. The full line represents the IRSN first guess profile. The dashed line represents the retrieved source term. Note that the retrieved profile coincides with the first guess profile for the first unobserved sequence of peaks, March 12 to March 14.

equal to zero. In our opinion, it particularly suits the context of the Fukushima Daiichi accident. Indeed the emission has been intermittent (strong emission peaks occur at the time of hydrogen explosions or during the ventings) and most of the time almost nothing was emitted. Additionally, it is not straightforward to obtain an accurate first guess from direct nuclear physics model, because even when an inventory of what has been generated in the core is hypothesized, it is difficult to know what has actually been released into the atmosphere and when. The related uncertainty of such an inventory would be barely known and in consequence difficult to use. That is the main reason why Nuclear Safety agencies (such as the Japanese Safety Agency or the French Institute of Radioprotection and Nuclear Safety (IRSN)) have provided first estimates constrained by the first observations (as in Chino et al. [2011] who used 17 observations to reconstruct a first estimate). But using these estimates as background terms in our inverse modeling process could be seen as an *inversion crime* and therefore should be avoided.

However, to highlight the sensitivity of the retrieval to such an approach, we have performed a reconstruction using a non-trivial cesium-137 background term provided by the IRSN, focusing on the Japanese observation set. The first guess and the retrieved source are plotted in Fig. 13. The total released mass retrieved is greater than the one retrieved without the use of the background term (1.9×10^{16} Bq instead of 1.2×10^{16} Bq). This is consistently explained by the fact that the analysis is strongly influenced by the first guess, specially in the early periods (mainly before March 15 and between March 16 and March 19 where it is impossible to distinguish between the first guess and the analysis profiles) where the observability was poor corresponding to a high uncertainty diagnosed in the retrieval with $\sigma_b = 0$. This retrieved mass of 1.9×10^{16} Bq, which is not anymore meant to provide a lower bound but a full estimation, is remarkably consistent with the mass retrieved using the complete observation set (1.9×10^{16} Bq).

3.6.3. Sensitivity to Physics Parametrization : Example of the Dry Deposition

Model errors are expected to plague comparison between the model and distant observations. In particular a biased removal process should generate a strong discrepancy between the model and distant observations. That is why we have studied the sensitivity of the inverse modeling results to a change in the dry deposition velocity. We chose 0.15 cm s^{-1} for ^{137}Cs (instead of 0.2 cm s^{-1}) and 0.3 cm.s^{-1} for ^{131}I (instead of 0.5 cm s^{-1}), which are the optimal parameter values found by Bocquet [2011] for the Chernobyl accident dispersion. As expected, the total retrieved released activities are lower: the maximum likelihood estimates for ^{137}Cs are 0.9×10^{16} Bq with the observations over Japan, and 1.6×10^{16} Bq with all observations. For ^{131}I , they are 1.6×10^{17} Bq with the observations over Japan, and 1.9×10^{17} Bq with all observations. If there is a gap reduction between retrieved activities for the two sets of observations and ^{131}I , the gap is maintained for ^{137}Cs .

4. Conclusion

In this article, we have proposed efficient inverse modeling methods to reconstruct the Fukushima Daiichi source term of radionuclides available for long-range transport. We have assimilated publicly released air concentration activity measurements, which represent 267 observations for ^{137}Cs and 428 for ^{131}I . Because of this small number of observations (compared to the 384 one-hour time-steps of the source terms) and the particular meteorological conditions leading to a poor observability of the release, the design of the priors is paramount. In particular, it is necessary to a posteriori estimate hyper-parameters that characterise the prior errors.

This analysis focused on two hyper-parameters: r the prior observation error standard deviation and m the background error standard deviation. Three methods have been

proposed to estimate these two parameters : (1) an L-curve method coupled with a χ^2 diagnosis, (2) Desroziers's scheme and finally (3) the maximization of the exact likelihood of the observation set as a function of r and m . We have also assumed that the background error statistics are Gaussian or semi-Gaussian (i.e. forcing the reconstructed source term to be positive). In this latter case and using all the available observations, a lower bound of the total activity released in the atmosphere between March 11 and March 26 is estimated to be between 1.0×10^{16} Bq and 1.9×10^{16} Bq for ^{137}Cs and between 1.9×10^{17} Bq and 7.0×10^{17} Bq for ^{131}I , which is comparable to the first estimates from the Nuclear Safety Commission of Japan (resp. 1.2×10^{16} Bq and 1.5×10^{17} Bq) but represents probably an underestimation of the real emission. Using the most reliable methods in this case and the Japanese observations only, the lower bound estimate is 1.2×10^{16} Bq for ^{137}Cs , and in the range $1.9 - 3.8 \times 10^{17}$ Bq for ^{131}I . They represent 5 to 10 times less emissions than the Chernobyl accident.

The uncertainty of the reconstruction and its reduction have also been analyzed. Under Gaussian assumption, the posterior error covariance matrix has been used (mainly through its diagonal terms), whereas under semi-Gaussian assumptions, we used a Monte-Carlo second-order sensitivity analysis to estimate the posterior error variance. In both cases, this analysis highlights the fact that some (between 2 and 3) time windows are well observed, during which the uncertainty is reduced, whereas some time windows are not well observed, with an uncertainty that remains high. These periods of high uncertainty on the reconstructed source correspond to configurations with a wind transporting the radioactive plume directly over the Pacific Ocean, the plume being eventually observed when arriving in North America, but with high uncertainties introduced by the dispersion model.

In addition, a non-zero first guess inversion has been performed in the case of cesium-137 with the robust maximum likelihood method assimilating the Japanese observations. The total retrieved activity was found to be consistently equivalent to the one of the inversion using all observations (1.9×10^{16} Bq). Indeed one can hope that some of the peaks of the non-trivial first guess unaccounted for by the Japanese set of observations will be explained by the additional observations.

We have identified several ways to improve these reconstructions. First, we shall improve the accuracy of the dispersion model, for example the deposition parameterizations (dry deposition and/or wet scavenging). In such long-distance simulations, the uncertainties introduced by the numerical model may be very important. For instance, the deposition velocity over the ocean is understood to be significantly lower than over land, but its value remains unknown. Physical parameter inverse modeling techniques such as the one developed in Bocquet [2011] could be useful.

It could also be interesting to implement inverse modeling methods that do not assume that observation error statistics are Gaussian (in addition to non-Gaussian background error statistics), such as in Winiarek et al. [2011]. Finally in order to increase the number of available observation data, it can be interesting to develop methods which can take into account different types of observations, such as fallout data and gamma dose rates. The amount of available data would then increase substantially, but the prior uncertainties introduced in the system will also significantly increase (at least about ten different species, uncertainties on the deposition processes, precipitation, rainfall, etc.).

Appendix A: Efficient computation of the semi-Gaussian integral

In the Semi-Gaussian case, the unapproximated value of the likelihood is given by Eq. (22):

$$p(\boldsymbol{\mu}|r, m) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}_b)^T(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}_b)}}{\sqrt{(2\pi)^d|\mathbf{H}\mathbf{B}\mathbf{H}^T+\mathbf{R}|}} \times 2^N \int_{\boldsymbol{\sigma}_{\geq 0}} \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_a)^T\mathbf{P}_a^{-1}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_a)}}{\sqrt{(2\pi)^N|\mathbf{P}_a|}} d\boldsymbol{\sigma}, \quad (\text{A1})$$

which can be seen as the product of the regular ‘‘Gaussian-case’’ term (the fraction term) by a correction taking into account the positivity of the source term (the integral term). Estimating the value of the integral of a truncated multivariate normal distribution is a challenging task, studied in econometrics or numerical analysis. An extensive review of different methods has been performed by Hajivassiliou et al. [1996]. We chose to use the Geweke-Hajivassiliou-Keane (GHK) simulator, which is found by these authors to be the most reliable method. We have found that, in our context, it was more efficient than Gibbs sampling, which was for instance used by Lauvernet et al. [2009]. The principle of the GHK simulator is to draw samples following the truncated multivariate (N variables) Gaussian distribution $\mathcal{TN}(\boldsymbol{\sigma}_a, \mathbf{P}_a, 0, +\infty)$.

To draw a sample $\boldsymbol{\sigma}_*$, the GHK simulator uses the Cholesky decomposition of the matrix $\mathbf{P}_a = \mathbf{L}^T\mathbf{L}$, with \mathbf{L} a lower-triangular matrix. It reduces the problem to N draws of scalar random variables $(u_i)_{1 \leq i \leq N}$ following an univariate truncated normal distribution $\mathcal{TN}(0, 1, \alpha_i, +\infty)$ with the lower bounds α_i being given by

$$\begin{cases} \alpha_1 = -\frac{[\boldsymbol{\sigma}_a]_1}{[\mathbf{L}]_{1,1}} \\ \alpha_i = -\frac{[\boldsymbol{\sigma}_a]_i}{[\mathbf{L}]_{i,i}} - \sum_{j=1}^{i-1} \frac{[\mathbf{L}]_{i,j}}{[\mathbf{L}]_{i,i}} u_j \quad \text{for } i = 2, \dots, N. \end{cases} \quad (\text{A2})$$

This method suggests that the draw of $\mathbf{u} = (u_i)_{1 \leq i \leq N}$ is made recursively. Firstly, α_1 is computed and u_1 is drawn from $\mathcal{TN}(0, 1, \alpha_1, +\infty)$, following the exponential accept-reject algorithm proposed by Robert [1995] if $\alpha_1 > 0.5$, a normal rejection algorithm if $\alpha_1 \leq 0$, and a half-normal rejection algorithm otherwise. Then $\alpha_2(u_1)$ is computed and u_2 is drawn from $\mathcal{TN}(0, 1, \alpha_2, +\infty)$ and so on. Then the desired sample $\boldsymbol{\sigma}_*$ is obtained via the transformation

$$\boldsymbol{\sigma}_* = \boldsymbol{\sigma}_a + \mathbf{L}\mathbf{u}. \quad (\text{A3})$$

In fact, this sampling scheme does not sample directly the target pdf, because of the dependence induced by the recursion. To correct the bias of the sample, one computes its importance sampling weight due to the difference between the proposal and the targeted distribution. The weight $w(\boldsymbol{\sigma}_*)$ is

$$w(\boldsymbol{\sigma}_*) = \frac{1}{2^N} \prod_{i=1}^{i=N} \text{erfc}\left(\frac{\alpha_i}{\sqrt{2}}\right). \quad (\text{A4})$$

Function erfc is the complementary error function defined by $\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{+\infty} e^{-t^2} dt$. The value of the integral is then estimated with n_* samples using the empirical mean corrected by the importance sampling weights: $\frac{1}{n_*} \sum_{k=1}^{k=n_*} w(\boldsymbol{\sigma}_*^{(k)})$. In the GHK simulator, the n_* samples are independent, so that the numerical computation of the integral can be massively parallelized. The value of the likelihood can then be easily screened for several values of (r, m) .

Acknowledgments.

This study has been supported by the IMMANENT project of Paris-Est University, and by the INSU/LEFE-ASSIM project ADOMOCA-2. We thank three reviewers for their helpful comments and suggestions. We gratefully acknowledge the help of Monika Krysta and Karine Sartelet with the data.

References

- Abida, R., Bocquet, M., 2009. Targeting of observations for accidental atmospheric release monitoring. *Atmos. Env.* 43, 6312–6327.
- Anderson, J. L., 2007. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A* 59, 210–224.
- Bocquet, M., 2005. Reconstruction of an atmospheric tracer source using the principle of maximum entropy. I: Theory. *Q. J. Roy. Meteor. Soc.* 131, 2191–2208.
- Bocquet, M., 2011. Parameter field estimation for atmospheric dispersion: Application to the Chernobyl accident using 4D-Var. *Q. J. Roy. Meteor. Soc.* 0, 0–0, in press.
- Bocquet, M., Pires, C. A., Wu, L., 2010. Beyond Gaussian statistical modeling in geophysical data assimilation. *Mon. Wea. Rev.* 138, 2997–3023.
- Brandt, J., Christensen, J. H., Frohn, M., 2002. Modelling transport and deposition of caesium and iodine from chernobyl accident using the dream model. *Atmos. Chem. Phys.* 2, 397–417.
- Chapnik, B., Desroziers, G., Rabier, F., Talagrand, O., 2004. Properties and first application of an error-statistics tuning method in variational assimilation. *Q. J. Roy. Meteor. Soc.* 130, 2253–2275.
- Chapnik, B., Desroziers, G., Rabier, F., Talagrand, O., 2006. Diagnosis and tuning of observational error in a quasi-operational data assimilation setting. *Q. J. Roy. Meteor. Soc.* 132, 543–565.
- Chino, M., Nakayama, H., Nagai, H., Terada, H., Katata, G., Yamazawa, H., 2011. Preliminary estimation of release amounts of I131 and Cs137 accidentally discharged from the Fukushima Daiichi nuclear power plant into the atmosphere. *Journal of Nuclear Science and Technology* 48, 1129–1134.
- Davoine, X., Bocquet, M., 2007. Inverse modelling-based reconstruction of the Chernobyl source term available for long-range transport. *Atmos. Chem. Phys.* 7, 1549–1564.
- Dee, D. P., 1995. On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.* 123, 1128–1145.
- Delle Monache, L., Lundquist, J. K., Kosovic, B., Johannesson, G., Dyer, K. M., Aines, R. D., Chow, F. K., Belles, R. D., Hanley, W. G., Larsen, S. C., Loosmore, G. A., Nitao, J. J., Sugiyama, G. A., Vogt, P. J., 2008. Bayesian inference and Markov chain Monte Carlo sampling to reconstruct a contaminant source on a continental scale. *Journal of Applied Meteorology and Climatology* 47, 2600–2613.
- Desroziers, G., Ivanov, S., 2001. Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Q. J. Roy. Meteor. Soc.* 127, 1433–1452.
- Elbern, H., Strunk, A., Schmidt, H., Talagrand, O., 2007. Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmos. Chem. Phys.* 7, 3749–3769.
- Enting, I. G., 2002. *Inverse Problems in Atmospheric Constituent Transport*. Cambridge, Atmospheric and Space Science Series.
- Hajivassiliou, V., McFadden, D., Ruud, P., 1996. Simulation of multivariate normal rectangle probabilities and their derivatives - Theoretical and computational results. *Journal of Econometrics* 72, 85–134.
- Hansen, P. C., 1992. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review* 34, 561–580.
- Hansen, P. C., 2010. *Discrete Inverse Problems: Insight and Algorithms*. SIAM.
- Krysta, M., Bocquet, M., Brandt, J., 2008. Probing ETEX-II data set with inverse modelling. *Atmos. Chem. Phys.* 8, 3963–3971.
- Lauvernet, C., Brankart, J.-M., Castruccio, F., Broquet, G., Brasseur, P., Verron, J., 2009. A truncated Gaussian filter for data assimilation with inequality constraints: Application to the hydrostatic stability condition in ocean models. *Ocean Model.* 27, 1–17.

- Li, H., Kalnay, E., Miyoshi, T., 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Q. J. Roy. Meteor. Soc.* 135, 523–533.
- Louis, J., 1979. A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteor.* 17, 197–202.
- Ménard, R., Cohn, S. E., Chang, L.-P., Lyster, P. M., 2000. Assimilation of stratospheric chemical tracer observations using a Kalman filter. Part I: Formulation. *Mon. Wea. Rev.* 128, 2654–2671.
- Michalak, A. M., Bruhwiler, L., Tans, P. P., 2004. A geostatistical approach to surface flux estimation of atmospheric trace gases. *J. Geophys. Res.* 109, D14109.
- Mitchell, H. L., Houtekamer, P. L., 1999. An adaptive ensemble Kalman filter. *Mon. Wea. Rev.* 128, 416–433.
- Pudykiewicz, J., 1989. Simulation of the Chernobyl dispersion with a 3-d hemispheric tracer model. *Tellus* 41B, 391–412.
- Quélo, D., Krysta, M., Bocquet, M., Isnard, O., Minier, Y., Sportisse, B., 2007. Validation of the Polyphemus platform on the ETEX, Chernobyl and Algeciras cases. *Atmos. Env.* 41, 5300–5315.
- Robert, C. P., 1995. simulation of truncated normal variables. *Statistics and Computing* 5, 121–125.
- Saïde, P., Bocquet, M., Osses, A., Gallardo, L., 2011. Constraining surface emissions of air pollutants using inverse modeling: method intercomparison and a new two-step multiscale approach. *Tellus B* 63, 360–370.
- Stohl, A., Seibert, P., Wotawa, G., Arnold, D., Burkhart, J.F., Eckhardt, S., Tapia, C., Vargas, A., Yasunari, T.J., 2011. Xenon-133 and caesium-137 releases into the atmosphere from the Fukushima Dai-ichi nuclear power plant: determination of the source term, atmospheric dispersion, and deposition. *Atmos. Chem. Phys. Discuss.* 11, 28319–28394.
- Villani, M. G., Bergamaschi, P., Krol, M., Meirink, J. F., Dentener, F., 2010. Inverse modeling of European CH₄ emissions: sensitivity to the observational network. *Atmos. Chem. Phys.* 10, 1249–1267.
- Vogel, C. R., 2002. *Computational Methods for Inverse Problems*. SIAM, *Frontiers in Applied Mathematics*.
- Wahba, G., 1990. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics 59. SIAM, Philadelphia.
- Winiarek, V., Vira, J., Bocquet, M., Sofiev, M., Saunier, O., 2011. Towards the operational estimation of a radiological plume using data assimilation after a radiological accidental atmospheric release. *Atmos. Env.* 45, 2944–2955.
- Wu, L., Bocquet, M., Lauvaux, T., Chevallier, F., Rayner, P., Davis, K., 2011. Optimal representation of source-sink fluxes for mesoscale carbon dioxide inversion with synthetic data. *J. Geophys. Res.* 116, D21304.
- Yee, E., Lien, F.-S., Keats, A., D’Amours, R., 2008. Bayesian inversion of concentration data: Source reconstruction in the adjoint representation of atmospheric diffusion. *Journal of Wind Engineering and Industrial Aerodynamics* 96, 1805–1816.

V. Winiarek, CEREa, joint laboratory École des Ponts Paris-Tech - EDF R&D, 6-8 avenue Blaise Pascal, Cité Descartes Champs-sur-Marne, 77455 Marne la Vallée Cedex 2, France. (victor.winiarek@cerea.enpc.fr)

M. Bocquet, CEREa, joint laboratory École des Ponts Paris-Tech - EDF R&D, 6-8 avenue Blaise Pascal, Cité Descartes Champs-sur-Marne, 77455 Marne la Vallée Cedex 2, France. (marc.bocquet@cerea.enpc.fr)

O. Saunier, Institute of Radiation Protection and Nuclear Safety, BP 17, 92262, Fontenay-aux-Roses, France.

A. Mathieu, Institute of Radiation Protection and Nuclear Safety, BP 17, 92262, Fontenay-aux-Roses, France.

Annexe F

Publication III

Article publié en 2014 dans Atmospheric Environment :

Winiarek, V., Bocquet, M., Duhanyan, N., Roustan, Y., Saunier, O., Mathieu, A., 2014. Estimation of the caesium-137 source term from the Fukushima Daiichi nuclear power plant using a consistent joint assimilation of air concentration and deposition observations. *Atmos. Env.* **82**, 268-279.

Dans cet article, le terme source de césium-137 consécutif à l'accident de Fukushima est estimé par des techniques de modélisation inverse en utilisant *simultanément* des mesures de concentrations d'activité dans l'air et des mesures de dépôts. L'équilibre entre ces différents jeux de données est réalisé par les méthodes d'estimation des paramètres caractérisant les statistiques des erreurs *a priori*, qui permettent, de plus, d'estimer l'incertitude *a posteriori* attachée à notre estimation.

Estimation of the caesium-137 source term from the Fukushima Daiichi nuclear power plant using a consistent joint assimilation of air concentration and deposition observations

Victor Winiarek^{a,b}, Marc Bocquet^{a,b}, Nora Duhanyan^a, Yelva Roustan^a, Olivier Saunier^c, Anne Mathieu^c

^a Université Paris-Est, CERE, joint laboratory École des Ponts ParisTech and EDF R&D, Champs-sur-Marne, France

^b INRIA, Paris Rocquencourt research centre, France

^c Institut de Radioprotection et de Sûreté Nucléaire (IRSN), PRP-CRI, SESUC, BMTA, Fontenay-aux-Roses, 92262, France

Abstract

Inverse modelling techniques can be used to estimate the amount of radionuclides and the temporal profile of the source term released in the atmosphere during the accident of the Fukushima Daiichi nuclear power plant in March 2011. In Winiarek et al. (2012b), the lower bounds of the caesium-137 and iodine-131 source terms were estimated with such techniques, using activity concentration measurements. The importance of an objective assessment of prior errors (the observation errors and the background errors) was emphasised for a reliable inversion. In such critical context where the meteorological conditions can make the source term partly unobservable and where only a few observations are available, such prior estimation techniques are mandatory, the retrieved source term being very sensitive to this estimation.

We propose to extend the use of these techniques to the estimation of prior errors when assimilating observations from several data sets. The aim is to compute an estimate of the caesium-137 source term jointly using all available data about this radionuclide, such as activity concentrations in the air, but also daily fallout measurements and total cumulated fallout measurements. It is crucial to properly and simultaneously estimate the background errors and the prior errors relative to each data set. A proper estimation of prior errors is also a necessary condition to reliably estimate the a posteriori uncertainty of the estimated source term. Using such techniques, we retrieve a total released quantity of caesium-137 in the interval 11.6 – 19.3 PBq with an estimated standard deviation range of 15 – 20% depending on the method and the data sets. The “blind” time intervals of the source term have also been strongly mitigated compared to the first estimations with only activity concentration data.

This article has been published in Atmospheric Environment with the reference: Winiarek, V., Bocquet, M., Duhanyan, N., Roustan, Y., Saunier, O., Mathieu, A., 2014. Estimation of the caesium-137 source term from the Fukushima Daiichi nuclear power plant using a consistent joint assimilation of air concentration and deposition observations. *Atmos. Env.* **82**. 268-279.

Keywords: Data assimilation, Atmospheric dispersion, Fukushima accident, Source estimation

1. Introduction

1.1. The Fukushima Daiichi accident

On March 11, 2011, 05:46 UTC, a magnitude 9.0 (M_w) undersea megathrust earthquake occurred in the Pacific Ocean and an extremely destructive tsunami hit the Pacific coast of Japan approximately one hour later. These events caused the automatic shut-down of

4 power plants in Japan. Diesel backup power systems should have sustained the reactors cooling process. In Fukushima Daiichi these backup devices were unfortunately inoperative mainly because of the damages caused by the tsunami.

The Fukushima Daiichi NPP has six nuclear reactors. At the time of the earthquake, reactor 4 had been defuelled and reactors 5 and 6 were in a cold shut-down for planned maintenance.

In the hours that followed, the situation quickly be-

Email address: bocquet@cerea.enpc.fr (Marc Bocquet)

came critical. Reactors 1, 2 and 3 experienced at least partial meltdown and hydrogen explosions. Additionally, fuel rods stored in pools in each reactor building began to overheat as water levels in the pools dropped.

All these events caused a massive discharge of radioactive materials in the atmosphere. The total quantities of released radionuclides as well as the time evolution of these releases have to be estimated in order to assess the sanitary and environmental impact of the accident.

Mathieu et al. (2012) used core inventories and in-situ data (pressure and temperature measurements in the reactors and γ -dose rates in the NPP) as well as first available observations over Japan to build a source term, which was used in Korsakissok et al. (2013) to assess the high γ -dose rates zones at a local scale. Chino et al. (2011) used a few activity concentration measurements in the air to calibrate the magnitude of some identified releases of ^{131}I and ^{137}Cs . Katata et al. (2012) and Terada et al. (2012) updated these estimations as new data were available. Stohl et al. (2012) performed an inverse modelling estimation using activity concentrations in the air, as well as deposition data at a global scale to estimate the release of ^{133}Xe and ^{137}Cs . However their estimation was shown to be quite sensitive to the prior information on the source used in the inversion. At the same time, Winiarek et al. (2012b) proposed a method to properly estimate the prior errors to perform inverse modelling using activity concentrations in the air. They applied it to estimate the source terms of ^{131}I and ^{137}Cs .

1.2. Objectives and outline

In Winiarek et al. (2012b) we emphasised the importance of a proper estimation of prior errors in the inverse modelling algorithm. We proposed new methods to estimate two hyper-parameters: the variance of observation errors and the variance of background errors, assuming that all observation errors have the same variance. This assumption is acceptable when using only one type of data in the algorithm. On the other hand the general lack of data in accidental situations stresses the need for algorithms that would use all available data in the same inversion. The first objective of this paper is to extend the methods proposed in Winiarek et al. (2012b) to the simultaneous estimation of prior errors when using different data sets in the inversion. The second objective is to apply these methods to the challenging reconstruction of the caesium-137 source term of the Fukushima Daiichi accident, as well as to obtain an objective uncertainty on this estimation.

In Section 2 we briefly recall the methodology for the inverse modelling of accidental releases of pollutants.

The algorithm is sensitive to the statistics of the errors so that we propose methods based on the maximum likelihood principle to estimate these errors. This general maximum likelihood scheme takes into account the positivity of the source and the presence of several types of data.

In Section 3 we apply this new method to the reconstruction of the atmospheric release of ^{137}Cs from the Fukushima Daiichi power plant. The inversions are computed using three data sets: activity concentrations in the air, daily measurements of deposited material and total cumulated deposition. The new source terms are discussed and compared to earlier estimations. The posterior uncertainties of the retrieved sources are also computed.

Results are summarised and conclusions are given in Section 4.

2. Methodology

2.1. Inverse modelling of accidental releases and estimation of errors

To reconstruct the source term of an accidental release of a pollutant into the atmosphere, inverse modelling techniques are a powerful alternative to a trial and error approach with direct numerical models, which is still widely used in such situations. Inverse modelling techniques can objectively estimate the source term using the information content of an observation set and a numerical model that simulates the dispersion event. The atmospheric transport model (ATM), which is linear in our case for gaseous and particulate matter, provides the relationship between the source term and the observation set through the source-receptor equation

$$\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\sigma} + \boldsymbol{\epsilon}, \quad (1)$$

where $\boldsymbol{\mu}$ in \mathbb{R}^d is the measurement vector, $\boldsymbol{\sigma}$ in \mathbb{R}^N is the source vector, and \mathbf{H} is the Jacobian matrix of the transport model which incorporates the observation operator as well. The vector $\boldsymbol{\epsilon}$ in \mathbb{R}^d , called the observation error in this article, represents the instrumental errors, the representativeness errors and a fraction of model error altogether.

In an accidental context the number of observations is very often limited. Specific meteorological conditions, like during the Fukushima accident, can also lead to a weak observability of a part of the source term. In these cases, the source-receptor relationship defined by Eq. (1) constitutes an ill-posed inverse problem (Winiarek et al., 2011).

One solution to deal with the lack of constraints is to implement parametric methods where the source to retrieve is reduced to a very limited number of parameters. The inverse problem is *de facto* regularised and it can be solved using different techniques, such as stochastic sampling techniques which allow to access the parameters posterior distributions (Delle Monache et al., 2008; Yee et al., 2008). Nevertheless if the true source does not match the parametric model, the inversion may fail or yield a meaningless solution.

Another option is the use of non-parametric methods to retrieve a general source field. This field can be discretised, for example to match the model discretisation, but the number of variables remains large and may still be larger than the number of observations in the data set. The non-parametric approach is robust and flexible since no strong a priori assumptions are made on the source but it has its own constraints. If Gaussian statistics are assumed for observation errors, the inversion relies on the minimisation of the cost function

$$\mathcal{L}(\boldsymbol{\sigma}) = \frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}), \quad (2)$$

where \mathbf{R} is the observation error covariance matrix: $\mathbf{R} = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]$, where $\boldsymbol{\epsilon}$ has been defined by Eq. (1). One simple choice is to neglect correlations between observation errors and to take \mathbf{R} diagonal. If all the observations are of the same type, one can even assume that $\mathbf{R} = r^2 \mathbf{I}_d$, r^2 being the variance of the observation errors (\mathbf{I}_d is the identity matrix in $\mathbb{R}^{d \times d}$). With a low number of observations and/or a poor observability of several source term parts, the minimisation of Eq. (2) gives infinitely many solutions. One needs to regularise the inverse problem, usually by adding a Tikhonov term in the cost function:

$$\mathcal{L}(\boldsymbol{\sigma}) = \frac{1}{2} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma})^T \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}) + \frac{1}{2} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_b)^T \mathbf{B}^{-1} (\boldsymbol{\sigma} - \boldsymbol{\sigma}_b). \quad (3)$$

The solution of the inverse problem is now unique, but two additional (vector and matrix) parameters have been introduced: $\boldsymbol{\sigma}_b$ is the first guess for the source (or background term) and \mathbf{B} is the background error covariance matrix: $\mathbf{B} = \mathbb{E}[(\boldsymbol{\sigma} - \boldsymbol{\sigma}_b)(\boldsymbol{\sigma} - \boldsymbol{\sigma}_b)^T]$.

In an accidental situation and particularly in the Fukushima accident context, the choice of $\boldsymbol{\sigma}_b = \mathbf{0}$ is relevant because (i) many of the parameters are likely to be zero (ii) it guarantees an independent estimate (iii) it avoids the risk of an *inversion crime*, since most of the first guess built by physics models are eventually calibrated using early observations. One can refer to

Bocquet (2005); Davoine and Bocquet (2007); Winiarek et al. (2012b) for extended discussions on this choice.

Still in an accidental context, off-diagonal terms in \mathbf{B} are often negligible. For example the γ -dose measurements made in the Fukushima nuclear power plant (NPP) indicate that the source term is probably composed of uncorrelated events¹. This is the reason why we take $\mathbf{B} = m^2 \mathbf{I}_N$ in this article.

As it was shown in the Chernobyl case in Davoine and Bocquet (2007) or in the Fukushima case in Winiarek et al. (2012b) or Stohl et al. (2012), the retrieved source is very sensitive to the matrices \mathbf{R} and \mathbf{B} , hence to the two hyper-parameters r and m . This is the reason why hyper-parameters estimation techniques are required. In Davoine and Bocquet (2007); Krysta et al. (2008); Saide et al. (2011), the L-curve technique of Hansen (1992) was successfully used to estimate the ratio r/m or both parameters. In the recent years several methodological developments in the data assimilation for weather forecast have focused on the estimation of the hyper-parameters of the prior errors. They are mostly based on either cross-validation technique or on the maximum likelihood principle (e.g. Mitchell and Houtekamer, 1999; Chapnik et al., 2004, 2006; Anderson, 2007; Li et al., 2009). These techniques have also been implemented in the context of atmospheric chemistry inverse modelling, using for example the χ^2 criterion (Ménard et al., 2000; Elbern et al., 2007; Davoine and Bocquet, 2007), the maximum likelihood principle (Michalak et al., 2004), or statistical diagnostics (Schwinger and Elbern, 2010).

In Winiarek et al. (2012b) we implemented several methods for the estimation of r and m . One of these methods relies on the L-curve coupled with a χ^2 criterion. Another one is based on the maximum likelihood principle and takes into account without approximation the positivity of the source. This study only exploited activity concentrations in the air. In the case where several types of data are used in the inversion algorithm, the number of hyper-parameters to estimate increases. One can for instance try to estimate one value of the variance of observation errors for each data set, that is an r^2 attached to each data set. This simultaneous estimation of prior errors is the methodological objective of this article.

¹<http://www.tepco.co.jp/en/nu/monitoring/index-e.html>

2.2. Prior errors statistics and cost function minimisation

The observation errors defined by Eq. (1) are assumed normally distributed, with a probability density function (pdf):

$$p_e(\epsilon) = \frac{e^{-\frac{1}{2}\epsilon^T \mathbf{R}^{-1} \epsilon}}{\sqrt{(2\pi)^d |\mathbf{R}|}}, \quad (4)$$

where $|\mathbf{R}|$ is the determinant of the observation error covariance matrix \mathbf{R} . \mathbf{R} is assumed diagonal (it is an approximation as model errors may introduce some correlations). Besides, it is assumed that the errors made on observations of the same type have the same variance, so that if \mathbf{R}_i represents the sub-block of \mathbf{R} relative to the data set i , one has $\mathbf{R}_i = r_i^2 \mathbf{I}_{d_i}$, where d_i is the number of observations in data set i ($\sum_{i=1}^{N_d} d_i = d$, N_d is the number of different data sets) and r_i^2 is the corresponding error variance. Note that due to the limited number of observations and in order to avoid a too severely underconstrained problem, it is essential to keep the number of hyper-parameters that define \mathbf{R} as small as possible (N_d hyper-parameters in our case: $(r_i)_{1 \leq i \leq N_d}$). For instance, an approach such as the one put forward by Desroziers et al. (2005) or Schwinger and Elbern (2010) is likely to be unaffordable in this accidental context.

As far as background errors are concerned, we could also use Gaussian statistics. The advantages of this choice would be analytical solutions for both the source term estimation and the related uncertainty based on the Best Linear Unbiased Estimator (BLUE) theory. However, such a Gaussian assumption could lead to negative values in the retrieved source term, because of the lack of sufficient observations to constrain the source term. To avoid non-physical results, truncated normal statistics for background errors are considered, enforcing the positivity of the retrieved source term. It is known to provide valuable information to the data assimilation system (Bocquet et al., 2010) and was successfully used in the context of the Fukushima Daiichi accident (Winiarek et al., 2012b). The corresponding pdf of this normalised truncated normal distribution reads in the general case

$$\begin{cases} \text{if } \sigma \geq \mathbf{0} & p(\sigma) = \left(\int_{s \geq \mathbf{0}} e^{-\frac{1}{2}(s-\sigma_b)^T \mathbf{B}^{-1}(s-\sigma_b)} ds \right)^{-1} \\ & \times e^{-\frac{1}{2}(\sigma-\sigma_b)^T \mathbf{B}^{-1}(\sigma-\sigma_b)} \\ \text{otherwise} & p(\sigma) = 0. \end{cases} \quad (5)$$

The normalisation factor can be simplified in our case where \mathbf{B} is diagonal and $\sigma_b = \mathbf{0}$ to yield the following

semi-Gaussian pdf:

$$\begin{cases} \text{if } \sigma \geq \mathbf{0} & p(\sigma) = \frac{e^{-\frac{1}{2}\sigma^T \mathbf{B}^{-1} \sigma}}{\sqrt{(\pi/2)^N |\mathbf{B}|}} \\ \text{otherwise} & p(\sigma) = 0, \end{cases} \quad (6)$$

where $|\mathbf{B}|$ is the determinant of the background error covariance matrix \mathbf{B} .

Bayes' rule helps to formulate the inference, after the acquisition of the measurement vector $\boldsymbol{\mu}$:

$$\begin{aligned} p(\sigma|\boldsymbol{\mu}) &= \frac{p(\boldsymbol{\mu}|\sigma)p(\sigma)}{p(\boldsymbol{\mu})} = \frac{p_e(\boldsymbol{\mu} - \mathbf{H}\sigma)p(\sigma)}{p(\boldsymbol{\mu})} \\ &\propto \exp \left\{ -\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\sigma)^T \mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{H}\sigma) \right. \\ &\quad \left. -\frac{1}{2}\sigma^T \mathbf{B}^{-1} \sigma \right\} \mathbb{I}_{\sigma \geq \mathbf{0}}, \end{aligned} \quad (7)$$

where $\mathbb{I}_{\sigma \geq \mathbf{0}}$ is equal to 1 when $\sigma_i \geq 0$, for every $i \leq N$. Otherwise its value is 0.

From this inference, the source term is estimated using the maximum a posteriori estimator (MAP), denoted σ_a :

$$\sigma_a = \underset{\sigma}{\operatorname{argmax}} p(\sigma|\boldsymbol{\mu}). \quad (8)$$

Maximising $p(\sigma|\boldsymbol{\mu})$ is equivalent to maximising $\ln p(\sigma|\boldsymbol{\mu})$, which is equivalent to maximising the term in the exponential under the constraint of positivity, which is ultimately equivalent to minimising cost function Eq. (3) under the constraint of positivity. Since there is no analytical solution to this problem, the positivity of σ should be enforced during the numerical minimisation which is performed with a bounded quasi-Newton algorithm (Byrd et al., 1995).

In practice, because of the low number of observations, the estimated source term is very sensitive to the matrices \mathbf{R} and \mathbf{B} , i.e. to the hyper-parameters $(r_i)_{1 \leq i \leq N_d}$ and m . This is the reason why we have to estimate these parameters rigorously. We propose to extend the methods developed in Winiarek et al. (2012b) to the use of several different data sets by simultaneously estimating the respective hyper-parameters.

2.3. Estimation of hyper-parameters

2.3.1. Unapproximated maximum likelihood values screening

The estimation of the prior errors' magnitude proposed in this section relies on the maximum likelihood paradigm (Dee, 1995). As the prior probabilities depend on the hyper-parameters, the likelihood of the observation set, which can be written

$$p(\boldsymbol{\mu}|\boldsymbol{\theta}) = \int d\sigma p(\boldsymbol{\mu}|\sigma; \boldsymbol{\theta}) p(\sigma|\boldsymbol{\theta}), \quad (9)$$

is a function of the hyper-parameters vector $\boldsymbol{\theta} = (r_1, \dots, r_{N_d}, m)^T$. The pdfs $p(\boldsymbol{\mu}|\boldsymbol{\sigma}; \boldsymbol{\theta})$ and $p(\boldsymbol{\sigma}|\boldsymbol{\theta})$ are the prior pdfs defined by Eq. (4) and Eq. (6). If it exists, the vector $\boldsymbol{\theta}$ that maximises $p(\boldsymbol{\mu}|\boldsymbol{\theta})$ is the most likely vector of hyper-parameters consistent with the observation set $\boldsymbol{\mu}$.

The most direct way to estimate these optimal hyper-parameters is to screen the likelihood function for a range of values of $\boldsymbol{\theta}$. With our hypothesis on the errors statistics, the covariance matrices and the first guess, the likelihood can be written:

$$p(\boldsymbol{\mu}|\boldsymbol{\theta}) = \frac{e^{-\frac{1}{2}\boldsymbol{\mu}^T(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1}\boldsymbol{\mu}}}{\sqrt{(2\pi)^d|\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}|}} \times \int_{\boldsymbol{\sigma} \geq \mathbf{0}} \frac{e^{-\frac{1}{2}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_{\text{BLUE}})^T \mathbf{P}_{\text{BLUE}}^{-1}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_{\text{BLUE}})}}{\sqrt{(\pi/2)^N |\mathbf{P}_{\text{BLUE}}|}} d\boldsymbol{\sigma}, \quad (10)$$

where $\boldsymbol{\sigma}_{\text{BLUE}}$ is the BLUE estimator, which in our case reads:

$$\boldsymbol{\sigma}_{\text{BLUE}} = \mathbf{B}\mathbf{H}^T (\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1} \boldsymbol{\mu}, \quad (11)$$

and \mathbf{P}_{BLUE} is the corresponding analysis error covariance matrix:

$$\mathbf{P}_{\text{BLUE}} = \mathbf{B} - \mathbf{B}\mathbf{H}^T (\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1} \mathbf{H}\mathbf{B}. \quad (12)$$

The integral term in Eq. (10) has no analytical solution, but can be numerically computed using a stochastic method, such as the GHK simulator from Hajivassiliou et al. (1996). Nevertheless, if the dimension of $\boldsymbol{\theta}$ is high, the size of the space to screen can lead to a costly computation.

For details about the calculation of the likelihood expression, and in particular the general case expression, and for details about the use of the GHK simulator to estimate the integral of a truncated normal distribution, one can refer to Winiarek et al. (2012b) and Winiarek et al. (2012a).

As a statistical consistent method, we considered it as our reference method, denoted ML in this study. Faster but approximate alternatives can nonetheless be considered and tested against this statistically consistent method.

2.3.2. Iterative scheme à-la-Desroziers

As an alternative to the costly computation of the likelihood, the use of an iterative scheme which would quickly converge to the maximum likelihood is relevant. In Winiarek et al. (2012b) we used such an iterative algorithm for the estimation of two hyper-parameters. This iterative algorithm was shown to converge to a fixed-point that corresponds, in the context of Gaussian

statistics, to the pair of hyper-parameters of maximum likelihood. This algorithm is only an approximation in the context of semi-Gaussian statistics but it yielded acceptable, through slightly different, results. Building on Desroziers and Ivanov (2001) we propose an extension of this online tuning scheme to the simultaneous estimation of several prior errors variances. The formulae read:

$$m^2 = \frac{2 J_b(\boldsymbol{\sigma}_a)}{N - \text{tr}(\mathbf{P}_{\text{BLUE}} \mathbf{B}^{-1})}, \quad (13)$$

$$r_i^2 = \frac{2 J_{oi}(\boldsymbol{\sigma}_a)}{d_i - \text{tr}(\mathbf{H}_i \mathbf{P}_{\text{BLUE}} \mathbf{H}_i^T \mathbf{R}_i^{-1})}, \quad (14)$$

where \mathbf{H}_i and \mathbf{R}_i are the sub-blocks of respectively the Jacobian matrix \mathbf{H} and the observation error covariance matrix \mathbf{R} related to data set i , whose observation vector is noted $\boldsymbol{\mu}_i$. \mathbf{P}_{BLUE} is defined by Eq. (12). J_b and J_{oi} are defined by:

$$J_b(\boldsymbol{\sigma}) = \frac{1}{2} \boldsymbol{\sigma}^T \boldsymbol{\sigma}, \quad (15)$$

$$J_{oi}(\boldsymbol{\sigma}) = \frac{1}{2} (\boldsymbol{\mu}_i - \mathbf{H}_i \boldsymbol{\sigma})^T (\boldsymbol{\mu}_i - \mathbf{H}_i \boldsymbol{\sigma}). \quad (16)$$

The source vector $\boldsymbol{\sigma}_a$ is obtained from the minimisation of the cost function:

$$\mathcal{L}(\boldsymbol{\sigma}) = \frac{J_b(\boldsymbol{\sigma})}{m^2} + \sum_{i=1}^{N_d} \frac{J_{oi}(\boldsymbol{\sigma})}{r_i^2} \quad (17)$$

under the constraint $\boldsymbol{\sigma} \geq \mathbf{0}$. These equations can be used in an iterative scheme which we shall call Desroziers' scheme later on. This algorithm quickly converges (3-4 iterations here) to a fixed-point giving estimated hyper-parameters and the related source term.

3. Applications to the Fukushima accident

3.1. Observations

In order to illustrate the proposed methods, observations from three different data sets will be considered:

- The activity concentration in the air over Japan as described and referred to in Winiarek et al. (2012b). This data set contains 104 observations.
- Starting from 18 March 2011, daily measurements of deposited ^{137}Cs in 22 prefectures, which represents a total of 198 observations².

²<http://www.mext.go.jp/english/incident/1305529.htm>

- Measurements of total deposited ^{137}Cs in an area near the NPP (approximately 100 km around). Among the 2180 deposition measurements provided by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) during May and June 2011³, 16 were filtered out because they are impacted by near-field effects that are not represented by larger scale ATMs, so that 2164 observations, that are distant enough from the NPP, were considered. As shown in Fig. 5(a), they are densely distributed in space, but on the downside there is no scale of time in these observations.

The distribution of the observation sites led to use a mesoscale domain approximately covering Japan. Because of the spatial extension and density of the sites, and because of the frequency of the observations, the ATM needs a rather limited simulation domain with high resolution in space and in time.

3.2. Modelling the atmospheric dispersion

3.2.1. Meteorological fields

ECMWF meteorological fields with a spatial resolution of $0.25^\circ \times 0.25^\circ$ and a temporal resolution of 3 hours are too coarse to be used for our need. That is the reason why we computed mesoscale meteorological fields with the Weather Research and Forecasting (WRF) numerical model (Skamarock et al., 2008). The main objective is to obtain meteorological fields with a spatial resolution of approximately 5 km and a temporal resolution of 1 h. The computed meteorological fields are inputs for the atmospheric transport model. Physical parametrisations as well as the design of simulation domains are summarised in Tab. 1 and the simulation domains are displayed in Fig. 1. One of the key feature is the use of several thousands of meteorological observations to constrain the meteorological fields through nudging techniques (Stauffer and Seaman, 1994).

As shown in Fig. 2, WRF simulations show a good ability to model the occurrence of rain episodes but have the general tendency to overestimate precipitation rates (Katata et al. (2012) observed the same behaviour using MM5). This bias is a severe drawback when looking at deposition processes as this study aims to do. To avoid overestimating wet deposition, we used ground observations of rain rates in Fukushima and Ibaraki prefectures to compute a global debiasing coefficient, that we

found equal to 2.4, by which were divided all precipitation rates computed by the WRF model. We have also tested local corrections of the precipitation fields using data assimilation techniques but we did not find them to be as robust and reliable as a simpler global debiasing correction.

3.2.2. Atmospheric transport model

The simulations of the dispersion of radionuclides from the Fukushima Daiichi nuclear power plant have been performed with the chemistry-transport model POLAIR3D, the Eulerian model of the POLYPHEMUS platform. It has been validated for the simulation of radionuclides transport on the European Tracer Experiment, on the Algeciras incident and on the Chernobyl accident (Quélo et al., 2007).

The model integrates the concentration field c of ^{137}Cs , following the transport equation

$$\frac{\partial c}{\partial t} + \text{div}(\mathbf{u}c) = \text{div}\left(\rho \mathbf{K} \nabla \left(\frac{c}{\rho}\right)\right) - \Lambda^s c - \Lambda^d c + \sigma \quad (18)$$

where ρ is the air density, Λ^s is the scavenging rate, Λ^d represents the radioactive decay and σ is the point-wise source. \mathbf{K} is the matrix of turbulent diffusion, diagonal in practice. The vertical component is given by K_z , computed with Louis parametrisation (Louis, 1979). The horizontal component K_H is taken constant. The boundary condition on the ground is

$$K_z \nabla c \cdot \mathbf{n} = -v^{\text{dep}} c \quad (19)$$

where \mathbf{n} is the upward oriented unitary vector, and v^{dep} is the dry deposition velocity of ^{137}Cs .

Two domains of simulation are considered. The finest domain of simulation is a mesoscale domain covering Japan, from 131.03°E to 144.53°E and from 30.72°N to 43.72°N with a spatial resolution of $0.05^\circ \times 0.05^\circ$. The number of grid points in this domain is 270×260 . Because of the small size of the domain there is a risk of re-circulation of the plume outside the domain so that the model could fail to account for radionuclides re-entries. To avoid such situation and in order to compute the boundary conditions to this domain by a nesting technique, another more extended domain of simulation with a coarser resolution is used. It covers a region from 115.03°E to 165.03°E and from 25.02°N to 60.02°N with a resolution of $0.25^\circ \times 0.25^\circ$. This configuration is displayed in Fig. 1. For both domains the POLAIR3D model is configured with 15 vertical levels ranging from 0 to 8000 m.

Caesium-137 is modelled as monodispersed passive particulate matter with a radioactive decay of 11000

³http://www.mext.go.jp/b_menu/shingi/chousa/gijyutu/017/shiryo/_icsFiles/afiedfile/2011/09/02/1310688_1.pdf

Table 1: Configuration and physical parametrisations of the WRF model. A two-way nesting technique is used between domain 1 and domain 2.

Domain	1	2
Spatial resolution	18 km	6 km
Number of grid points	340 × 250	241 × 241
Number of vertical levels	27	27
Numerical time-step	60 s	20 s
Output time-step	3600 s	3600 s
Planetary boundary layer	Yonsei University	Yonsei University
Micro-physics	Kessler	WRF Single Moment 3
Cumulus physics	Grell-Devenyi	Grell-Devenyi
Longwave radiation	RRTM	RRTM
Shortwave radiation	Dudhia	Dudhia
Surface layer	MM5 similarity	MM5 similarity
Land surface	Noah LSM	Noah LSM
Nudging	Grid nudging	Grid nudging

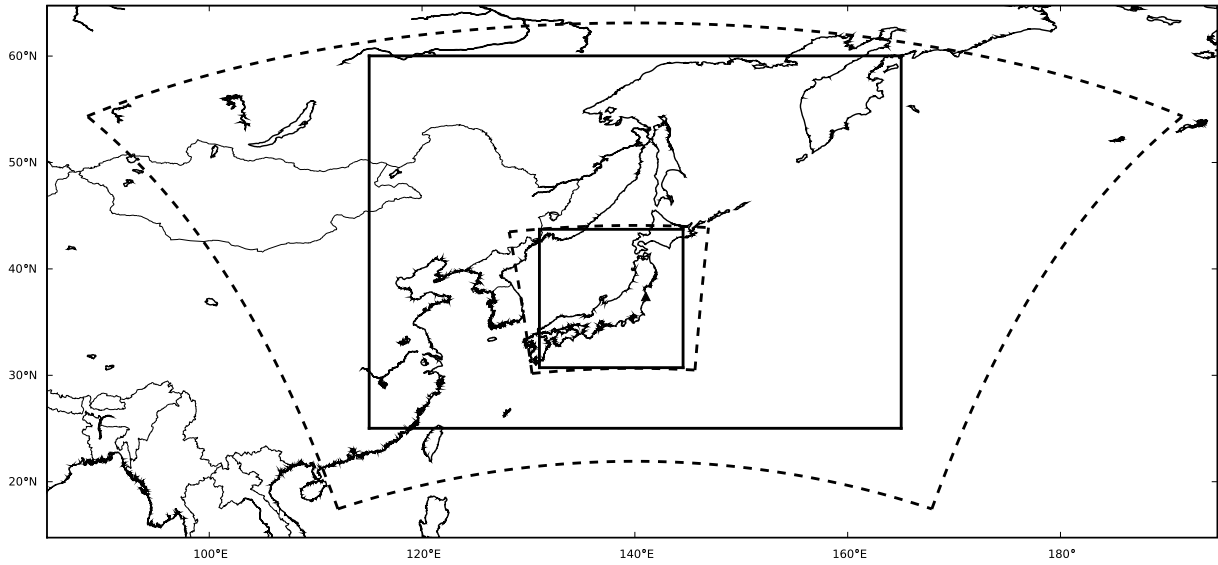


Figure 1: Map of the simulation domains used in WRF (dashed-line domains) and in POLAIR3D (full-line domains). Two-way nesting is used in the WRF simulations and one-way nesting in the POLAIR3D simulation. A triangle marks the location of the Fukushima Daiichi nuclear power plant.

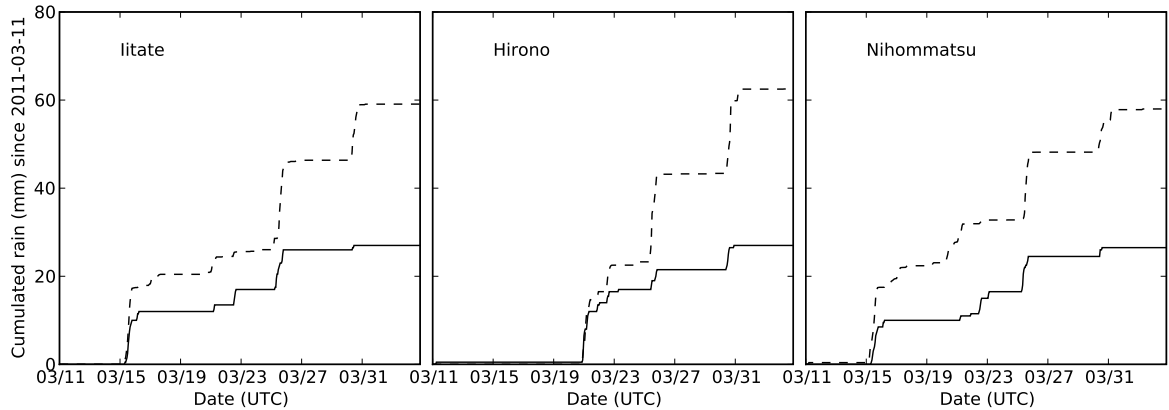


Figure 2: Comparison between WRF calculations (dashed lines) and observations (full lines) for cumulated precipitation (in mm) at some surface weather stations in Fukushima prefecture.

days. Dry deposition is modelled using a simple scheme with a constant deposition velocity: $v^{\text{dep}} = 0.15 \text{ cm.s}^{-1}$ (Bocquet, 2012) over land, and $v^{\text{dep}} = 0.01 \text{ cm.s}^{-1}$ over the ocean (see Estournel et al. (2012) and references within). As far as wet deposition is concerned, the parametrisation used in this study is a below-cloud scavenging scheme of the form:

$$\Lambda^s = a \left(\frac{p}{p_0} \right)^b \quad (20)$$

where p stands for the precipitation rate and $p_0 = 1 \text{ mm.h}^{-1}$, a and b being two constants respectively equal to $8.4 \cdot 10^{-5} \text{ s}^{-1}$ and 0.79 following Maryon et al. (1991).

The advection is implemented thanks to a third-order direct space-time scheme, with a Koren-Sweby flux limiter function. Because of the plume sharp gradients, it is important that such a limiter is used. The diffusion scheme is integrated through an implicit second-order Rosenbrock scheme, with a three-point spatial scheme, and directional splitting.

This study attempts to reconstruct the source term from March 11 to April 1 (which represents $N = 504$ one-hour time-steps). However the simulations run over a longer period (from March 11 to April 5) in order to exploit the information content of later observations. A total of 504 direct simulations are thus performed to fill the Jacobian matrix \mathbf{H} column by column, and no adjoint model is needed in this process following Abida and Bocquet (2009); Winiarek et al. (2011).

3.3. Inverse modelling results only using activity concentrations in the air

The proposed methods can of course be applied to inverse modelling using only one data set, such as activity concentration in the air data (the first data set described in Section 3.1 and which contains 104 observations). In this case, the formulae are rigorously equivalent to the ones presented in Winiarek et al. (2012b). It is nonetheless interesting to perform such inversion. It allows to check and confirm the consistency of the model, even though the meteorological fields and the removal processes parameterisations are different.

3.3.1. Estimation of parameters and total released activity

The estimated hyper-parameters and the estimated total released activities are reported in Tab. 2 together with results found in Winiarek et al. (2012b) using the same data set. The estimated hyper-parameters as well as the total released activities in the case of the ML method are close and confirm the consistency of this approach.

As discussed in Winiarek et al. (2012b) Desroziers' scheme is based on Gaussian assumptions and thus has to be considered an approximation in the case of semi-Gaussian assumptions for background errors statistics. In the situation where only few data are available it could yield results different from the maximum likelihood estimation. Nevertheless the gap here is smaller than it was in Winiarek et al. (2012b) with the same data. It may be caused by the use of a mesoscale model with a better resolution in space and in time which increases the general consistency in the system. The clear decrease of r , that represents the magnitude of the observational error, including representativeness error and

part of model error, was to be expected and may be due to a better resolved model hence reducing the impact of model error in the inversion.

Using the maximum likelihood estimation, the total released activity of ^{137}Cs is estimated to be 1.1×10^{16} Bq. This is consistent with other estimations: 1.2×10^{16} Bq for Winiarek et al. (2012b) and Chino et al. (2011), 1.3×10^{16} Bq for Terada et al. (2012), 1.6×10^{16} Bq for Saunier et al. (2013), 2.1×10^{16} Bq for Mathieu et al. (2012) or 3.6×10^{16} Bq for Stohl et al. (2012).

3.3.2. Temporal profile and uncertainty reduction

Due to the meteorological conditions which have often transported the radionuclides towards the Pacific Ocean, and to the low number of activity concentration data, the observability of the plume is reduced. That is why inverse modelling methods are only able to reconstruct the source term in some specific time intervals. Consequently the estimated total released activities have to be considered as lower bound estimates of the actual releases.

The reconstructed source term and its uncertainty are displayed in Fig. 3(a). The posterior uncertainty has been computed from a Monte Carlo analysis, where the observations and the background term are perturbed using their prior errors definition and the hyper-parameters estimates (2×10^4 draws and inversions are performed). Then the standard deviation of the estimators ensemble is used to estimate the posterior uncertainty of the reconstructed source term. The uncertainty on the total released activity is around 65%.

The time intervals of observability are clearly visible in the shape of the uncertainty. Three periods are particularly well observed: the first one lies approximately from 14 March to 15 March, the second one lies from 19 March to 22 March and the last one from 24 March to 26 March. Once again the temporal profiles of the source term and its uncertainty are very consistent with the inversion made in Winiarek et al. (2012b).

Compared to source term estimates constructed from in-situ events monitoring and core inventories (Mathieu et al., 2012), a few events are not retrieved: the first hydrogen explosions in Unit 1 on 12 March, the ventings of Unit 3 on 13 March and the events concerning Unit 2 and 3 on 16 March and 18 March. On the other hand, the multiple events of 14 March and 15 March are present in the reconstructed source term, even though in an incomplete way since the last release, probably between 7:00 UTC and 12:00 UTC on 15 March, is missing. This release is partly accounting for the north-west

pattern on the deposition map, but no activity concentration observation is available to help reconstruct this event. The releases of 20 March and the ones from 21 March to 23 March are also retrieved. As far as the release around 25 March is concerned, there seems to be a slight offset of 12 hours in the reconstruction. The releases reconstructed on 19 March are not mentioned by these inventories, but seem compatible with the in-situ measurements of γ -dose rates from operator TEPCO⁴, for example on the northside of main office or near the west gate of the NPP. They are also retrieved by Stohl et al. (2012). They precede the attempts of emergency cooling with Tokyo Fire Department means. Finally, the releases around 30 March, not mentioned by the former inventories, are also retrieved by Terada et al. (2012).

The scatter plot of all observations is displayed in Fig. 4(a). The simulation using the source term reconstructed with activity concentration data only does not show any systematic bias when estimating the deposited activities. Again, this shows the consistency of the model and in particular of the removal processes (wet and dry deposition).

3.4. Results of the inverse modelling using the three raw data sets

The total deposited ^{137}Cs measurements offer no information about the time of deposition as they are measurements performed a posteriori, in May and June 2011. When only these measurements are used to attempt reconstructing the source term, the total released estimated activity seems consistent (between 1.1×10^{16} Bq and 1.2×10^{16} Bq). However, the temporal profile, displayed in Fig. 3(b), is highly doubtful. Only releases on 23 and on 25 March are clearly visible. Indeed, the system has too much freedom to fit the data. The only constraint is the background term in the cost function, which in the case where the first guess is taken null (as we do) only defines a scale of amplitude. Therefore, even if these observations are abundant with a good spatial distribution, we propose in the next sections to jointly use them with other measurements with a good temporal resolution, such as activity concentrations in the air and daily measurements of deposited material. Consequently we propose to use the three data sets described in Section 3.1 in the same inversion. In this aim the prior errors have to be estimated simultaneously for the background and the three data sets ($N_d = 3$).

⁴<http://www.tepco.co.jp/en/nu/monitoring/index-e.html>

Table 2: Estimation of parameters and corresponding reconstructed released activity for caesium-137 source reconstruction using only activity concentration observations in the air.

parameter	method	Regional scale model Winiarek et al. (2012b)	Mesoscale model This study
r (Bq m^{-3})	Desroziers' scheme	5.4	2.1
	Maximum likelihood	3.3	1.9
m (Bq s^{-1})	Desroziers' scheme	5.3×10^{10}	8.9×10^{10}
	Maximum likelihood	2.0×10^{11}	1.6×10^{11}
Released activity (Bq)	Desroziers' scheme	3.3×10^{15}	7.2×10^{15}
	Maximum likelihood	1.2×10^{16}	1.1×10^{16}

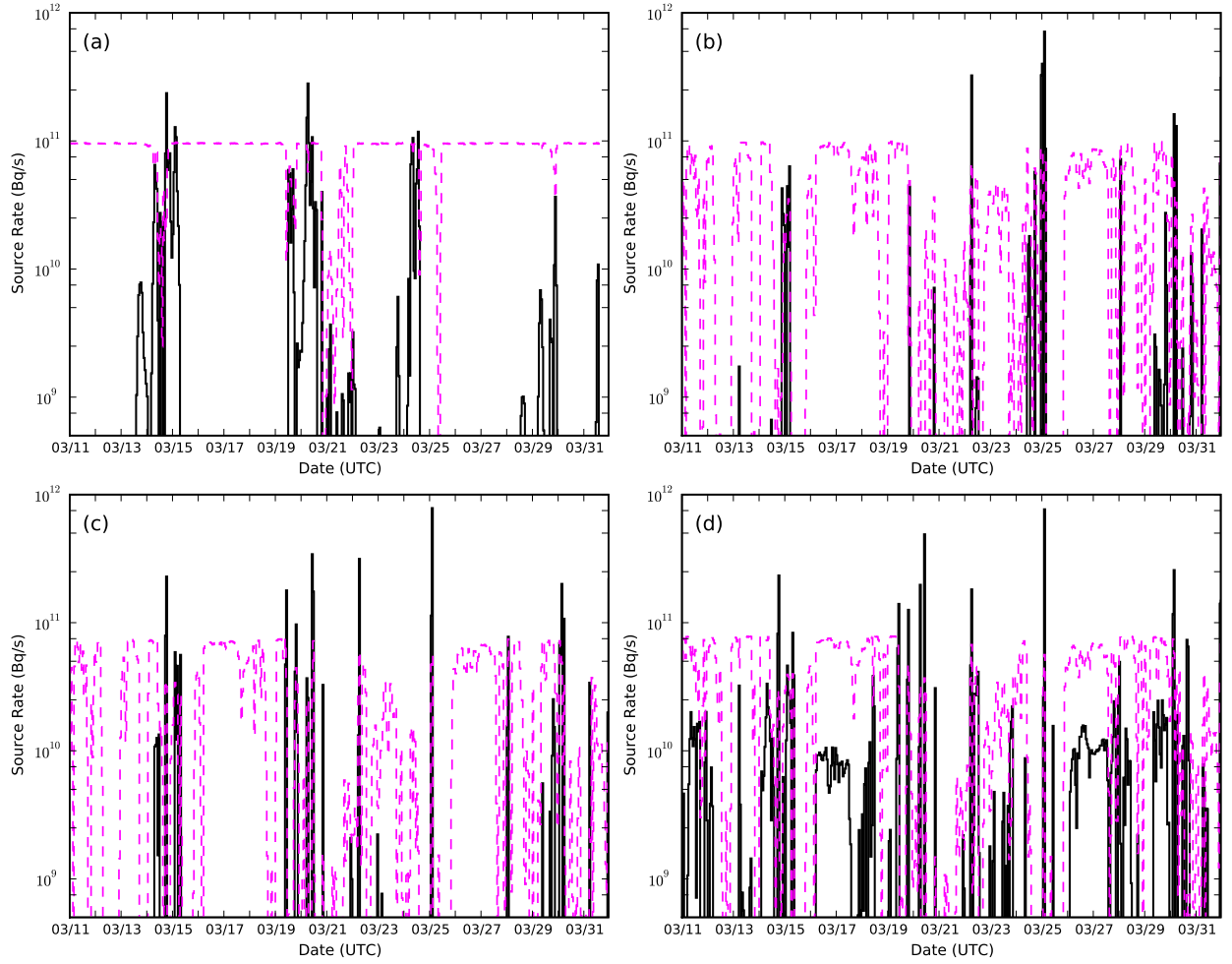


Figure 3: Full line: Temporal profile of reconstructed caesium-137 source term using (a): activity concentration of ^{137}Cs measurements in the air only, (b): total cumulated deposited ^{137}Cs measurements only, (c): the activity concentration in the air, the daily fallout measurements and the total cumulated deposition raw measurements, (d): the activity concentration in the air, the daily fallout measurements and the super-observations computed from the total cumulated deposited ^{137}Cs . Dashed lines: posterior uncertainty of the source term computed using a Monte Carlo simulation.

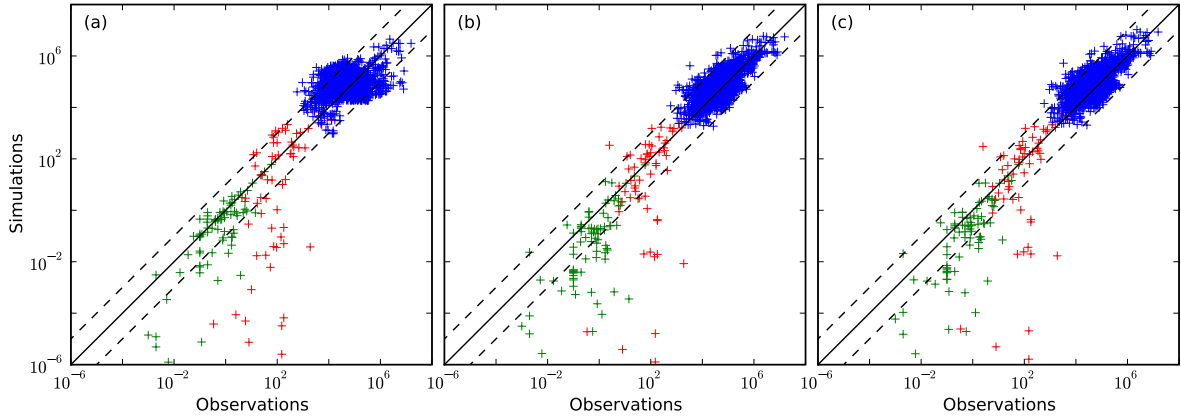


Figure 4: Scatter plots of ^{137}Cs activity concentration in the air (green crosses in Bq m^{-3}), daily measurements of deposited ^{137}Cs (red crosses in Bq m^{-2}) and total cumulated deposited ^{137}Cs (blue crosses in Bq m^{-2}). (a) inversion only using activity concentrations in the air. (b) inversion using the three raw data sets. (c) inversion using the three data sets including super-observations. The dashed lines represent a misfit of a factor 10 between observations and modelled values.

3.4.1. Estimation of parameters and total released activity

As expected, when the number of available observations increases, the results yielded by Desroziers' scheme tend to get closer to the maximum likelihood estimation.

The estimated hyper-parameters as well as the estimated total released activities are reported in Tab. 3. Excepted for the daily measurements hyper-parameter, the values of estimated hyper-parameters do not differ much when using Desroziers' scheme or the maximum likelihood method.

Using these data sets, the total released activity of ^{137}Cs is estimated to be between 1.2×10^{16} Bq and 1.3×10^{16} Bq.

3.4.2. Temporal profile and uncertainty reduction

The reconstructed source term and its uncertainty are displayed in Fig. 3(c).

The observability of the accident is improved compared to the inversion using only activity concentrations in the air. The time windows where the uncertainty is reduced are much larger, specially before 14 March and between 19 March and 26 March. The uncertainty on the total released activity has been reduced from around 65% when using only activity concentration to around 25% in this case.

The events that were retrieved in the first inversion are still reconstructed with this data set. The time offset on the release of 25 March has disappeared. The system is now able to reconstruct the late release of 15 March, from 7:00 UTC to 9:00 UTC and hence to model the

north-west pattern of deposited ^{137}Cs . The maps of observed and simulated total deposition are displayed in Fig. 5(a,b,c,d).

3.5. Results of the inverse modelling using super-observations computed from the three raw data sets

As mentioned above, measurements of the third data set are very densely distributed. This proximity certainly induces correlations in the observation errors (mainly through model errors) that are not taken into account in our system. One way to take into account these correlations would be to use a non-diagonal \mathbf{R} matrix. This method would increase the computation cost as the inverse of \mathbf{R} is needed for the minimisation of the cost function Eq. (3). Besides, this would introduce additional hyper-parameters in the correlation model which would have to be properly estimated as well. Such an estimation could be a difficult task in a situation with a low number of observations. Another route that can be chosen to deal with densely distributed data is the thinning of observations which consists in the optimal selection of observations. These techniques are already used in the assimilation of satellite data in the weather forecast community (Liu and Rabier, 2002). Yet another method would consist in computing super-observations by averaging all the observations contained in a model grid cell (the model resolution is $0.05^\circ \times 0.05^\circ$), in order to mitigate the spatial correlation in the errors. This is the method that we chose. From the 2180 measurements initially in the third data set, we computed 523

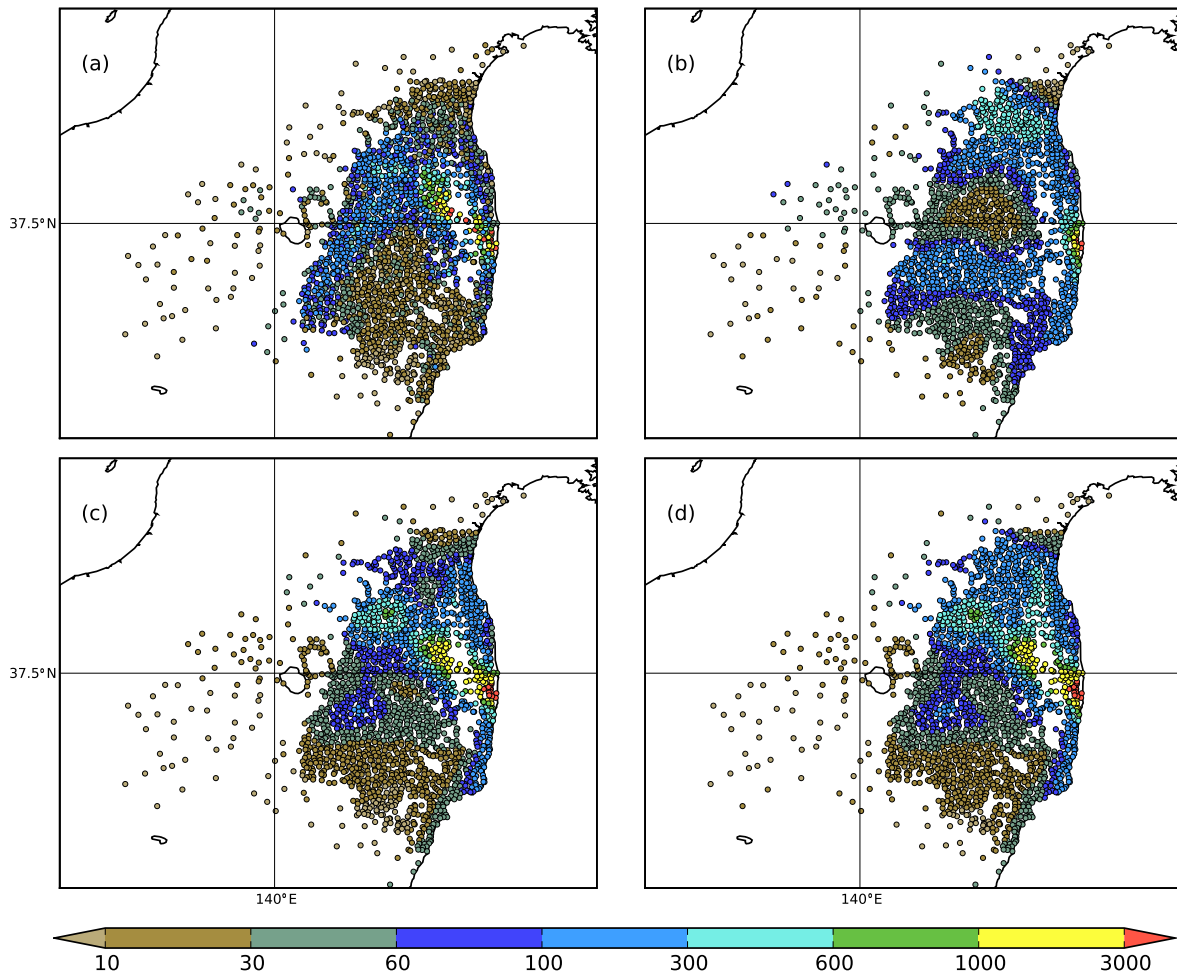


Figure 5: Map of observed and simulated deposited ^{137}Cs . Upper left (a): measurements from MEXT. Upper right (b): simulation with source reconstructed using only ^{137}Cs activity concentrations in the air. Lower left (c): simulation with source reconstructed using ^{137}Cs activity concentrations in the air, daily measurements of deposited ^{137}Cs and measurements of total cumulated deposited ^{137}Cs . Lower right (d): simulation with source reconstructed using ^{137}Cs activity concentrations in the air, daily measurements of deposited ^{137}Cs and super-observations computed from measurements of total cumulated deposited ^{137}Cs . The displayed values are in kBq m^{-2}

Table 3: Estimation of parameters and corresponding reconstructed released activity for caesium-137 source using activity concentration, fallout daily measurements and total deposition data. r_1 , r_2 and r_3 represents respectively the errors variances of these three data sets.

parameter	method	with deposition raw data	with deposition super-observations
r_1 (Bq m ⁻³)	Desroziers' scheme	3.8	2.9
	Maximum likelihood	3.3	2.3
r_2 (Bq m ⁻²)	Desroziers' scheme	580	540
	Maximum likelihood	210	240
r_3 (Bq m ⁻²)	Desroziers' scheme	325000	240000
	Maximum likelihood	320000	230000
m (Bq s ⁻¹)	Desroziers' scheme	1.2×10^{11}	1.3×10^{11}
	Maximum likelihood	1.0×10^{11}	1.0×10^{11}
Released activity (Bq)	Desroziers' scheme	1.3×10^{16}	1.9×10^{16}
	Maximum likelihood	1.2×10^{16}	1.8×10^{16}

super-observations from which we eliminated 4 super-observations located too close to the NPP, hence leaving 519 super-observations to be used as the new third data set. Note that it is not necessary to estimate the reduced variance of a super-observation since this is implicitly accounted for in the related hyper-parameter estimation.

3.5.1. Estimation of parameters and total released activity

The estimated hyper-parameters as well as the estimated total released activities are reported in Tab. 3. As expected, the estimated standard deviation of the error in the cumulated fallout (r_3) is significantly decreased by about 28%.

Using these data sets, the total released activity of ¹³⁷Cs is estimated to be between 1.8×10^{16} Bq and 1.9×10^{16} Bq.

3.5.2. Temporal profile and uncertainty reduction

The reconstructed source term and its uncertainty are displayed in Fig. 3(d). The uncertainty on the total released activity is estimated to be around 15% from a Monte Carlo study of 2×10^4 draws. However, it is the same absolute standard deviation, of about 3 PBq, as in the raw data sets inversion.

All the previously mentioned events are now retrieved with these data sets:

- Around 12 March: identified as the hydrogen explosion in Unit 1.
- On 13 March: identified as the ventings on Unit 3.
- On 14 and 15 March: multiple ventings and hydrogen explosions mainly concerning Unit 2 and

Unit 3. The late release on 15 March is now retrieved (from 7:00 to 9:00 UTC) even if its magnitude might appear weak (see section 3.6.1 for a discussion on the magnitude of the retrieved peaks).

- On 16 March: unidentified events but which correspond to pressure drops in Unit 2 and Unit 3.
- On 18 March: unidentified events probably related to Unit 3.
- On 19 March: unidentified events which correspond to an increase in several in-situ γ -dose rate measurements. The attempts of emergency cooling with the Tokyo Fire Department means began just after these events.
- On 20 March: unidentified events concerning at least Unit 2 and Unit 3.
- From 21 March to 23 March: events corresponding to smokes emitted from Unit 2 and Unit 3.
- On 25 March: unidentified event possibly concerning Unit 2. The magnitude of this peak might appear over-estimated (see section 3.6.1 for a discussion on the magnitude of the retrieved peaks).
- On 30 March: unidentified event.

From the scatter plots of Fig. 4(b,c), it is clear that the simulated cumulated deposition data are comparable when using the source term built from raw data or from super-observations. This is also visible on the deposition maps in Fig. 5(c,d). On the other hand the system's ability to simulate the two other data sets is slightly improved when using the source term retrieved from the super-observations. The assumption that the observation errors are uncorrelated led the system to give too

much weight to the third raw data set in the inverse modelling algorithm. This over-confidence may be corrected when using super-observations so that the information content in the system is better balanced.

3.6. Discussion about the retrieved sources

These inversions offer an objective estimation using data assimilation techniques of what can be extracted from the data sets and the numerical model. The assimilated observations may or may not suffice to constrain the source term parameters. We showed that the cumulated deposition data alone are not sufficient to offer a satisfying chronology of the source term and that the joint assimilation of the three data sets helped to better constrain the chronology. Yet, the magnitude of the identified peaks remains questionable. The data may or may not be able to constrain them well enough, although the estimate of the total released activity seems robust.

3.6.1. On the magnitude of the peaks

It is generally thought that the main releases have occurred around March 15. These releases probably contributed the most to the north-west pattern of the deposition map. Our system, where the cumulated deposition data do not contain any information in time, may not be constrained enough by air concentration observations or daily measurements of fallout (specially before March 18) to precisely balance the contributions of the releases of March 14-15 (around 15% of the total retrieved activity), March 20 (around 13%) and March 25 (around 15%).

The peaks of March 20 and March 25 are not only reconstructed with the help of deposition measurements. They are both additionally explained by measurements of activity in the air from a particular monitoring station, located in Fukushima city, which measured an activity concentration of about 32 Bq.m^{-3} around March 20 and about 14 Bq.m^{-3} around March 25. It is possible that these measurements are caused by a re-suspension of previously deposited caesium-137. But resuspension is not implemented in our model, so that resuspension events are likely to be accounted for by fictitious releases in the source term. We tried and performed a new inversion using the same data as in Section 3.5 but without the measurements of this station. Indeed, the results from the super-observations showed an increase of the released activity on March 14 and 15, from 2.6 to 3.0 PBq (from 13% to 16% of an unchanged total emitted activity of 19 PBq), and a decrease on March 20, from 2.6 to 2.0 PBq (from 13% to 11% of the total emitted activity). Yet, no consequences were observed on the magnitude of the peak on March 25.

We also performed inversions using a non-null first guess inferred from independent γ -dose measurements (Saunier et al., 2013), which indicates to the system that most of the releases occurred before March 18. As a result, the total estimated released activity increased by about 15%, but the peak on March 25 still remained almost at the same level.

As a more drastic test and to force the system to reconstruct higher releases before March 18, we reduced the inversion window to this period. As a consequence the releases of March 15 increased (from 0.6 PBq to 2.6 PBq using the three raw data sets, and from 0.9 PBq to 3.9 PBq using the three data sets with super-observations as the third data set), but the total estimated releases decreased (respectively from 12 PBq to 5 PBq, and from 18 PBq to 11 PBq). It is consistent with the inversion using only activity concentration in the air where approximately 6 PBq were estimated to be released after this date. At the same time the reanalysis of deposition map has been degraded especially in the central area of the north-west pattern which is the most contaminated. This shows that: (i) considering only this shorter time window, the system can not properly reconstruct the deposition pattern shown in Fig. 5(a). (ii) Releases probably occurred after March 19 and may have contributed to the north-west pattern of the deposition map, but their magnitude is still difficult to estimate because of the lack of observations with a temporal information (such as activity concentration in the air) in this area. (iii) This may highlight the difficulty to model the removal processes and specially the wet deposition. It is also possible that the cumulated deposition map, whose measurements have been made several months after the accident, is not anymore faithful to the deposition events of the accident.

We also tested a different reconstruction resolution. Instead of reconstructing a source term with a one-hour time-step ($N = 504$), an inversion was carried out on a source term with a three-hour time-step ($N = 168$). It is generally admitted that the inversion is sensitive to the resolution of the control space (Bocquet et al., 2011) and that the system cannot generally provide reliable information at a too precise resolution (for instance the model resolution). Nevertheless, estimating the prior errors allows to compensate the impact of the control space resolution by tuning regularisation in the inversion, so that the dependence in the resolution should be mitigated. The estimated source term, using the same data sets as in Section 3.5, as well as its uncertainty are displayed in Fig. 6. The total released activity is now estimated to be $1.6 \times 10^{16} \text{ Bq}$ with an uncertainty of 18%. The estimated released quantities were mainly

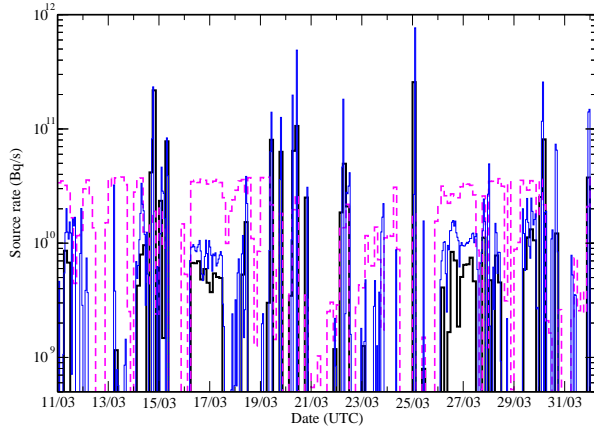


Figure 6: Full black line: Temporal profile of reconstructed caesium-137 source term with a reconstruction resolution of 3 h using the activity concentration in the air, the daily fallout measurements and the super-observations computed from the total cumulated deposited ^{137}Cs . Dashed line: posterior uncertainty of the source term computed using a Monte Carlo simulation. Thin blue line: Temporal profile of reconstructed caesium-137 source term with a reconstruction resolution of 1 h using the same data.

reduced on March 30 (from 2.1 PBq to 1.3 PBq), on March 20 (from 2.6 PBq to 2.1 PBq) and on March 25 (from 2.9 PBq to 2.8 PBq), but remain at a high level. At the same time, the estimated released quantities increased on March 14-15 from 2.6 PBq to 4.1 PBq. Many of the low magnitude peaks with a high uncertainty have been reduced or have disappeared.

3.6.2. On the difference between the sources reconstructed with raw data and super-observations

Comparing the source profile obtained from the super-observations (Fig. 3(d)) with the source profile obtained from the inversion of the raw data sets (Fig. 3(c)), 2 PBq of caesium-137 are found to increase already existing peaks.

Moreover, when using super-observations instead of raw data, new release episodes appear in the source reconstructed term. Compared to the raw data sets inversion, 4 PBq of caesium-137 are found in new time slots. Yet the related uncertainty is still high and even above the retrieved peaks. Nonetheless, the new releases episodes that appear in the inversion do not seem to be artefacts of the inversion algorithm. They do correspond to observed events in the NPP and to events reconstructed by other studies (Stohl et al., 2012; Saunier et al., 2013).

Judging from the r_i , the errors of the super-observations as well as their use by the model is di-

agnosed as more reliable than the use of the raw data sets. By contrast, the Tikhonov regularising term is less constraining and new peaks can more easily form in the reconstructed source term.

On a physical level, the appearance of the new peaks results from the smoothing of the deposition observations that in turn leads to a smoothing in the retrieved source. This is probably why the only peak which is reduced when using super-observations is the highest one on March 25. This peak is mainly induced by the ^{137}Cs deposit observations of the north-west very thin pattern. Hence, a smoothing of this pattern can lead to a new balance of the peaks (March 15, March 20 or March 25).

4. Conclusion

In order to reconstruct source terms of accidental pollutant releases into the atmosphere, we have proposed an inverse modelling algorithm able to use all available data in the same inversion (concentrations in the air, measurements of fallout, integrated measurements, etc.). The algorithm relies on the relationship provided by the atmospheric transport model between the source vector and the observation set and on prior errors introduced in the system: the background errors and the observation errors. To properly balance the information content in the system a proper estimation of these prior errors is crucial.

In this aim, we proposed two methods relying on the maximum likelihood principle that we applied to the challenging reconstruction of the ^{137}Cs source term released during the Fukushima Daiichi nuclear power plant accident in March 2011. Three data sets were used in the inversion process: activity concentrations in the air, daily measurements of fallout and total cumulated deposition data. Consequently, the prior estimation concerned 4 hyper-parameters: the variance of the background errors and the variance of each observation set errors. Averaged observations (called in the article super-observations) have also been considered for the third data set in order to reduce correlation in errors that had been neglected in the algorithm. Posterior uncertainty related to the estimated releases have also been estimated through a Monte Carlo analysis. Such an estimation is only possible with properly estimated prior errors.

With these methods and without super-observations, the total released activity is estimated to be between $1.2 - 1.3 \times 10^{16}$ Bq with a related uncertainty around 25%. When using super-observations instead of raw fallout data, the total released activity is estimated to be between $1.8 - 1.9 \times 10^{16}$ Bq with a related uncertainty

around 15%. From our estimations the main ^{137}Cs contamination over Japan results from releases on 14-15 March, 19-20 March, 25 March and 30 March. Nevertheless, the uncertainty of each retrieved peak still remains high. Consequently the exact magnitude of the retrieved peaks has to be handled carefully. Some other significant releases might also have occurred when the wind was blowing directly towards the Pacific Ocean and are thus not totally reconstructed by our method using only data over Japan.

With the given data sets, the reconstruction of the source term could be improved on the condition that model error be better constrained. Two main influential sources of error were identified in the course of this study. Firstly, the reconstruction was found to be highly sensitive to the precipitation fields of the meteorological model. Even if the spatial distribution and chronology of the precipitation events were matching independent precipitation observations, we found it difficult to properly estimate the magnitude of those events. Secondly, with the given precipitation fields, the reconstruction remains sensitive to the physical process parameterisation in the ATM. One promising route to better constrain those processes is the inverse modelling of physical parameters (Bocquet, 2012).

Finally it seems promising to develop methods able to simultaneously reconstruct source terms of several radionuclides using all available data including γ -dose rates (Saunier et al., 2013). The number of observations, and in particular observations with an information of time, would then increase substantially. But the involved methods would certainly be more complex owing to correlations in observation errors and higher prior uncertainties in the data assimilation system.

5. Acknowledgements

This study has been supported by the IMMANENT project of Paris-Est University, and by the INSU/LEFE-ASSIM project ADOMOCA-2.

References

Abida, R., Bocquet, M., 2009. Targeting of observations for accidental atmospheric release monitoring. *Atmos. Env.* 43, 6312–6327.
 Anderson, J. L., 2007. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A* 59, 210–224.
 Bocquet, M., 2005. Reconstruction of an atmospheric tracer source using the principle of maximum entropy. I: Theory. *Q. J. R. Meteorol. Soc.* 131, 2191–2208.
 Bocquet, M., 2012. Parameter field estimation for atmospheric dispersion: Application to the Chernobyl accident using 4D-Var. *Q. J. R. Meteorol. Soc.* 138, 664–681.

Bocquet, M., Pires, C. A., Wu, L., 2010. Beyond Gaussian statistical modeling in geophysical data assimilation. *Mon. Wea. Rev.* 138, 2997–3023.
 Bocquet, M., Wu, L., Chevallier, F., 2011. Bayesian design of control space for optimal assimilation of observations. I: Consistent multiscale formalism. *Q. J. R. Meteorol. Soc.* 137, 1340–1356.
 Byrd, R. H., Lu, P., Nocedal, J., 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing* 16, 1190–1208.
 Chapnik, B., Desroziers, G., Rabier, F., Talagrand, O., 2004. Properties and first application of an error-statistics tuning method in variational assimilation. *Q. J. R. Meteorol. Soc.* 130, 2253–2275.
 Chapnik, B., Desroziers, G., Rabier, F., Talagrand, O., 2006. Diagnosis and tuning of observational error in a quasi-operational data assimilation setting. *Q. J. R. Meteorol. Soc.* 132, 543–565.
 Chino, M., Nakayama, H., Nagai, H., Terada, H., Katata, G., Yamazawa, H., 2011. Preliminary estimation of release amounts of I131 and Cs137 accidentally discharged from the Fukushima Dai-ichi nuclear power plant into the atmosphere. *Journal of Nuclear Science and Technology* 48, 1129–1134.
 Davoine, X., Bocquet, M., 2007. Inverse modelling-based reconstruction of the Chernobyl source term available for long-range transport. *Atmos. Chem. Phys.* 7, 1549–1564.
 Dee, D. P., 1995. On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.* 123, 1128–1145.
 Delle Monache, L., Lundquist, J. K., Kosovic, B., Johannesson, G., Dyer, K. M., Aines, R. D., Chow, F. K., Belles, R. D., Hanley, W. G., Larsen, S. C., Loosmore, G. A., Nitao, J. J., Sugiyama, G. A., Vogt, P. J., 2008. Bayesian inference and Markov chain Monte Carlo sampling to reconstruct a contaminant source on a continental scale. *Journal of Applied Meteorology and Climatology* 47, 2600–2613.
 Desroziers, G., Berre, L., Chapnik, B., Poli, P., 2005. Diagnosis of observation, background and analysis error statistics in observation space. *Q. J. R. Meteorol. Soc.* 131, 3385–3396.
 Desroziers, G., Ivanov, S., 2001. Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Q. J. R. Meteorol. Soc.* 127, 1433–1452.
 Elbern, H., Strunk, A., Schmidt, H., Talagrand, O., 2007. Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmos. Chem. Phys.* 7, 3749–3769.
 Estournel, C., Bosc, E., Bocquet, M., Ulses, C., Marsaleix, P., Winiarek, V., Osvath, I., Nguyen, C., Duhaut, T., Lyard, F., Michaud, H., Auclair, F., 2012. Assessment of the amount of cesium-137 released to the Pacific Ocean after the Fukushima accident and analysis of its dispersion in the Japanese coastal waters. *J. Geophys. Res.* 117, C11014.
 Hajivassiliou, V., McFadden, D., Ruud, P., 1996. Simulation of multivariate normal rectangle probabilities and their derivatives - Theoretical and computational results. *Journal of Econometrics* 72, 85–134.
 Hansen, P. C., 1992. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review* 34, 561–580.
 Katata, G., Terada, H., Nagai, H., Chino, M., 2012. Numerical reconstruction of high dose rate zones due to the Fukushima Dai-ichi Nuclear Power Plant accident. *J. Environ. Radioactivity* 111, 2–12.
 Korsakissok, I., Mathieu, A., Didier, D., 2013. Atmospheric dispersion and ground deposition induced by the Fukushima Nuclear Power Plant accident: A local-scale simulation and sensitivity study. *Atmos. Env.* 70, 267–279.
 Krysta, M., Bocquet, M., Brandt, J., 2008. Probing ETEX-II data set with inverse modelling. *Atmos. Chem. Phys.* 8, 3963–3971.
 Li, H., Kalnay, E., Miyoshi, T., 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble

- Kalman filter. *Q. J. R. Meteorolog. Soc.* 135, 523–533.
- Liu, Z.-Q., Rabier, F., 2002. The interaction between model resolution, observation resolution and observation density in data assimilation: A one-dimensional study. *Q. J. R. Meteorolog. Soc.* 128, 1367–1386.
- Louis, J., 1979. A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteor.* 17, 197–202.
- Maryon, R. H., Smith, F. B., Conway, B. J., Goddard, D. M., 1991. The U.K. nuclear accident model. *Progress in Nuclear Energy* 26, 85–104.
- Mathieu, A., Korsakissok, I., Quélo, D., Groëll, J., Tombette, M., Didier, D., Quentric, E., Saunier, O., Benoit, J.-P., Isnard, O., 2012. Atmospheric dispersion and deposition of radionuclides from the Fukushima Daiichi nuclear power plant accident. *Elements* 8, 195–200.
- Ménard, R., Cohn, S. E., Chang, L.-P., Lyster, P. M., 2000. Assimilation of stratospheric chemical tracer observations using a Kalman filter. Part I: Formulation. *Mon. Wea. Rev.* 128, 2654–2671.
- Michalak, A. M., Bruhwiler, L., Tans, P. P., 2004. A geostatistical approach to surface flux estimation of atmospheric trace gases. *J. Geophys. Res.* 109, D14109.
- Mitchell, H. L., Houtekamer, P. L., 1999. An adaptive ensemble Kalman filter. *Mon. Wea. Rev.* 128, 416–433.
- Quélo, D., Krysta, M., Bocquet, M., Isnard, O., Minier, Y., Sportisse, B., 2007. Validation of the Polyphemus platform on the ETEX, Chernobyl and Algeciras cases. *Atmos. Env.* 41, 5300–5315.
- Saïde, P., Bocquet, M., Osses, A., Gallardo, L., 2011. Constraining surface emissions of air pollutants using inverse modeling: method intercomparison and a new two-step multiscale approach. *Tellus B* 63, 360–370.
- Saunier, O., Mathieu, A., Didier, D., Tombette, M., Quélo, D., Winiarek, V., Bocquet, M., 2013. An inverse modeling method to assess the source term of the Fukushima nuclear power plant accident using gamma dose rate observations. *Atmos. Chem. Phys. Discuss.* 13, 15567–15614.
- Schwinger, J., Elbern, H., 2010. Chemical state estimation for the middle atmosphere by four dimensional variational data assimilation: A posteriori validation of error statistics in observation space. *J. Geophys. Res.* 115, D18307.
- Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., Huang, S.-Y., Wang, W., Powers, J. G., 2008. A description of the advanced research WRF version 3. Tech. rep., NCAR.
- Stauffer, D. R., Seaman, N. L., 1994. On multi-scale four dimensional data assimilation. *J. Appl. Meteor.* 33, 416–434.
- Stohl, A., Seibert, P., Wotawa, G., Arnold, D., Burkhardt, J. F., Eckhardt, S., Vargas, A., Yasunari, T. J., 2012. Xenon-133 and caesium-137 releases into the atmosphere from the Fukushima Dai-ichi nuclear power plant: determination of the source term, atmospheric dispersion, and deposition. *Atmos. Chem. Phys.* 12, 2313–2343.
- Terada, H., Katata, G., Chino, M., Nagai, H., 2012. Atmospheric discharge and dispersion of radionuclides during the Fukushima Dai-ichi Nuclear Power Plant accident. Part ii: verification of the source term and analysis of regional-scale atmospheric dispersion. *J. Environ. Radioactivity* 112, 141–154.
- Winiarek, V., Bocquet, M., Saunier, O., Mathieu, A., 2012a. Correction to "Estimation of errors in the inverse modeling of accidental release of atmospheric pollutant: Application to the reconstruction of the cesium-137 and iodine-131 source terms from the Fukushima Daiichi power plant". *J. Geophys. Res.* 117, D18118.
- Winiarek, V., Bocquet, M., Saunier, O., Mathieu, A., 2012b. Estimation of errors in the inverse modeling of accidental release of atmospheric pollutant: Application to the reconstruction of the cesium-137 and iodine-131 source terms from the Fukushima Dai-ichi power plant. *J. Geophys. Res.* 117, D05122.
- Winiarek, V., Vira, J., Bocquet, M., Sofiev, M., Saunier, O., 2011. Towards the operational estimation of a radiological plume using data assimilation after a radiological accidental atmospheric release. *Atmos. Env.* 45, 2944–2955.
- Yee, E., Lien, F.-S., Keats, A., D'Amours, R., 2008. Bayesian inversion of concentration data: Source reconstruction in the adjoint representation of atmospheric diffusion. *Journal of Wind Engineering and Industrial Aerodynamics* 96, 1805–1816.

Annexe G

Collaborations

Chazette, P., Bocquet, M., Royer, P., Winiarek, V., Raut, J.C., Labazuy, P., Gouhier, M., Lardier, M., Cariou, J.P., 2012. Eyjafjallajökull ash concentrations derived from both lidar and modeling. *J. Geophys. Res.* **117**, D00U14.

À la suite de l'éruption du volcan Eyjafjöll en Islande en avril 2010, un panache constitué de vapeur d'eau, de gaz volcaniques et de cendre a recouvert une grande partie de l'Europe, perturbant pendant plusieurs jours les transports aériens sur le continent. Des mesures de concentrations de particules par LIDAR (LIght Detection And Ranging) ont été réalisées, aussi bien par des instruments au sol que par le satellite CALIPSO (Cloud Aerosol Lidar and Infrared Pathfinder Satellite Observations). Parallèlement, nous avons réalisé des simulations de la dispersion du panache à l'échelle européenne en utilisant le modèle eulérien POLAIR3D. Nous avons pour cela utilisé une méthode "simulation/correction" pour estimer la taille caractéristique des particules qui influence la vitesse de sédimentation, impossible à négliger dans cette situation. Nous avons trouvé un très bon accord entre la simulation et les mesures, notamment en terme de chronologie et de quantités retrouvées.

Estournel, C., Bosc, E., Bocquet, M., Ulses, C., Marsaleix, P., Winiarek, V., Osvath, I., Nguyen, C., Duhaut, T., Lyard, F., Michaud, H., Auclair, F., 2012. Assessment of the amount of cesium-137 released into the Pacific Ocean after the Fukushima accident and analysis of its dispersion in Japanese coastal waters. *J. Geophys. Res.* **117**, C11014.

Dans le cadre de l'estimation des conséquences de l'accident de Fukushima sur l'environnement, des simulations ont été réalisées pour calculer l'évolution spatio-temporelle de la pollution dans l'Océan Pacifique. Deux modes d'entrée des polluants radioactifs dans l'océan co-existent : une entrée directe sur la côte japonaise par écoulement de l'eau contaminée après son utilisation pour refroidir les différents éléments des réacteurs qui en ont besoin et une entrée indirecte par dépôt des polluants présents dans le panache transporté à grande échelle. Nous avons estimé, par des simulations avec POLAIR3D, la distribution spatio-temporelle de cette composante, qui par la suite a été fournie comme donnée d'entrée au système SIROCCO simulant la dispersion des radionucléides dans l'Océan Pacifique.

Saunier, O., Mathieu, A., Didier, D., Tombette, M., Quélo, D., Winiarek, V., Bocquet, M., 2013. An inverse modeling method to assess the source term of the Fukushima nuclear power plant accident using gamma dose rate observations. *Atmos. Chem. Phys.* **13**, 11403-11421.

Comme nous l'avons déjà évoqué dans ce manuscrit et en particulier dans la conclusion,

dans le cas d'un rejet dans l'atmosphère de polluant radioactifs, les mesures de débits de dose γ sont généralement disponibles en grand nombre. Mais comme ces mesures rendent compte de l'influence de plusieurs radio-éléments, qu'ils soient présents dans l'air ou déjà déposés au sol, elles restent difficiles à assimiler pour estimer le terme source. Dans cet article, une méthode de modélisation inverse, inspirée des méthodes présentées dans le chapitre 3 et adaptée à ce problème spécifique, est proposée pour réaliser cette estimation et permet d'estimer les profils temporels des rejets d'une dizaine d'espèces dans le cas de l'accident de Fukushima.